Universiteit
Leiden
The Netherlands

Why Jesus and Job spoke bad Welsh : the origin and distribution of V2 orders in Middle Welsh

Meelen, M.

Citation
Meelen, M. (2016, June 21). *Why Jesus and Job spoke bad Welsh : the origin and distribution of V2 orders in Middle Welsh*. *LOT dissertation series*. LOT, Utrecht. Retrieved from https://hdl.handle.net/1887/40632

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/40632

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/40632 holds various files of this Leiden University dissertation.

Author: Meelen, M.
Title: Why Jesus and Job spoke bad Welsh : the origin and distribution of V2 orders in Middle Welsh
Issue Date: 2016-06-21

Why Jesus and Job spoke bad Welsh

The origin and distribution of V2 orders in Middle Welsh

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
e-mail: lot@uu.nl
http://www.lotschool.nl

ISBN: 978-94-6093-206-9
NUR: 616

Copyright © 2016 Marieke Meelen. All rights reserved.

Why Jesus and Job spoke bad Welsh

The origin and distribution of V2 orders in Middle Welsh

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus Prof. mr. dr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 21 juni 2016
klokke 15.00 uur

door

Marieke Meelen

geboren op 10 september 1986
te Maastricht

Promotores:	Prof. Dr. Lisa Lai-Shen Cheng Prof. Dr. Alexander Lubotsky
Promotiecommissie:	Prof. Dr. Maarten Mous Prof. Dr. Susan Pintzuk (University of York) Dr. David Willis (University of Cambridge)

to my family

Contents

Acknowledgements

In accordance with Leiden tradition, I'll skip the words of gratitude dedicated to those who helped me become an academic researcher, from the first drafts to the final approval of this thesis. Needless to say, there would be no thesis or academic researcher without them.

First of all, I would like to thank Karel Jongeling, who inspired me to work on the topic of Welsh word order from the first Welsh class I took as an undergraduate. His stories of Wales made me fall in love with the country and people of Cymru. The title of this thesis honours his impact on my academic choices, because no one knows more interesting facts about both the Welsh and Hebrew Bible than he.

Many thanks to all others who helped me learn their beautiful language at *Cymraeg i Oedolion* and to Patrick Sims-Williams for inviting me to the *Datblygiad yr Iaith Gymraeg* meetings funded by the British Academy in Oxford, Cardiff, Utrecht, London and Cambridge. This thesis benefited greatly from the comments and discussions I had there with all these great Welsh scholars!

Then a word of thanks to my 'Award family', all those wonderful people I've met since I started volunteering for the International Award for Young People in 2004. You have never failed to show me the importance of the world outside Academia, an invaluable view that put my work on Welsh linguistics in the right perspective (*Stelling 8* is for you). Special thanks to Paul van Berlo, for introducing me to this wonderful programme and Wim van der Laan for always finding ways to convince me it was worthwhile to spend all my free time volunteering for the Award. Thanks to all members of the Board of the Dutch NAO, especially Pascalle, Franck, Sophie, Janieke, and Barry, for teaching me so many practical skills I needed as a Secretary, from designing web shops to risk assessment in third-world countries as well as difficult legislative procedures and methods on hiring, firing and managing volunteers. Thanks to all Award holders, participants, volunteers and ambassadors I've met, trained, assessed and worked with during numerous Adventurous Journeys and (Residential) Projects we did. Amir, Amy, Fleur Jana, Jassin, Jenny, Helen, Kirsten, Melinda, Sameer and many others working

or volunteering for the Award: thanks to you I feel I can travel anywhere in this world meeting friends.

A special thanks to Luc Bartholomée who organised Project Nepal with me from 2012 onwards and to our wonderful partners at Himalayan Care Hands and Himalayan Leaders: Albert and Fer for the great organisation and meetings in the Netherlands; Indira, Jagat, Gopal, Prakash and Toran for their inspiring work in Nepal. Finally, I want to thank the EMAS office and in particular Deirdre, Dale and David for letting me be part of the most wonderful team delivering teacher training courses in Geneva, Abu Dhabi, Cyprus, Dubai and Madrid in the past few years. Adriana, Anna, Dragos, Gasper, Lina, Ludek, Mihai, Milan, Ruxi, Steve and Tasmin: you're the best!

Turning back to my academic life, many thanks to my fellow members of the PhD council who kept me sane every time we had to fight for our rights - something we had to do way too often: Andreea, Bora, Daan, Elly and Viktorija. Thanks to many others in our department for all the great discussions and lunch breaks: Allison, Aliza, Aniko, Enrico, Giuseppe, Laura, Leticia, Sima, Victoria and Yang. I'm grateful to Bastien, Bobby, Olga and Stella for teaching me how to do 'pretty things' in LateX and R statistics. And to Erwin Komen and Barend Beekhuizen for their endless patience and support whenever I had any computational questions, especially Barend for helping me code some crucial scripts in Python. This thesis could not have been done without Erwin's dedicated software Cesax and CorpusStudio and his patience while teaching me how to use it.

The people I've met during my stay abroad in Cambridge in both the Celtic and Linguistics departments (made possible by a LUF scholarship): many thanks to David Willis, Ian Roberts, Theresa Biberauer, Jenneke van der Wal and Paul Russell for giving me a warm welcome there! My fellow PhDs in Cambridge, Desirée, Kathrin, Myriah and Silva: thanks for always letting me stay and for the most amazing cake! Without the writing marathons at Queens' and John's, I'd never have started writing in the first place: Faye, James, Liz and Sazana, I miss you terribly - may the ducks and drakes never die! Apart from Cambridge, I also had the wonderful opportunity to spend some time in Aberystwyth and Nant Gwrtheyrn to learn more Welsh, because of a generous allowance from the Paolo Pistoi Scholarship. Many thanks as well to Erich Poppe and Elena Parina for inviting me to Marburg to talk more about annotating Middle Welsh texts and using the database I created for this thesis. A very special thanks to Elena's family for letting me stay and taking me up all those pretty hills in Hesse!

Finally, a few words to some of my friends and family outside academia. First of all my paranymphs, Olivia and Andreea. You were there for me whenever I needed to pour my heart out: I cannot thank you enough for listening, understanding and making me smile when I needed it the most, you're like my closest family!

Thanks to my wonderful climbing friends (*Stellingen 10 & 11* are for you). Daniel for exploring the Lake District with me in Winter; Scott for teaching me all about ice climbing in Snowdonia; Christian, Edo, Jessie and Lena for conquering the walls in Haarlem, Leiden and The Hague; and Constantijn, for our incredible

adventures in the Cairngorms, Norway and Austria, and most of all, for reaching the summit of Mt Blanc with me in the Summer of 2012, when I needed a victory like that more than ever.

My dearest Genie, Tingilya and 'big sis', I am more grateful to you than I can ever express in words for 'bringing me back' during the second year of my PhD when a combination of very unfortunate circumstances made me lose all faith in myself. Without your endless patience, comforting words and warm hugs, even from afar, I would never have found the energy and inspiration again to continue working. You reminded me I'm actually a very positive person by showing me how Matterhorn whiskey glasses can be more than half full, even when they're empty. 'Nothing's forgotten', as I learnt from Robin through Polverello, 'Nothing is ever forgotten'.

This book is dedicated to my family because they've always supported me, whatever crazy adventure I was about to undertake (and there were many over the past few years: majoring in Welsh historical syntax was definitely not the worst!) and whatever unexpected choices I made without giving any logical explanation. I have the utmost respect for never failing to believe in me without questioning or understanding even what I do and why I do it. It's exactly because of this early-given independence, I learnt how to take good care of myself and those around me: a more valuable lesson than anything I could ever write in this or any other book.

Abbreviations and Primary Text Editions

Works of Reference and Primary Text Editions

AM	*Audacht Morainn*, ed. Fergus Kelly, Dublin: Dublin Institute for Advanced Studies, 1976.
B	*The Bulletin of the Board of Celtic Studies*, Cardiff.
b1588	*Y Beibl Cys-segr-lan. Sef yr Hen Destament, a'r Newydd*, translated by William Morgan, Imprinted at London by the Deputies of Christopher Barker, Printer to the Queenes most excelent Maiestie, 1588. Available online from the National Library of Wales digital manuscripts `www.llgc.org.uk/big/index_s.htm` and in the Historical Corpus of Welsh by David Willis: `http://people.ds.cam.ac.uk/dwew2/hcwl`.
BBC	*The Black Book of Camarthen*, ed. J. Gwenogvryn Evans, Pwllheli, 1906.
BD	*Brut Dingestow*, ed. Henry Lewis, Caerdydd: Gwasg Prifysgol Cymru, 1942.
BM	*Breuddwyd Maxen*, ed. Ifor Williams, Bangor: Jarvis a Foster, 1908.
BMer	*The Life of St. Meriasek*, ed. Whitley Stokes, London, 1872.
BR	*Breudwyt Ronabwy*, ed. Melville Richards, Caerdydd: Gwasg Prifysgol Cymru, 1948.
Branwen	*Branwen Uerch Lyr*, ed. Derick S. Thomson, Dublin: Dublin Institute for Advanced Studies, 1961.
BT	*Facsimile and Text of the Book of Taliessin*, ed. J. Gwenogvryn Evans, Llanbedrog, 1910.
CA	*Canu Aneirin*, ed. Ifor Williams, Caerdydd: Gwasg Prifysgol Cymru, 1938.
CF	'The Computus Fragment', ed. Ifor Williams, *Bulletin of the Board of Celtic Studies*, 3 (1927), 245-72.

Chad	'The Welsh Marginalia in the Lichfield Gospels Part I', ed. Dafydd Jenkins and Morfydd E. Owen, *Cambridge Medieval Celtic Studies*, 5 (1983), 37-66.
CLlH	*Canu Llywarch Hen*, ed. Ifor Williams, Caerdydd: Gwasg Prifysgol Cymru, 1935.
CLlLl	*Cyfranc Lludd a Llefelys*, ed. Brynley F. Roberts, Dublin: Dublin Institute for Advanced Studies, 1975.
CO	*Culhwch ac Olwen: An Edition and Study of the Oldest Arthurian Tale*, ed. Rachel Bromwich and D. Simon Evans, Cardiff: University of Wales Press, 1992.
DB	*Delw y Byd (Imago Mundi)*, ed. Henry Lewis and P. Diverres, Caerdydd: Gwasg Prifysgol Cymru, 1928.
DGVB	*Dictionnaire des gloses en vieux breton* (containing glosses from MS Angers 477), ed. Léon Fleuriot, Paris: Klincksieck, 1964.
Gereint	*Ystorya Gereint Uab Erbyn*, ed. R.L. Thomson, Dublin: Dublin Institute for Advanced Studies, 1997.
GPC	*Geiriadur Prifysgol Cymru*, ed. R.J. Thomas, Caerdydd: Gwasg Prifysgol Cymru, available online: `http://geiriadur.ac.uk/gpc/gpc.html`.
HGC	*Hen Gerddi Crefyddol*, ed. Henry Lewis, Caerdydd: Gwasg Prifysgol Cymru, 1931.
HGK	*Historia Gruffud vab Kenan*, ed. D. Simon Evans, Caerdydd: Gwasg Prifysgol Cymru, 1977.
Juv.	'Naw Englyn y Juvencus', ed. Ifor Williams, *Bulletin of the Board of Celtic Studies*, 6 (1932), 205-24.
Laws	*Machlud Cyfraith Hywel*, Llawysgrif BL Add. 22356 (S), ed. Christine James, available from `www.cyfraith-hywel.org.uk/en/machlud-cyf-hyw.php`, last accessed on 5 April 2016.
LL	*The Text of the Book of Llan Dav*, eds. J. Gwenogvryn Evans and John Rhys, Oxford, 1893.
MAV	*Marvailhou ar Vretoned*, a collection of Breton tales, mostly from the 19th century, Brest, 1941.
MBJJ	*Ma Beaj Jeruzalem*, ed. L. Le Clerc, Saint-Brieuc, 1903.
MC	OW glosses in the Corpus Christi College (Cambridge) MS. of the *De Nuptiis Philologiae et Mercurii* by Martianus Capella, ed. Whitley Stokes, in *Archaeologia Cambrensis*, 1-21, 1873.
N	*La Vie de Sainte Nonne*, Middle Breton mystery play, ed. E. Ernault in *Revue Celtique* 8, (1887), pp. 230-301.
Nl	*An Nouelou ancien ha devot*, Middle Breton Christmas hymns, ed. Hersart de la Villemarqué from a book of 1650, in *Revue Celtique* 10-13.
O	*Ordinale de Origine Mundi*, ed. Edwin Norris,in *The Ancient Cornish Drama* 1, 2ff, Oxford, 1859.
Owein	*Owein or Chwedyl Iarlles y Ffynnawn*, ed. R.L. Thomson, Dublin: Dublin Institute for Advanced Studies, 1968.

Ox	'Glosau Rhydychen [The Oxford Glosses]: Mesurau a Phwysau', ed. Ifor Williams, *Bulletin of the Board of Celtic Studies*, 5 (1930), 226-48.
Peredur	*Historia Peredur vab Efrawc*, ed. Glenys Witchard Goetinck, Caerdydd: Gwasg Prifysgol Cymru, 1976.
PKM	*Pedeir Keinc y Mabinogi*, ed. Ifor Williams, Caerdydd: Gwasg Prifysgol Cymru, 1930.
Pwyll	*Pwyll Pendeuic Dyuet*, ed. R.L. Thomson, Dublin: Dublin Institute for Advanced Studies, 1957.
R	*Ordinale de Resurrexione Domini Nostri Jhesu Christi*, ed. Edwin Norris, in *The Ancient Cornish Drama* 2, 2ff, Oxford, 1859.
RIG	*Recueil des inscriptions gauloises (XLVe supplément à "GALLIA")*, ed. Paul-Marie Duval et al. 4 volumes, Paris: CNRS, 1985-2002.
RM	*The Text of the Mabinogion...from the Red Book of Hergest*, ed. John Rhys and J. Gwenogvryn Evans, Oxford, 1887.
SG	*Y Seint Greal, Selections from the Hengwrt MSS.*, volume 1, ed. Robert Williams, London, 1876.
SM	'The Surexit Memorandum', ed. John Morris-Jones, *Y Cymmrodor*, 28 (1918), 268-79; 'The Welsh Marginalia in the Lichfield Gospels Part II: The "Surexit" Memorandum', ed. Dafydd Jenkins and Morfydd E. Owen, in *Cambridge Medieval Celtic Studies*, 7 (1984), 91-120.
T	*Gwaith Talhaiarn*, ed. John Jones, Llanrwst, 1869.
Trip	*The Tripartite life of Patrick, with other documents relating to that saint*, ed. Whitley Stokes, MRIA, 1887.
WM	*Llyfr Gwyn Rhydderch*, ed. J. Gwenogvryn Evans with introduction by R.M. Jones, Caerdyd: Gwasg Prifysgol Cymru, 1973; new edn. of *The White Book Mabinogion*, ed. J. Gwenogvryn Evans, Pwllheli: J. Gwenogvryn Evans, 1907.
Wz	*Thesaurus palaeohibernicus: a collection of Old-Irish glosses, scholia, prose, and verse*, volume 1: *Biblical glosses and scholia*, Cambridge: Cambridge University Press, 1901.
YCM	*Ystorya de Carolo Magno*, ed. Stephen J. Williams, Caerdydd: Gwasg Prifysgol Cymru, 1930.
YBH	*Ystorya Bown de Hamtwn*, ed. Morgan Watkins, Caerdydd: Gwasg Prifysgol Cymru, 1958.
YSG	*Ystoryaeu Seint Greal*, ed. Thomas Jones, Caerdydd: Gwasg Prifysgol Cymru, 1992.
YMTh	*Ymddiddan Myrddin a Thaliesin*, A.O.H. Jarman, Caerdydd: Gwasg Prifysgol Cymru, 1951.
YT	Elis Gruffydd, *Ystoria Taliesin: The Story of Taliesin*, ed. Patrick K. Ford, Cardiff: University of Wales Press, 1991.

Grammatical Glosses

1S etc.	first-person singular inflection/clitic
1P etc	first-person plural inflection/clitic
4	impersonal inflection
REL	relative form of the verb
F	Feminine
FUT	future-tense verb form
IMPERS	impersonal verb form
IMPF	imperfect-tense verb form
IPV	imperative verb form
INF	infinitive
M	Masculine
NEG	negative marker
PERF	perfect verb form/perfect aspect marker
PLUPERF	pluperfect verb form
PRED	predicative marker
PRES	present-tense verb form
PROG	progressive aspect marker
PRT	preverbal particle
REDUP	reduplicated pronoun
SUBJ	subjunctive form of the verb
VN	verbal noun
*	(in syntax) ungrammatical form; (in historical phonology) reconstructed form
#	pragmatically infelicitous form

Verbs unglossed for tense are present tense.

Other Abbreviations

ACC	Accusative
AgrSP	Agreement Subject Phrase
A*y*VSO	Adjunct-*y*-Verb-Subject-Object
AdvP	Adverb Phrase
AdvVSO	Adverb-Verb-Subject-Object
AM	Aspirate Mutation (of initial consonants)
AspP	Aspect Phrase
AuxSVO	Auxiliary-Subject-Verb-Object
C	Copula
CG	Common Ground
ContrP	Contrast Phrase
CP	Complementiser Phrase
df	degree of freedom (in statistical tests)
DP	Determiner Phrase

EF	Edge Feature
F(-score)	Harmonic mean of precision and recall
FamP	Familiar Topic Phrase
FocP	Focus Phrase
ForceP	Force Phrase
FinP	Finite Phrase
G1,G2	Grammar 1 (in historical syntax G2 can replace G1)
GEN	Genitive
GroundP	Ground Phrase
HT	Hanging Topic
ID	Identity
IND	Indicative
IS	Information Structure
LD	Left-dislocated Topic
LF	Logical Form
MBT(g)	Memory-Based Tagger (generator)
ModW	Modern Welsh
MP	Minimalist Program
MW	Middle Welsh
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NM	Nasal Mutation (of initial consonants)
NP	Noun Phrase
NT	New Testament
Op	Operator
OSWB	Old South-West British
OT	Old Testament
OVS	Object-Verb-Subject order
OW	Old Welsh
P	Predicate
PF	Phonetic Form
PoS	Part-of-Speech
PP	Prepositional Phrase
PredP	Predicate Phrase
ShiftP	Shift Topic Phrase
SM	Soft Mutation (of initial consonant)
Spec	Specifier
SVO	Subject-Verb-Object order
TP	Tense Phrase
t	trace
UG	Universal Grammar
v̬P, VP	little *v*, Verb Phrase
VSO	Verb-Subject-Object order
V1,2,3,4	Verb-first, -second, -third, -forth word order
XP	Any phrasal constituent

CHAPTER 1

Introduction

> *"From the words which are called parts of speech, is a sentence formed. There are two kinds of sentences; a perfect sentence, and an imperfect sentence. That is a perfect sentence, in which a noun and a verb are placed properly together."*
> (Williams ab Ithel, 1856:174)

1.1 The Middle Welsh word order puzzle

Middle Welsh word order has been a "vexed" problem for a very long time (cf. MacCana (1973)). It was obvious to nineteenth-century Welsh grammarians that finite verbs preceded their subjects in most forms of their language, but this was clearly not what was preached at Sunday Schools. In the Welsh Bible translations, dating from the late Middle Welsh period, subjects and even other constituents such as objects or adjuncts could appear before the finite verb. To many people in Wales it was utterly embarrassing to hear "Jesus and Job speaking 'bad Welsh' " (D. S. Evans, 1990).

This 'bad' impression led to the introduction of the term 'Abnormal Order'. In this prevalent 'Abnormal' word order in the Middle Welsh period (until the 16-17th century) the verb occupied the second position in the sentence, following its subject, direct object or even adjuncts. It was 'abnormal' from a Modern Welsh preferred VSO point of view. This puzzling change in word order had, however, not received much attention from scholars before the 19th century. Syntax had never really been the focus of research of historical linguists. In addition to that, Welsh had always been 'the Cinderella of the Celtic languages' (D. S. Evans, 1990), mainly because

the corpus of available Old and Middle Welsh texts and manuscripts is considerably smaller than, for example, that of Old and Middle Irish.

W.O. Pughe and (from 1900 onwards) the Oxford Welsh reformers led by J. Morris-Jones and O.M. Edwards put this problem of the 'Abnormal Sentence' on top of the Welsh research agenda. Discussions on 'the real Welsh language' (the literary or the spoken varieties) were mixed with a general aversion to any possible influence from the English (SVO) language. Henry Lewis's lecture to the British Academy in 1942 about 'The Sentence in Welsh' aimed to solve the same issue. Shortly after the appearance of scholarly editions and translations of the most important Middle Welsh texts, dozens of papers on word order were published, most notably Proinsias MacCana's (1973) analysis of the Abnormal Sentence. In the 1991 collection of papers on Brythonic Word Order, Fife & King describe the then current state of research as follows: "If the question of abnormal order was 'vexed' at the time of MacCana's article, by now it is positively tormented." (Fife & King, 1991:81). Much progress has been made since then, but nonetheless, even today there still seems to be some kind of syntactic variation in Middle Welsh that "frustratingly defies easy explanation" (Poppe, 2014:73).

The present study aims to shed more light on this intricate syntactic variation in Middle Welsh and the origin of the Abnormal Sentence by combining new insights from different subfields of linguistics. First of all, recent developments in computational and corpus linguistics are employed to create a consistently annotated database of the most important Middle Welsh texts. The very detailed part-of-speech annotation and the shallow syntactic parse not only provide solid information of the exact type of variation, but they also allow us to determine which possible syntactic, pragmatic and/or extra-linguistic features can influence word order. In addition to this, a clear and consistent methodology for the annotation of information-structural factors proves to be indispensable for a comprehensive analysis of Middle Welsh. Finally, the most recent developments and tools in the field of (generative) diachronic syntax as well as syntactic reconstruction are employed to answer the questions on how the Abnormal Sentence could have developed in Brythonic, why it developed the way it did in Middle Welsh and how and why it disappeared again in Early Modern Welsh.

1.2 Introduction to Welsh

Welsh is a Brythonic language most closely related to Breton and Cornish. It belongs to the Insular-Celtic branch of the Indo-European language family. The other branch of Insular-Celtic languages, the Goidelic branch, consists of Irish, Manx and Scots Gaelic. Continental Celtic languages like Celtiberian and the limited inscriptions in Lepontic do not share specific Insular-Celtic innovations, most notably for this study, they do not exhibit verb-initial word order that has become prevalent in both Modern Welsh and Goidelic. The parent language of Welsh, Breton and Cornish is usually referred to as 'Common Brythonic' or, to indicate its reconstructed form 'Proto-British'. This was the language spoken across most of

Britain until the Anglo-Saxon invasions in the 6th century AD. According to Koch (1992), Schmidt (1990) and other proponents of the 'Gallo-Brittonic hypothesis', Common Brythonic and the continental Celtic language Gaulish share some linguistic characteristics that are not found in the Goidelic languages. Evidence for this mainly comes from shared sound changes like $*k^w > *p$ in Brythonic and Gaulish. From a morpho-phonological point of view, Common Brythonic shares with Goidelic the phenomenon described as initial consonant mutation (though exact morphophonological details differ in the two branches). In particular in the earlier manuscripts, however, the often inconsistent orthography did not reveal consonants that changed according to these complex rules (first purely phonetic, but later lexicalised and grammaticalised to occur in very specific contexts). The lack of overt reflection of consonant mutation in an already inconsistent orthography can lead to ambiguity in the case of pronominal elements and a wide range of grammatical particles that were rendered monosyllabic (and often consisting of one single letter) after the loss of final syllables. For the sake of clarity and convenience, I only explicitly mark mutation triggers in the present study if it is relevant for the present argument. Forms that superficially look ambiguous like the masculine and feminine possessive pronouns *e* triggering soft and aspirate mutation respectively, are simply disambiguated by providing detailed glosses '3MS' (third-person masculine singular) or '3FS'.

1.2.1 Attestations and descriptions

The first attestations of Welsh are glosses and some poems written in the margins of Latin manuscripts dated around 800 AD. The period from the loss of final syllables through apocope around 550 AD until then is referred to as 'Early Welsh'. There are some further glosses in a Brythonic dialect called Old South-West British (OSWB), the predecessor of Middle Breton and Middle Cornish. The amount of prose of the Old Welsh period, from 800-1150 AD, is extremely limited. From the 12th century onwards, historical writings and narrative literature - both translated and native tales - were written down in various manuscripts. The earliest text I used for the present corpus study is a law text. The early Welsh laws are found in a variety of manuscripts copied (in different versions) throughout the Middle Welsh period, but the legal nature of these texts suggests at least certain passages preserve older stages of the language as well.

The *White Book of Rhydderch* and the *Red Book of Hergest*, both dating from the 14th century, contain the most famous collection of Middle Welsh native literature: the *Mabinogion*. All extant tales of the *Mabinogion* (11 in total) are used here to represent the narrative prose of the Middle Welsh period of the language, from c. 1150-1500 AD. In the Early Modern Welsh period, between 1500 and 1600, we find some chronicles and translations from Latin and other European languages, including the first Bible translation and the chronicle of St David. The first full translation of the Bible in 1588 contributed to the standardisation of the written literary language.

The majority of Welsh literature in the following centuries was religious in

nature, although some early grammars appeared as well (by William Salesbury in 1550 and Siôn Dafydd Rhys in 1592). From 1600 onwards, the language enters the stage that is called Modern (literary) Welsh. This literary register in present-day Wales differs significantly from the spoken dialects. The proportion of Welsh speakers in the population declined rapidly in the nineteenth century with the large-scale immigration of Irish and English industrial workers, mainly to South Wales (cf. Borsley, Tallerman, and Willis (2007:3)). The Welsh Language Act of 1967 guaranteed the right to use Welsh and further acts led to a growth in Welsh-medium education on primary, secondary and university level. Welsh is nowadays spoken by around 25% of the population in Wales, but there are also small communities of Welsh speakers in other parts of the UK (mainly London) and even in Patagonia (the result of a small colony of Welsh settlers there).

The language of the medieval period is described and analysed in detail by, among others, D. Simon Evans (*A grammar of Middle Welsh*, Evans (1964)). The Middle Welsh lexicon consists of items that can, on the basis of comparative evidence from other Brythonic and also Goidelic languages be reconstructed for Common Celtic. From a very early age, however, Latin loan words are incorporated into the language. First a typical influx of trade vocabulary, but at a later stage when most of Britain was Romanised various other loan words appear as well. From a phonological point of view, Brythonic is characterised as a 'P-Celtic' language referring to the above-mentioned sound change $*k^w > *p$ as opposed to 'Q-Celtic' languages like Irish, in which this phonological innovation did not take place (cf. Irish *mac* vs. British *mab* 'son, boy').

Case morphology was lost already in Middle Welsh (although some archaic remnants remained). Verbal morphology is synthetic. With multiple tenses and moods (Future, Past, (Plu)perfect, Imperfect, Present Indicative, Subjunctive and Conditional) and seven different person-number suffixes each, written Welsh has a "rich Romance-like" morphological inflection (cf. Roberts (2010)). Furthermore, in Welsh, just as in Irish or Breton, prepositions can also be inflected for person, number and gender.

Syntactic characteristics of Welsh include a strong head-initial preference in all phrase types. Verbs, nouns, adjectives and prepositions all precede their complements. Adjuncts typically follow the head they modify, although some variation occurs in particular in the verbal domain. Adjectives mainly follow their nouns, but just as in, for example, French, Welsh has a specific set of adjectives that can appear before the noun they modify. The unmarked word order in Modern Welsh is VSO (or AuxSVO, see Chapter 7). Middle Welsh, on the other hand, as explained in the introduction above, exhibits a verb-second word order preference that was, according to Willis (1998) an integral part of the grammar of the spoken language as well (and thus not merely a literary phenomenon as argued by, among others, MacCana (1991) and Fife and King (1991)).

The 'basic word order' of Old Welsh has been subject of much debate amongst Welsh traditional grammarians. In the scarce material available, many sentences show verb-initial word order, but sentences with V2 or V3 orders are found as well.

The central problem I address in the present study is the status of the V2 orders in Middle Welsh (in particular from the point of view of interaction between syntax and information structure) as well as the origin of the V2 orders in the history of the Brythonic languages.

1.2.2 The Middle Welsh corpus: texts and manuscripts

Almost all material used in the present study is drawn from an annotated corpus of Middle Welsh (> 9,000 positive declarative main clauses) especially created for this purpose. The texts chosen for this first annotated historical Welsh corpus include the most important Middle Welsh narrative tales (the *Mabinogion*), excerpts of the Early Welsh Laws, the late Middle Welsh chronicle *Buched Dewi* 'The Life of St David' and various narrative tales from the first full Welsh Bible translation (d. 1588).

The Middle Welsh *Mabinogion* is a collection of tales and bits of traditional lore. Continuous narrative passages are interspersed with dialogues set in Wales and Ireland and presented as (pseudo-)history with some magical interventions. These tales (of unknown authorship) were part of an oral literary tradition and were only put down in writing centuries later.

The tales of the *Mabinogion* can be divided into several subsections. The first four tales are also known as the *Pedeir Keinc* 'Four Branches'. These include the narratives concerning four leading characters: Pwyll, Branwen, Manawydan and Math. Then there are the three Arthurian Romances about Peredur, Owain and Gereint. Arthurian literature of this kind featuring the same protagonists is found in other European languages as well, e.g. Chrétien de Troyes's French versions. These might have influenced the Welsh tales, but they are not direct translations. These Romances are found together in the *White* and (slightly later) *Red Book* manuscripts with three further native tales: *Culhwch and Olwen*, *Breudwyt Macsen* 'The dream of Macsen' and *Breudwyt Rhonabwy* 'The dream of Rhonabwy'. Finally, one tale of the *Mabinogion* collection I added to the corpus appears in two different manuscripts that contain very different genres: the tale of *Llud and Llefelys*. By adding both of these to the corpus, the literary and historical manuscripts can be compared systematically.

For this initial annotated corpus, only the (older) *White Book* (c. 1350) version was used. Syntactic variants have, however, been checked against the later *Red Book* (c. 1385) version of the tales as becomes clear from various examples in the present study. High-definition photographs of both of these manuscripts are available online via the websites of the National Library of Wales (`www.llgc.org.uk` - *White Book* Peniarth 4-5) and Jesus College Oxford (`www.image.ox.ac.uk` - *Red Book* Jesus College 111). The *White Book* manuscript, *Llyfr Gwyn Rhydderch* (Peniarth MSS 4 and 5), is one of the most important Welsh manuscripts (cf. Gwenogvryn Evans (1898-1910) and Huws (1991)). According to Daniel Huws, Keeper of the Manuscripts at the National Library of Wales, it was a coherent manuscript, written by five different scribes for Rhydderch ab Ieuan Llwyd of Parcrhydderch in Strata Florida Abbey (Ceredigion, Mid-Wales). The tales of the

Mabinogion are all written by scribes D and E in the last part of the book (quires 15-21 and 23-26) (cf. Huws (1991)). The rest of the *White Book* contains translations or retellings of mainly French (religious) tales, like *Can Rolant* 'Song of Roland' and *Purdan Padrig* 'Patrick's Purgatory'.

Although most tales of the *Mabinogion* were not written down until the 14th century,[1] the texts were undoubtedly of earlier origin. How early exactly is still a matter of much debate among Welsh scholars. The remark by S. Davies (1998) cited again by Rodway (2013:1) inadvertently describes this wide range like this: "it is probably safe to assume that they [the *Mabinogion tales*] were written down some time between the end of the eleventh and the beginning of the fourteenth centuries" (S. Davies, 1998:134).

The excerpts of the Early Welsh laws are from the BL Add. 22356 (S) manuscript, one of the most important manuscripts in the tradition of the Welsh Laws of Hywel. It is dated from the mid-15th century, but the texts go back centuries. The latest edition is accessible online via `www.cyfraith-hywel.org.uk`. The content of the excerpts used for the present corpus study focusses on the laws of the country and women. The rights and duties of women both married and unmarried are discussed in detail and as in all law texts, penalties and compensation fees for any possible crime are described related to the victim's *wynebwerth* lit. 'face-value'.

This particular genre differs from the narrative tales in style. The range of vocabulary is limited to specific legal terms and there are many enumerations and repetitions of particular verbs. The section on divorce, for example, contains a list of items each of the partner receives after the marriage is ended, e.g. 'The wife gets the salted meat; the husband gets the unsalted meat. The wife gets the pots and pans; the husband gets the knives.' To present a more balanced view of the law texts, excerpts from various parts of the laws were chosen to avoid a long list of one particular word order type of that formulaic nature.

Buched Dewi or 'The Live of St David' is one of many versions of a description of the saint's life found in the late fourteenth-century *Red Book of Talgarth* (NLW Llanstephan 27, 62v-71v). It is written in the hand of Hywel Fychan, who also wrote parts of the *Red Book of Hergest* for Hopcyn ap Thomas in the late 14th century. *Buched Dewi* belongs to the genre of historical writing consisting of a mix of chronicle and narrative styles. St. David was a Welsh bishop of Menevia during the 6th century AD. As with most 'biographies' of saints' lives in those days, many details like the exact date of his birth remain uncertain and stories of 'historical events' are often presented as a series of miracles.

The excerpts taken from the 1588 Bible translation are narrative passages from both the Old and the New Testament. They include Joseph's and David's tales (Genesis 37-45 and 1 Samuel 16-18), fragments of the gospels (Matthew) and Paul's letters to the Corinthians. The style of Paul's letters differs somewhat from the narrative prose found in the other excerpts: sentences are longer and the content is more dramatic with the intention of converting the audience to Christianity. The

[1] There are some fragments of individual texts found in earlier manuscripts, further written evidence has not survived.

texts were translated directly from the Hebrew and Greek originals. No significant difference between the Old and New Testament have so far been noted specifically due to translation from each of these languages, but a thorough comparative investigation of this kind is still a desideratum.

1.3 Methodology & working framework

Investigating word order variation in historical sources poses significant challenges. Some of those are inherent to historical linguistic research in general, such as the limited availability of data and the gaps in knowledge about a text's philological background (see also Poppe's remark on Tuija Virtanen's methodological reminders, cf. Poppe (2014:72)). In addition to those, there are some specific challenges looking at variation in historical data, in this particular case word order variation. Poppe (2014) furthermore reminds us that looking for reflexes of textual and pragmatic considerations on word order patterns based on the hypothesis that such reflexes exist "may in the end find what it looks for, and support its own initial hypothesis" (Poppe, 2014:94). When investigating historical pragmatic factors in particular, we thus have to be very careful not to end up with such circular argumentation.

Before we can say anything about when, how and why Welsh word order changed before and after the Middle Welsh period, we need an excellent understanding and thus comprehensive synchronic description of Middle Welsh. If we want to make any adequate generalisations about the syntax of this stage of the language, we need a large amount of consistently analysed data. A historical corpus, with part-of-Speech as well as phrase- and information-structural annotation can provide exactly what we are looking for. Since no such annotated corpus was available for Middle Welsh, I conducted pilot studies on individual texts of different historical periods, evaluated the results and subsequently extended the number of texts to produce a corpus that included the most important Middle Welsh literature. Building on recent studies in the field, I furthermore developed the methodological tools necessary for annotating and analysing Information Structure. Combined with the detailed morpho-syntactic annotation, this allows us to study all possible factors that can influence superficial word order patterns in a systematic way. The synchronic and diachronic results concerning syntactic changes were finally analysed within the framework of generative grammar.

1.3.1 Building an annotated corpus

One of the great challenges for anyone working with historical linguistic data is the fact that we are limited to work with 'what we have'. There are no native speakers of the Medieval period who can tell us what the language sounded like or whether a particular construction is at all possible. The linguist is solely confined to the corpora at hand. And more often than not, these are not 'at hand' at all. When it comes to Welsh manuscripts in particular, they are conserved in the main libraries

in England and Wales. There are digital photographs of the manuscripts available on the website of Jesus College Library in Oxford and the National Library of Wales (`http://image.ox.ac.uk` and `www.llgc.ac.uk`), but not all of those have been converted to searchable (online) corpora yet.

The only way to do historical linguistic research is by relying on the distribution of the different forms and constructions that are attested in the corpora. When analysing larger corpora, linguists need to be extremely consistent in their approach. Doing all this manually would take an enormous amount of time. Furthermore, especially when investigations last longer, they are prone to error. Therefore, it is useful to employ methods from the field of Natural Language Processing (NLP) and the tools created by Computational Linguists. Because of their computational nature, these tools are designed to consistently deal with large amounts of data in a very short period of time. The results are objective and can then be made readily available for any (Welsh) linguist.

Having said this, however, as a highly inflected language without standardised orthography, Middle Welsh poses some specific challenges for detailed morpho-syntactic tagging. One way to overcome these is by using specific NLP tools like memory-based part-of-speech taggers. The Memory-based tagger (MBT) designed by Daelemans, Zavrel, Van den Bosch, and Van der Sloot (2010) in particular yielded good results in terms of automatically assigning morpho-syntactic tags to this challenging dataset. For this study the words were automatically tagged on the basis of their specific characteristics and the context in which they occur. To facilitate more detailed linguistic queries for languages with rich inflection, the UPenn tagset, originally designed to annotate the English historical corpora (see, among others, Kroch (2000)) was systematically extended to include person, number and gender inflection for verbs and prepositions as well as additional tags for pronouns, adjectives and functional particles. The PoS-tagged texts in the corpus were then manually corrected. These so-called gold standards were subsequently used to add phrase-structure annotation as well.

The Natural Language Toolkit (NLTK) provides a rule-based chunk or shallow parser that can combine tagged words into larger constituents. I designed a rule-based phrase-structure grammar for Middle Welsh that automatically created the basic phrase types such as noun phrases, determiner phrases, prepositional phrases and verb phrases. With a python script that let the parser run through the data multiple times, hierarchical structures (NP in DP in PP, for example) could be created. Finally, the results of this automated shallow parse were manually corrected again and subordinate clause structure was added as well. This combination of morpho-syntactic and phrase-structure annotation was then converted to XML format to make various types of syntactic and information structural queries possible (see also Meelen and Beekhuizen (2013) for technical details of the evaluation and application of this). This is a first step in the process of creating a full historical treebank for Welsh, like the ones created for historical corpora in English (Kroch, 2000) and, for example, Old Icelandic (Wallenberg, Ingason, Sigurdsson, & Rögnvaldsson, 2011). In Chapter 2, I discuss the necessity and processes involved

in this type of corpus linguistics in more detail.

1.3.2 Factors determining word order

If we want to find out if information-structural factors played a role in word order variation in Middle Welsh, we first need to establish a base line and ask ourselves which factors have the potential to influence the observed word order patterns in the first place.[2] Broadly speaking the type of factors we can imagine can be divided into language-internal and language-external factors. Internal factors include any linguistic domain, such as phonology, morphology and core grammatical or syntactic features such as tense/aspect/mood, transitivity, diathesis, etc. The exact place of Information Structure in the grammar of language is still a matter of some debate (see Introduction to Chapter 3), but the fact that it includes the information status of constituents and how this relates to the rest of the sentence and the preceding and following context is well-established. Since languages differ in how they treat information-structural notions such as focus, topic or givenness (e.g. via special prosody or word order), this may also be seen as a language-internal factor.

Factors external to language in a historical context include, for example, philological tradition and textual transmission. The text we find in manuscripts today can be the result of multiple copying by scribes we do not know, in a place we have no (linguistically relevant) information about. The date of origin as well as the author are often obscure, which significantly hampers detailed diachronic studies of the language. A further general limitation of (historical) corpus data is that we often cannot be sure to what extent the written corpus text represents any given stage of the spoken language as well. This finally leads us to some usage-based considerations.

Usage-based factors lie somewhere in between purely internal and external factors that could possibly have a linguistic effect (in this case, determining the word order). These include anything related to how language is used and why in this particular way and/or context. Examples are different genres and text styles that belong to specific genres. The syntax of narrative prose, for example, often differs from that of elevated poetry. Other socio-linguistic factors such as register can play a role as well. Stylistic factors within texts (such as differences in passages with direct or indirect speech) can also result in variation.

When comparing different texts from different stages of the language, we should always bear all these factors in mind. Ideally we create a perfectly balanced corpus with extensive metadata about the philological background of both the manuscript and textual tradition. In practice, however, at least for Middle Welsh,

[2]Note that 'factors influencing superficial word order patterns' is meant to be a broad notion covering direct and indirect ways of influence. Strictly speaking there could be various forms (registers/dialects/genres) of Middle Welsh that each have a different grammar and thus a different range of possible word order patterns. External factors in particular are likely to influence the choice of a specific form of Middle Welsh, which, in turn, exhibits a particular grammar with certain word order patterns. In this way they 'influence word order' indirectly. I do not mean that external factors interact directly with syntactic features of the grammar resulting in different possible word order patterns.

much information about the exact date and place of origin is beyond our reach. For the present study I nonetheless aim to keep all language-external and usage-based variables constant, e.g. by only taking into account narrative prose. As for the language-internal factors, I systematically examined the role and distribution of the most important morpho-syntactic features over the different word order types found in Middle Welsh. Consistently controlling for each of these variables then allows us to establish the actual influence of the information-structural factors like topic, focus or givenness we are interested in for the present study. Chapters 4 and 5 extensively discuss these factors and their interaction with the wide range of possible word order patterns in Middle Welsh.

1.3.3 Syntactic analysis

Syntax is more than just word order. Words are combined to form constituents and these constituents in turn can again be combined to form even larger constituents. These groups of constituents are called phrases and indicated by the first letter(s) of their categorial heads: noun phrases are NPs, verb phrases are VPs, etc. Linear order of the kind XP preceding YP (regardless of any intervening material) is not relevant to the interpretation of a sentence like (1) (an old example by Chomsky, discussed again in Chomsky (2013:39)):

(1) Can eagles that fly swim?

When questioning an ability of eagles with *can*, native speakers of English (or those who are sufficiently fluent in the language) know that we are not questioning their flying skills, even though the verb *fly* is linearly closer to the questioning modal auxiliary *can*. Similarly, in example (2b) below, the subject *a large friendly gorilla* is linearly even closer to the gerund *moving* that it relates to than its equivalent in (2a). This linear adjacency, however, is equally insufficient to explain why it is perfectly possible to say (2a) in English, but not (2b) (examples from W. D. Davies and Dubinsky (2004:98)):

(2) a. Near the fountain, a large friendly gorilla sat without moving.
b. *Near the fountain (there) sat a large friendly gorilla without moving.

Even if the linear word order and each individual lexical item in a clause is the same, the meaning can be different. Chomsky (1986) gives the following example in which the pronoun *them* in sentence (3a) cannot have the same reference as *them* in (3b) (coreferentiality is indicated by the subscript index):

(3) a. I wonder who [the men$_i$ expected to see them$_i$].
b. The men$_i$ expected to see them$_j$.

In addition to these puzzling contrasts with similar word order patterns, some examples show there must be more (words, elements) than we see. There is nothing in the word order, phonology or morphology that explains why the examples with contraction are possible in (4), but not in (5).

(4) a. Who do you want to kiss? Who do you wanna kiss?
b. I'm going to go. I'm gonna go.

(5) a. Who do you want to kiss the puppy? *Who do you wanna kiss the puppy?
b. Who do you want to win? *Who do you wanna win?

The grammatical function of a core argument, e.g. the subject or object of a clause, is also important. Children that are exposed to the variants with and without the complementiser *that* in example (6) can easily conclude that the complementiser is optional. Crucially, however, they know that in very similar sentences as in (7), the second option with *that* is impossible.

(6) a. Who do you think that Peredur will kiss first?
b. Who do you think Peredur will kiss first?

(7) a. Who do you think will kiss Rhiannon first?
b.*Who do you think that will kiss Rhiannon first?

Each of the examples above shows in one way or the other that we need more than just the surface linear order of words we see or hear. These puzzling facts led to the crucial insight that language has *hierarchical structure*: there is more than the 'superficial' order of words in the sentence. Within the framework of Generative Grammar, this idea of syntactic structure is inherently linked to a further puzzle referred to as *The Poverty of Stimulus* or *Plato's Problem*. Plato's Problem is the phenomenon Noam Chomsky (mainly in Chomsky (1986)) referred to in an attempt to explain the origin of knowledge. He made reference to the Socratic dialogue *The Meno*, in particular the passage in which a boy is able to understand some mathematical concepts of the Pythagorean theorem without prior instruction. Socrates explains this is possible because of his *a priori* knowledge that has been "aroused through questioning" (86a).

In the context of language and grammar or syntax in particular, the question is on the one hand how children are able to understand and produce sentences they have never heard before. On the other hand, the input children get is not only limited but also filled with 'noise'. Utterances in speech are often incomplete or contain false starts (speech/performance errors). Children might even be exposed to two or more languages (or dialects and registers) at the same time. In other words, the spoken language around them (the Primary Linguistic Data or PLD) is neither a complete nor a perfect reflection of the grammar they nonetheless learn almost perfectly in such a short period of time. How is that possible on the basis of such limited evidence? Chomsky (backed later by acquisition studies by J. A. Fodor (1966) and others) answered this question along the same lines as Socrates: some essential 'knowledge' about the grammar of language must already have been present. After some exposure to a particular language (the 'input experience'), this knowledge about the grammar "is aroused" to become practical knowledge about the language the child can start to apply. This intrinsic capacity in human beings to learn language is often referred to as 'Universal Grammar' (UG). Within the framework of Generative Grammar, Plato's Problem is thus solved by a specific

architecture for the human linguistic cognitive capacity, a learning bias that restricts or structures the child's range of choices so that convergent learning is possible. One of the main goals of the generative enterprise has been to identify these biases, or, in other words, understand and define these UG principles through the study of individual languages and language variation. Since Universal Grammar and/or an 'innate language faculty' has received much criticism from opponents of Generative Grammar, let us pause a moment to address some of these core issues.

First of all, despite the name, Universal Grammar (UG) has nothing to do with Greenberg's typological language Universals. The assumption is not that all languages are 'underlyingly the same'. UG does not imply universal patterns or require rules that manifest in every single language. The 'universality' refers to the types of possible Grammars, i.e. the *kinds* of rules and principles they have. The assumption is thus that there is one set of principles governing all human languages and that individual languages may vary from those principles, but - crucially - they only vary in constrained ways. Discussion of what principles exactly are postulated to be part of UG and how their function has changed over the years and is still ongoing. The research is cumulative: new insights are continuously built on previous work to develop and refine the theory.

Then there is the question of *how* UG helps children to become fluent in their mother-tongue in such a short period of time. Ambridge, Pine, and Lieven (2014) hold the most critical view claiming that UG principles can in fact not account for language acquisition at all, because of three main problems: linking, data coverage, and redundancy (innate representations do not help general learning mechanisms that are already known) (Ambridge et al., 2014:e54-e55). Let us briefly look at each of these in turn. The linking problem refers to the question of what mechanisms help the learner to link innate representation to the input language. Assuming a set of universal principles in the form of learning biases does not solve that problem, they argue. As Beekhuizen, Bod, and Verhagen (2014:e92) rightly point out, however, to solve this particular problem we need to be extremely explicit about the mechanisms (to the extent it is mechanistically testable) and furthermore, we need a proper way to evaluate how the system operates as a whole. Many generative studies on acquisition indeed focus on individual empirical cases, making it difficult to establish their effect on the overall acquisition process. This is, however, due to practical challenges in experimental research in first-language acquisition, not limited to researchers advocating generative grammar. Proponents of usage-based (or any other linguistic) approaches to acquisition have equally failed to meet both requirements and thus solve the 'linking problem' (Beekhuizen et al., 2014:e92-e94). A way forward would be to include computational models to properly test and evaluate proposed systems and mechanisms. Examples of this new direction are found in both usage-based (e.g. Beekhuizen (2015)) and generative approaches (e.g. Pearl (2014) or various studies by Charles Yang, e.g. C. D. Yang (2000) and C. D. Yang (2002)).

The second problem Ambridge et al. (2014) have with UG is that the innate representations that are proposed yield incorrect empirical predictions. This type

of criticism touches on recent more general claims that large-scale typological studies of descriptive grammars would yield better results than hypothesis-driven approaches. N. Evans and Levinson (2009) and Levinson and Evans (2010) in particular go out of their way to divide the field into 'C-linguists' ('Chomskyan linguists') and 'D-linguists' ('the rest', mainly characterised as 'Diversity-' and 'Data-driven')[3]. For the present thesis, the strict division based on opposite stances in central issues they formulate (Levinson & Evans, 2010:2734-2735) is irrelevant because these 'opposite stances' can actually come together on various levels. First of all, the use of a large amount of data available in a systematically annotated corpus and a statistic analysis thereof (issues 1, 4 and 5) are addressed in Chapters 2 and 5 of this thesis respectively. Secondly, the use of insights from related (sub)fields like pragmatic/functional and historical approaches to linguistics and psychology/neuroscience (issues 3 and 7) are discussed and incorporated in Chapters 3, 5, 6 and 7. Finally, the way the thesis is organised, starting from a proper description and analysis of the language on its own (Chapters 2-6 of this thesis) before moving on to cross-linguistic comparisons (the reconstruction part of Chapter 7) should 'solve' the second issue they mention.[4] Despite the fact that six out of seven issues Levinson and Evans raise are at least *also* addressed from a 'D-linguistic' perspective here, the present thesis is based on 'C-linguistic' assumptions. These 'data/diversity-driven' aspects in 'C-linguistic' research are not new or unique, as shown by numerous generative studies on languages far removed from English (cf. Legate (2002), M. Baker (2008), Preminger (2011) among many others) and all comparative work specifically focussed on language diversity within the 'Rethinking Comparative Syntax' project at Cambridge University (`www.recos.cam.ac.uk`). Levinson and Evans finally state that "[a] theory should be responsible for a wide range of predictions across data types, and it should be possible to disconfirm it with primary data." (Levinson & Evans, 2010:2736).

It could be argued that when working with historical data, it is impossible to make any falsifiable predictions and that therefore (going back to Ambridge et al.'s original point) hypothesis-driven approaches based on UG are not appropriate. Since we have no access to negative evidence, historical data are certainly more limited than studies of contemporary languages when it comes to defining the exact characteristics of individual languages and possible principles of UG. This is, however, exactly why generative studies of contemporary languages are so beneficial to the historical linguist. Not only do they provide us with a well-tested set of tools and methodology, they also systematically limit our hypothesis space. In other words, when we are trying to describe earlier stages of a language as accurately as possible, information about which types of grammars are possible or impossible is extremely valuable (see also recent studies on the significance of 'what hasn't happened' on changes that did not take place in historical syntax and

[3] For a comprehensive overview of recent literature on what N. Evans and Levinson (2009) call the 'Myth of language universals' see a series of responses cited and addressed again by Levinson and Evans (2010), most notably M. C. Baker (2009), Longobardi and Roberts (2010) and Harbour (2011)).

[4] The final issue they raise concerns models of culture-biology coevolution, which goes far beyond the present research on Middle Welsh word order.

why by Biberauer and Roberts (2015)). As Davis, Gillon, and Matthewson (2014) show with a wide range of examples from lesser-studied languages of a diverse background, hypothesis-driven research is very important in this domain as well, because for many of these languages statistical analysis of large-scale corpora is unavailable.

If predictions based on innate representations and learning principles of UG are not borne out by (new) empirical data, we need a better understanding of the old and new data, a reformulation of our generalisations and from there we can redefine our initial hypotheses. This type of theory-internal development does not imply we need to reject any kind of innate constraints on linguistic representation (UG). Many empirical findings in fact defy easy (or any) explanation without a UG component that is part of a successful learning strategy (cf. studies on parasitic gaps illustrated by Adger (2013a) or syntactic islands by Pearl (2014) and Schütze, Sprouse, and Caponigro (2015)).

This then touches on the final problem of UG Ambridge et al. (2014) raise: that UG principles are 'redundant' in that they have nothing to add to general learning strategies and cognitive capacities we are already familiar with. Schütze et al. (2015) show, however, that established cross-linguistic constraints on A-bar dependencies cannot be explained by independently motivated non-syntactic factors. In a further attempt to convince generativists that island constraints are not purely syntactic, Goldberg (2006) provides the following usage-based alternative: "It is pragmatically anomalous to treat an element as at once backgrounded and discourse prominent." (Goldberg, 2006:135). To the extent that this is a useful and concrete alternative tool to those employed by generative syntacticians working on island constraints, it actually makes the wrong empirical predictions. One key counter-example that is relevant for the present study on information structure shows that focus in backgrounded contexts is actually perfectly possible in a sentence like (8) (taken from Lidz and Williams (2009:184)):

(8) I certainly did not read the book that CHOMSKY recommended.

In Chapter 4 of this thesis I will outline a methodology of detecting the core notions of information structure, showing the exact same thing. 'Pragmatic anomaly' as a criterion can thus not make any useful predictions about grammar. In Chapter 7 I furthermore explain in detail that another usage-based concept of 'Motivation' as applied to Early Modern Welsh data faces the same problem.

Alternative syntactic frameworks like Construction Grammar (CxG), Lexical-Fuctional Grammar (LFG) and Head-driven Phrase-Structure Grammar (HPSG) mainly differ in that they do not employ silent lexical items (in particular traces or copies: they are non-transformational). Goldberg (2006) (working within a usage-based CxG approach) assumes this kind of 'surface-approach' facilitates processing. Lidz and Williams (2009:185) argue, however, that "[t]here are no decisive demonstrations that any of these assumptions necessarily simplify processing or learning". Another basic assumption of CxG is the direct association of meaning with structure, whereas generative grammar associates meaning with

lexical items. Essentially, this is an issue of compositionality: can meaning always be derived from the meanings associated with the components of those structures or not? According to Adger (2013a), the functional heads that project structure as assumed in a Minimalist framework (e.g. Tense, Topic, Complementisers) solve this potential problem: abstract structure with a particular grammatical form is thus associated with meaning. These abstract functional categories then are not different in this respect from the constructions proposed in CxG. Within generative grammar, cartographic approaches (e.g Cinque (1999)) assume that there is an elaborate hierarchy of functional categories that is always present (and thus part of UG). But most recent Minimalist studies within the generative framework prefer to postulate a particular functional category *only* if a language shows evidence for it. The newly developing 'emergentist approach' to syntactic variation (cf. Wiltschko (2014), Biberauer (2015) and Van der Wal (2015)) states that certain functional categories, e.g. Tense, are actually part of a broader notion 'anchoring an event in the world'. Only this latter notion is stipulated to be part of our language capacity, specific functional categories need not be. Along the same lines, as I point out in Chapters 6 and 7, I will start from the very basic assumption that there is only one generic projection in the left periphery of the clause (only a generic Complementiser Phrase, not necessarily divided into subcategories indicating specific kinds of Topics or Foci). Only when there is evidence for more structure, this is postulated (e.g. the added Force Phrase in Chapter 7 based on evidence from auxiliary-initial phrases in Middle Welsh).

To conclude, UG is rejected by proponents of CxG and others because innate processes of social cognition, categorisation and statistical learning are assumed to be sufficient for the child to learn her first language. If that is indeed the case, we need concrete evidence that a representational bias for learning grammar can in itself be statistically induced. In addition to that, these non-language-specific learning strategies would have to be able to account for the empirical data. So far, the above-mentioned studies on syntactic islands and parasitic gaps (to mention just two syntactic phenomena) do need more than purely probabilistic learning approaches. A final problem arises if we only adopt general cognitive learning strategies. As Adger (2013b) points out, this leaves the hypothesis space unconstrained in the sense that anything could have an effect on linguistic phenomena. This makes it even harder for linguists to explain any grammatical effects.

Adopting Generative Grammar as a working framework for the final part of this thesis (Chapters 6 and 7 concerning the syntactic analysis) thus has various advantages. A transformational theory with a UG component meets all three required levels of adequacy. Its tools and mechanisms help us ask the right questions leading to important **observations** (for example, in work on lesser-known languages as Davis et al. (2014) point out). The highly consistent way of finding generalisations in addition to the growing amount of comparative research within the generative framework furthermore provides adequate **descriptions** of phenomena in a wide range of languages. Specific language-learning biases or principles of UG on the one hand constrain the otherwise too large range of options, on the other, they

allow us to make predictions and thus **explain** the observations in a systematic way. An additional, very practical reason for adopting a Chomskyan approach for the syntactic analysis in the final chapters of this thesis is the wide range of literature on the linguistic phenomena we are interested in. The analysis on the interaction of information structure and syntax in Chapter 6 benefits greatly from generative studies on similar phenomena in other languages with V2 word order. In Chapter 7 I furthermore show that generative tools fare better than other approaches when it comes to explaining how exactly and why certain grammatical changes in the history of Welsh took place the way they did.

Syntactic assumptions for the present study

For the syntactic analysis of the present study, I therefore adopt the generative syntactic framework developed in the context of the Minimalist Program (cf. Chomsky (1995), Chomsky (2000) and later). I thus assume a transformational approach to grammar including a UG component that consists of (i) a cognitive capacity used to create recursive structures via the operation called Merge, and (ii) a capacity connecting these structures to both sounds and signs and systems that involve internal computations such as thinking, planning, etc. (cf. Adger (2013b)). The goal of the present study, however, is not to investigate the 'Strong Minimalist Thesis'. This idea by Chomsky (2000:96) stipulates that language is an optimal solution to legibility conditions. Although I adopt the rationale behind the Minimalist Program, the present study is not meant to contribute further evidence supporting that idea in any way. I merely use the results and tools of other Minimalist studies to achieve a better understanding of the research questions concerning Middle Welsh word order.

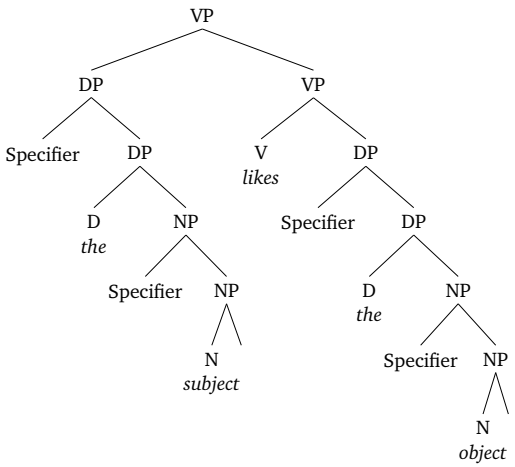
The two core operations of the Minimalist Program are Merge and Agree. Merge is the main structure-building operation that simply takes two syntactic objects α and β and forms a new object $\gamma = \{\alpha, \beta\}$ (Chomsky, 2001:3). The syntactic items can be drawn from the set of items in the Numeration (the set of lexical and functional items that will eventually make up the sentence), but they can also be drawn from parts of the structure that are already built (so-called 'internal Merge', which is in effect a refined statement of traditional cases of transformations or movement (Chomsky, 2005:12)). The operation Agree "establishes a relation ... between an LI [lexical item - MM] α and a feature F in some restricted search space (its *domain*)" (cf. Chomsky (2000:101)). Examples of features are familiar notions in the nominal domain such as person, number or gender features, that are combined under the umbrella-term φ-features, but also more abstract clause-type features such as Tense and Negation or information-structural notions like Topic or Focus.

Features can enter the derivation (the build-up of the structure of the sentence) in two ways: they are either interpretable or uninterpretable. Uninterpretable features cannot be interpreted by the conceptual-intentional ('logical form' (LF)) and sensorimotor domains ('phonetic form' (PF)) responsible for semantic interpretation and externalisation in the form of sound and/or signs respectively. If features

are uninterpretable, they must be checked by entering into an Agree relation with an equivalent *in*terpretable feature in the derivation. I assume this type of checking to be a process of valuation (Chomsky, 2001): an uninterpretable Tense feature (indicated as *u*Tense) can be checked by an interpretable Tense feature (*i*Tense) that for example has a specific value indicating future tense. I use the cross-out notation ~~u~~ to indicate such an Agree relation is established with the added value (if this is relevant), e.g. [~~u~~Tense:future]. Agree between an uninterpretable feature (the Probe) and an interpretable feature (the Goal) may trigger Internal Merge (or movement) of elements to the phrase of the Probe as well.

Lexical and functional items 'project' to form phrases that are labelled according to the heads (the specific item) rendering the simplified structure for the noun phrase 'the subject' as shown in (9). A noun 'N' projects a Noun Phrase (NP) that can be the complement of a determiner (e.g. a definite article) 'D', which in turn can project to form a DP. Only phrases can appear in Specifier (Spec) positions. I assume all parts of speech can project phrases in this way, e.g. adjectives 'A' render APs, verbs 'V' render VPs, etc. Apart from these lexical items, I assume a set of functional items, like Tense (T) and Aspect (Asp).[5] I follow the standard hierarchy of projections for the clause starting with the Complementiser Phrase (CP), followed by Tense (TP) and then the verb phrase (VP) and, if present, an aspectual phrase (AspP) in between TP and VP. I adopt the common assumption that the verb is first merged with its complement, the direct object and the subject is merged in the specifier position of the verb phrase. The first stage of the derivation of a sentence thus looks like (9):

(9)



One type of feature that is especially relevant in the present study is the so-called 'Edge Feature' on the C-head that triggers internal merge (movement) of a particular

[5] I furthermore assume the verbal domain has an additional functional layer indicated by 'little v' called *v*P, although arguments for this are not relevant in the present thesis and therefore not discussed in detail.

phrase to the Specifier of the CP resulting in the observed verb-second patterns in Middle Welsh. Any further syntactic assumptions related to information structure and diachronic changes are specified in the introductions to Chapters 6 and 7 respectively.

1.4 Overview of the thesis

The main aim of this thesis is to shed more light the Middle Welsh word order puzzle outlined in the introduction by taking synchronic and diachronic evidence from syntax and information structure into account. I therefore address two main questions:

1. How can we explain the distribution of the various word order patterns in Middle Welsh? (In other words: which factors determine the 'choice' of using subject-initial order, rather than object-, adjunct- or verb-initial?)

2. Where do the various verb-second orders (including those with and without subject-verb agreement) come from?

This complex puzzle requires a thorough investigation of the independent pieces representing various subfields of (Welsh) linguistics: corpus linguistics, Information Structure, Welsh word order studies, synchronic and diachronic syntax and syntactic reconstruction. All of these elements are organised in separate chapters in this thesis. Each of these chapters contain a detailed introduction to the subject matter and relevant literature so that no prior knowledge of these linguistic subfields is required. In this way, I aim to make the present study accessible to scholars of various fields with a particular interest in, for example, the creation of an annotated historical corpus, information structure in Middle Welsh or methods in diachronic syntax. This thesis thus makes contributions to each of the subfields, but as a whole, it also provides an overall methodology for approaching word order puzzles taking historical syntax and information structure into account.

In Chapter 2, I first of all describe the necessary steps in creating an annotated corpus of Middle Welsh and how and why this is useful for syntactic studies. Guidelines for detailed part-of-speech (PoS) tags are presented building on the tagsets used for the historical corpora of English and Icelandic. The corpus was then chunk-parsed to create basic phrase structure and furthermore enriched with information-structural annotation. In Chapter 3 I present a systematic way of analysing information-structural notions so that they can add useful information to the annotated corpus.

Chapters 4 and 5 focus on Middle Welsh word order. In Chapter 4 I first give a detailed description of all possible word order patterns found in the corpus. Chapter 5 then systematically analyses which language-internal and -external factors can influence this wide variety of word orders with particular emphasis on the role of information-structural notions such as Givenness, Topic and Focus.

In Chapter 6 I discuss the intricate interaction of information structure and word order from a synchronic perspective: how does information structure work in the syntax of Middle Welsh? How are topics or focalised elements encoded? Does the referential status of constituents play a role in the syntax? Furthermore, I present a formal syntactic analysis of the puzzling verb-second patterns with and without subject-verb agreement in Middle Welsh.

Chapter 7 finally turns to the question of the origin of these and other patterns. The focus lies on diachronic syntax and syntactic reconstruction on the basis of comparison of other languages closely related to Welsh like Breton and Cornish. On the basis of two Case Studies related to syntactic change and information structure, I provide a detailed overview of processes of grammaticalisation and reanalysis in the history of Welsh. I furthermore reflect again on the role of information structure in syntax, focussing on the diachronic aspects and what implications this might have for studies of diachronic change in general.

CHAPTER 2

Creating an annotated corpus of historical Welsh

"The corpus linguist says to the armchair linguist,
'Why should I think that what you tell me is true?',
And the armchair linguist says to the corpus linguist,
'Why should I think that what you tell me is interesting?' "

(Fillmore, 1992:35)

2.1 Introduction

Any scholar who ever took the challenge can confirm that creating a linguistically annotated corpus of historical texts is a daunting task. Nelson (2010) is certainly right to start his guide to compiling written corpora with the question: "Do I have to do this?". As the first part of the methodological considerations of this thesis (the second part of the methodology concerning Information Structure is described in Chapter 3), this chapter addresses Nelson's question by closely examining the *nature* of the evidence necessary to answer the research questions. A brief history of creating corpora, along with their advantages and disadvantages, will illustrate why I indeed 'had to do this' for the present study.

This chapter furthermore aims to give a detailed answer to all further questions this conclusion entails: How to compile a corpus? How to annotate the data? How to query that data? and, finally, How to analyse the results?

2.1.1 What is an annotated corpus?

Although the Latin *corpus* 'body' had already undergone a semantic shift to 'collection of facts or things' in the classical period, this meaning was not attested in English before Ephraim Chambers published his *Cyclopaedia* in the 18th century (Chambers, 1728). According to the *Oxford English Dictionary*, it was W.S. Allen who first used the term in his 1956 paper in the *Transactions of the Philological Society* as a 'body of written or spoken material upon which a linguistic analysis is based' (OED 2014 full online edition s.v. *corpus*).

Corpora may vary in size, composition and purpose, but corpus linguists agree that good corpora are never just random collection of texts (cf. among others Biber, Conrad, and Reppen (1998:246) and Meyer (2002:xi)). Most corpora are electronically available these days and contain metalinguistic data about the background and context of the texts. Depending on the specific purpose of the corpus, textual markup and further linguistic annotation can be added to facilitate various types of research (morphological, syntactic and, in case of spoken corpora, also phonetic, to mention just a few).

Irrespective of the *type* of corpora, the content always remains the output of *performance*. As such, an annotated corpus has therefore certain limitations: it cannot give direct evidence of speakers' language *competence* (see for discussion section 2.3.2 below). With the creation of in particular annotated digital corpora, however, an invaluable source was added to the linguistic toolbox.

2.1.2 Why create an annotated corpus?

On the necessity of *more* data...

This thesis is mainly concerned with word order change in Welsh. In any language there are many different factors determining the word order or 'surface structure'. The way the speaker or writer chooses to convey the information in a particular context, paragraph, genre or register can result in different word order patterns. The syntax of a language or dialect, however, limits the seemingly endless possibilities of putting words together to form a sentence. When investigating the different word order patterns in a single text from one particular time period, all these factors have to be taken into account. Even within the syntactic limits of a language variant, there are numerous ways to form novel sentences. It is thus very unlikely to find the right context for *every* possible word order pattern in one single text. And if a particular context is not attested in the one text under investigation, it is impossible to trace its history.

Research in the field of comparative and historical syntax, word order and information structure crucially differs from investigations in the related fields of historical phonology and morphology in two respects. The focus of scholars in these respective fields is different to begin with: the first are concerned with clauses or whole sentences in their context, the latter investigate individual sounds and phoneme and morpheme inventories. Even the large phoneme inventories of

Caucasian or Khoisan languages are very small compared to the endless possibilities combining words into clauses and sentences. This means that the chance of the particular phoneme under investigation occurring is very high, even in one single paragraph. A certain phoneme can furthermore only occur in a relatively limited number of 'contexts', i.e. phonological environments, exactly because of the limited number of phonemes in a language. These environments or conditions are crucial for the concept of *Ausnahmslosigkeit* ('exceptionlessness'[1]) of sound laws in the Comparative Method of historical phonology. For example, Proto-Indo-European (PIE) short *o in non-final open syllables *always* becomes a long *ā* in Indo-Iranian (Brugmann's law (first proposed in Brugmann (1876)), there should be no exceptions.[2].

Environments or conditions in which certain word order patterns occur and/or change are, however, not as easy to ascertain. Again, there are many factors potentially influencing the surface structure and not all of those superficial word order patterns have the same underlying syntactic structure. The Comparative Method propagated in the field of historical phonology cannot be applied to syntax in the exact same way (cf. Walkden (2009) and Willis (2011a) for a comprehensive overview of the Comparative Method and attempts to transfer it to the field of historical syntax). Especially when comparing clauses and word order patterns, a single text is hardly ever long enough to contain all the possible options. To be able to compare a particular word order pattern (with one particular information structure and underlying syntax) occurring in a specific context in the thirteenth century with a different word order pattern occurring in the same context in the sixteenth century, a large and well-designed corpus is needed.

On the efficacy of digitising the data...

More data mean more work. Not only the quantity, but also the *type* of work that is required is important here. It may take a person days, weeks or maybe a few months to conduct a study of the phonology or morphology of one single text, going through it word by word, carefully annotating all peculiarities and regularities. It may take a year, a decade or even a lifetime to do a thorough syntactic analysis of the necessary *collection* of texts in the same way. Human beings tend to have difficulties dealing with large volumes of data and are horribly inaccurate and inconsistent without going through it twice at the very least (cf. Kennedy (1998:5)).

[1]The importance of the distribution of phonemes in establishing systematic correspondences and sound changes was already noted by the main philologists of the 1870s: 'alle Wörter, in denen der Lautbewegung unterworfene Laut unter gleichen Verhältnissen erscheint, werden ohne Ausnahme von der Änderung ergriffen' (Osthoff & Brugmann, 1878:xiii).

[2]For an illustration of the importance of this principle of *Ausnahmslosigkeit*, consider the 'dramatic history' of Brugmann's Law (Lubotsky, 1997). There were, in fact, many apparent exceptions to this sound law especially before the discovery of laryngeals at the end of a thereby closed syllable (cf. H. Hirt (1913) for a list of 67 items and his famous 'Das Gesetz [i.e. Brugmann's Law - MM] ist tot' (H. Hirt, 1921:19)). Although famous scholars like De Saussure and Osthoff accepted it at first, the exceptions forced Brugmann to withdraw his Law (Lubotsky, 1997:55). Attempts to modify the conditions and thus rehabilitate it were later done by, among others, Kuryłowicz (1927) and Volkart (1994). Cf. Beekes (1995:138) and for a longer discussion Lubotsky (1990) and Jamison (1983).

Computers do exactly what their name suggests: they count routinely, rapidly and, unlike humans, tirelessly. A search through millions of words that would take a month by hand can be done by a computer in a matter of seconds, with fewer[3] mistakes (cf. Curzan (2008:1091) and Scott (2010:136)). Moreover, computers are better at multitasking and recognise novel patterns by considering multiple factors in large numbers of sentences and texts simultaneously (cf. Conrad (2010:234) and Hunston (2010:154)). Since a consistent analysis of all potential factors is exactly what we need in the study of word order change, the use of computers and a digitised corpus is indispensable.

2.1.3 Chapter overview

In this chapter, I first give a very brief overview of the history of creating corpora (section 2.2). Then I discuss the most important challenges and criticisms of corpus-based research (section 2.3), and in the next section the most important advantages of using an annotated corpus (Section 2.4). In sections 2.5 and 2.6, I will elaborate on the compilation of the historical Welsh corpus, focussing on the tools from Natural Language Processing I used and the specific linguistic annotation. Section 2.7 is concerned with the technical details of getting the data required to answer the research questions, including exact formulation of the queries to facilitate replicability and future research. Finally, in section 2.8, methodological issues concerning analysing and interpreting the data are discussed.

2.2 History of creating corpora

> Dr. Samuel Johnson (presenting his long-awaited dictionary to the prince):
> *'Here it is, sir: the very cornerstone of English scholarship.*
> *This book, sir, contains every word in our beloved language.'*
> Prince Regent George: *'Hmm.'*
> Edmund Blackadder: *'Every single one, sir?'*
> Johnson (confidently): *'Every single word!'*
> Edmund: *'Oh, well, in that case, sir, I hope you will not object if I also offer the Doctor my most enthusiastic contrafribularities...'*
> - dialogue from BBC's Blackadder III, Episode 2: Ink & Incapability

2.2.1 Early text-based linguistic traditions

Collections of texts have been important sources for linguists since the first structured analyses and descriptions of languages. Pāṇini based his grammar of Sanskrit

[3] Although routine computations *should* give the correct result all the time, there are some famous examples of computational mistakes, in particular rounding errors, with unfortunate results in areas ranging from rocket science (e.g. the very short flight of the first Ariane 5 cf. Lions (1996)) to German politics (e.g. the change of Parliament makeup after automatically counting the votes cf. Weber-Wulff (1992))

(ca. 4th century BC) on the language of the Vedic texts instead of describing 'Classical' Sanskrit, the language spoken around his time (Meyer, 2008:3). Similar grammatical descriptions appeared later in Europe, based on the Greek epics (e.g. by Dionysus Thrax and Aristonicus of Alexandria) or Latin literature (cf. grammars by Donatus and Priscian, respectively in the 4th and 6th centuries AD). At the back of an early grammar of the Welsh language (written in Latin), John Davies similarly gives a list of names of poets from whose works the given examples in his grammar were taken (J. Davies, 1621[1809]).

In the late 19th century, linguists like Otto Jespersen and Hermann Paul also preferred linguistic descriptions based on examples found in real texts (*Sprachdenkmäler*, 'language monuments', (Meyer, 2008:4)). This textual data supported evidence about the present-day dialects and the language history studied by the Neogrammarians (Lüdeling & Kytö, 2008:vi). Around the same time the first dialect maps and collections of dialect expressions were compiled systematically according to a well-defined set of criteria. These efforts can be seen as a precursor to the field of modern corpus linguistics (Lüdeling & Kytö, 2008:vii).

The tradition of systematically compiling corpora is firmly rooted in the work of concordances, indexers and lexicographers. Already in the Middle Ages there was a practical need for good biblical concordances. These concordances specified words in the Bible along with citations of important passages, starting with Anthony of Padua's twelfth-century *Concordantiae Morales* based on the fifth-century Vulgate and Cardinal Hugo's monumental word index compiled in 1230 with the help of 500 Dominican monks (Bromiley, 1997:757). Concordances of literary works, such as Chaucer or Shakespeare, followed later.

The aim of many modern corpus linguists to collect the maximum amount of data possible (in order to capture even the rarest forms of usage) stems from early lexicographers. Dr Samuel Johnson's dictionary, first published in 1755, contained 150,000 quotations,[4] the result of writing down samples of usage on slips of paper for ten years (O'Keeffe & McCarthy, 2010). The OED project turned into a massive three-million-slip corpus of attested words: "It was estimated that the project would be finished in approximately ten years. Five years down the road, when Murray [one of the first main editors - MM] and his colleagues had only reached as far as the word 'ant', they realized it was time to reconsider their schedule." (OUP, 2014). The final volume of the OED published in 1928 was the culmination of 71 years of work by many different editors and thousands of volunteer contributors (Kennedy, 1998:14).

2.2.2 The dawn of electronic linguistic corpora

When American structuralists in the early twentieth century put real language data at the core of linguistic study (Lüdeling & Kytö, 2008:viii) and the Prague School of

[4]Johnson planned to use examples from before the Restoration only (Meyer, 2008:7), because the English language after that period was (in his words from the preface to the first edition) "gradually departing from its original Teutonick character, and deviating towards a Gallick structure and phraseology..." (S. Johnson, 1755).

linguists started focussing on quantitative studies of frequencies (Krámsky, 1972), modern corpus linguistics was born. Teachers of English became more and more interested in using corpora to create textbooks containing 'the most frequently used words of the English language' (e.g. Thorndike and Lorge (1944) and West (1953)). This trend of finding useful applications for corpus data grew rapidly after George Zipf's groundbreaking discovery that in a given corpus the frequency of any word is inversely proportional to its rank in the frequency table (cf. Zipf (1935) and Zipf (1949)).

The first systematically compiled linguistic corpus was the Survey of English Usage (SEU) Corpus, started by Randolph Quirk in 1959. Quirk aimed to go beyond the grammatical descriptions found in regular grammars (e.g. Jespersen's *Modern English Grammar on Historical Principles* (1909-1949)) by carefully choosing texts, balancing size and genres in both written as well as spontaneous spoken material. Quirk's principles for the design of a balanced corpus are still used in the creation of corpora today (Meyer, 2008:10-13).

Around the same time, Roberto Busa started building the first machine-readable corpus and automated concordance of the works of St Thomas Aquinas, the *Index Thomisticus* (Busa, 1992). These types of first-generation concordances were usually held on one mainframe computer (McEnery & Hardie, 2012a:37). Major advances in technology, the 'revolution of software and hardware' in the 1980s and 1990s, allowed for large-scale digitisation of the electronic corpora we know today (cf. Kennedy (1998) and McEnery and Hardie (2012a)). At Brown University in Rhode Island, Nelson Francis and Henry Kučera started compiling a large corpus of written American English. This Brown Corpus (Francis & Kučera, 1964) is still very much in use. With the foundation of the Unicode Consortium, allowing encoding and reliably representing various writing systems on screen, digital corpora could finally be created for any language.

2.2.3 From synchronic to diachronic and other corpora

In 1978, the Brown Corpus found its British English counterpart in the Lancaster-Olso-Bergen (LOB) Corpus (Johansson, Leech, & Goodluck, 1978). Other languages followed the 'Brown tradition', i.e. the choice of balanced text samples that are as representative as possible, amongst which the *Lancaster Corpus of Mandarin Chinese* (McEnery & Xiao, 2004a) and the *Cronfa Electroneg o Gymraeg* 'Electronic Corpus of Welsh' by Ellis, O'Dochartaigh, Hicks, Morgan, and Laporte (2001) (this corpus, however, contains only Modern Welsh data and is as such not nearly sufficient to answer the historically-focussed research question of the present thesis).

It was not until the late 1980s that the first diachronic corpora were developed consisting of over 400 samples (over 1.5 million words) of continuous text from Old to Early Modern English (c. 750-1700 AD) (cf. Kytö (1991) and Kytö and Rissanen (1992)). ARCHER, 'A Representative Corpus of Historical English Registers', covers the subsequent period up to 1990 for both British and American English (Lee, 2010:113). As the basis for a new dictionary of Old English, a comprehensive corpus of all 3,022 Old English texts was compiled at the University of Toronto in

1981 (cf. Kennedy (1998:38)).

In the following years, other specialised corpora were developed for various purposes, such as the study of first and second language acquisition (CHILDES (MacWhinney, 2000) and ICLE/LCLE (Granger, 2003) respectively), (old) regional varieties (e.g. the *Helsinki Corpus of Older Scots (1450-1700)* and the *Corpus of Irish English* (Rissanen, 2008:60)) and corpora of sign languages (cf. Johnston (2010) and Marriott, Meyer, and Wittenburg (1998)). Many of the above-mentioned corpora now (also) contain some sort of linguistic annotation to facilitate language-specific or cross-linguistic research.

2.2.4 Treebanks

Treebanks are corpora including grammatical analyses of each sentence, named (by Geoffrey Leech) after a common way of representing syntactic structure. The small Swedish Gothenburg corpus was one of the first corpora to be annotated syntactically (Teleman, 1974). This was done by hand, since in the 1970s there were no automatic parsers available. Although the level of detail and theory-(in)dependency of the annotation varies widely, the construction of treebanks always requires significant effort (cf. Nivre (2008:226) and Wallis (2008:738)). It took years to parse (and manually correct) the historical corpora of Old, Middle and Early Modern English (cf. Kroch and Taylor (2000), Pintzuk and Plug (2002), Taylor, Warner, Pintzuk, and Beths (2003), Kroch (2000), Kroch, Santorini, and Delfs (2004) and Kroch, Santorini, and Diertani (2010)).

In their paper on quality assurance and sustainability in the handbook of corpus linguistics, Zinsmeister, Hinrichs, Kübler, and Witt (2008:760) conclude that "[i]t is fair to say that the Penn Treebank has served as a model of best practice for the creation of treebanks for many other languages.". This will therefore be the model for the annotated historical corpus of Welsh as well (see section 2.6 below).

2.3 Challenges in corpus linguistic research

In section 2.1.2, I briefly mentioned some strengths of digital corpora and computers: compared to humans, they are fast in dealing with loads of data, they do not get tired or bored and they make virtually no mistakes in routine tasks. It is, however, at the same time important to be aware of their limitations.

2.3.1 Where humans are better than computers

Computers first of all do not *notice* what they are doing. They can recognise and, if necessary, count recurrent patterns in the data, but unless given explicit input and instructions, these repetitions are meaningless to them. Scott (2010) exemplifies this lack of intuition problem as follows. When (in for example a restaurant setting) a man and a woman sit down at adjacent tables and the woman asks the man to pass the salt & pepper not just once, but over and over again, the man may

be led to the conclusion: "she fancies me" (Scott, 2010:139). A computer could obviously never reach that conclusion on the basis of multiple requests to spice up the woman's food.

Since without extra input, computers cannot interpret any *meaning*, they can also not judge the results or answers they find in a query. In an experimental setting, even mice exhibit a preference for one side or the other (cf. Brown (1988) or Takahashi et al. (1997) among many others). Computers on the other hand, do not and, crucially, cannot care about the results they find. A final general limitation worth mentioning before turning to implications for linguistic research is a computer's incapability of guessing the answer. Again, unless given specific further instructions, it is impossible for a computer to guess the meaning of, for example, words that are abbreviated in various ways.

2.3.2 Limitations in the context of linguistic research

The above-mentioned shortcomings of computers lay at the basis of most of the critiques on corpus linguistics. Initially, many scholars in other subfields of linguistics had a somewhat disparaging outlook on linguistic findings based on corpus research alone. Their concerns focussed around two main questions: to what extent do corpora represent the 'real' language (if at all) and how useful are statistical analyses of, for example, certain frequency patterns? Both of these issues will be discussed in this section.

"God's truth fallacy" and Competence vs. Performance

> *"It is crucial to distinguish* langue *from* parole*, competence from performance. (...) Performance can provide evidence about competence, as use can provide evidence about meaning. Only confusion can result from failure to distinguish these separate concepts."*
>
> (Chomsky, 1969:65)

The difference between *langue*, the abstract system of a language, and *parole*, the individual, practical acts of speech, was already pointed out by Ferdinand de Saussure in the beginning of the twentieth century (cf. De Saussure's posthumously published lecture notes by Bailly and Séchehaye, *Cours de linguistique générale*, (De Saussure, Bailly, & Séchehaye, 1916)). In the light of this distinction, corpora first of all represent the output of language *performance*, not *competence*. When larger corpora provide ample linguistic evidence, it is very tempting to identify those findings with the language itself. Failing to see this distinction is therefore what Rissanen (2008:65) called the "God's truth fallacy".

This immediately begs the question: if corpora are supposed to *represent* the output or performance, to what extent are they actually *representative* of that language? Furthermore, if your research question is merely concerned with a certain aspect of language *competence*, how useful is corpus data, or, as Fillmore (1992:35) tentatively sketched (quoted above): to what extent, if at all, is it interesting? As a native speaker of a language, you can call on your own competence to make up any

example of a particular grammatical pattern you want: how could a finite corpus of texts ever compete with this infinite source?

Corpora cannot always tell much about grammaticality; only intuition can provide that insight in a person's individual grammar. But analyses based on corpora consisting of non-elicited linguistic performance are still important, because they can shed light on what *many* people consider acceptable sentences or constructions (cf. Meyer and Tao (2005) and Conrad (2010:237)). A related problem for those interested in language competence is the fact that it remains unclear if sentences attested in a corpus are considered grammatical by the speaker/writer or if they were, in fact, simply a mistake. Aarts (1991) lists many (and very frequently occurring) examples of 'ungrammatical' sentences in a corpus, or rather, sentences "that do not conform to what is represented in intuition-based descriptions of what is possible" (Kennedy, 1998:272). Collections of texts can, however, never be large enough to contain examples of *all* possible constructions under investigation (Fillmore, 1992:35).

A full discussion of the apparent dichotomy between linguistic subfields interested in either language competence or language performance goes beyond the scope of the current thesis (see the numerous discussions on this topic, e.g. Fillmore (1992:35), Leech (1992:107), P. Baker (2006:6-9), Sampson (2007), Lüdeling and Kytö (2008:viii), Bonelli (2010), McEnery and Hardie (2012a:25-26) as well as József Andor's interview with Noam Chomsky (Andor, 2004)). Although 'naturalistic' corpus data differs from the results of controlled experiments, theoretical insights on language competence can be tested against those corpora, simply because they contain an abundance of usage data (Wasow, 2002:163). Although "Chomskyan" and corpus-based linguistic research typically exhibit different goals and/or foci of study, "the two approaches can be seen as complementary rather than conflicting." (Kennedy, 1998:271). In other words, "a corpus linguistics perspective on grammar has not made human judgements superfluous; it has actually expanded the judgements and interpretations that are made." (Conrad, 2010:229).

Regardless of its size and no matter how well-balanced the corpus is in terms of representing different genres, text types and registers, the language under investigation will always remain a 'corpulect': a cross-section of actual language performance at the very most (Komen, 2013:15). Examples from this 'corpulect' can represent decontextualised data (Widdowson, 2000:7) and a bottom-up approach is always required (Swales, 2002) (see section 2.4 and, among others, P. Baker (2006) and Handford (2010) for further discussion and solutions to these problems).

(Im)possible statistical analyses

"Corpora are quantitative number-crunching tools." (Handford, 2010:255) is a frequently-cited criticism of corpus research. But the obvious new path of research opportunities paved by the emerging (digital) corpora lay in frequency data. Words, collocations and grammatical structures could now be counted systematically. As Biber et al. put it: "The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers."

(Biber, Conrad, & Cortes, 2004:376).

Merely counting many words or patterns under investigation, however, cannot establish frequency: there is no invariable value associated with 'frequent', it remains a relative judgement (cf. McEnery and Hardie (2012a:49)). If corpora can only contain samples of the infinite number of possible sentences in a language, it becomes much harder to answer the question: relative to what? When a certain construction does not appear in a corpus, it does not imply this particular construction *never* appears in the language and/or is per definition ungrammatical. Its absence *could* suggest it is infrequent, but the corpus could also be inadequate, not well-balanced or simply not representative enough of the particular language (variant) under investigation (Kennedy, 1998:272).

Statistical analysis is needed to establish the relative or normalised frequency of occurring patterns (McEnery & Hardie, 2012a:49). However, since corpora are never just a random selection of texts representative of a language, standard statistical techniques cannot always be applied (Komen, 2013:17). To make sure the frequency patterns found in the corpus are not just a matter of coincidence, tests for statistical significance can be used. A serious drawback of most of these is nonetheless that they can only point to significant *differences*: "[t]hey cannot tell us how significant one point in our data is" (Komen, 2013:17). "The mystery of vanishing reliability" (Rissanen, 2008:65) is connected to this problem. If certain patterns exhibit a low frequency overall, they are likely to be too low for any reliable conclusions when various factors such as occurrence per text, genre, chronological period or any sociolinguistic variables are taken into account.

Observing frequency patterns alone will thus never be sufficient to describe grammar. Frequency data can nonetheless identify certain interesting patterns that require explanation and thus further investigation (Biber et al., 2004:76). Section 2.8 will go into more detail as to *how* statistics can indeed help linguistic analyses.

2.3.3 Challenges with (written) historical corpora

There are some additional challenges working with historical corpora. First of all, historical corpora (covering the period up to the invention of tape recorders) necessarily contain written material only. Written texts are possibly even further removed from the speaker's language competence, because the process of writing is much slower than spontaneous speech. Moreover, there are possible effects of standardisation of the language or literary stylistic features that surface in carefully crafted texts. Finally, especially when working with older manuscripts, there may be distortions due to repeated copying by various different scribes.

This last problem relates to what Rissanen called 'the philologists' dilemma' (Rissanen, 1989), focussing on the issue of the 'slow' work of philologists and whether that had become irrelevant with the rise of digital corpora. Evidently, not just the corpus compilers but also their users can only draw meaningful conclusions from their corpus data if they understand the philological background of the consulted texts. Not all necessary metalinguistic data such as the social and cultural

background of the text or even its author and exact date and place of origin is available, however.

In general, "a historical corpus can only be as thorough as the available texts" (Curzan, 2008:1098). This lack of availability may be the result of historical events. Examples of this in 'Celtic history' include viking raids, the dissolution (and destruction) of the monasteries where manuscripts were kept, unfavourable wet climate causing rapid decay of codices, etc. A striking exception to this is the recently discovered Fadden More Psalter in a peat bog in County Tipperary, Ireland (Kelly & Sikora, 2011). However, even in that case it is difficult to ascertain the original text considering the fact that the actual pages are mostly gone and only the pieces with ink have survived, resulting in a mixed-up soup of letters.

Present-day copyright considerations can finally cause problems for the distribution of texts. Annotated corpora very often rely on the availability of modern *edited* versions of the historical manuscript versions. Only with intensive collaboration between philologists and the corpus linguists can these old texts be made available for linguistic scholars.

2.4 Benefits of annotated corpora

In the years following the creation of the first digital corpora, the new 'corpus linguists' managed to address many of the above-mentioned issues. The computer's main shortcomings (their lack of typical human intuition or ability to guess and reason based on meaning) were partly overcome by increasingly good software solutions, including integrated lists of words, names and abbreviations, morphological stemmers to recognise various word forms automatically and even elaborate semantic tools to recognise word meanings. This section will focus on what corpora *can* do, what new research opportunities they brought along and why they are excellent tools for linguists in various subfields, including historical syntax.

2.4.1 What corpora *can* do

Once the difference between language competence and performance and its importance in corpus linguistics is acknowledged, an entirely new field of research opens up. Corpus data may be far removed from the abstract grammar of one particular language theoretical linguists are interested in, but one single text written by one single person still has a grammar. The writer in question may have employed a specific literary style that may be very different in nature from his/her daily speech, but that does not render the quest for the text's internal grammar futile. Even if the author was code-switching between his literary grammar and his spoken language, both are worth investigating as long as the researcher is aware of this distinction and acknowledges that the corpus text is never direct evidence of language competence.

A similar reasoning applies to 'dubious statistics' and 'number crunching'. When used with care, numerous research opportunities open up with the availability of

more easily accessible language data than ever before. According to P. Baker (2006), it is exactly the quantitative evidence of patterns that helps researchers find (or not overlook) certain patterns in the language. Aberrant (e.g. both surprisingly high or low) frequencies cannot be ignored: they need to be explained and are thereby creating new research questions that had not even occurred to scholars in the field.

Elena Bonelli argues that frequency of occurrence might be indicative of frequency of use: "[t]he corpus, in fact, is in a position to offer the analyst a privileged viewpoint on the evidence, made possible by the new possibility of accessing simultaneously the individual instance, which can be read and expanded on the horizontal axis of the concordance, and the social practice retrievable in the repeated patterns of co-selection on the vertical axis of the concordance." (Bonelli, 2010:20). Moreover, reliable estimates of frequencies of use are very difficult to make, not only by native speakers but also by linguists who spent years studying the language (Alderson, 2007).

Apart from these advantages of investigating the frequency of words or patterns, the very fact that *only* computers can do such systematic and complex studies in large collections of texts cannot be discarded (Conrad, 2010:228). The complexity mainly lies in frequencies of patterns found in combinations of possible factors such as different contexts, genres, periods of time, etc. (exactly what is needed in historical investigations into word order and information structural change). Traditional linguistic variables can be measured in relation to one another, but the more text there is available for analyses, the more likely it is that new patterns or even new linguistic variables will be discovered (cf. Kennedy (1998:70) and Wright (1993)).

Biber (1988)'s 'multifactor' analysis used in his investigation into variation in different registers of English is a good example (cf. section 2.8 for this and other statistical methods that are worthwhile when interpreting corpus data). Another good example is Leech's chapter on modals in his work on the meaning of English verbs (Leech, 2004a): the 2004 edition that appeared more than 30 years after its original publication was substantially revised, because of new evidence found in large corpora of English usage (McEnery & Hardie, 2012a:28). Especially in large annotated corpora with well-documented and detailed metalinguistic data for each text, statistical analyses can be very useful uncovering hitherto hidden rules and patterns of language use. Finally, frequencies and probabilities in themselves are making their way in more theoretic research as well (cf. Nivre (2008:236) and contributions in Bod, Hay, and Jannedy (2003)).

2.4.2 On testing hypotheses

Another way to take advantage of digitised corpora is by using them to test hypotheses. If a hypothesis predicts that certain forms are grammatically *im*possible in a language, the occurrence of one or more examples of that particular form could lead to the rejection or reformulation of the afore-mentioned hypothesis. Digitally annotated historical corpora can even help to verify hypotheses about connection, causation and development in time.

Roberts (1997), for example, proposed that there was a connection between case and word order in Old and Middle English. He argued that there was a direct causal connection between the loss of OV orders and the loss of the rich system of case marking. Pintzuk (2002), however, showed on the basis of historical corpus data, that this was not so straightforward. Richness of case is not directly linked to word order facts, because the grammar is sensitive to properties of individual words as well: only a case system as a whole could affect the entire language. Moreover, Pintzuk (2002) found that English was already shifting to VO by 950 and the case system was still intact at the end of the eleventh century. Without an annotated corpus, Roberts (1997) could conclude the two events roughly coincided; with a corpus, Pintzuk (2002) could go into far more detail discovering there was, at the very least, no direct causal relation, if the two phenomena were connected at all (cf. McFadden (2014)).

The verifiability of certain linguistic hypotheses has thus increased with the coming of well-annotated corpora. This process, related to what Leech (1992:112) described as 'total accountability', must, however, be relative to the used dataset, not the language as a whole. But the bigger the corpus, the more data we can account for. The likelihood of falsification[5] and the replicability of the results of other scholars in the field has thus improved tremendously with the coming of corpora and good tools to annotate and query them in a systematic way (cf. Rissanen (2008:54-64) and McEnery and Hardie (2012a:16)).

Exactly because of this, "corpus linguistics has the potential to reorient our entire approach to the study of language" (McEnery & Hardie, 2012a:1). The next section will provide a brief overview of these new applications and opportunities.

2.4.3 New applications and research opportunities

Annotated corpora with well-designed and easy-to-use query software can thus be very useful tools in linguistic research (McEnery & Hardie, 2012a:28) (see section 2.7 for a discussion of the most common options). But apart from testing existing hypotheses, new opportunities were created for functional and cognitive linguistic research based on language 'as it is used' in particular (cf. Gries and Stefanowitsch (2007) and McEnery and Hardie (2012a:171)). Grammars of languages could now, according to O'Keeffe, McCarthy, and Carter (2007), not only be described in structural, but also in probabilistic terms.

Especially in the field of second language acquisition, access to typical social and discourse circumstances associated with certain words, idioms or grammatical patterns is highly beneficial for language learners and their teachers (cf. Kennedy (1998:280), Hoey (2005:150) and Conrad (2010:228)). But also computational linguistics and applications in the field of natural language processing (NLP) could be further developed by corpus data. Computational models and NLP techniques in their turn played a big role in the creation of better tools for annotating and

[5] Note that it is the likelihood of falsification, not the logical issue of falsifiability in itself: *verifiability* of hypotheses increased dramatically, not their *falsifiability* (cf. Popper (1935) on the difference between verifiability and falsifiability and the latter's crucial role in scientific methodology).

querying corpora (cf. Church and Mercer (1993), Kennedy (1998:277), Handford (2010) and McEnery and Hardie (2012a:203-205)). Tasks traditionally based on paper concordances, could with the digitisation of corpora now be extended to large searches for multi-word units, phrases and n-grams from which, for example, machine learning and 'translation' tools could be developed (cf. Greaves and Warren (2010)), such as 'Google Translate', which does not translate in fact, but finds n-gram parallels.

Other fields of applied linguistics such as discourse analysis, forensic linguistics, pragmatics and speech technology benefit from larger accessible amounts of language data as well. Examples of discourse-related research based on corpora come from, among others, Sinclair (2004) and P. Baker (2006). Pragmatically annotated corpora are now also available (cf. the Michigan Corpus of Academic Spoken English (MICASE) Maynard and Leicher (2007) and, for a general overview, Rühlemann (2010)).

Overall, the coming of digitally annotated corpora has impacted many subfields of linguistics. Regardless of the discussion between corpus-based or corpus-driven scholarship and of the question whether corpora are merely useful tools, 'corpus linguistics' and the methodology of designing, building, annotating and querying corpora has become a field of its own (see, for example, McEnery and Hardie (2012a:6 &157-162) and references there for a full discussion).

2.4.4 Corpora in formal & historical linguistic research

Although the usefulness of corpora might seem less obvious in formalist approaches, there are various examples of corpus-based studies in this field as well (e.g. in relation to first-language acquisition by Bloom (1990), Déprez and Pierce (1993), MacWhinney (2000) and various publications by Charles Yang, e.g. C. D. Yang (2000) and C. D. Yang (2002)). In the study of language change, corpora can be invaluable tools as well. Apart from testing old hypotheses (as described above in section 2.4.2), new generalisations and effects were found and tested in the growing corpus data, for example, Anthony Kroch's "Constant Rate Effect" ("when one grammatical option replaces another with which it is in competition across a set of linguistic contexts, the rate of replacement, properly measured, is the same in all of them." (Kroch, 1989:200) and Chapter 6).

Language change can be caused by internal or external processes. In the latter case, corpora with well-documented metadata can take possible extralinguistic factors into account at the same time (see section 2.8 below and, among others, Rissanen (1998:400) and Rissanen (2008:59)). The transmission or implementation problem Weinreich, Labov, and Herzog (1968) described can be tackled more easily with the availability of more historically-annotated corpora containing digitised texts from various regions and periods in time (Curzan, 2008:1092).

The easy access to digitised forms of the text also aids philologists. Collaboration between (corpus) linguists and philologists is thus not only indispensable to make any sound generalisations about the history of the language, it can also be valuable in the field of philology. Text editing, linguistic reconstruction and

phylogeny benefit greatly from wide range of easy-accessible data in the digitised corpora (Rissanen, 2008:54).

To conclude this section, (historical) corpora offer a great variety of new possibilities to scholars in many different subfields of linguistics and beyond (Curzan, 2008:1105). There are some limitations in some corpus-based research, in particular when language competence and performance are not kept apart. But research questions concerned with language change over longer periods of time combining many different grammatical and information-structural variables cannot be addressed properly *without* a well-annotated historical corpus.

2.5 Compiling the corpus

For the present study, I built a partial corpus of Middle Welsh, including the most important narrative literature from the medieval period. This partial corpus can be used as a starting point to build a fully annotated treebank of historical Welsh. In this section I describe the necessary steps in the process of creating an annotated corpus in greater detail. Language-specific decisions concerning any type of annotation can be found in the Annotation Manual in the Appendix.

As pointed out above, ideally any corpus is well-balanced in terms of text type, length, origin etc. When working with historical data, however, the choices are often limited. For the present annotated corpus I decided to include the most important narrative native prose: all extant tales of *The Mabinogion*. In addition to this, I chose to include a contemporary version of the Welsh Laws, two versions of the tale of *Llud and Llefelys* and *Buched Dewi*, the story of the life of St David. Finally, various narrative passages from the 1588 Bible translation were selected to reflect the stage of the language at the very end of the Middle Welsh period.

Future extension of the corpus should include alternative manuscript versions of each of these texts. In addition to that, it would be useful to extend the corpus to include more texts from different genres such as the historical chronicles of the kings and princes, but also translations and retellings of further Arthurian literature from the same period.

2.6 Annotating the data

As argued above, a well-annotated historical corpus is extremely useful for linguists investigating earlier stages of the language. Because manual annotation is very time-consuming, we should make as much use as possible of automated methods and tools from the field of Natural Language Processing (NLP) to facilitate this task. Before we can apply these tools, however, we need to prepare or 'preprocess' our dataset to ensure it is in the right format for any further NLP tasks. A properly preprocessed version can then be tagged automatically by a Part-of-Speech tagger. For Middle Welsh, no such tagger was available, so I furthermore describe the

process of training a Memory-Based Tagger here that could subsequently be used to assign morpho-syntactic tags to the Middle Welsh data. For this purpose, decisions have to be made concerning the tagset. A very detailed tagset facilitates more (and different types of) research. When working with a corpus of limited size, however, too many different tags leads to low frequencies and many hapaxes, which in turn complicates the automatic tagging task. In this section I describe these challenges and furthermore offer some solutions that are not only useful for those working on Middle Welsh, but for anyone working with similar complex historical data.

2.6.1 Preprocessing

There are various orthographical peculiarities in the White Book version of the *Mabinogion* (cf. Huws (1991)). For the present study, the texts were not extensively preprocessed, because there was no stemmer available yet for Middle or Early Modern Welsh. Detailed photographs of the White Book of Rhydderch are available on the website of the National Library of Wales (`www.llgc.org.uk`).

Utterance boundaries in the form of `<utt>` were added to the transcribed text with regular expressions following full stops (that were added manually if they did not appear sentence-finally in the manuscript). The only punctuation that was removed were the full stops preceding and following numbers, e.g. '.11.' was turned into '11' to facilitate automatic tagging. Tokenisation (the isolation of word-like units) was done automatically by the PoS-tagger on the basis of word spacing and full stops at the end of an utterance.

As became clear from the initial pilot, the huge amount of orthographical variation complicates the PoS-tagging task tremendously. The Memory-Based Tagger (MBT, see below), however, could filter those out on the basis of the context most of the time. In this way, there was no real need for time-consuming preprocessing of the text in terms of splitting merged tokens. Some tokens, however, were particularly challenging for the automated tagger, since very few generalisations could be made from the small training set (cf. Meelen and Beekhuizen (2013)). To overcome some of those very specific orthographical challenges, combined words with nasalising prepositions like *yn* 'in', were split, e.g. *ymwyt* > *y** + *mwyt* 'in food'.

There is still a large amount of homophony, but the tagger was often able to distinguish between up to five different possible meanings of, for example, Middle Welsh *y* 'the, his, her, to, to his/her, in' etc. on the basis of the preceding and following context.

2.6.2 Part-of-Speech tagging

The standard UPenn annotation scheme (cf. `www.ling.upenn.edu`) does not always provide enough information to answer certain research questions, mainly queries concerning agreement patterns and change in Information Structure. To enable further research in these and other areas, I have extended the Part-of-Speech tagset. Starting from the already extended tagset used for the Icelandic corpus (cf.

Wallenberg et al. (2011)), I have examined the features of Middle Welsh grammar and systematically added dash-tag features, mainly in the verbal domain. A full overview of the tagset is given in the Appendix.

Establishing the morpho-syntactic tagset

Verbal inflection in Welsh occurs as a suffix to the verbal stem. Inflected verbs in the UPenn tagset are tagged VB. Past tense is indicated by the regular English past-tense ending in *-ed*, resulting in VBD. For Welsh, I kept the VBD for the preterite tense. In the same way, I added tags for present (-P), future (-F) and pluperfect (-G, for Welsh *gorberffaith* 'pluperfect'), imperative (-I) and imperfect (-A, for Welsh *amherffaith* 'imperfect') etc. Finally, I added the distinction between indicative (-I) or subjunctive (-S) mood for the tenses in which that is relevant. This results in insightful systematic combinations like VBPI (present indicative), VBAI (imperfect indicative), VBG (pluperfect) etc. The same letters were systematically added to irregular verbs, resulting in for example DOPI (present indicative of the verb *gwneuthur* 'to do'), HVI (imperative of the verb *cael* 'to get') or BEAS (imperfect subjunctive of the verb *bod* 'to be').

Apart from these more-detailed tense-aspect-mood markers, I added further information about the inflection to indicate person and number. Following standard glossing practices, person and number were represented as -1SG (first-person singular), -2PL (second-person plural) etc. Welsh has a further inflectional suffix for the 'impersonal' form of the verb that can be used in true impersonal contexts meaning 'one' or underspecified 'they', but also as a passive ending. I used the number 4 for this specific suffix and added it to the verbal tags like the other personal endings, e.g. VBPI-4 (impersonal present indicative) or DOAI-4 (impersonal imperfect indicative of the verb *gwneuthur* 'to do').

Inflected and combined prepositions

Another feature of the grammar, specific to Welsh and other Celtic languages (but also seen in for example Semitic languages like Arabic or Hebrew), is inflected prepositions. Middle Welsh had a specific set of prepositions that could be inflected for person, number and gender (in third-person singular only). There are also 'uninflected' prepositions in Welsh, but the inflected set includes very common prepositions like *i* 'to', *ar* 'on' and *yn* 'in'. Middle Welsh *iddi* 'to her' is for example tagged as P-3SGF 'preposition third-person singular feminine'.

Welsh also allows for some combined prepositions: a combination of a preposition plus a grammaticalised noun. If the object of this type of preposition is a pronoun, it can appear in between the two prepositions as a possessive pronoun, e.g. *yn eu herbyn* 'against/towards them' (PKM 65.6-7) from *yn* 'in' + *eu* 'their' + *erbyn* 'opposition'.

There are two possible ways to annotate constructions that are changing in historical corpora: we can annotate the original structure and form or the new construction as a whole. Since the exact date of grammaticalisation is often difficult

to determine, it is not always easy to choose one or the other. As long as the construction is tagged consistently in one text (or one period of the historical corpus) and the annotation manual is clear, this should not be a problem. In that case future researchers will always be able to find and, if necessary, to change the annotation again. A full annotation manual is presented in the Appendix. In this particular case of combined prepositions, a more conservative annotation scheme, acknowledging the nominal origin of the construction yielding the tag sequence 'P 3P N' (preposition - third-person plural possessive - noun) was preferred to facilitate rule-based chunk-parsing.

Prepositions in Welsh could also be combined with other prepositions, e.g. *y dan* 'under, below' from *y* 'to' + *tan* 'under'. These complex prepositions were tagged PSUB + PSUB, so they could be recognised as separate, but also as combined prepositions. A further advantage of this is that the automatic tagger looking at the tags preceding and following the focus word, will not encounter the rare sequence of two prepositions. A disadvantage remains, of course, that the tagset is further extended and there are more homophonous forms that could render worse results if the complex preposition in question does not frequently occur in the training set. For combined conjunctions, a similar extension was used: *o + herwydd* CONJSUB + CONJSUB meaning 'because'.

Distinguishing different types of prononimal forms

Another part of grammar in which the tag set was extended significantly is pronominal forms. Since Welsh has various sets of pronouns for different (grammatical) contexts, a more fine-grained distinction here could enhance research not only in the pronominal domain, but also in Information Structure. Conjunctive pronouns, for example, (see table 6.1 above) are used in contexts of topic switch, meaning 'but I', 'I, then,' etc. Reduplicated pronouns like *tydi* 'you', on the other hand, are only used in focussed contexts. Separate tags for those are thus useful for finding the focus domain of sentences.

A further distinction is made between possessive pronouns and object pronouns. Following the extensions of the tagset for the Icelandic parsed corpus, these pronouns receive case endings like *fy* 'my' ⇝ PRO-G, or *e* 'him' ⇝ PRO-A. Since the infixed versions of these pronouns often exhibit the exact same form, a more fine-grained distinction in the tagset facilitates syntactic research here as well.

Further extensions of the tagset

Further extensions of the tagset include ADJQ for equative constructions, e.g. *cochet* 'as red' (PKM 1.24) (from *coch* 'red' + equative *-et*) and ADJPL for plural adjectives, e.g. *gueisson ieueinc* 'young servants' (PKM 4.8). More detailed tags like these are helpful to syntacticians looking at the structure and agreement patterns of noun phrases.

As described above, Welsh employs a wide range of particles. These too were

tagged separately according to their function (e.g. PCL-QU, PCL-FOC, PCL-NEG) to help distinguish different types of clauses. Aspectual particles like *yn* 'progressive' (PROGR) or *wedi* 'perfective' (PERF) were also distinguished from their homophonous prepositions (P) and predicative particles (PRED).

The verbal noun category so specific for Celtic was tagged VN for regular verbs. Irregular verbs with verbal nouns that have specific functions in Welsh, e.g. *cael* 'get', also used for the passive, received specific verbal noun tags. The -N was added systematically to their base forms, e.g. HV- 'have, get' > HVN 'verbal noun of the verb *cael* 'to get'. The verbal noun of the verb 'to be' was kept separate and tagged as 'BOD', since it can also appear in this form in many other syntactic contexts, e.g. as a complementiser.

Finally, some additional lexical items with specific functions were tagged separately. An example of this is the petrified form *sef* (tagged 'SEF') that was used in earlier stages of the language to focus identificational copular sentences. During the Middle Welsh period, it grammaticalised further until it became an adverbial element used in apposition to noun phrases meaning 'that is' (cf. Latin *id est* still used as the abbreviation *i.e.* in English).

Combined tags

With a 'hands-off' diplomatic transcription of one single manuscript, tokenisation forces decisions on splitting certain merged combinations found in the transcription, like *yr* 'to the' and *ae* 'and his'. This works as long as there is a logical boundary (e.g. *yr* can be split up in *y* 'to' and *r* 'the'). For some fused forms, however, it poses more difficulties, e.g. *y* (from *y* + *y*) 'to his, her'. This problem is further complicated by the fact that *y* in Middle Welsh can have a variety of meanings, ranging from the definite article to the preposition 'to' and various pronominal forms. Preprocessing will thus have to be done manually, to be able to take the full context into consideration. Or, - and this is less time-consuming - these forms need to be checked manually after automatic PoS-tagging when creating gold standards. Alternatively, combined tags can be used (e.g. *y* (< *y+y*) 'to his' as P-PRO-G). This, however, significantly expands the tagset and thus yields worse results in the evaluation. Especially because this usually concerns short words that have various meanings and/or functions already, I chose to manually split these forms when correcting the automatically tagged texts.

This then, appears to be the limit of useful extension of the tagset. Expanding the training set can improve the results of the tagger as well, but only slightly. If more combined tags are used the results of the memory-based tagger would need to be improved by either more rigorous preprocessing (e.g. regularisation of the orthography and more splitting of tokens), manual correction = and/or adding rule-based techniques (e.g. or, for example develop a reliable Middle Welsh stemmer).

Tagging with the MBT

The technical details concerning the generation of the PoS-tagger are discussed in the Appendix. Once the Middle Welsh tagger is generated, the settings file of the tagger is then used to assign PoS-tags to a new part of the corpus (presented as a tokenised text file). Based on the training set, the MBT divides the new text in need of annotation into 'known' and 'unknown' words. Depending on the exact parameter settings, the tagger will then assign a tag to each word.

As mentioned above, in Welsh the inflection appears as a suffix (on verbs or prepositions). When the tagger finds an unknown word like *arnaf* 'on me', for example, it can compare the last three characters to known words with assigned tags in the training set. An example of this could be another inflected preposition, like *ohonaf* 'of me' with the PoS-tag P-1SG ('Preposition + first person singular ending). The exact same final characters (in combination with the other tags in the preceding and following context) lead the MBT to assign the same tag 'P-1SG' to *arnaf*, which would be the correct tag.

Known words are easier if there are no homophones with different tags. If there are, for example for the above-mentioned Middle Welsh word *y*, the context in which it appears is crucial. In between an adverb (ADV) and an inflected verb (VB*), *y* is undoubtedly the preverbal particle following sentence-initial adjuncts, like in (1a). In front of verbal nouns, however, like at the end of (1b), *y* could be the preposition 'to' or a possessive pronoun (masculine, feminine or third-person plural), as in (1).

(1) a. *Tranhoeth y deuthant y 'r llys.*
next.day PRT come.PAST.3P to the court
'The next day they came to the court.' (CO 595)

b. *a dyuot yn y uryt ac yn y uedwl uynet y hela*
and come.INF in 3MS mind and in 3MS thought go.INF to hunt.INF
'and he was minded to go and hunt' (PKM 1.3-4)

The output file of the tagging process is a text file consisting of a word + TAG and an indication whether this word was known or unknown from the training set. A full list of tags can be found in the Appendix.

MBT allows for different settings according to features of the words themselves or the context in which they appear (see Appendix for further details). To obtain the maximally reliable tags, I tried a wide range of parameter settings concerning those features. The Global Accuracy of the classifier was then evaluated to get the best parameter settings. The optimal settings for Middle Welsh are (see the MBT manual for further details Daelemans et al. (2010)):

```
-p dfa -P sssdFawchn -M 200 -n 5 -% 5 -O +vS -F Columns

-G K: -a 0 U: -a 0 -m M -k 17 -d IL
```

For Middle Welsh, the corrected gold standard of one text was subsequently used to annotate other texts of the *Mabinogion* automatically with greater accuracy. Each of those texts was in turn manually corrected as well.

In order to estimate the quality of the PoS-tagger and obtain optimal parameter settings, I evaluated on the manually annotated data by a ten-fold cross-validation, i.e. taking 90% of the data, training the model on that subset and then testing it on the other 10%, repeating this procedure for ten 90%/10% splits. Because the ten percent that the model is tested on is manually annotated, we can see how often the model assigns the correct tag to a word, as well as obtain insightful statistics about the over- and undergeneralisations of some tags. The above-mentioned settings gave the following results for the 59k Middle Welsh corpus:

Global accuracy: 90.4%

Global accuracy seen words: 93.3%

Global accuracy unseen words: 63.3%

The results are split between seen (Figure 2.1) and unseen (Figure 2.2) words as well. Looking at the results for the largest categories of tags for seen words, we find high results for simple tags like N 'noun' or CONJ 'conjunction' that occur extremely often. As expected, Precision and Recall for tags occurring only once or twice is extremely low. These tags are often combined tags or forms of verbs that occur very infrequently with irregular endings.

I calculated the Precision (percentage of system-provided tags that were correct), Recall (percentage of tags in the input that were correctly identified by the system) and F-score (weighted harmonic mean of recall and precision).

For the individual categories, Precision and Recall give more insight in the degree to which the model over- or undergeneralises certain tags. The genitive (possessive) pronoun category (PRO-G), for instance, is correct in 86% of the cases where it is applied, but out of all actual possessive pronouns, only 67% is recognised. This is understandable, because the possessive pronoun usually consists of only one letter that is homophonous with the object infixed pronoun. The model thus undergeneralised that category in particular.

On the other hand: 94% of the actual conjunctions are recognised as such, whereas when an item is classified as a conjunction, the model is correct in only 92% of the cases. This category is thus slightly overgeneralised. As expected, the F-score for frequently occurring tags is considerably higher than that for tags and tokens occurring only once or twice in the corpus. The extremely fine-grained tagset (cf. Appendix) can thus only reach an acceptable Accuracy in a large corpus.

Category	Precision	Recall	F-score	n
N	0.95	0.96	0.96	5413
CONJ	0.92	0.94	0.93	4411
P	0.86	0.85	0.86	4404
PCL	0.92	0.93	0.92	3211
D	0.79	0.95	0.86	3062
VN	0.97	0.97	0.97	2070
PRO	0.98	0.99	0.99	2026
PRO-G	0.86	0.67	0.75	1593
NPR	0.98	0.96	0.97	1204
ADJ	0.92	0.93	0.93	981
ADV	0.96	0.95	0.96	886
VBPI-3SG	0.89	0.96	0.92	883
DEM	0.99	0.99	0.99	827
PSUB	0.89	0.85	0.87	767
PCL-NEG	0.99	0.97	0.98	692
VBD-3SG	0.98	0.99	0.99	660
P-3SGM	1	1	1	565
PROC	1	1	1	514
NPL	0.96	0.95	0.95	513
PRED	0.85	0.74	0.79	430
PCL-QU-NEG-PRO-A	1	1	1	2
HVPI-1PL	0	0	0	2
HVG-3SG	1	1	1	2
DOI-1PL	0	0	0	2
DOAI-2SG	1	1	1	2
BED-1SG	1	1	1	2
BEI-2SG	0.5	1	0.67	1
VBG-3PL	0	0	0	1
VBAS-1PL	0	0	0	1
VBAI-2SG	0	0	0	1
PCL-FOC	0	0	0	1
PCL-A	0	0	0	1
HVPS-3SG	0	0	0	1
HVD-4	0	0	0	1
HVAS-2SG	0	0	0	1
DOAS-3SG	0	0	0	1
CONJ-PRO-G	0	0	0	1

Table 2.1: Sample of the results for seen words - Precision (P), Recall (R) and F-score (F), as defined by Manning & Schütze (1999) and Jurafsky & Martin (2009:489)

Category	Precision	Recall	F-score	n
N	0.63	0.75	0.68	1570
NPR	0.83	0.75	0.79	535
VN	0.67	0.69	0.68	526
ADJ	0.64	0.53	0.58	421
NPL	0.69	0.67	0.68	364
VBD-3SG	0.62	0.76	0.68	168
VBPI-1SG	0.78	0.87	0.82	112
VBAI-3SG	0.6	0.71	0.65	105
VBD-3PL	0.66	0.87	0.75	75
VBI-2SG	0.4	0.24	0.3	59
ADV	0.39	0.27	0.32	59
VBPI-3SG	0.24	0.2	0.22	51
VBPI-2SG	0.71	0.65	0.68	49
ADJS	0.68	0.62	0.65	48
ADJQ	0.65	0.3	0.41	43
VBD-4	0.37	0.55	0.44	40
BEPI-3SG	0	0	0	1
BEPI-1SG	0	0	0	1
BEI-3SG	0	0	0	1
BEI-2SG	0	0	0	1
BEG-3SG	0	0	0	1
BEF-2SG	0	0	0	1
BED-3SG	0	0	0	1
BED-3PL	0	0	0	1
BED-2SG	0	0	0	1
BEC-3SG	0	0	0	1
BEAS-2SG	0	0	0	1
BEAI-3PL	0	0	0	1

Table 2.2: Sample of the results for unseen words - Precision (P), Recall (R) and F-score (F), as defined by Manning & Schütze (1999) and Jurafsky & Martin (2009:489)

Middle Welsh presents a good test case for PoS-tagging a historical corpus of a language with rich verbal and prepositional inflection and non-standardised orthography. Further challenges in assembling this corpus lie in the availability of good diplomatic or critical text editions. More collaboration with scholars specialised in the philological background producing these editions can help syntacticians make the right decisions, both in terms of selecting the right texts and editions for the corpus, but also in preprocessing and tokenisation in particular.

Adding person and number features for verbal suffixes and thus expanding the tagset does not yield a significantly lower Global Accuracy using the Memory-Based Tagger (MBT) by Timbl (cf. Daelemans and Van den Bosch (2005)). This

tagger showed robust results and flexibility with the highly variable orthography of minimally preprocessed Welsh texts (see Meelen and Beekhuizen (2013)). The parameter settings of MBT allow for focus on the context and the last 3 letters of unknown words. Since Literary Welsh verbal endings usually consist of 2/3-letter suffixes (reflecting tense, mood, aspect, person and number combined), it is not difficult for the tagger to predict the right form (e.g. *gwel-ais* "I saw" as VBD-1SG denoting 'preterite-1sg'). Other parameter settings like an additional focus on the first 3 letters of the word proved to be less helpful for a language like Welsh with initial consonant mutation. This might, however, improve the results for languages with a strong prefixing preference, like for example Navajo (Young & Morgan, 1980:103,107). A full overview of the morpho-syntactic tagset can be found in the Appendix.

2.6.3 Chunkparsing

In order to facilitate syntactic queries, I used the PoS-annotation to develop hierarchical phrase structure. A full parse would require a detailed Context-Free Grammar or Dependency Grammar. Developing this would go beyond the scope of the present study, however. Instead, I modified the rule-based chunkparser available in the Natural Language Toolkit (NLTK via `www.nltk.org`) in such a way that not only phrasal chunks, but also hierarchical structure could be added.

Designing the rule-based grammar

The NLTK rule-based chunkparser is a regular expression parser: it systematically combines PoS-tags as defined in a grammar that allows regular expressions to create more (specific) options. Frequently-used regular expressions include:

? ⇒ for optional preceding items
| ⇒ 'or'

The combination of words with their PoS-tags into phrases is achieved with the following sample pattern of commands:

```
NP: {<N|NPL|NPR>}
DP: {<D><NP>}
PP: {<P><NP|DP>}
```

According to the above rules, a noun phrase (NP) can be formed of words with one of three different PoS-tags: a noun (N) *or* a plural noun (NPL) *or* a proper noun (NPR). The order in which this rule-based grammar operates is important. The DP-rule above must follow the NP-rule to find the label <NP>. In this way single-layered hierarchical structures (NPs within DPs) are created. Similarly, a further layer can be created resulting in a PP containing a DP containing an NP, as long as they are called in the right order.

This is all straightforward in a language with extremely simple noun phrases and/or with a very limited amount of PoS-tags. Middle Welsh noun phrases, however, present some problems in this respect. First of all some adjectives either follow or precede the noun they modify, with different meanings in the two positions. In addition to this, possessive pronouns and quantifiers can be part of the noun phrase as well. Furthermore, demonstratives must follow the noun (and its modifying adjectives) and they are also obligatorily accompanied by the definite article preceding the noun phrase. Finally, Welsh numerals above ten can be split to occur before and after the noun phrase. In addition to that, phrases with numerals can also employ the preposition *o* 'of'. Examples of these various kinds of DPs that potentially present problems for simple rule-based grammars are given below:

(2) a. *y cathod mawr*
the cats big
'the big cats'
b. *yr hen gathod*
the old cats
'the old cats'
c. *yr hen lyfr mawr hwn*
the old book big this.M
'this big old book'

(3) a. *dau hen lyfr*
two.M old book
'two old books'
b. *y chwe chath newydd*
the six cat new
'the six new cats'
c. *tair merch ar ddeg*
three.F girl on ten
'13 girls'

(4) a. *un mlynedd ar ddeg*
one year on ten
'11 years'
b. *pob yn ail fis*
every PRED second month
'every other month'
c. *yr holl broblemau*
the all problem
'all the problems'

(5) a. *tair o ferched*
three.F of girls
'three girls'
b. *tri o bobl eraill / newydd*
three.M of people other.P / new
'three other / new people'

Complex noun phrases can also consist of two juxtaposed nouns in a so-called 'genitive construction'. In these constructions, the definite article only appears before the second noun, but the whole construction is definite.

(6) a. *dyn y siop*
man the shop
'the man of the shop'

b. *cŵn y cymdogion*
dogs the neighbours
'the neighbours' dogs'

The above types of complex noun phrases require a very detailed rule-based grammar that includes all possible phrases, including some phrases with special labels to facilitate further syntactic queries, e.g. phrases with verbal nouns (that can function as infinitives or nouns). The full rule-based grammar I designed can be found in the appendix.

2.6.4 Manual correction

No automatic NLP task is 100% correct. The rule-based chunkparsers performs very well with simple matrix clauses, but subordinate clauses and some complex DPs in particular need some correction. I manually corrected the entire corpus using CesaX. CesaX is a special software package developed by Erwin Komen to facilitate corpus-linguistic research (cf. Komen (2013)). The chunkparsed .psd-files can be converted to xml-files. These files can then be queried using CorpusSearch or the XML-based XQuery language. Manual correction in Cesax is quick and easy, because of its graphic representation of the tree structures. Alternatively, the bracket representation shown in figure 2.1 below, can also be edited manually if needed.

```
(S
 (DP (NP (N taryan))(ADJP (ADJ eur))(NP (N grwydyr)))
 (VP (PCL a)(VBD-3PL dodassant))
 (PP (P dan)(DP (PRO-G y)(NP (N penn))))
 (, ,))
```

Figure 2.1: Bracket representation provided per clause in Cesax

The above output from the automatic chunkparser reflects the following example:

(7) *Taryan eur grwydyr a dodassant dan y penn*
shield gold enamelled PRT put.PAST.3P under 3MS head
'They placed a gold enamelled shield under his head'(BM 1.18-19)

2.6.5 Annotating Information Structure

Information-structural features were added semi-automatically. In CorpusStudio (cf. Komen (2013)), various features can be automatically added. Information for these features can be derived from the PoS-tags of the specific words, from the phrasal structure or from the context in which it occurs. Since personal pronominal subjects usually convey 'Old' information, with some simple XQuery commands the referential status of these subject pronouns can be automatically labelled 'Old' (or, more specifically according to the Pentaset I adopt in Chapter 3, they will receive the 'Identity' label 'ID'). Other specific features of the clause such as the tense,

aspect or mood of the verb or the person-number inflection can be derived from the detailed set of PoS-tags in the same way.

Further information-structural notions such as topic or focus are not as easy to detect automatically. If special focus words or particles are used, the focus domain or articulation can be labelled accordingly. In addition to this, Constituent Focus in Middle Welsh could be indicated by a (reduced) cleft and a verb with default third-person singular inflection. Whenever there are pronominal subjects in the first or second person or plural full DPs, these structures can be automatically detected as well. When it comes to labelling the exact type of topic (e.g. familiar, aboutness or contrastive) or focus, much more manual annotation is required. These specifications were thus done at the very end using the strategies laid out in Chapter 3 taking the context into account.

All additional features (including the information-structural ones discussed here) are added at the matrix clause level. In practice, this means a list of features with automatically derived values (by querying the PoS-tags) and open values (to be adjusted manually) is available for every matrix clause. These features include:

- Focus Articulation, e.g. Constituent focus
- Focus particle/word, e.g. *hefyd* 'also'
- Point of Departure, e.g. Temporal clause 'At that moment...'
- Information flow, e.g. unmarked
- Referential State Subject, e.g. Old Information labelled 'ID'
- Referential State Object, e.g. New Information
- Diathesis, e.g. Impersonal verb
- Tense/Aspect, e.g. Preterite
- Mood, e.g. Indicative
- Semantic roles (in order), e.g. agent-patient
- Animacy & definiteness subject, e.g. definite-animate
- Animacy & definiteness object, e.g. indefinite-inanimate

2.7 Querying the data

There are various online tools available for corpus research, e.g. the search interface for the British National Corpus. Search interfaces provide easy access to the data, because no prior knowledge of specific search algorithms is necessary to get any results. The relevance and accuracy of these results can be questionable, however: these types of searches are often limited to the level of individual words or simple Part-of-Speech labels. If we want to gain a deeper insight in our linguistic data, we need a more thorough way of searching for the right information.

2.7.1 CorpusStudio and Cesax

CorpusSearch is an example of an application that can retrieve the detailed linguistic data relevant to syntacticians. It enables queries in the treebank or labelled bracketing format (.psd described above). A further way to retrieve detailed syntactic information is by converting the (parsed) files to XML-format (with the accompanying application CesaX, (Komen, 2013)) and query them with the usual search function for xml-databases: XQuery. Erwin Komen developed a wrapper around CorpusSearch2 (Randall, Taylor, & Kroch, 2005) and XQuery to facilitate these searches: CorpusStudio (Komen, 2009b). CorpusStudio not only simplifies the task of formulating search queries, it also provides easy ways to organise them along with the corpus data and research logs documenting your goals, subqueries, definition files and any emendations while gathering the right data.

2.7.2 Search queries for the present study

The main question in the present study concerns the word order of the sentence. The chunk-parsed files provide enough information to retrieve the main constituent order of all matrix clauses in the corpus automatically. This task is mainly one of categorisation: the possible word order types of Middle Welsh were described first. The query then systematically searched for the VP and the sentence-initial constituent (conjunctions and complementisers excluded). The order of queries for the different types of word order is of crucial importance. First the word order types with overt markers like sentences with focus markers or wh-question words need to be defined. Then sentences with periphrastic constructions can be distinguished from copular clauses (both using forms of the verb *bod* 'to be' with specific PoS-tags starting with 'BE'). After this, VP-initial clauses (however few in Middle Welsh) can be singled out, dividing them in their subcategories (Complementiser-V1, Conjunct-V1, Particle-V1 or absolute verb-initial). After this, the verb-second patterns can be categorised based on the phrase label of the sentence-initial constituent, e.g. sentence-initial PP or AdvP followed by a VP will be categorised as an adjunct-initial word order patterns. If the sentence-initial constituent is a pronoun or a noun, it will be categorised as an argument-initial order. It can further be specified as 'subject-initial' if it is a pronoun, because sentence-initial object pronouns do not exist in Middle Welsh. If the VP contains an inflected form of the verb *gwneuthur* 'to do' and the sentence-initial constituent contains a verbal noun, the sentence will be categorised as the specific periphrastic verb-second order with 'to do'. Finally, we can automatically detect sentences without VPs and categorise them as either 'non-verbal' or 'absolutive', if they contain the conjunction *a(c)* and are followed by a DP and DP/PP. The full search query can be found in the Appendix.

2.8 Interpreting the data

> *"Variation in grammatical choices exists not only through lexical, grammatical, discourse and situational context, as described in this chapter, but also for stylistic reasons (...). Speakers and writers are also creative with language (...). Given this complexity, if a rare choice is attested in a corpus, how are we to determine whether it is just a rare choice or an error?"*
>
> (Conrad, 2010:237)

2.8.1 On errors, examples and evidence

Conrad (2010) makes a valid point that has been discussed in philological literature over and over again. Errors are made in both speech and writing. If they end up 'uncorrected' in a manuscript we use as a source for our annotated corpus, how do we know if the peculiar form or pattern we find really existed? And even if it did, we can often not be sure why it only occurs once. In fact, we are unable to exclude the possibility that a particular form or pattern that does not occur at all in the corpus also never existed.

Before we can use examples from the corpus as 'evidence' for or against a certain hypothesis, it is important to be aware of the philological background of the specific text and manuscript. Theoretical syntacticians could thus benefit tremendously from close cooperation with philological experts when investigating historical stages of the language. Careful philological studies of scribal errors and emendations can be invaluable to the historical linguist as well when they help to estimate the date of origin of a particular text. A more accurate date of the texts can for example be established by comparing scribes of manuscripts of unknown dates with texts that refer to specific historical events. Scribal errors are furthermore indispensable in many cases, as succinctly put by Paul Russell in the context of the Welsh philological tradition: "the perfect scribe, who can standardise his orthography and not make errors, is the least useful for our purposes" (Russell, 1999:84).

Another important factor in what constitutes good evidence in corpus linguistics is a thorough understanding and description of the linguistic examples we find. A simplified example related to the present study on word order would be the following sentence with verb-initial word order:

(8) *Dos titheu ar Arthur y diwyn dy wallt.*
go.IPV.2S you to Arthur to cut.INF 2S hair
'Go to Arthur to cut your hair.' (CO 58)

The question is whether we could use this example to argue Middle Welsh had verb-initial word orders. The verb is clearly the first constituent in this sentence, so in principle we could. The statement would only be meaningful, however, if we are more specific. In this case, the imperative form of the verb is important, for example. In many languages with various types of basic word orders (e.g. Present-day English SVO, German and Dutch V2, Modern Welsh VSO, etc.), imperative verbs always

occupy sentence-initial positions. If we observe the same thing in this Middle Welsh sentence, it is first of all not surprising. More importantly, cross-linguistic evidence suggests that the fact that imperatives appear in sentence-initial position does not tell us much (if anything) about the 'basic word order' of the language (see Chapter 4 for a discussion of this notion).

Related to this is the issue of extrapolation in general: to what extent is an example we find in a corpus representative of the spoken language at a particular time. We can never know this with 100% certainty. Therefore, it remains important for anyone making claims about historical stages of a language to bear in mind that a 'corpulect' we work with can differ in various ways from the spoken language we try to describe. As discussed at length in the introduction about corpus linguistics above, this does not mean studying corpora is a futile endeavour or that we cannot trust our data or make any interesting observations. On the contrary, the very fact that we are taking a large amount of data into account (instead of studying one particular text) means that we can employ several statistical methods that can give us various kinds of new insights.

2.8.2 The use of statistics

'Statistics' are both loved and hated in the field of linguistics, not in the least, because the field is exceptionally broad and encompasses an incredible amount of research methods. It is important to bear in mind that statistics is a field of study in itself with its own developing theories and researchers advocating and/or aiming to disprove specific results, tools or methodologies. The historical corpus linguist already manoeuvring between philological expertise and modern linguistic theories, should also consult statisticians to evaluate their research outcomes properly.

Statistics can be used to estimate how likely it is that something would happen in a particular way. In the context of our word order research, for example, we could ask ourselves how likely it is that imperatives are found in sentences with verb-initial word order, compared to sentences with V2 or V3 orders. Statistical tools can furthermore help to establish and investigate certain correlations. Does an increased frequency of verbs with preterite tense inflection correlate with an increased frequency of a particular word order pattern, for example? If this is the case: what does that *mean*? Correlation does not equal causation, but observed correlations can give us useful information about the exact questions we need to ask to arrive at meaningful conclusions taking all possible variables into account. The use of statistics finally allows us to make inferences from a small sample of items to the large system they came from. Since we have no access to negative evidence in historical sources, this last part - if done properly - can be of great use for historical linguists.

Descriptive statistical methods

According to McEnery and Hardie (2012b:49), in most studies in corpus linguistics, only descriptive statistics are used. This type of statistics differs from inferential

statistics in that it does not test for significance. Frequencies are reported in absolute numbers or in a normalised way (often noted in percentages). The type-token ratio is furthermore often employed in corpus statistics. A token (any instance of a particular form/pattern in a text) is compared to the number of types of tokens (a particular unique form/pattern). This can for example be used to measure how large (in percentages) a range of vocabulary is used in a text. When comparing type-token ratios across different texts or corpora, the size must remain constant because it can affect the ratio (cf. McEnery and Hardie (2012b:50)).

Inferential statistical methods

Inferential statistical methods, on the other hand, do look for significance. This can be used to find out if the results we find (e.g. a certain number of examples of type X) are likely to happen under certain assumptions or not. Starting from the assumption that things are normal (the null hypothesis), we look at the collected results and calculate the probability that things would have happened that way by chance, if the null hypothesis is correct. The probability is a value between 0 and 1: the p-value. If the p-value is lower than a pre-agreed-upon threshold (usually 0.05 in the Social Sciences and Linguistics, but often 0.01 in Medical or Pharmaceutical Studies), the results are characterised as 'statistically significant' meaning that the null hypothesis is likely to be incorrect. In other words, the results we observe are probably not due to mere coincidence. This does not necessarily mean the results are in any way meaningful or interesting, it just shows we should reject our null hypothesis that says 'things are normal'.

This type of statistics for instance allows us to look at differences in the frequency of a construction in two different contexts and see whether it is significant. If it is, it would indicate that there is a connection between the two. The same can be done for constructions in two different time periods to provide evidence for change. This type of reasoning could also be extended to find the significance of 'negative evidence' (cf. McFadden (2014:14-15)): can we explain the fact that we do not observe a certain construction at all, because it is infrequent and the corpus is not sufficiently large or not?

In the present study, apart from descriptive statistics in the form of word order frequencies, I only employ two types of statistical tests: Chi-square and Fisher's exact test (the latter is used for low frequencies). The results of those merely serve as an indication of which factors should be looked at more carefully. To gain a better understanding of the distribution of word order types in Middle Welsh in various contexts, a Chi-square test can be used. This is a test specifically designed for qualitative data testing how likely it is that observed distributions are due to chance. This so-called "goodness-of-fit" statistic measures how well the distribution we observe fits the expected distribution if both variables are independent. The Chi-square test is thus specifically designed to analyse counted data divided into categories. The categories can vary in type: in this case the variables, for example, are the different types of word order and their distribution in the various texts in the corpus. But apart from that, I also check what other

possible factors have a significant interaction with the choice of word order type, e.g. information-structural factors such as referential status of the subject or object, but also grammatical factors such as tense, aspect or mood.

The null hypothesis in these cases is that these variables are independent. If the test renders a significant result, this is an indication that there is a possible interaction. It does not tell us why, but it does indicate that this is a fruitful direction for further investigation. If it is not significant, it indicates we do not have to control for this particular value making further comparisons: we do not necessarily have to keep that factor constant to gain a good insight in what is going on.

The formula of the Chi-square test (originally designed by Karl Pearson in 1900, cf. Plackett (1983)) compares the number of actual observations (O) to the expected frequencies (E). For each result, the chi-square value (χ^2) and the degree of freedom (df) is presented alongside the p-value[6]. Yates's continuity correction of -0.5 was added for contingency tables of 2x2 (cf. Yates (1934)) resulting in the following formula:

$$\chi^2_{Yates} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Figure 2.2: Chi-square formula with Yates's continuity correction

A disadvantage of the chi-square test is that it presupposes a normal distribution of the data, i.e. if most values cluster around a mean value to give a bell-shaped curve. Qualitative linguistic data is, however, usually not normally distributed: word frequencies, for example are typically positively skewed with a few high-frequency words and very many low-frequency words producing a long tale (cf. McEnery and Hardie (2012b:51-52)). This might lead to slightly inaccurate results. A somewhat more complex log-likelihood test (Dunning, 1993) does not make such an assumption and could therefore be a good alternative to the chi-square test. Another alternative (in particular when frequencies are low) is Fisher's Exact Test (McEnery, Xiao, & Tono, 2006). The formula for a 2x2 contingency table as shown below in Table 2.3 (with cells a, b, c and d and a total of N) for Fisher's Exact test is:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

Figure 2.3: Formulae for Fisher's Exact test

[6]All calculations were done with R statistics.

	V2	VSO	Total
Middle Welsh	a	b	a+b
Modern Welsh	c	d	c+d
Total	a+c	b+d	N

Table 2.3: Contingency Table

Since I mainly use statistics here to show potential interesting factors that interact with word order (see Chapter 5 for a complete overview), I only give the results of Chi-square and Fisher's Exact test here and leave the Log-likelihood tests for future research.

2.9 Conclusion

Building a linguistically annotated corpus is a tremendous task. This chapter first of all provides a thorough introduction to corpus linguistics focussing on the specific benefits of using well-annotated corpora in historical syntactic research. Exactly because the amount of extant data is extremely limited, we must try and retrieve the most information we possibly can. This can be achieved by first of all providing very detailed partof-speech tags. This elaborate morpho-syntactic annotation helps to automatically extract information about all kinds of grammatical and information-structural features.

In the latter part of this chapter I described each step in the process of creating an annotated corpus in detail, from selecting and preprocessing the texts to training a PoS-tagger for Middle Welsh to assign morpho-syntactic tags automatically. These annotated texts were manually corrected and prepared for chunkparsing with the NLTK rule-based regular expression parser. With an extremely detailed grammar and a double loop, hierarchical structures could be created to facilitate the syntactic queries concerning word order patterns. These automatic parses were again manually corrected and subsequently converted to bracketing formats to enable searches via CorpusSearch of XQuery. Samples of queries for word order patterns and feature values were also presented. A full annotation guide can be found in the Appendix.

I finally described some further benefits in terms of statistical analysis. For the present study, I only use a range of descriptive methods indicating the frequencies of word order patterns over time and two specific inferential methods: the Chi-square test and Fisher's Exact test. These options are fully explored in Chapter 5.

CHAPTER 3

Coding features relevant for Information Structure

If we want to determine to what extent - if at all - Information Structure (IS) relates to word order (change), we first need an adequate description of IS and its relevant notions in the grammar of historical Welsh. Although IS is a relatively new subfield of pragmatics (cf. Meurman-Solin, López-Couso, and Los (2012:3)), there is a vast literature on IS-related phenomena in a great number of languages. A general consensus on the exact definition of most information-structural notions expressed in the grammar is, however, still lacking.

Apart from defining information structure and its place in linguistic research, this chapter aims to provide an overview of those interpretive notions that are considered to be information-structural primitives. The grammar of a language has several means at its disposal to express information structure, but only those relevant to the present diachronic research will be discussed in detail.

Although recent overviews by Krifka (2008), Ritz, Dipper, and Götze (2008), Traugott and Pintzuk (2008) and, in particular, Götze et al. (2007) are insightful, there is no generally accepted or standardised way of coding IS features systematically yet. In this chapter I argue that any good description of the information structure of a language at the very least contains a detailed overview of how the grammar of the language expresses the core notions of **givenness**, **topic-comment** and **focus-background** (cf. section 3.3). I furthermore provide step-by-step guidelines on the procedures of coding those IS features. I conclude this chapter with a methodological note on the strategies implemented in the rest of this thesis to find the right mappings of information-structural primitives to the expressed word order types.

3.1 What is Information Structure?

> *"Terminological profusion and confusion, and underlying conceptual vagueness, plague the relevant literature to a point where little may be salvageable (...) In addition there is reason to think that the whole area may be reducible to a number of different factors (...)."*
>
> (Levinson, 1983:x)

The whole field of information structure (or, in fact, 'confusing' terminology like 'topic/comment' or 'theme/rheme') belongs to a long list of topics Stephen Levinson chooses *not* to discuss in his textbook on pragmatics. Some ten years later, Knud Lambrecht proposes a new theory of sentence formation, because there "still is disagreement and confusion" about information structure, a term he borrows from Halliday (1967) for a "grammatical component" of language. Another decade passes and Kruijff and Duchier (2003) are still concerned with the 'proliferating terminologies', to the extent that they find it necessary to add an insightful diagram to their paper visualising the 'terminological profusion and confusion' that seems to have haunted the field since the 1980s.

The *profusion* is indeed partly responsible for the enduring *confusion*. Using two (or three or even more) terms for one and the same phenomenon is often misleading. Employing just one of those terms to describe different phenomena at the same time is downright ambiguous. From that perspective, Vallduví and Vilkuna's *kontrast* with a *k*, no matter how well-argued for, perfectly illustrates the field's confused history (cf. Vallduví and Vilkuna (1998)).

Difficulty in *defining* information structure other than 'a subfield' (of pragmatics or semantics) contributed to the afore-mentioned confusion as well. Most collections of papers describing IS phenomena in various languages that bother to give a definition, resort to explaining what IS *does* or what it is *not*, rather than what it *is*. Examples of those information-structural effects include "encoding of the relative salience of the constituents of a clause" (Foley, 1994:1678), "presentation of information as old and new" (De Swart & De Hoop, 1995:3) and "packaging of information" (cf. Féry and Krifka (2008:2) following Chafe (1976)). Other common 'definitions' actually aim to identify the place of IS in relation to various linguistic notions, cognitive domains or as an in-between 'interface issue' (Mereu, 2009:2).

This brief introduction does not solve any issues in information-structural theory, it merely serves to illustrate the difficulty in choosing the right terminology on the one hand, and the necessity to give a detailed overview of the methodological considerations on the other. I use Zimmermann & Féry's definition of IS mediating "between the modules of linguistic competence in the narrow sense, such as syntax, phonology, and morphology, and other cognitive faculties which serve the central purpose of the fixation of belief by way of information update, pragmatic reasoning, and general inference processes." (Zimmermann & Féry, 2010:1). This notion is fully compatible with the Communicative model of Common Ground, which I use as a starting point for the present overview of the IS annotation guidelines (see section 3.2).

3.1.1 Brief history of IS research

The systematic study of the pragmatic organisation of discourse has its origin in the theory of the 'Functional Sentence Perspective' by the Prague Linguistic Circle initiated by Vilém Mathesius (1882-1945) (cf. Nekula (1999) and Mereu (2009)). His work on functional linguistics (Mathesius, 1929 [1983]) showed that the presentation of given material (the theme) and new material (the rheme) plays an important role in the structure of a language. Later scholars of the Prague School like Firbas (1964) employed the gradient notion of Communicative Dynamism (CD) to account for information structural phenomena, arguing that CD is responsible for the linear arrangement of syntactic constituents. Elements in the sentence with 'least CD' (i.e. the theme or topic or that which is contextually known) precede those with 'more CD' (i.e. those conveying new or unlinked information) (cf. Erteschik-Shir (2007:2)).

The notion of Common Ground (CG) was introduced by Paul Grice in the William James lectures of 1966-1967 as a term for the presumed background information or 'the context' of a conversation (cf. Stalnaker (1974), Grice (1989), Stalnaker (2002) and 3.2 below). Chafe (1976) first discussed semantic distinctions used in 'information packaging' (adopted in a formal context by Vallduví (1992)). Typological research in the late 1970s and 1980s by Li and Thompson (1976) and Mithun (1987) distinguished subject- and topic-oriented, or syntactically- or pragmatically-based languages. Givón (1984:204) argued that word order variation is "controlled by discourse-pragmatic considerations pertaining to new vs. old, topical vs. non-topical, discontinuous vs. disruptive information".

Following this, various researchers in the late 1980s and 1990s investigated focus structures (Abraham & de Meij, 1986) or topic structures (cf. Reinhart (1982), Lambrecht (1994), É.Kiss (1995), Dik (1997) and Büring (1997)) or a hierarchy of both topic and focus, see Payne (1987), Choi (1999), Frascarelli (2000) and Mereu (2009) for an overview).

In 2003, researchers from the universities of Potsdam and Berlin founded the 'Collaborative Research Center (Sonderforschungsbereich / SFB 632)' on Information Structure. Between 2003 and 2015, a grand total of 19 projects and 53 researchers aimed to formulate integrative models of information structure in various disciplines of linguistics and human cognition. They defined information structure as 'the structuring of linguistic information, typically in order to optimise information transfer within discourse.' (see the project description on their website `www.sfb632.uni-potsdam.de`). Research output of this centre focusses on the interaction of the relevant formal linguistic levels, general cognitive processing of information structure and finally on a cross-linguistic typology of information structural devices.

In an attempt to provide an insightful overview of what has by now become a (linguistic) field of its own, the *Handbook of Information Structure* will be published by Oxford University Press in the course of 2016 (Féry & Ishihara, 2016).

3.1.2 Where is information structure?

Information structure is usually mentioned as a subfield of pragmatics within the field of linguistics (cf. Meurman-Solin et al. (2012:3)), because it is related to language use: the relation of signs to those who interpret the signs. According to Kruijff and Duchier (2003:249), both utterance-internal (IS) as well as utterance-external semantic devices interact to provide the discourse context. IS is thus closely related to discourse analysis and semantics.

The question 'Where is information structure?' *in language*, rather than *in the field of linguistics* is far more interesting, but also more difficult to answer. Is it a 'grammatical component' as Lambrecht (1994:xiii) suggested? Is it part of (or encoded in) syntax, semantics or phonology? Or do IS phenomena operate on the interfaces of all of those (cf. Mereu (2009:2))?

Functional theories of language focus on what information structure contributes to the grammar (cf. Kuno (1987) and Dik (1997)). In a similar way, Role and Reference grammar, as employed by, among others, Van Valin (1993b), stores grammatical structures as constructional templates with specific sets of morphosyntactic, semantic and pragmatic properties, so that they are naturally linked (cf. Erteschik-Shir (2007:4-5)). Jackendoff (1972) and Horvath (1981) formalised discourse-semantic notions in structural relations, paving the way for discourse-configurational approaches (e.g. É.Kiss (2001)) in which topic and focus are linked to particular structural positions and thus part of the syntax. Further within Generative Grammar then, in particular in Rizzi's Cartography (cf. Rizzi (1997) and Rizzi (2004)), information structural features surface as separate projections in the sentence peripheries.

However, if information structure plays a role in semantics and phonology as well as in syntax, these representations make it difficult to express IS notions in a unified and systematic way. Alternatives to cartographic approaches by, among others, Neeleman and Van de Koot (2008) and Kučerová and Neeleman (2012) aim to solve this by mapping the syntax to the information structure at the interfaces. Multi-layered theories like lexical-functional grammar (LFG), head-driven phrase structure grammar (HPSG) or combinatory categorial grammar (CCG) take a different approach by formalising information structure in a way equal to the status of the other components of grammar (cf. Erteschik-Shir (2007:4)).

3.1.3 Main questions in IS research

As is clear from the above introduction, there are still many questions in information-structural research left unanswered. Even the exact object or unit of investigation varies from study to study. It is clear that IS phenomena can be observed by studying sentences in their context, but is that the only way? Can certain IS-related expressions also occur on the sentence or clause level, or possibly even on lower ranks of syntactic structure (cf. Kruijff and Duchier (2003:251))?. Information structure seems multi-modular and multi-levelled: an exhaustive investigation of IS phenomena in a language thus requires input from various aspects of the grammar

(syntax, semantics, morphology and phonology), but also from interacting cognitive domains (pragmatic reasoning, the fixation of belief and the update of information states, (cf. Zimmermann and Féry (2010:2)). In this chapter, I relate all coded IS notions to their grammatical markings as well as the way they function in our brain.

What are the basic notions or dimensions of IS?

Information-structural phenomena can be found in various parts of the grammar of a language, but what is it exactly that we are trying to find? The 'profusion' of terminology mentioned in the introduction hardly makes it easier to define the basic notions of IS. Recent IS literature, however, has not only described certain phenomena in a particular language, but also aimed to find the core dimensions or primitives of information structure. Kruijff and Duchier (2003:251) identify two recurrent patterns: "topic/comment" or "theme/rheme" and "background/kontrast" or "given/new". Zimmermann and Féry (2010:1) separate the second notion and claim that there are three basic concepts of IS:

— focus vs. background
— topic vs. comment
— given vs. new

Kučerová and Neeleman (2012:1) agree stating "these notions may require refinements and subdivisions, but there does not seem to be a substantial case in the literature for extending the set.". In other words, there seem to be no languages that, for example, have a separate class for elements that are neither new nor given with a specific syntactic distribution.

There is one important notion of IS, however, that has not been mentioned so far, namely 'contrast' (or 'kontrast', following Vallduví and Vilkuna (1998)). Intuitively, contrast is associated with an element of rejection or correction. Contrastive focus often emphasises one particular alternative. Repp (2010:1338) points out, however, that "contrast does not necessarily involve an element of rejection". In an earlier paper, Krifka (1999) already pointed out that contrastive focus can also be additive and furthermore, that contrast does not have to be associated with focus structures, because contrastive topics can also be found (cf. Krifka (2008)).

I therefore do not treat contrast as an IS primitive, but rather discuss the contrastive examples as they occur in one of the above-mentioned dimensions. These three dimensions will form the basis of my methodological analysis and IS annotation scheme.

How can IS be expressed in the grammar?

Knowing what to look for is one thing, knowing what it looks *like* in a language is a very different question. The great number of publications on IS phenomena is partly due to the many ways in which IS can be expressed. Examples can be found in a wide variety of languages in one or more of the following grammatical components:

- **Phonology**, in particular prosodic devices like pitch accent, deaccenting, and, as an extreme form of deaccenting, complete phonological reduction or ellipsis. Intonational phrases can also be used to indicate topics in English, German or Japanese (Krifka & Musan, 2012:34).

- **Morphology**. Some languages have special suffixes to mark, for example, VP focus, such as the perfective *-go* on the verb in Chadic (cf. Hartmann and Zimmermann (2007)) or the *no/gon* morphemes in Tsez (cf. Kučerová and Neeleman (2012:2)).

- **Syntax** can express IS phenomena in different ways: particular positions or word order patterns (e.g. fronting), agreement or the lack thereof (e.g. in a language like Tsez, cf. Kučerová and Neeleman (2012) or Middle Welsh, see Chapter 5) and specific constructions, such as cleft or pseudo-cleft sentences that are well-known in English.

- **Lexical items** related to certain IS phenomena come in various kinds: specific topic or focus particles, adverbials or determiners or anaphoric expressions.

What are the mapping rules between IS dimensions and expressions?

There are still many questions about the exact relation between information structure and the above-mentioned components of grammar. Some generalisations can be clearly formulated when it comes to IS and phonology: there seem to be no languages, for example, in which "old material must be stressed and new material de-stressed", which, according to Kučerová and Neeleman (2012:19), can hardly be a coincidence. The extent to which, and how exactly, IS is integrated in syntax and semantics is still an open question too, although "[T]here appears to be general agreement in the field that it would be more desirable for information structure and semantics to be part of the same system" (Kučerová & Neeleman, 2012:18).

The present thesis is concerned with the interaction of information structure and word order change. Therefore, although some elements of other grammatical components are coded, the syntactic way(s) of expressing IS in Welsh will be the main focus of my analysis. How this is implemented exactly will be discussed in Chapter 5.

3.1.4 Why study Information Structure?

Information structure is an integral part of human language, making the study of it invaluable in any effort to fully understand and describe the grammar and underlying mechanisms of a language. IS research can in particular shed light on variation and 'free' alternations found in languages, such as OV/VO word order, particle verbs (*He carried out the instructions. vs. He carried the instructions out.*) and the well-known dative alternation (*He give Sarah the book. vs. He gave the book to Sarah.*). Upon closer look at their information-structural status, these subtle

alternations very often turn out to be less 'free' than previously thought. Better insight in IS mechanisms can therefore be useful in the more applied field of L2 acquisition, designing textbooks and grammars that help learners to gain the much-desired native-speaker fluency (cf. Hannay and Mackenzie (2002) and Lozano (2006)).

But variation is also frequently encountered (and not always sufficiently explained) in diachronic data. Again, studying the information-structural properties of the specific alternations might shed more light on why changes in the language occurred, and, even more interestingly, why changes developed in one way and not the other. The information-structure background of the change in Welsh word order therefore serves as an excellent example.

3.1.5 Information Structure in diachronic data

Studying IS in diachronic data also has its limitations. Most of these have their origin in limited access to the data, which in turn, is only available in a limited form, i.e. only written sources survived. An additional problem for at least some of these sources is that we cannot always be sure to what extent they represent the language as it was used in a particular time or place (if, in fact, we know when and where that was in the first place) (cf. Meurman-Solin et al. (2012:10)). Is the manuscript version that survived merely a rendition of a story that clearly belonged in an oral tradition? If so, to what extent was it reworked - if at all - to fit the written medium? There is a clear stylistic difference between written and spoken language, so how can we evaluate any variation we encounter if we are not sure to which broad genre the text belongs in the first place? In general, the lack of information that may convey crucial IS differences such as intonation, is problematic. If prosody played an important role in marking IS patterns in the language, its impact is difficult to ascertain (although some research on prosodic phrases and stress patterns in historical data has been carried out (cf. Speyer (2008) and Hinterhölzl (2009)). Finally, the lack of native speaker judgments or possibility to run psycholinguistic experiments means traditional tests for specific IS patterns cannot be carried out. Certain particles or questions testing the scope of focus constructions, for example, such as *What happened?* or *Who did you see?* are simply not always available in the data (Traugott & Pintzuk, 2008:63).

We thus have to work with the data we have, limited as it may be, and a certain amount of caution is necessary in drawing far-reaching conclusions from results based on data with an uncertain philological background. As long as we are aware of what the data *can* tell us, studies of IS in diachronic data form an invaluable contribution to the description of older stages of the language and how it developed. Starting from the Common Ground, the rest of this chapter provides an overview of the most important notions of IS discussed above to describe the annotation scheme used for the historical Welsh database. The IS notions are discussed in relation to the two important elements of Zimmermann and Féry's (2010) definition of IS: their cognitive reality and the way they can be expressed or marked in the grammar.

3.2 Information Packaging & Common Ground

> *"Once upon a time there was a man who went to cut firewood in the forest above his village in the depths of winter. As he was cutting branches from a tree on the edge of a cliff he missed his footing and fell into the gorge, and resigned himself to a certain death on the rocks below. As it happened, there was a hibernating dragon in the gorge, and it opened its jaws in a great yawn just in time to catch the falling woodcutter."*
>
> (Ramble, 2013:75)

Storytelling, like any other act of discourse (reading a book, talking to a friend, listening to the radio, etc.[1]), involves the transfer of information. Successful communication of coherent discourse (making the reader/listener understand) depends at least partly on the optimisation of this information transfer, relative to the temporary needs of interlocutors (cf. Krifka (2008:15)).[2]

Stalnaker (1974) and Karttunen (1974) used the Gricean concept of Common Ground (CG) "as a way to model the information that is mutually known to be shared and continuously modified in communication" Krifka (2008:15). According to Krifka, the CG contains both a set of mutually accepted **propositions** as well as a set of **entities** that have been introduced into the CG before. As the discourse develops, the CG changes continuously and therefore the information has to be 'packaged in correspondence with the CG at the point at which it is uttered' (cf. Krifka (2008:16) following Chafe (1976)'s "Information Packaging"). As Stalnaker (2002) points out, the Common Ground is not necessarily the same as our Common Belief, i.e. the presuppositions of speakers, listeners, readers and writers. The Common Ground defines the context only, irrespective of whether the propositions uttered in a particular context are true or believed to be true.

There can be a divergence between the assumed context or Common Ground and people's actual beliefs. This is seen in Von Fintel's example of a daughter informing her father she is getting married with the words: "O Dad, I forgot to tell you that my fiancé and I are moving to Seattle next week" (Von Fintel, 2000:9). Even though the proposition about the engagement is new to her father, her daughter has decided to present it as old news in the context, because, for example, she does not want to discuss it further. Her father can then choose to grant his daughter's wish by accepting this context along with its subtext (i.e. she does not want to talk about it), even though their initial common beliefs about the daughter's relationship status were very different. Stalnaker (2002:716) therefore points out

[1] Note that I use the term "discourse act" in the sense of any piece of communication, both oral and written (cf. Di Eugenio (2003)). This linguistic interpretation does not include the Foucauldian sense of 'discourses of knowledge', which usually does not involve any textual analysis (cf. Fairclough (1992) and Bucholtz (2008)). Its use here is broader than just 'Conversation Analysis' in sociocultural linguistics.

[2] In spoken direct discourse like conversations, optimal communication is based on the cooperative principle of the four Gricean maxims of quantity, quality, relation and manner (Grice, 1989). Since the current study investigates historical data, I only focus on written texts in the rest of this discussion on discourse structure.

that the common ground "should be defined in terms of a notion of *acceptance* that is broader than the notion of belief". In the following section, I turn to how this kind of model of the Common Ground relates to text comprehension in our brain, this concerns the *accepted* context, irrespective of whether this corresponds to the parties' actual common beliefs.

3.2.1 Text comprehension in our brain

How do we interpret any form of discourse in the first place? The main reason we can understand the opening paragraph of the (originally Tibetan) woodcutter's tale cited in the beginning of this section is because we know the meaning of the individual words and because of the coherence between the sentences. Coherence between sentences (the systematically structured passages of discourse) is one "of the most fundamental characteristics of texts" (Schmalhofer, Friese, Pietruska, Raabe, & Rutschmann, 2005:1949). There are various ways in which textual coherence can be established, e.g. Schmalhofer et al. (2005:1949):

(1) a. anaphora resolution (cf. Glenberg, Meyer, and Lindem (1987))
    b. identifying overlaps in arguments of different propositions (cf. Kintsch and Van Dijk (1978))
    c. memory processes resonating for words with closely related meanings (cf. O'Brien, Rizzella, Albrecht, and Halleran (1998))
    d. inference processes driven by a search for meaning (cf. Graesser, Singer, and Trabasso (1994))

Psycho- and neurolinguistic experiments can provide insights on how our brain works when we are reading a text. Brain imaging techniques such as electroencephalography (EEG) measuring electrical activity of brain waves and functional magnetic resonance imaging (fMRI) can be used to shed more light on the processes mentioned in (1) (cf. Ferstl and von Cramon (2001) and Hagoort, Hald, Bastiaansen, and Petersson (2004)). Event-related potentials or 'ERP effects', in particular, are useful in linguistic research, because they are the results of the electrical activity of brain waves in relation to the event of interest (a word/sentence/construction etc) measured by EEG (cf. Luck (2005) and Sprouse and Lau (2013)). Negative and positive peaks in this EEG activity can indicate mismatches in particular linguistic domains. A problem in anaphora resolution, for example, yields a sustained negative offset after 300ms: the 'Nref effect' (cf. Van Berkum, Koornneef, Otten, and Nieuwland (2007:160) and Komen (2013:27)). To illustrate this, consider the first two sentences of the woodcutter's tale again in (2):

(2) a. *Once upon a time there was a man who went to cut firewood in the forest above his village in the depths of winter.*
    b. *As he was cutting branches from a tree on the edge of a cliff*
    c. *he missed his footing and fell into the gorge,*
    d. *and resigned himself to a certain death on the rocks below.*

When reading a sentence like (2a), we hold as much information as possible in our working memory. However, instead of trying to store the separate words we read, we try to extract the ideas they represent (cf. Kintsch (1989)). Following Komen (2013:28), I call the representational form of the linguistic expression we build in our mind a 'mental entity'. The syntactic phrase *a man who...* is the linguistic expression that first of all refers to this created mental entity. The mental entity in its turn refers to "real-world concepts or to imaginary ones" (Komen, 2013:28) or its denotation (cf. Krifka (2008)) (in this case a man who is cutting firewood). Zwaan and Radvansky (1998) show that we dynamically transform every part of the discourse into a "situation model" consisting of a set of participants (the mental entities) and a set of propositions (actions or relationships involving these mental entities) (cf. Van Dijk and Kintsch (1983) and Kintsch and Rawson (2005) on propositional representations in the situation model, or the similar "mental model" as it is called by Craik (1943), Johnson-Laird (2013)). Figure 3.1 shows a schematic representation of Mental Entities in the Situational Model applied to our woodcutter's tale.

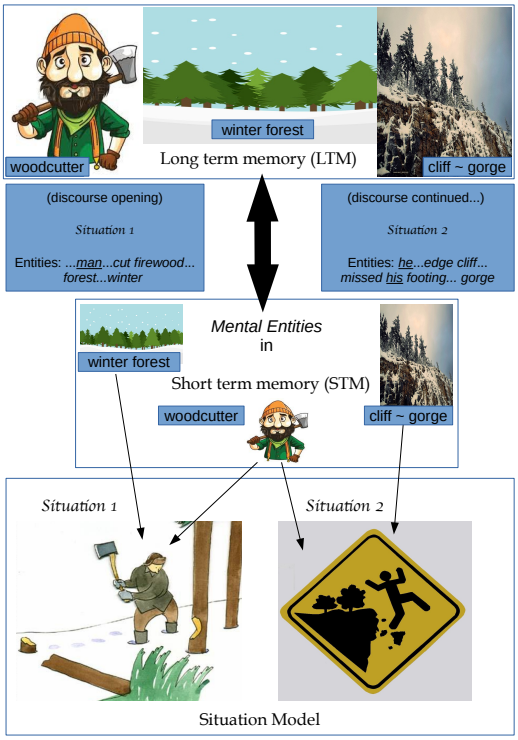


Figure 3.1: Schematic representation of entities in Long and Short-term memory

When we continue to read (2b), we dynamically update the model we built in our working memory describing the situation in which a particular mental entity, *a man*, is involved with certain propositions: he is cutting wood in a forest, the forest is above his village, it is the middle of winter, etc. As we parse (2b), we create a new mental entity in our working memory of the first linguistic expression we encounter, the pronoun *he*. Since pronouns are anaphors, we start a process of reference resolution (process (1a) above) in which we try to determine whether this mental entity matches with an already existing mental entity in the "situation model" (cf. Komen (2013:30)). In this case there is a perfect match with the mental entity we created to refer to *a man* in the previous sentence, so the features/characteristics of the phrase are added to the existing entity. Note that if (2b) were to have continued with *As she...*, we would have encountered a mismatch in gender (in English, *a man* cannot be referred to as *she*) resulting in the above-mentioned Nref effect in an experimental setting (as shown in various contexts by, among others, Van Berkum et al. (2007:160)). When we continue reading we further update our model with the propositions concerning the fact that the man is now cutting branches from a tree and that this tree is on the edge of a cliff, etc. Since the story goes on to relate how the same man *who went out to cut firewood* is now, in fact, *cutting branches from a tree*, there is a clear overlap in the arguments (see processes (1b) and (1c) above). The *edge of a cliff* in (2b) and the *gorge* in (2d) are another good example of this overlap in meaning. When parsing the rest of the sentence, we continue updating our model by adding and matching new mental entities and propositions. These propositions are not necessarily all found in the text itself: we can also access propositions that are stored in our long-term memory. We may for example associate *the depths of winter* in (2a) with a lot of snow, which in turn may result in a dangerous situation when you are busy working *on the edge of a cliff*.[3] We fully understand the following dramatic events in (2c), because we could make the right inferences (see process (1d) above) from the preceding context (i.e. working on the edge of a cliff in winter may be dangerous). We have just created a situation model in which the woodcutter is headed for a certain death, because he is falling into the rocky gorge. But now we continue to read this:

(3) *As it happened, there was a hibernating dragon in the gorge, and it opened its jaws in a great yawn just in time to catch the falling woodcutter.*

The scenario in which the man does *not* die was not part of our situation model: we did not expect this to happen especially not after the man himself pictured his 'certain death'. The developments in (3) are new and unexpected and we will have to create a new situation model containing the possibility of the man surviving the fall, or, at the very least, of the man not dying because he hit rock bottom, but because he was eaten by a dragon. According to Johnson-Laird (1989), it is easier to comprehend passages that lead unambiguously to a single model than

[3] For the potential audience of this particular tale, the inhabitants of the Tibetan plateau, this association will be even more accessible than for those living in much warmer areas of the world, but this only proves the point of 'optimal communication' in discourse.

passages that lead to multiple models. Again, we see that we do not just rely on the text to find the meaning of the passages, we also incorporate it in a broader context containing our knowledge of the physical, social and cultural world in which the discourse is presented. Bearing this in mind, the passage in (3) might be more accessible for the potential audience, the inhabitants of the Tibetan plateau (where dragons feature in many stories). Since we can only hold a limited number of models in our working memory at any given time (Johnson-Laird, Byrne, & Schaeken, 1992), we will soon discard the incorrect models to make room for new ones. With the next passage in the woodcutter's tale, we can finally reject the scenario involving the man's certain death:

(4) *The man survived the winter in the warmth of the sleeping dragon's maw, sustaining himself on the edible jewels that lay about the place in abundance.*

3.2.2 The Common Ground in our brain

The processes involved in text comprehension described in (1) were investigated in a combined ERP and fMRI study by Schmalhofer et al. (2005). The results allowed them to distinguish separate brain processes such as memory resonance (see (1c) above) and situational constructions (like the creation of situation models from mental entities, propositions and inferences, (1d) above). Later behavioural studies by, among others, C. L. Yang, Perfetti, and Schmalhofer (2007), point to the same results, separating the ERP-effects in even more detail. There is thus psycholinguistic evidence for the cognitive situation model as described above.

The *communication* model of the Common Ground (CG) discussed before contains both entities and mutually accepted propositions (cf. Krifka (2008)). The Common Ground is constantly updated: new entities and propositions are introduced as the discourse moves along. The propositions are not only derived from the discourse, but can also stem from common belief and world knowledge the interlocutors or readers have stored in memory. What Krifka (2008) describes as the Common Ground thus closely resembles the descriptions of the mental entities and propositions we use to build the situation or mental model in our brain, as we saw in the previous section. If this is indeed the case, the communicational model of the Common Ground has a cognitive correlate and at least some processes involved in information packaging, such as anaphora resolution (Van Berkum et al., 2007), foregrounding of information (Zwaan & Radvansky, 1998), topic identification (Kintsch & Rawson, 2005) or focus structures (Cowles, Walenski, & Kluender, 2007) can be measured by non-invasive studies of the brain.

The present study aims to describe Welsh information-structural processes and how they interact with the observed word order variation. As such, psycho- and neurolinguistic experiments that could further investigate the suggested correlation are beyond the scope of the present research. More detailed studies of IS and the Common Ground in many different languages can, however, certainly provide both inspiration and specific guidance concerning experimental settings that could show precisely how the communicational model of the Common Ground functions in our

brain.

3.2.3 CG content vs. CG management

So far we have mainly focussed on the *content* of the Common Ground, the set of entities and propositions that are known to and shared by the interlocutors or readers. Apart from this notion of CG content, Krifka (2008) introduces 'CG Management' for the way the CG content should develop. The CG management too is shared, but the responsibility for it "may be asymmetrically distributed among participants" (Krifka, 2008:17). This distinction between CG content and CG management can be observed in two different kinds of focus constructions that are called semantic or pragmatic focus respectively (cf. Krifka (2008:21)). Semantic focus is concerned with the factual information of the CG content; it can thus affect the truth-conditional content. Pragmatic forms of focus constructions serve the communicative goals of the participants and do not immediately influence the truth conditions. In section 3.3.4, I will get back to this division of the Common Ground with further explanation and examples of both types.

3.3 Coding Information Structure

In the previous chapter I discussed the technical side of developing an annotated database of historical Welsh. The texts are first of all digitised, PoS-tagged and chunkparsed and converted to xml-files to facilitate any queries into morphological or syntactic aspects. In addition to that, any information that could be relevant to information structure is added to each clause in the form of features rendering attribute-value pairs that are searchable as well (cf. Chapter 2). The following sections are concerned with these coded IS features. Which features were coded? Why those features and not others? And, finally, how were they coded? Which possible values belong to the feature attributes and how did I decide for one value or the other?

This chapter does not aim to provide an exhaustive overview of all IS terms and how they are used in the literature. Instead, it describes the strategies and definitions used in the present historical investigation of Welsh information structure. As a starting point, I assume that the information structure of every clause can be described as one of the following 'focus domains' or 'focus articulations' (cf. Lambrecht (1994) and Komen (2013), among others):

(5) a. THETIC focus (containing thetic and presentational sentences)
b. PREDICATE focus ('wide focus', 'information focus' or 'topic-comment' structure)
c. CONSTITUENT focus ('narrow focus' or 'identificational focus')

Lambrecht (1994) built on work by Gundel (1974) and Prince (1981) arguing that languages can focus three domains: the whole clause, the predicate of the clause or just a single constituent. In thetic sentences, both the subject and the predicate are

in focus (cf. Bailey (2009) and section 3.3.2). Predicate focus is the most frequently found focus domain, especially in narratives. It provides (new) information on an already established topic and is therefore often called 'topic-comment' structure (see section 3.3.3). Finally, in constituent focus one constituent is selected to be put against the background that forms the rest of the clause. The numerous ways of doing this will be discussed in section 3.3.4 below.

Both the referential state of the core arguments (see section 3.3.1) as well as syntactic and text-organisational (see Chapter 2) features help define the focus domain of the clause (cf. Komen (2013)). Two further pragmatic phenomena interact with each of the above-mentioned focus domains: the point of departure (or 'delimitation' or 'frame setting') and the principle of natural information flow (see section 3.3.6). Since the core arguments of copular clauses have a different syntactic configuration, I will discuss their information structural status separately in section 3.3.5.

The suggested IS annotation scheme thus covers multiple levels ranging from the referential state of the core arguments to the focus articulation of a clause, frame setting on a sentence level and discourse development in terms of cohesion of multiple sentences and paragraph/episode boundaries.

3.3.1 Given vs New: Referential State

> *"The origin of bees is from paradise and because of the sin of man they came thence; and God conferred his grace on them, and therefore the mass cannot be sung without the wax."*
>
> (Translation of *The Laws of Hywel Dda* by Wade-Evans (1909))

As we have seen in section 3.2 above, when we read a story we continuously add new entities and propositions to the Common Ground (cf. Chapter 2 of Komen (2013)). The to-be-added entities are first matched with whatever is part of the Common Ground already. If there is a perfect match with an existing entity in the CG, the features of the new phrase will be added to the existing mental entity, which is considered to be exactly identical. In the above fragment of a Welsh law text, for example, *bees* are introduced as a new entity and added as such to the CG. The third-person plural pronoun *they* a bit further on refers to the exact same entity as the *bees* that are just mentioned so they form a perfect match. The proposition in which the pronoun *they* occurs, the fact that *they came thence*, is now added to the mental entity we already created in the CG for *bees*.

But what about *paradise*, *the sin of man* and *God*? Neither of those were mentioned in the previous context, but we know nonetheless what they refer to. These entities are not identical to anything we previously added to the Common Ground. There is no textual antecedent; in other words, the denotations are assumed to be part of the 'world knowledge' of those living in a Christian society at least. Therefore they are stored in our long-term memory. This is exactly why the definite article can be used in the phrase *the sin of man*. We are not talking about a random sin. This is *the* sin everyone knows about: the reason man and, according to this Welsh

law, also bees, had to leave paradise. The definite article in *the mass* is there for the same reason: this concept is assumed to be known by the reader and is therefore not a completely new piece of information. The final phrase *the wax*, however, is not necessarily part of the assumed Christian model in our minds. Furthermore, when we try to match this with the existing entities in the Common Ground, we fail to find an exact match. The first entity we added (*bees*), however, evoked a link to a model of bees that we store in our long-term memory (e.g. bees are insects, they fly and buzz, they make honey, etc.). We can easily infer the existence of *wax* from the *bees* we already have in our Common Ground, so *the wax* in this example does not convey completely new information either.

This brief interlude about the importance of bees in Welsh laws serves as an introduction to one of the most crucial dimensions of information structure: givenness. From the early days of research into information structure, 'givenness' in its various forms has played a crucial role. The degree of Communicative Dynamism, as Firbas (1964) called it, is what pushes communication forward. Chafe's (1976) cognitive theory distinguishing degrees of givenness was extended by Yule (1981), among others. And in more recent literature, 'givenness' is (the extent to which a particular phrase is) 'existentially entailed by the context' (cf. Zimmermann and Féry (2010:2) following Schwarzschild (1999)). Krifka (2008:37) defines it in relation to its presence in the Common Ground, and/or the degree to which the particular referent is present. The same gradient notion we already encountered identifying some constituents as 'not completely new' in the introductory Welsh law text is found in the definition by Traugott and Pintzuk (2008:64): "the degree to which a referent is represented as identifiable by the addressee/reader and is "hearer/addressee-old". Gregory and Michaelis (2001) distinguish givenness from what they call 'anaphoricity', which is concerned with textual reference only, rather than the hearer's cognitive status.

In theory, the givenness or information/referential state of any kind of discourse referent can be assessed, but for the purpose of the present thesis only the core arguments of the sentence will be annotated. The 'information status', as Götze et al. (2007) call it, reflects the retrievability of the referent: how difficult is it to find an antecedent? Is there an identical match, can we infer or assume its existence? or is the noun phrase we are currently adding to the Common Ground not linked to anything at all? As we have seen in the introductory analysis of the bee fragment, there must be more than a simple binary option of given vs. new.

To capture this gradience a wide variety of taxonomies and hierarchies were developed over the years: Prince's (1981) taxonomy of given-new information or information states of noun phrases elaborated and refined by Birner (2006) into discourse and hearer old-new distinctions, Riester, Lorenz, and Seemann (2010)'s detailed set combined with semantic information, Ariel (1999)'s accessibility marking scale, Gundel, Hedberg, and Zacharski (1993)'s givenness hierarchy or the tag sets for PROIEL (Haug, 2009) or Cesac's Pentaset (Komen & Los, 2012:21,23) (see Komen (2013:133-154) for a detailed overview and evaluation of each of those).

Komen (2013) shows that a combination of syntactic annotation and a small set of five referential state primitive suffices to capture all relevant degrees of givenness. In this thesis, I employ this same 'Pentaset' to enrich the core arguments in the Welsh historical database. Komen's primitives are very similar to the PROIEL tag set (Haug, 2009), Birner's discourse/hearer distinctions (Birner, 2006) and to those suggested by Götze et al. (2007) in their Linguistic Information Structure Annotation (LISA) guidelines, although the latter is unable to capture certain subtle differences concerning anchoring (see below).

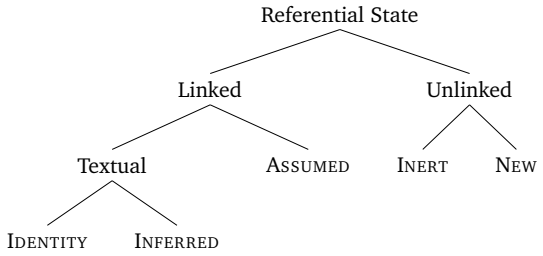
Taylor and Pintzuk (2014) test the effect of various annotation systems on Old English pre- and post-verbal objects. They find three significant differences: (i) between elaborating and bridging inferables, (ii) between specific new referents and short-term discourse referents and (iii) between short-term referents and semantically incorporated objects (Taylor & Pintzuk, 2014:72). As for (i), only Birner (2006) makes this distinction directly. In the Pentaset, however, the most-frequent cases of elaborating inferentials (the ones with inalienable possession) are marked with an Identity anchor (see discussion in the next section) and can thus be distinguished from bridging inferables. The next significant difference found between specific new referent, short-term referents and incorporated objects (numbers (ii) and (iii) above) fall in the Inert category in the Pentaset. They can be distinguished from other inert categories on the basis of their syntax and further featural annotation only. For the present study I used the Pentaset labels, because it makes more precise and clearer distinctions than the PROIEL or LISA annotations guidelines. In future research, it would be interesting to test Birner's (2006) distinctions on the Welsh dataset as well to see if there are similar significant results as the ones found for Old English object position by Taylor and Pintzuk (2014).[4] The main strength of the Komen's system is its ability to *derive* topic and focus structures from the IS and syntactic annotation combined. No additional assumptions have to be made to detect the right focus domain of a clause and it can even be extended to investigate copular clauses (the IS analysis of which is by my knowledge not specifically discussed elsewhere). In sections 3.3.3 and 3.3.4 below, I further develop the IS annotation system so that it can cover even more specific IS concepts such as the many different types of focus Krifka (2008) discusses, but also contrastive topics.

The Pentaset of referential state primitives

The referential state primitives that make up the pentaset are the minimal labels necessary to derive any other taxonomies or topic or focus domains (see Chapter 5 of Komen (2013) for a detailed overview). In this section, I provide definitions and examples for each of those five primitives. I furthermore point out subtle differences with the LISA guidelines by Götze et al. (2007). This is the Pentaset hierarchy (after Figure 11 in Komen (2013:144)):

[4] Taylor & Pintzuk's test results were published when the annotation with the Pentaset of the Middle Welsh database was already done.

(6)



The Pentaset is couched in the situation model (or Common Ground) discussed in section 3.2.1 above. The system first of all distinguishes noun phrases *with* an antecedent ('Linked') from those *without* ('Unlinked'). If there is a phrase (NP_i) referring to a certain mental entity $\texttt{MEnt}(NP_i)$ and there is another phrase (NP_j) that refers to the exact same mental entity of NP_i and NP_j linearly precedes NP_i, there is a perfect match with an already existing mental entity in our situation model. In this case, NP_i will receive an IDENTITY label, because its mental entity is identical to the mental entity of NP_j that already existed in our model. An example of this is a pronoun referring back to the mental entity created by a previously-mentioned NP. The formal definition of the IDENTITY label is, according to Komen (2013:144):

(7) **Identity**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ has the referential status "Identity" if there is an NP_j with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$.

The *bees* in the introduction that were matched by the pronouns *they* and *them* further on are a clear example of this. Götze et al. (2007) further divide this category, which they call 'given' into 'active' and 'inactive' referents. 'Active' referents are those that are referred to "within the last or in the current sentence" (Götze et al., 2007:154). There indeed seems to be a difference in terms of accessibility the further you move from the antecedent. The sentence boundary, however, is a somewhat arbitrary notion. In many medieval manuscripts, for example, it may be hard to divide the text into clear sentences in the first place. Clause boundaries are easier to define, but there can be multiple subordinate clauses in one sentence, so cutting off at one, two or three clauses or even one matrix clause remains a random decision. It remains unclear, however, whether "one sentence" is meaningful as an IS notion here.

Looking at the last-mentioned possible antecedent could be a more meaningful distinction, but even that may vary from language to language. Grammars can act differently if they have no (rigid) gender or number marking in the nominal system, for example, from those with 'rich' morphological paradigms of pronouns and demonstratives. I leave this as an open question for now, because for the present investigation, this particular distinction is not relevant. In the present thesis I will stick to the simple IDENTITY label for any referent that has an exact match with a mental entity that is referred to in the previous context. In long narratives

featuring the same main characters over and over again, I will furthermore indicate whether the specific referent occurred in the same *scene* or not. A change in location or setting is a clear indication of a scene boundary. If the hero of the story disappears for a while, for example, because the narrative changes its focus for a few paragraphs, we replace the model we created in our mind. The same hero can then be identified later on, but the scene has changed so the particular noun phrase will receive an additional label: IDENTITY - CHANGE OF SCENE as the subject *y mab* 'the boy' in this example following a scene in which the father of the boy gives his son advice on how to find Olwen:

(8) *Mynet a oruc y mab ar orwyd penlluchlwyt...*
go.INF PRT do.PAST.3S the boy on steed gleaming-grey-head...
'The boy went off on a steed with a gleaming grey head...' (CO 60)

Antecedents can occur in the text, but they can also be part of the general 'world knowledge' stored in our long-term memory. Entities in our long-term memory can be evoked and become part of the Common Ground. When this type of link to an entity in long-term memory can be created, the referential state of the mental entity that is added to the situation model is ASSUMED. Komen (2013:147) gives the following formal definition of the categorie ASSUMED:

(9) **Assumed**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ is "Assumed" if
a. there is no NP_j with j<i, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$
b. nor such that $\texttt{MEnt}(NP_j)$ can be inferred from $\texttt{MEnt}(NP_i)$, but
c. there exists an $\texttt{MEnt}(NP_{LTM})$ (in long-term memory), such that $\texttt{MEnt}(NP_{LTM}) = \texttt{MEnt}(NP_i)$

We have seen examples of this in the fragment on the origin of the bees above: *God*, *paradise*, *the mass* and even *the sin of man* do not need a textual antecedent to be meaningful to a reader who is familiar with at least the basic background of the Christian faith. This is considered 'world knowledge', just as much as we all know the sun, moon and stars exist. Situational knowledge about the speaker, hearer, the book that is being written or the setting in which the sentence is uttered, also belongs in this category. Imagine, for example, a conversation over lunch where one person points to the box on the other side of the table and asks:

(10) "Could you pass the chocolate sprinkles, please?"

The noun phrase *the chocolate sprinkles* has an antecedent, even though it was not mentioned in previous discourse. The other person can see the box of chocolate sprinkles on the table, so the new mental entity of the noun phrase will match the referent in the currently relevant situation. The referential state of the phrase *the chocolate sprinkles* is thus ASSUMED. In the extended tag set of the LISA guidelines, Götze et al. (2007) create a special label for referents that are part of the discourse situation such as *the chocolate sprinkles*: 'accessible-situative'. Since it is unclear if

and why matching with something in our long-term memory or with the situation at hand would make a difference, I will stick to the Pentaset label for referents whose information status is ASSUMED.

It can also be the case that there is no direct match with a textual antecedent or an antecedent in the current situation or long-term memory, but the information referred to is not completely new either. In the introductory fragment, *the wax* was an example of this, because we could establish a link with the afore-mentioned *bees* via the model concerning bees in our long-term memory that was evoked as soon as we read about them. In other words, we could infer the existence of *the wax* from the *bees*. When this form of logical reasoning is necessary to establish an entity, Komen (2013:146) defines it as INFERRED:

(11) **Inferred**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ has the referential status "Inferred" if
(i) there is no NP_j with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$, but
(ii) there is an NP_k with $k<i$, such that:
a. $\texttt{MEnt}(NP_i) \in S_x$
b. $\texttt{MEnt}(NP_k) \in S_y$
c. there exists a *direct set relation* between set S_x and S_y.

A direct set relation, as used in this definition can for example occur in the form of a subset, a part-whole relation or as an entity-attribute relation as in the following examples:

(12) a. Deryn hates working close to the microwave. *The noise* is distracting.
b. Asiye loved the Turkish chocolates. *Their flavour* was so soothing.

The italicised noun phrases in examples (12a) and (12b) create mental entities that are not *identical* to anything in our situation model or in long-term memory. We can, however, create a link to the existing entities, *the microwave* and *the Turkish chocolates*, because there exists a direct set relation: microwaves make a lot of noise and chocolates have flavours. If there is no antecedent in the context (IDENTITY), in our long-term memory or direct situation (ASSUMED) and if we cannot infer the existence of the referent from anything previously mentioned (INFERRED), the referential state of the phrase is 'Unlinked'. The Pentaset further differentiates the 'Unlinked' category: referential phrases that could serve as an antecedent in the following discourse are labelled NEW (Komen, 2013:150):

(13) **New**
A constituent NP_i with mental entity $\texttt{MEnt}(NP_i)$ is "New" if
a. there is no $\texttt{MEnt}(NP_j)$ with $j<i$, such that $\texttt{MEnt}(NP_j) = \texttt{MEnt}(NP_i)$,
b. nor such that $\texttt{MEnt}(NP_j)$ can be inferred from $\texttt{MEnt}(NP_i)$, but
c. it **is** possible that there exists an NP_k with $k>i$,
such that $\texttt{MEnt}(NP_k) = \texttt{MEnt}(NP_i)$.

New entities are usually introduced as indefinite noun phrases or phrases with postmodifiers, as in the following examples:

(14) a. *Ac yno ti a wely lwyn.*
and there you PRT see.2S grove
'And there you will see a grove.' (Peredur 294)
b. *A ffon yssyd idaw o hayarn*
And stick be.3S to.3MS of iron
'And he has an iron stick.' (WM 228.23-24)

There are also phrases that can *not* be referred to in the following context. Usually, they function as attributes of other entities. Götze et al. (2007) do not have a specific label for these expressions in the LISA guidelines, exactly because of this reason: they do not annotate "NPs or PPs that don't refer to discourse referents". Examples of non-referential expressions are expletives or parts of idiomatic phrases or attributes as in:

(15) a. Mabon son of Modron is here in *prison*; and none was ever so cruelly imprisoned in a prison house as I.
b. Maxen Wledig was emperor of Rome, and he was *a comelier man*.

In example (15a), the prisoner Mabon is shouting from within his confined space in a cry for help. The noun phrase *prison* in the first part of the sentence refers to the general concept of his confinement. The phrase is INERT: it cannot serve as an antecedent for the following discourse. Similarly, in example (15b), the noun phrase *a comelier man* cannot be picked up later on. A following sentence starting with *The man went hunting.* sounds odd at the very least (cf. Johnson-Laird (1983) and Komen (2013)).

To sum up this section, I give a full analysis of the referential status of the most important noun phrases in the following fragment from the translation of *Culhwch ac Olwen*, the oldest Arthurian tale. The immediately preceding context relates how Arthur was hunting a wild boar called Twrch Trwyth. The boar has fled to Ireland and Menw tried to capture it, but failed, upon which Twrch Trwyth destroyed a large part of the country. There is a brief intermezzo about a magic cauldron and then...

(16) ***Arthur*** *came to* ***Esgeir Oerfel*** *in* ***Ireland,***
to the place where ***Twrch Trwyth*** *was,*
and ***his seven young pigs*** *with* ***him****.*

Arthur is one of the main characters of the tale and is also mentioned in the immediately preceding context. The referential state is thus IDENTITY. *Esgeir Oerfel* on the other hand, is NEW in this context. It was mentioned once or twice in the beginning of the tale, but since there were many different scenes in between and this place does not play any significant role in the tale, it is unlikely that this is still in our situation model. If this was a famous place in Ireland, a medieval Welsh audience might have stored it in their long-term memory, rendering its referential

state ASSUMED. As for *Ireland* itself, this too was mentioned in the immediately preceding context, so this, just like *Twrch Trwyth* and *him* at the end of the first sentence, is labelled IDENTITY. Finally, *his seven young pigs* bring a new entity into our situation model, because these pigs were not mentioned before. The phrase is linked to the wild boar Twrch Trwyth in two ways. First of all, we can establish an inferential relation between pigs and wild boars, because boars (can) have pigs. The referential state will thus be INFERRED. But there is another element in the phrase linking it to this wild boar in particular: the possessive pronoun *his*. Following Prince (1981) and Komen (2013), I call this an "identity anchor". This anchor can be added independently: the full referential state of *his seven young pigs* will thus be INFERRED + IDENTITY ANCHOR.

(17) a. ***Dogs*** *were let loose at* ***him*** *from* ***all sides****.*
b. ***That day*** *until evening* ***the Irish*** *fought with him; (...)*
c. ***His men*** *asked* ***Arthur*** *what was* ***the history of that swine****, and he told them:*
d. *'He was* ***a king****, and for* ***his wickedness God*** *transformed him into a swine.'*

When we continue reading, we find *dogs* in (17a), which forms a NEW mental entity in our situation model, as opposed to the pronoun *him*, which forms a perfect match with the wild boar we have seen before and is thus labelled as IDENTITY. The phrase *all sides* is not linked to anything either, but this phrase does not add an entity to our Common Ground, because it is non-referential. It cannot serve as an antecedent in the following discourse, so we will label it as INERT. The first phrase in (17b), *that day* is ASSUMED, because it is part of the current situation. We can infer the existence of *the Irish* from the previously-mentioned *Ireland*: countries have inhabitants, a country called 'Ireland' has inhabitants that are called 'the Irish'. Its label is INFERRED. In (17c), *his men* form a new mental entity, because these men just come to the scene. There is a possessive pronoun, however, that links this phrase to *Arthur*. Therefore it will get the label NEW + IDENTITY ANCHOR, making it more accessible than new entities without any form of anchoring in the previous context. The same goes for *the history of that swine*: the *history* is NEW, but the *swine* is already well-established in our model, so this too gets the label NEW + IDENTITY ANCHOR. This is also the case for the phrase *his wickedness* in (17d). The phrase *a king* is not linked to anything, but again, it is very difficult to see how this phrase could serve as an antecedent. It is a clear example of an attributive indefinite noun phrase in the complement position of an equative clause and therefore INERT. *God*, finally, is an entity that we can link to a concept in our long-term memory and it is therefore labelled as ASSUMED.

3.3.2 Presentational or Thetic structures

Once we have annotated the morphology (see PoS-tagging in Chapter 2), basic syntactic structure (see Chunkparsing in Chapter 2) and the referential state (see section 3.3.1), we can derive the focus domain from this combined information

(Komen, 2013). Presentational or thetic sentences focus both the subject and the predicate. If the sentence merely consists of a comment and there is no core argument constituent that could be the topic, the sentence is called 'thetic' (cf. Krifka (2008:43) following Marty (1884)). I therefore label the clause's focus domain as THETIC FOCUS. Krifka (2008:43) gives the following example of a sentence without a topic constituent in a situation where somebody is running towards you in a panic, for example, shouting:

(18) [*The HOUSE is on fire.*]$_{\text{Comment}}$

There still is a topic *denotation* in this clause, the sentence is still 'about' something. But there is no constituent expressing this, because the entire sentence consists of a comment explaining someone's panic. Both the subject and the predicate, convey new information. Another example of a thetic statement is:

(19) [It is raining]$_{\text{Comment}}$.

The topic of sentences like (19) is also called a "stage topic" (cf. Gundel (1974) and Sasse (1987)), because it predicates about the 'here and now'.

Presentational sentences are similar in that they also contain subjects and predicates that are NEW. They are relatively easy to recognise, because they introduce a new entity into the discourse. Very often, these sentences occur at the beginning of narratives:

(20) *In the days when Maelgwn Gwynedd was holding court in Castell Deganwy, there was* ***a holy man named Cybi*** *living in Môn.*

In this opening passage of *Ystoria Taliesin* from the 16th-century Chronicle of the World by Elis Gruffudd, a new entity is introduced, namely *a holy man named Cybi*. The preceding prepositional phrase *In the days...* functions as a point of departure or 'frame setting' (see section 3.3.6), but the focus domain is determined by the rest of the clause in which a new entity is introduced as the subject. The focus domain of this clause comprises the subject and the predicate and it is thus labelled THETIC FOCUS as well.

Other examples of thetic focus will be discussed in section 3.3.5 below. For now it suffices to say the thetic focus domain can be detected when the subject contains NEW information and the predicate is also part of the focus domain. Komen (2013:42) furthermore adds that thetic focus can be overridden by constituent focus. This means that if the subject is, for example, providing the value for a variable that has just been raised, the sentence does not belong to the thetic focus domain, but receives the label of CONSTITUENT FOCUS (see also section 3.3.4). The example Komen (2013:42) gives is the following dialogue:

(21) a. "Who would want to listen to you?"
b. "*An educated man* will read my books!"

The italicised noun phrase in (21b) provides the value for the variable created by the question in (21a). The predicate *read my books* in (21b) furthermore does not contain completely new information, because the verb *read* can be inferred from *listen* in (21a) (cf. Komen (2013:42). In this case, the clause in (21b) is thus an example of a CONSTITUENT FOCUS domain. Apart from CONSTITUENT and THETIC FOCUS, there is a third type of focus domain called PREDICATE FOCUS for topic-comment structure. This domain is discussed in the next section.

3.3.3 Topic vs. Comment

Consider the following fragment from the tale of Branwen, the second branch of the *Mabinogion*, translated by Lady Charlotte Guest and try to think of what this passage is *about*:

> *"In Ireland none were left alive, except five pregnant women in a cave in the Irish wilderness; and to these five women in the same night were born five sons, whom they nursed until they became grown-up youths. And they thought about wives, and they at the same time desired to possess them, and each took a wife of the mothers of their companions, and they governed the country and peopled it. And these five divided it amongst them, and because of this partition are the five divisions of Ireland still so termed. And they examined the land where the battles had taken place, and they found gold and silver until they became wealthy."*
>
> (Guest, 1849)

The most logical answer is that it is about five sons who grew up to 'people' Ireland: that is the topic of this piece of discourse. There is a vast literature on different kinds of topics including various definitions, functions and ways to express them. In this section, I discuss only those notions relevant for the present thesis. Starting with a definition of topic by Krifka (2008) (following Reinhart (1981)), I continue to characterise the most frequently found focus domain called PREDICATE FOCUS that consists of the basic topic-comment structure and whose frequent occurrence in narratives makes sense from a cognitive point of view. Finally, I describe different kinds of topics in sentences and discourse and how they can be marked in the grammar.

Topics and the Predicate focus domain

Krifka (2008:41) defines topic constituents in the following way:

(22) "The topic constituent identifies the entity or set of entities under which the information expressed in the comment constituent should be stored in the CG content."

The content of the Common Ground thus plays a crucial role. The propositions in the CG are stored under certain entities just like the file card system proposed by Reinhart (1981) and Vallduví (1992). Other definitions of 'topic' containing 'subject' (cf. Chafe (1976)) or 'theme' conflated with 'old information' (cf. the Prague

School, e.g. Daneš (1970)) should according to Krifka (2008) and Zimmermann and Féry (2010) be avoided, because they are not necessarily grammatical subjects or inferable from the preceding context.

There are various ways to find topics described in the literature. Gundel (1988:210)'s definition comprising the speaker's intention "to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to" the topic of the sentence is an intuitive working definition, but it does not give any concrete guidance on how to identify topics. Götze et al. (2007:165) formulate three conditions identifying aboutness topics X in sentence S if:

(23) a. S would be a natural continuation to the announcement: "Let me tell you something about X."
b. S would be a good answer to the question: "What about X?"
c. S could be naturally transformed into the sentence "Concerning X, S'." or into the sentence "Concerning X,S'," where S' differs from S only insofar as X has been replaced by a suitable pronoun.

Eckhoff and Haug (2011) are more precise and formulate an algorithm that ranks constituents that are possible topic candidates according to parameters such as their referential status, animacy, morphosyntactic realisation, saliency, syntactic relation, word order and antecedent properties. The strength of this algorithm lies in the combination of those features yielding 90% agreement between the outcomes of their algorithm and that of human intuition. In a similar way, the Cesac application (Komen, 2009a) attempts to detect topics based on the type of NP and their grammatical function (subject, object, etc.). Centering theory finally, (cf. Grosz, Weinstein, and Joshi (1995), and in particular the OT type of centering discussed by Beaver (2004)), is according to Komen (2013) a particularly successful way to find the topic of a sentence. It ranks the topic candidates according to their category (e.g. demonstrative, pronoun, definite noun phrases, etc.), the referential state of the phrase (linked or not) and their grammatical role (e.g. subject or object).

In the present research, all these notions (and more) are annotated in the database to facilitate the search for topics in each sentence, separating them from the rest of the clause that makes up the comment. In terms of focus domains, this topic-comment structure differs from the above-mentioned THETIC sentences in the sense that the latter always contain subjects (and predicates) conveying new information: both subject and predicate are in focus. In topic-comment structures, the focus domain is the predicate that conveys the NEW information. This is also called 'wide' or 'information focus' (e.g. É.Kiss (1998)), but following Lambrecht (1994) and Komen (2013), I label this domain PREDICATE FOCUS.

Why exactly is this type of focus domain the one we find most frequently in narratives? Psycholinguistic experiments (e.g. Gernsbacher (1990)) have shown that from a processing perspective, the predicate focus domain with the topic-before-comment structure is likely to be the most commonly used, since language

is processed in a largely incrementally way. It furthermore makes sense to present linked information before unlinked information in the predicate. In this respect, it is also interesting to look at VOS and, in particular, VSO languages and their topic-comment distribution, because the verb in the latter case is fronted leaving the direct object behind and thus the focussed predicate is split up (more on Modern Welsh VSO is discussed in Chapter 5). As Cowles (2012) puts is: "when we encounter or produce a sentence we begin to process it right away, at the beginning, without waiting for the entire sentence to be available for either production or comprehension." (Cowles, 2012:290). This first information to be processed is very often the given referent, but there is also evidence from German that sometimes new information may be ordered first (cf. Cowles (2012)). For the present research, it suffices to say that two IS notions that seem particularly relevant in topic-comment structures, namely givenness (see section 3.3.1) and accessibility (see Chapter 2) are annotated separately. If topic-status is, as these production studies indicate, indeed assigned at the pre-linguistic message level, we need to investigate how this can be encoded in the grammar in general. In this thesis I show how this can be done in earlier stages of the Welsh language and how this changed over time.

Finding the focus domain

PREDICATE FOCUS is the most frequently found focus domain in narratives, as we have seen in the introductory fragment about the five sons. Every predicate of the following sentence adds new information, a new file-card if you will, to the existing entity: the sons want wives, get married to each other's mothers, govern the country, etc. We can find this focus domain of the sentence by following a decision-making tree based on the combined syntactic and referential state information of the core constituents of the matrix clause. It is also possible to determine the focus domain of subordinate clauses (see Chapter 2), but here we try to determine the focus domain of matrix clauses first.

First of all, we make sure we are not dealing with a thetic or presentational sentence by asking the following questions:

(24) Is there a topic constituent?
  (i) Yes ⇝ Move on to (25)
  (ii) No ⇝ Are both subject and predicate new?
    (i) Yes ⇝ THETIC FOCUS
    (ii) No ⇝ Start over (something went wrong).

(25) Is there a new entity introduced into the story?
  (i) Yes ⇝ THETIC/PRESENTATIONAL FOCUS
  (ii) No ⇝ Move on to (26)

After ruling out the domain of THETIC FOCUS, we check if we are dealing with a copular clause (see section 3.3.5). If this is not the case we continue to ask whether the sentence forms part of a dialogue with a whole set of further questions to rule out various types of CONSTITUENT FOCUS (see section 3.3.4 below). If the sentence

is not part of a dialogue, we first of all see if this is a case of a contrastive topic (see section on Types of Topic below). Finally, we distinguish between the domains PREDICATE and CONSTITUENT focus by asking whether there are relevant alternatives for any of the constituents in the clause, based on Krifka (2008)'s definition of focus (see section 3.3.4 below). If this is not the case, we are almost certainly dealing with a topic-comment structure and label it PREDICATE FOCUS. We can furthermore test this by finding the topic (combining different pieces of information as described above) and establishing the referential state of the predicate. If the predicate adds new information to the topic in a file-card manner, we are indeed dealing with the most commonly found focus domain: PREDICATE FOCUS. Schematically, this procedure looks as follows:

(26) Is it a copular clause?
  (i) Yes ⇝ Go to copular clauses (see section 3.3.5)
  (ii) No ⇝ Is it part of a dialogue?
    (i) Yes ⇝ Go to dialogue options (see section 3.3.4)
    (ii) No ⇝ Is there a contrastive topic?
      (i) Yes ⇝ PREDICATE FOCUS + CONTRASTIVE TOPIC
      (ii) No ⇝ Are there relevant alternatives for one of the constituents?
        (i) Yes ⇝ CONSTITUENT FOCUS (see section 3.3.4)
        (ii) No ⇝ PREDICATE FOCUS

The type of CONSTITUENT FOCUS will be specified in section 3.3.4 below. But with the above decision making tree, we can determine the domain of focus in every clause: THETIC, PREDICATE or CONSTITUENT FOCUS.

Types of topics

Topics come in different kinds and shapes. In the previous section, we zoomed in on the most common type, the 'aboutness topic'. This is also the kind of topic that is usually meant in IS literature (although it differs from the 'syntactic topic' in studies of the information structure of Old English, which denotes the first constituent of the sentence, cf. Traugott and Pintzuk (2008:64)). Götze et al. (2007) furthermore have a special label in the LISA guidelines for what they call 'frame-setting' topics that "constitute the frame within which the main predication of the respective sentence has to be interpreted." (Götze et al., 2007:167) and they give the following example:

(27) *Körperlich geht es Peter sehr gut.*
Physically goes it Peter very well.
'Physically, Peter is doing very well.' (German)

The frame setter in this sentence is the adverb *körperlich* 'physically', but the sentence also has an aboutness topic, namely *Peter*. Götze et al. (2007) choose to annotate both topics in this case, one as an 'aboutness' topic and the other as a 'frame-setting topic'. I chose to treat these frame setters differently labelling

the sentence as having POINT OF DEPARTURE (cf. Komen (2013:44-46) and section 3.3.6 below), because these frame setters interact with all three types of focus domains and do not exactly function like the 'aboutness' topics. According to Krifka (2008:46), for example, frame setters can indicate "the general type of information that can be given about an individual". He interprets frame setters as delimitators restricting the notions that can be expressed to the indicated dimension of a clause, e.g. as for his physique / physically, in example (27). The crucial point of frame setters is the possibility of alternatives, which makes them always focussed in a sense, following from Krifka (2008)'s definition of focus (see section 3.3.4 below). There would be no need for a frame setter in the first place, if there is no alternative perspective: they imply that "there are other aspects for which other predications might hold" (Krifka, 2008:46). As such, they behave similarly to what Büring (2003) and Krifka (2008) have called "contrastive topics". Contrastive topics are "topics with a rising accent" representing "a combination of topic and focus" (Krifka, 2008:44). Just like frame setters, they can take a complex issue and split it into sub-issues. Consider first Krifka's (2008) example from an English dialogue in (28):

(28) A: What do your siblings do?
B: [My [SISter]$_{\text{Focus}}$]$_{\text{Topic}}$ [studies MEDicine]$_{\text{Focus}}$,
and [my [BROther]$_{\text{Focus}}$]$_{\text{Topic}}$ is [working on a FREIGHT ship]$_{\text{Focus}}$.

The two topics are contrastive in (28), but they really function as the topic with new information added in the focussed predicate. The rising accent indicated with the capital letters furthermore denotes some sort of focus to show the contrast as a strategy of incremental answering in the CG management. In Middle Welsh, we do not have the necessary information about accents, but we do find examples that look very similar. The first example is found in a passage in the Welsh Laws describing the rights of the officers of the court; the second is from the Middle Welsh Arthurian tale *Culhwch ac Olwen*:

(29) a. *[Brenhines]*$_{\text{Topic}}$ *a geif [trayan gan y brenhin]*$_{\text{Focus}}$ *(...), ac velly*
queen PRT get third by the king (...) and so
y dyly [sswydogion y vrenhines]$_{\text{Topic}}$ *[y trayan gann swydogion*
PRT entitled officers the queen the third by officers
y brenhin]$_{\text{Focus}}$.
the king
'The queen will get a third from the king (...), and so the officers of the queen are entitled to a third from the officers of the king.'
(Cyfreithiau Hywel Dda yn ôl Ll. BL Add. 22356, 5.11)

b. *[Y trywyr]*$_{\text{Topic}}$ *a [ganant eu kyrn]*$_{\text{Focus}}$*, a ['r rei ereill*
the three.men PRT play.3P their horns and the some others
oll]$_{\text{Topic}}$ *a [doant y diaspedein]*$_{\text{Focus}}$
all PRT come.3P the outcry
'The three men shall play their horns, and all the others will come to make outcry.'
(CO 743-744)

In such cases where there is a clear contrast between two aboutness topics in one sentence that has another focus (e.g. in the predicate in the above examples), I label them as CONTRASTIVE TOPIC.

This extra focus outside the topic, also holds for the frame setters. In an attempt to capture this delimitating function of both frame setters and contrastive topics, Krifka (2008:48) characterises these structures as follows:

(30) A Delimitator α in an expression [...α...β_{Focus}...] always comes with a focus *within* α that generates alternatives α'. It indicates that the current informational needs of the CG are not wholly satisfied by [...α...β_{Focus}...], but would be satisfied by additional expressions of the general form [...α'...β'$_{\text{Focus}}$...].

This definition allows for more types of delimitators than the two mentioned here, contrastive topics and frame setters. It might, however, be too strict to include examples like (29a) and (29b). Without access to prosodic information, it is hard to establish whether there would be a rising accent, for example, and thus focus on the topics *brenhines* 'queen' and *sswydogion y vrenhines* 'the officers of the queen'. In order to let them count as real examples of **Delimitation**, according to Krifka (2008), we would have to assume the CG is not 'wholly satisfied' without the second part of the sentence. It is not altogether clear whether this is the case, because 'The queen will get a third from the king' could make perfect sense in itself in a law text that describes the legal rights of the queen. If there is evidence to the contrary, e.g. because from the context it is clear that the sentence is not complete without the second clause, example (29a) would indeed count as a Delimitator under Krifka's definition.

In the context preceding example (29b), the giant Ysbadadden Pencawr lists a number of men and beasts that are required to hunt the wild boar, Twrch Trwyth (see also example (16) above). He then specifies what the three men will do: they will blow their horns. All the others he mentions will then come and cry out. Here too, we could argue that we expect the second part of the sentence: we do not just want to know what the three men of the long list will do, we also want information about the others.

Since it seems difficult to apply the general notion of Delimitation in historical data where we have no access to prosodic information, I have annotated examples like (29a) and (29b) and those with explicit frame setters on the basis of what we *can* detect from the sentence and the context. Frame setters will receive a POINT OF DEPARTURE label with a further specification according to their function (see section 3.3.6 below); topics that are contrasted with a topic in the following clause, with separate focus structures in the predicate as we have seen above, are labelled CONTRASTIVE TOPICS. I leave aside the question here whether contrastive topics are aboutness topics as well. Evidence from parallel (gapping) structures indicates that this is not necessarily the case (cf. Repp (2010)). This distinction is, however, not relevant for the present thesis.

In some historical studies (e.g. Frascarelli and Hinterhölzl (2007) and Walkden (2014)), a further distinction is made between 'Aboutness' and 'Familiar' Topics.

FAMILIAR TOPICS are D-linked topics (i.e. linked to an antecedent in the preceding discourse) that occupy a lower position in the left periphery of the clause than ABOUTNESS TOPICS. In Middle Welsh, only one argument can occupy a pre-verbal position. Only if the ABOUTNESS TOPIC acts as a frame setter (e.g. a temporal or locational phrase), can we find a second topic that could be labelled as the FAMILIAR TOPIC. In Chapter 7, I discuss this difference further in the context of the Middle Welsh Abnormal Sentences.

Topics in discourse

> *"The maiden came inside. 'Maiden,' he said, 'are you still a maiden?' 'I know no reason why I should not be.' Then he took the magic wand and bent it. 'Step over this,' he said, 'and if you are a maiden, I will know it.' Then she stepped over the magic wand, and in that step she dropped a large boy with curly yellow hair. What the boy did was give a loud cry. After the boy's cry, she made for the door, and in the process a little something dropped from her."*
>
> (Parker, 2007)

As we have seen in the fragment about the five sons in Ireland in the previous section, aboutness topics can be the center of attention for a longer period, extending beyond one single sentence to paragraphs, texts or complete conversations. This is not the case, however, in the above fragment from Math (the fourth branch of the *Mabinogion*), because first we focus on the maiden (and her virginity test; the Welsh text uses the same word for 'maiden' and 'virgin' here, hence this translation by Parker). After that we switch to the boy that dropped out of her, only to go back to the maiden again when she is making for the door.

In the field of discourse studies, much work has been done on identifying "topic chains" or "focus chains" (Erteschik-Shir, 2007:3). Topics can be derived or introduced in three ways: a) from the topic of the previous clause ("topic chain"), b) from the rheme of the previous clause ("focus chain") or c) from a hypertheme (cf. Daneš (1974)). Topic chains or 'topic persistence' is simply the continuation of the same topic in the following sentence(s). Traugott and Pintzuk (2008:70) distinguish this from "Subsequent Mention". Subsequent Mention requires that the topic constituent is referred to again, as opposed to "Topic Persistence" indicating a continuity of pragmatic/aboutness topics. In the above fragment, *the magic wand* is brought up and subsequently mentioned in the next sentences, but the *maiden* is the topic of the following sentence where she steps over the wand, not the magic wand itself. The topic chain is broken up by the boy that dropped out of her while she steps over the magic wand. From the rheme or focussed part of this sentence, the boy is taken as the topic of the next sentence where he gives a loud cry, thus forming a "focus chain".

According to Daneš (1974), a topic can also be derived from a "hypertheme". This hypertheme consists of a set of elements restricted by the discourse. Erteschik-Shir (2007:3) gives the following example:

(31) I'll tell you about my friends, *John, Paul, and Mary*. *John* is an old friend from school, *Paul* I met at college, and *Mary* is a colleague at work.

The topics in examples (31) above can be derived from a hypertheme that explicitly mentions all members of the set, as in (31), or it can describe the set, as long as its members are obvious. The distinction between topic and focus chains could be derived from the annotated historical Welsh data automatically. Hyperthemes are not marked as such, but the referential state INFERRED of a particular entity indicates a set relation nonetheless. The final part of this section concludes the discussion on PREDICATE FOCUS with an overview of how topics can be marked in the grammar of a (written) language.

Marking topics

Topics can be marked in various ways. Since the use of specific lexical items to mark topics is not relevant in the Welsh language, I will not discuss this option further here. Prosody and intonational patterns are notoriously difficult to investigate in historical sources. If the boundaries of prosodic phrases consistently coincide with syntactic phrases and if we know more about stress and metrics, we can start looking at prosodical patterns relevant for information-structural categories. This has been done, for example, for Old High German by Hinterhölzl (2009). Since our knowledge of this in Middle or Early Modern Welsh is still limited, for now I focus on those IS markings we *can* observe in our data, for example, the word order.

Word order and 'fronting' in particular has received much attention in the literature about information structure and topicalisation. 'Fronting' is a general term for the leftward movement of a constituent that is 'topicalised', i.e. put in a position where it is interpreted as the topic of the sentence. In West-Germanic languages like German, Dutch (dialects) or Frisian with a verb-second constraint in matrix clauses, topicalisation can be implemented in three ways: movement of a constituent (an NP or even an entire clause) (see (32)), left dislocation (see (33)) or as a hanging topic (see (34)):[5]

(32) **Movement**

a. *Diesen Mann habe ich noch nie gesehen.*
this.ACC man have I yet never seen
'I have never seen this man.' (German)

b. *De zon in oew leve kan ik oe nie geve.*
the sun in your life can I you not give
'I cannot give you the sun in your life.'
(Brabantish, from *Lieke vur Mariken* by Gerard van Maasakkers)

[5] According to Ross (1986:253n18), the term 'left dislocation' was coined by Maurice Gross. The term 'hanging topic' was, according to Cinque (1977:406) coined by Alexander Grosu.

(33) **Left Dislocation**

a. *Den Hans, den kenne ich seit langem.*
the.ACC Hans this.ACC know I since long
'Hans I've known for a long time.' (German, Cardinaletti, Cinque, and Giusti (1988:9))

b. *Di lieke, da zing ik vur jou.*
this song that sing I for you
'I sing this song for you.'
(Brabantish, from *Lieke vur Mariken* by Gerard van Maasakkers)

(34) **Hanging Topic**

a. *Der Hans - ich kenne diesen Kerl seit langem.*
the.NOM Hans - I know this.ACC guy since long
'Hans - I've known this guy for a long time.' (German, Nolda (2004:424))

b. *Skulpen, troch de ieuwen hinne hawwe minsken dy al sammele.*
shells through the centuries through have people them already collected
'Shells, throughout the centuries people have collected them.'
(Frisian, from `http://pers.tresoar.nl/bericht.php?id=377`)

The main difference between sentences like (32) labelled 'movement' and sentences with left dislocation of a constituent or a 'hanging topic' can be detected from the prosodic structure: in (33) and (34) the commas clearly indicate a pause separating the fronted constituent from the rest of the sentence. A further difference between (33) and (34) can be observed in languages with morphological case marking like German. Sentences with hanging topics are therefore also called 'nominativus pendens'.

According to Willis (1998), Middle Welsh also had a verb-second constraint. Consider the following example with a fronted direct object:

(35) *Ac ystryw a wnaeth y Gwydyl*
and trick PRT made the Irish
'And the Irish played a trick.' (Middle Welsh, PKM 44.11)

Why is the direct object constituent fronted in (35)? What is its exact referential status? What is the information structure of this clause and how does it fit in the context? One of the main research questions of the present thesis is concerned with the variation in word order and to what extent, if at all, this relates to information-structural features. To investigate this properly, we have to take all possible IS features into account. The syntactic and clause type features were discussed in Chapter 2, all other IS notions and their annotation are discussed in this chapter.

In Chapter 4 and 5, I zoom in on the historical Welsh data and the main generalisations concerning the interaction of IS and word order. One important question is, for example, if all above-mentioned 'fronting' or topicalisation strategies

are found in Middle and Early Modern Welsh, what their exact IS status is, and possibly how and why this changed over the centuries. Middle English had a verb-second rule with topicalisation strategies, but this is no longer found in present-day English (cf. Holmberg (2013)). Middle Welsh and closely related Middle and Modern Breton have a verb-second constraint, but the word order of Modern Welsh (VSO) is very different from present-day English (SVO). These issues and their interaction with topicalisation strategies are discussed in the following chapters.

3.3.4 Focus vs. Background

> *Gwen Cooper: 'That was your last chance!'*
> *Lyn Peterfield: 'Yeah? What are you going to do about it? If you're the best England has to offer, God help you!' [Silence while Gwen gets up.]*
> *Gwen Cooper: 'I'm WELSH.' [And Gwen punches her out.]*
> (scene from BBC's Torchwood, season 4, episode 2)

Focus is as much an intuitive notion as it is a linguistic one. Intuitively, or generally, we are inclined to associate 'focus' with 'contrast' as in the above dialogue, or 'emphasis' of some sort. This latter part is exactly what makes focus so difficult to define linguistically. A definition of focus comprising 'emphasis' requires a strict definition or a description of 'emphasis' at the very least. In an attempt to capture all different types of focus, linguistic notions vary from a general 'new' (versus 'given', 'background' or 'presupposed') information to more specific contrastive (versus non-contrastive) information. The notion of contrast is, however, not necessarily limited to focus constructions, because topics can be contrastive as well (cf. Krifka (2008) and Repp (2010)). Komen (2013:33) gives the following definition of focus:

(36) Focus is the part of the sentence that should be understood as most highlighted or salient by the addressee, because it is new with respect to the current mental model, or contrasts with presupposed information, or is unpredictable, non-recoverable or of high communicative interest.

This is a very intuitive and practical definition capturing a wide variety of possibilities, but it still contains some gradient notions that remain undefined. What exactly is unpredictable or when exactly is something of 'high' communicative interest? Krifka (2008) has furthermore shown that there does not need to be a correlation between given or well-established information (getting a linked label IDENTITY, INFERRED or NEW) and the distribution of focus: even well-established phrases with an IDENTITY label like pronouns can be focussed:

(37) Mary only saw [HIM]. (Krifka, 2008:39)

The capital letters in example (37) denote a stressed, rising accent and thus a focus on the pronoun. This example is perfectly fine in English, even though the referential state of the focussed pronoun is IDENTITY and thus linked or 'given'. In semantics, a constituent that is selected from a set of alternatives is understood to

be focussed (cf. Rooth (1985) and Zimmermann and Féry (2010)). Krifka's (2008) exact definition is as follows:

(38) A property F of an expression α is a Focus property iff F signals
  (a) that alternatives of (parts of) the expression α or
  (b) alternatives of the denotation of (parts of) α
  are relevant for the interpretation of α.

As long as 'relevant' is not further defined, this too leaves some room for subjective interpretation. If we want to investigate the information structure of a language we should not be distracted by possible phonological, morphological or syntactic *expressions* of IS. A high pitch accent, for example, may be used to focus a constituent in one language, but it does not necessarily have the exact same effect in another language. Nevertheless, there are certainly some cross-linguistic generalisations on the way IS is expressed. Ideally, we try to go *beyond* the surface expression to find its IS status first before we make the association between, e.g. high pitch and contrastive focus, or fronted constituents and topicalisation. Krifka's definition in (38) allows the separation of the way IS is expressed from what the IS status (referential state, focus domain, etc.) is. I therefore use the definition in (38) as a guideline to recognise focus constructions, or, in particular the domain that I generally label CONSTITUENT FOCUS. CONSTITUENT FOCUS can be marked in various ways, just like the topicalisation structures we noted above (see the sections on different types of focus and their markings below). Again, however, I can only discuss those forms of focus marking that can be detected in historical, written documents. Birch and Clifton (1995) showed in their experiments with *it*-clefts and *there*-insertions that structural positions can also make focus stand out in sentence comprehension tasks.

From a cognitive perspective, constituent focus structures play an important role in directing attentional focus in our brains. They also influence the availability of information in our memory and the degree to which it continues to be activated (Cowles, 2012:298). From psycholinguistic experiments we know that auditory cues like the pitch accents mentioned above can be helpful to identify focussed constituents (Cutler & Fodor, 1979). There is no consensus yet about a one-to-one mapping between prosody and information status (cf. Cowles (2012:293) and Hedberg and Sosa (2007)), but there is further evidence of these focussed structures from ERP studies. In some of those experiments, for example, N400 effects were detected when participants heard sentences with focus-violations (cf. K. Johnson (2003) for English and Hruska, Alter, Steinhauer, and Steube (2000) for German). The N400 effect, consisting of a characteristic change in brain wave activity 400 milliseconds after the stimulus, is associated with lexical and semantic processing (Kutas, Van Petten, & Kluender, 2006). This effect suggests focus anomalies influence the semantic processing of the word. Later studies on reading tasks with focus constructions by Bornkessel, Schlesewsky, and Friederici (2003), however, suggested that focus modulates information integration, indexed by a late positivity effect, instead of the N400 (cf. Cowles (2012)). Whichever it

turns out to be, it is clear that reading or hearing a focussed constituent results in a measurable effect in our brain. Although more experimental research is needed, a different focus domain, like THETIC FOCUS (see section 3.3.2) or PREDICATE FOCUS (see section 3.3.3) where the whole predicate is focussed instead of just one constituent, clearly gives the listener or the reader very different options.

Types of Constituent Focus

In section 3.3.3 on finding the right focus domain, we went through several steps to detect THETIC FOCUS and PREDICATE FOCUS. CONSTITUENT FOCUS, or 'narrow' or 'identificational' focus, as it is also called (cf. É.Kiss (1998)) can be found when there are alternatives of a certain expression that are relevant for the interpretation of the particular clause (see definition of Focus by Krifka (2008) above). Figure 3.2 shows the three focus domains, including the subtypes that can be detected in the domain of CONSTITUENT FOCUS:

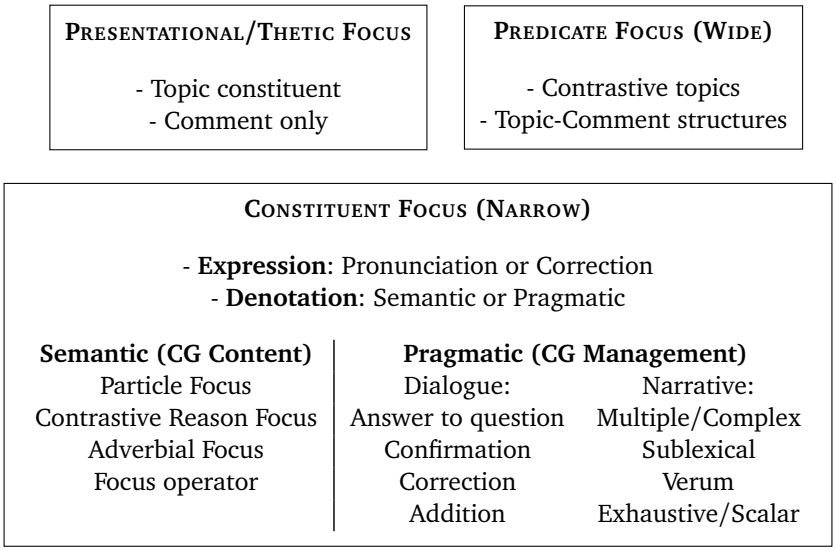


Figure 3.2: Focus Domains with subtypes

When we find relevant alternatives in a dialogue, we proceed to find out if the constituent is part of a question or answer. If it is not, we try and detect whether a constituent (or even part of it, a sublexical item) functions as a confirmation, correction or parallel structure. If this is the case, the clause will get the label of CONSTITUENT FOCUS with an addition: CONFIRMATION, PARALLEL and CORRECTION or another form of CONTRASTIVE FOCUS. If not, we are simply dealing with a topic-comment structure and thus label it PREDICATE FOCUS. (39) shows the schematic procedure just described. Examples (following Krifka's examples, unless indicated otherwise) of these types of focus are given in (40), (41), (42) and (43):

(39) Is it a question-answer dialogue?
- (i) Yes ⇝ Go to (45)
- (ii) No, did the speaker confirm information?
  - (i) Yes ⇝ CONFIRMATION FOCUS
  - (ii) No, did the speaker correct information?
    - (i) Yes ⇝ CORRECTION FOCUS
    - (ii) No, did the speaker use parallel structures?
      - (i) Yes ⇝ PARALLEL FOCUS
      - (ii) No, is there an explicit contrast?
        - (i) Yes ⇝ CONTRASTIVE FOCUS
        - (ii) No ⇝ PREDICATE FOCUS

(40) CONFIRMATION FOCUS

A: Siriol ate the last biscuit.
B: Yes, [SIRIOL] ate the last biscuit.

(41) CORRECTION FOCUS

A: Siriol ate the chocolate.
B: No, [ASIYE] ate the chocolate.

A: Theofiel?!
B: *Nee, Theo[DOOR] is mijn naam.*
No Theodoor is my name
'No, Theo[DOOR] is my name!'
(Dutch, from *De Texasrakkers*, Suske & Wiske 124)

(42) PARALLEL FOCUS

A DUTCH football fan talked to a ENGlish football fan about the world cup.

(43) CONTRASTIVE FOCUS

Martha: Woah, Nelly! I know for a fact you've got a wife in the country.
Shakespeare: But Martha, this is [TOWN].
The Doctor: Come on! We can have a good flirt later.
Shakespeare: Ooo, is that a promise, Doctor? *[winking at him]*
The Doctor: Oh, [FIFty-seven academics] just punched the air!
(from *Doctor Who*, series 3, episode 2)

The contrastive focus can be an explicit antonym or an alternative from a restricted set, as in the example above where *country* and *town* are contrasted. The contrast can also be implicit. The *fifty-seven academics* further on, for example, are raising their fists in victory, because they were just proven right: the phrase implies a contrast with all the other English literary scholars who do not think that Shakespeare was bi- or homosexual (referring to sonnet 57, which is about a relationship with a young man). If knowledge of English literary history is part of the world knowledge stored in the long-term memory of the reader, this contrast is obvious. Another example of implicit contrast is found in the following dialogue between someone hosting a workshop at a conference in Sydney and HRH the Earl of Wessex:

(44) A: May I invite you to join us for drinks, Sir?
B: Yes, why not? [In SYDney], I can safely go out.

The contrast in this utterance is obvious to those who know the British royal family and have dealt with their protocols before. In the UK, the Earl could never accept an invitation to go for drinks, because people will recognise him. In Australia, however, this is not the case.

Focus in dialogue
If we *are* dealing with a question-answer dialogue, we need to investigate the type of question: is it a wh-question and if so, does it extend over the entire VP or not? If it is not a wh-question, several other options remain: parallel answers (similar to parallel focus sentences above), delimitation focus and closed or open set answers. Consider the following continuation of the decision tree and the examples (after Krifka (2008), unless indicate otherwise):

(45) Is there a delimitation?
(i) Yes ⇝ DELIMITATION FOCUS
(ii) No, is it a simple wh-question?
(i) Yes ⇝ Go to (46)
(ii) No, is there a parallel answer?
(i) Yes ⇝ PARALLEL ANSWER
(ii) No, go to (46).

(46) Does focus extend over the entire VP or a NP/PP?
(i) Entire VP ⇝ VP WH-ANSWER
(ii) NP or PP, is it a closed or open set?
(i) Closed ⇝ CLOSED NARROW FOCUS
(ii) Open ⇝ OPEN NARROW FOCUS

(47) VP WH-ANSWER
A: What is Rhys doing?
B: He is [climbing Snowdon].

(48) PARALLEL ANSWER
A: Who ate what?
B: SIriol ate the BIScuit and ASiye ate the CHOcolate.

(49) DELIMITATION FOCUS
Which sister loves what?
a. As for ASiye, she loves CHOcolate.
Who do YOU think stole the chocolate?
b. In MY opinion, ASiye stole the chocolate.

(50) OPEN NARROW FOCUS
A: What would you like to drink?
B: I'd like some TEA, please.

A: Who is climbing Snowdon?
B: RHYS is climbing Snowdon.

A: How do you tell the story of pain?
B: You don't: you tell the story [of how], after everything falls apart, [you slowly rebuild].
(after `http://itellstories.com`, d.d. 31-12-12, *Twentytwelve*)

(51) CLOSED NARROW FOCUS
A: What would you like to drink, tea or coffee?
B: I'd like TEA, please.

Expression vs. Denotation Focus
If the clause under investigation is not part of a dialogue, the next question we ask is whether we are dealing with expression or denotation focus (cf. Krifka (2008:19-20)). Expression focus affects aspects like the choice of words or pronunciation; they do not have to involve meaningful units like constituents. When it affects the pronunciation, I label it PRONUNCIATION FOCUS. Another example of expression focus is found in corrections, e.g.:

(52) EXPRESSION FOCUS
Grandpa didn't [kick the BUcket], he [passed aWAY].

(53) PRONUNCIATION FOCUS
A: They live in BERlin.
B: They live in BerLIN.

Denotation focus is the most common form of focus outside dialogue situations. The first question here is whether we are dealing with semantic or pragmatic focus. According to Krifka (2008), pragmatic focus does not immediately influence truth conditions, but semantic focus *does* affect the truth-conditional content of the Common Ground. Contrastive focus is one of the best-studied cases of this type of focus. Semantic focus constructions are often clearly marked by semantic operators, such as focus-sensitive particles or adverbs like English *only*, *even*, *also* or *fortunately*, but this is not necessarily the case. The annotation procedure continues with the following decision-making tree:

(54) Is there an explicit lexical item as a semantic operator?
(i) No, go to (58).
(ii) Yes, are there more focussed constituents?
(i) Yes, go to (55).
(i) No, is there an adverbial focus operator?
(i) Yes ⇝ ADVERBIAL FOCUS
(ii) No, is it a negation or a particle?
(i) Negation ⇝ NEGATION FOCUS
(ii) Particle ⇝ PARTICLE FOCUS

(55) Are there two expressions introducing two different sets of alternatives?
(i) Yes ⇝ MULTIPLE FOCUS
(ii)No ⇝ COMPLEX FOCUS

Consider the following examples with more than one focussed constituent in (56) and (57) (from Krifka (2008:31-32)):

(56) MULTIPLE FOCUS
John only introduced BILL only to SUE.

(57) COMPLEX FOCUS
John only introduced BILL to SUE.

Example (56) contains two expressions introducing alternatives that are exploited in two different ways. The first *only* has scope over the second, reflected by a stronger accent on *Bill* than on *Sue*. This is not the case in (57) that only has one single focus on the pair <Bill, Sue>. If there is no overt semantic operator, we continue with (58):

(58) Is there a contrast with something in the CG?
(i) Yes ⇝ CONTRASTIVE FOCUS
(ii)No, is there a reason clause or variation of counterfactual?
(i) Yes ⇝ REASON CLAUSE FOCUS
(ii) No, start over (see Appendix for full procedure)

Krifka (2008) mentions (59) as an example of focus that I label REASON CLAUSE FOCUS:

(59) REASON CLAUSE FOCUS
a. **Clyde had to marry [BERtha]** in order to be eligible.
b. **Clyde had to [MARry]** Bertha for the inheritance.

Examples of CONTRASTIVE FOCUS can be found in many constructions and many different languages. Just like in the dialogue examples above, the contrast can be made explicit by repeating the same lexical item with a different modification (see (60) and (61)) or by using its antonym (or a close resemblance, see (63) and (62)). But it can also be implicit, contrasting the expected meaning of the items (as in (64)):

(60) The average pencil is [seven inches] long, with just a [half-inch] eraser, in case you thought optimism was dead. (Robert Brault)

(61) *Sans toi, les [émotions d' aujourd'hui] ne seraient que la peau morte*
without you the emotions of today NEG would ONLY the skin dead
des [émotions d' autrefois]
of.the emotions of past
'Without you, today's emotions would only be the dead skin of the emotions of the past.' (French, from *Amélie*)

(62) It is not enough for us to *believe* that what we do makes a difference - we must *prove* that it does, and be accountable to everyone we serve.
(from *Measuring the Award's impact*, B. Hirt (2012))

(63) *Wir vermögen [mehr], als wir glauben. Wenn wir das erleben, werden wir*
we can.do more than we think when we that realise will we
uns nicht mehr mit [weniger] zufrieden geben.
us not more with less satisfied give
'We are all better than we think. If (only) we can be brought to realise this we will never again be prepared to settle for anything less.'
(German, from Kurt Hahn)

(64) That's the whole problem with science. You've got a bunch of [empiricists] trying to [describe things of unimaginable wonder].
(from Calvin & Hobbes)

(65) When I meet you, in that moment, I'm no longer a part of [your future]. I start quickly becoming part of [your past]. But in that instant, I get to share [your present]. And YOU, you get to share MINE. And that is the greatest present of all. (from *Hiroshima* by Sarah Kay)

There is a wide variety of semantic operators that can indicate focus structures in different languages. Contrast can also play a role here, depending on the type of particle. Consider the following examples in Present-Day English and Welsh:

(66) PARTICLE FOCUS

a. *Dim ond gofyn am fenthyg sgriwdreifar ro'n i, nid adrodd hanes fy*
only ask.INF about borrow.INF screwdriver was i not relate story my
mywyd.
life
'I was only asking to borrow a screwdriver, not to relate the story of my life.'

b. *Dydyn nhw ddim yn gwneud dim byd eu hunain, dim ond dwyn*
are they NEG PROGR do.INF nothing themselves only steal.INF
oddi wrth eraill maen nhw.
from others are they
'They don't do anything themselves, they only steal from others.'
(from *Y rhyfel oeraf*, Baxendale (2009:43 and 89))

c. One of the great things about going to high school with people from 60 different countries was that we were all forced to see things, *even* the small, everyday things we all took for granted, from different perspectives.

d. I sincerely hope the results of our impact research framework will *not just* prove the value of this remarkable youth achievement award, but *also* convey the emotional effect.
(HRH The Earl of Wessex KG GCVO in *Measuring the Award's impact*, B. Hirt (2012))

Finally, there are some other types of focus we have not discussed yet. One further

question we can ask concerns the size of the constituent: is the entire constituent focussed or just part of it? Note that according to É. Kiss (1998), 'Identification Focus' (our CONSTITUENT FOCUS) can be distinguished from 'Information Focus' (the 'new information' often found in the topic-comment structures that I labelled PREDICATE FOCUS above) by the fact that only the latter can be smaller or larger than an XP as in (67):

(67) SUBLEXICAL FOCUS
Let me exPLAIN, exPOUND, exPAND and exPOSIT.
(from *A discussion on Language* in BBC's 'A bit of Fry & Laurie')

Strictly speaking, SUBLEXICAL FOCUS (see example (67)) cannot be part of 'Identification Focus' in her system. If we want to equate 'Identification' and 'Constituent' Focus domains, É. Kiss's categorie of 'Identification Focus' should be slightly expanded to ensure that it can capture every form of focus. Krifka (2008) furthermore mentions an extreme focus on the truth value of a sentence, VERUM FOCUS (see example (68) after Krifka (2008)).

(68) VERUM FOCUS
Asiye DOES like chocolate, why do you think she wouldn't?

There are furthermore two types of contrastive focus that we have not discussed: EXHAUSTIVE and SCALAR FOCUS (after Krifka (2008)):

(69) EXHAUSTIVE FOCUS
It's [ASIYE and ELANOR] that saved us.

(70) SCALAR FOCUS
Wild HORses wouldn't drag me there.

Example (69) is exhaustive in the sense that all possible candidates who could have 'saved us' were listed: Asiye and Elanor. Example (70) is scalar because it implies that there are more forces that could possibly 'drag me there', but even animals as strong as wild horses would not be able to do so (because I have made up my mind and really don't want to go). These last examples conclude a long section about many different types of CONSTITUENT FOCUS. In the next section, I discuss some ways to *mark* these focus structures.

Marking Constituent Focus

Evidence of CONSTITUENT FOCUS in historical data first of all comes from detecting possible alternatives relevant for the context. Once these possible alternatives have been found, we need to describe how they can be marked. As we have seen in topic marking above, in historical data we can only work with morphology, word order patterns, lexical items and, possibly, underlying syntactic structure. In the previous section, I already showed some examples of focus particles and other operators.

(71) **Focus Particles**

a. The leaves change colors in the fall. [People] change colors in the fall, **too**.
(from `http://itellstories.com`, d.d. 31-12-12 and 18-08-14)

b. *(...) y dywedir nad yw 'n rhewi* ***hyd yn oed*** *mewn*
... PRT said.IMPERS NEG.FOC is PROGR freeze.INF even in
gaeaf caled.
winter hard
'... it is said that it doesn't freeze, not even in a hard winter.'
(Modern Welsh)

c. *Does* ***dim ond*** *eisiau dechrau*
NEG.is only need begin.INF
'You only need to begin'
(Modern Welsh, from a poem by Ceiriog)

Special constructions like clefts are also commonly used in languages to mark focussed constituents:

(72) **Clefts, pseudoclefts and inverted pseudoclefts**

a. *Fi sydd ar fai am hynny.*
I is.REL on blame for that
'I am the one to blame for that.'
(Modern Welsh, Baxendale (2009:89))

b. *ma Se-rut hayta ze nexmada*
what that-Ruth was.F Z.M nice.F
'What Ruth was was nice.'
(Hebrew, Heller (1999:47))

c. There'll be days like this (...) when you step out of the phone booth and try to fly and the very people you want to save are the ones standing on your cape.
(from *Point B* by Sarah Kay via `www.kaysarahsera.com`)

Answers to questions furthermore often exhibit different word order patterns, depending on the type of question (yes/no, wh, broad/narrow focus, etc.):

(73) **Questions and answers**

a. *Wyt ti ffansi mynd am wibdaith fach 'te? Ydw, plis.*
are you fancy go for trip small TAG am please
'Do you fancy to go on a short trip then? I do, please.'
(Modern Welsh, Baxendale (2009:46))

b. *Pam mae 'r graig hon yn gynnes, tybed? Oherwydd nad craig*
why is the rock this PRED warm you-think because NEG.FOC rock
yw hi.
is it
'Why is this rock warm, you think? Because it is not a rock.'
(Modern Welsh, Baxendale (2009:92))

c. *Felly beth sy 'n digwydd nawr? Mae hi 'n amser mynd adref.*
So what is PROGR happen.INF now is it PRED time go home
'So what's happening now? It is time to go home.'
(Modern Welsh, Baxendale (2009))

In traditional grammars of Middle Welsh, focus structures are usually called 'mixed order': "[w]hen a part of the sentence other than the verb is to be emphasised, this is placed at the beginning of the sentence, preceded by a form of the copula and followed by a relative clause." (D. S. Evans, 2003 [1964]:140). Some examples he gives are (with his translation):

(74) **Mixed Order**

a. *Ys mi a 'e heirch.*
it-is me PRT her search.3S
'it is I who seek her' (Middle Welsh, WM 479.29)

b. *Oed maelgun a uelun i n imuan.*
was Maelgwn PRT saw.IPF.1S I PROGR fight.INF
'It was Maelgwn that I could see fighting.'
(Middle Welsh, YMTh 57.5)

In a later stage of the language, this sentence-initial copula was lost "before the emphasised word or phrase" (D. S. Evans, 2003 [1964]:141). Compare the following examples (again with Simon Evans's translation):

(75) **Mixed Order**

a. *Mi a 'e heirch.*
I PRT her search.3S
'(it is) I who ask for her' (Middle Welsh, WM 479.24)

b. *Mi yd wyt yn y geissaw.*
I PRT are.2S PROGR 3MS search
'(it is) I whom thou art seeking' (Middle Welsh, WM 138.21)

In these examples of the 'mixed order' there is no agreement between the subject and the verb. There is a very similar word order pattern in Middle Welsh, however, that does show agreement, but is not a focus structure:

(76) **Abnormal Order**

a. *Gwydyon a gerwys yn y blaen.*
Gwydyon PRT travelled.3SG in the front
'Gwydion travelled in the forefront' (not: 'It was Gwydyon who...')
(Middle Welsh, PKM 90.27)

b. *Mi a wn dy hanuot o 'm gvaet.*
I PRT know.1S 2S be.INF from 1S blood
'I know you are from my blood.' (Middle Welsh, CO 167)

This 'abnormal order' is often referred to as a topicalisation device (cf. Poppe (1991) and Willis (1998) among others). The first slot in this 'verb-second' construction can be filled by the subject, object or adjunct phrase (as we have seen in example (35) above). Finding the information-structural and syntactic constraints of these various word order patterns and how they change is the main research question of the present thesis. Chapter 4 presents a detailed description of IS in different stages of the Welsh language. The syntactic analysis of the various word order patterns in Chapter 5 sheds more light on the interface issues. For now it suffices to say that word order and syntactic relations interact with information structure in Welsh, so those above-mentioned markings of focus (and topic) structures will be investigated in more detail.

3.3.5 Focus domains of copula clauses

The three focus domains discussed above can also be found in copular clauses. Since the syntactic structure of copular clauses differs, I discuss the procedure of detecting the focus domains of these clauses separately. Komen (2013:164-170) gives a detailed overview of focus domains in copular clauses in English. In this section I propose a similar way of deriving the focus domain of copular clauses in Welsh, combining the coded syntactic and IS information, in particular the referential state of the core arguments. The focus domain is derived via a number of questions in a decision-making tree:

(77) Is it an equative clause?
  (i) Yes, move on to (79)
  (ii) No, is the subject NEW?
    (i) Yes ⇝ CONSTITUENT FOCUS as in (78a)
    (ii) No ⇝ PREDICATE FOCUS as in (78b)

(78) a. *Y mae Arthur yn gefnder iti.*
PRT be.PRES.3S Arthur PRED cousin to.2S
'Arthur is a cousin of yours.' (CONSTITUENT FOCUS - Modern Welsh)
  b. *Cauall oed y enw.*
Cafall be.PAST.3S 3MS name
'His name was Cafall.' (PREDICATE FOCUS - Gereint 399)

(79) Is the equative NP complement an Adjectival Phrase?
  (i) No, move on to (81)
  (ii) Yes, is the subject NEW?
    (i) No ⇝ PREDICATE FOCUS as in (80a)
    (ii) Yes ⇝ THETIC FOCUS as in (80b)

(80) a. *Roedd pawb yn 'gwybod' mai Jyrman Sbei oedd hi.*
was all PROGR know.INF that German spy was she
'Everyone knew that she was a German spy.'
(PREDICATE FOCUS - Modern Welsh)
  b. The world is wonderful. (THETIC FOCUS)

(81) Is the equative NP complement INERT?
(i) No, move on to (83)
(ii) Yes, Is the subject NEW?
(i) No ⇝ PREDICATE FOCUS as in (82a)
(ii) Yes ⇝ THETIC FOCUS as in (82b)

(82) a. *Ac Ioseph ydoedd fab deng mlwydd ar hugain pan...*
and Joseph be.PAST.3S lad ten year on 20 when...
'And Joseph was 30 when...' (PREDICATE FOCUS b1588 - Gen. 41.46)
b. In the next year Marius was consul. (THETIC FOCUS - Komen (2013:166))

(83) Is it a case of variable identification?
(i) Yes ⇝ CONSTITUENT FOCUS as in (84)
(ii) No, is the subject NEW?
(i) Yes ⇝ THETIC FOCUS as in (85)
(ii) No, is the subject INFERRED or ASSUMED?
(i) Yes, move on to (87)
(ii) No, is the subject INERT?
(i) Yes ⇝ PREDICATE FOCUS as in (86)
(ii) No, go to (87)

(84) CONSTITUENT FOCUS
a. *Y TARDIS yw hwn.*
the TARDIS is that
'That is the TARDIS.' (answer to: 'What's that?') (Baxendale, 2009:46)
b. (Last week, part of the Pont Des Arts in Paris collapsed. It collapsed, quite literally, under the weight of aspirations and expectations of everlasting love;) the Pont Des Arts was one of the famous bridges upon which young lovers would affix locks to signify the foreverness of their affection.
(from `http://itellstories.com`, d.d. 18-06-14, *Love locks*)

(85) *Maxen Wledic oed amherawdyr yn Ruuein*
Maxen Wledig be.PAST.3S emperor in Rome
'Maxen Wledig was emperor in Rome.' (THETIC FOCUS - BM 1.1)

(86) What is the weather in Siberia? In the winter, it is cold.
(PREDICATE FOCUS - Komen (2013:166))

(87) Is the complement NEW?
(i) Yes ⇝ CONSTITUENT FOCUS as in (88a)
(i) No ⇝ PREDICATE FOCUS as in (88b)

(88) a. *Gwidonot Kaer Loyw ynt.*
witches Gloucester be.3P
'They are the witches of Gloucester.' (CONSTITUENT FOCUS - Peredur 29.18-19)
b. The driver of that car is from Finland.
(PREDICATE FOCUS - Komen (2013:165))

3.3.6 Additional IS factors

As mentioned above, there are at least two further information-structural factors that can interact with each of the three focus domains: delimitation strategies or frame setters (see section 3.3.4 above) and the 'principle of natural information flow'. For every sentence we can detect one of the three focus domains, but we should further annotate these two notions to provide a comprehensive description of all IS facts.

Delimitation and Point of Departure

> *"**When you've told your love what you're thinking of***
> *things will be much more informal;*
> ***Through a sunlit land** we'll go hand-in-hand,*
> *drifting gently back to normal.*
> *(...)*
> ***With your hand in mine,** idly we'll recline*
> *amid bowers of neuroses,*
> ***While the sun seeks rest in the great red west***
> *we will sit and match psychoses".*
>
> (fragment from *The Passionate Freudian* by Dorothy Parker)

Delimitation strategies or 'points of departure' like the bold-faced phrases in the above poem by Dorothy Parker were already discussed in the section on topics (see section 3.3.3), because they are also called 'frame setting topics' (cf. Götze et al. (2007)). Krifka (2008) uses the term 'delimitation' for any expression (both frame setters and contrastive topics) that "always comes with a focus" generating alternatives (Krifka, 2008:48). This definition allows for more than just frame setters, e.g. (from Krifka (2008:48)):

(89) [An [inGEnious] mathematician]$_{\mathrm{Delim}}$ he is [NOT]$_{\mathrm{Focus}}$.

Komen (2013:44) gives the following definition of what they call 'Point of Departure' (PoD):

(90) **Point of Departure**
A point of departure is a constituent fulfilling the following conditions:
i) It is placed at the beginning of a clause or sentence;
ii) It expresses a change in the point of view in the discourse;
iii) It anchors to something that is accessible to the addressee (either from the preceding linguistic context or through shared knowledge)

I will label constituents that meet the requirements in (90) POINT OF DEPARTURE, because their presence can influence the IS status of the entire sentence. A sentence without a PoD is not as tightly linked to the previous context or content of the current Common Ground as sentences *with* a PoD. These types of frame setters occur very often in Middle and Early Modern Welsh (cf. Poppe (1991) where it is

called 'Situationskulisse'). To ensure all possible IS variables are covered, I make a further distinction between the functions of the PoDs. In this way if we encounter word order variation in different sentences, we could determine whether or not this is due to the different function of the PoD. Consider some examples of sentences with different PoDs below:

(91) PoD: LOCATIONAL

a. (I cycled to the office in the morning and worked all day.) **From the office**, I went straight to BodyCombat training.

(92) PoD: TEMPORAL

a. ***Et quand tu seras consolé*** *(...), tu seras content de m' avoir connu.*
and when you will.be consoled (...) you will.be happy of me have known
'And when you'll be comforted (...), you will be happy to have known me.'
(French, from *Le petit prince* by De Saint-Exupéry)

b. ***Om half 10*** *begint de handbalwedstrijd.*
at half 10 starts the handball game
'At half past nine, the game will start.' (Dutch)

(93) PoD: CIRCUMSTATIAL

a. **With an incredible amount of effort**, he managed to convince her.
b. **Healthwise**, my friend is fine.

(94) PoD: SITUATIONAL
As they had been friends for a long time, he expected her to help him.

(95) PoD: REFERENTIAL
That battery, however, continued its fire.

All of the above sentence-initial 'points of departure' contain information stored in the current CG: they all either refer back to something that was mentioned in the text or that is accessible as 'world knowledge' from our long-term memory. They set the frame or limit the space in which the following proposition holds. They can be added to clauses with any of the three focus domains: THETIC FOCUS, PREDICATE FOCUS or CONSTITUENT FOCUS.

Principle of Natural Information Flow

Another IS phenomenon that can interact with each of the three focus domains is what Comrie (1989), Kaiser and Trueswell (2004) and others have called the "Principle of natural information flow" (cf. Komen (2013:43-44)). This principle concerns the degree of 'givenness' of constituents: established information precedes less established information. If the syntactic structure of the language allows for alternatives, some constituents can be reordered changing the 'information flow' of the sentence. We can see the principle in presentational constructions in English (cf. Komen (2013:44)):

(96) UNMARKED INFORMATION FLOW
Once upon a time there was **a handsome prince**.

The referential state of the phrase *a handsome prince* is NEW and it is thus placed at the very end of the sentence. In the English Dative Alternation we also see a clear example of this principle:

(97) UNMARKED INFORMATION FLOW
a. Rhys gave the student **a book**.
b. Rhys gave the book to **a student**.

Both examples in (97) abide by the principle of information flow, because in both cases (as the definite article shows), the first constituent following the verb conveys 'more established' information than the second constituent. Note that the opposite word order in English with the same noun phrases is odd or even impossible:

(98) MARKED INFORMATION FLOW
a. Rhys gave a book to the student.
b. Rhys gave a student the book.

In some constructions in English, however, putting the least-established constituent before the rest has a special effect, for example, to focus the place in the Locative Inversion or the direct object that has been the centre of attention of the entire lecture, as in example (99a) and (99b):

(99) MARKED INFORMATION FLOW
a. **Up, up, up the stairs** we go!
(from *The Lord of the Rings* by JRR Tolkien)
b. Sir William Jones and John James Jones both worked tirelessly to bring to a world far distant in time and place **some of the wealth of ancient Indian culture**.
(from a lecture on JJ Jones and the *Mahavastu* by Silk (2014:439))

The Principle of Natural Information Flow can occur with any of the three focus domains. All clauses are annotated as MARKED (unlinked before linked) or UNMARKED (linked before unlinked) for this in the Welsh database.

3.4 Conclusion

In this chapter I gave an introduction to Information Structure and its place in the field of linguistics. I discussed three core information-structural notions in greater detail: Givenness, Topic (vs. Comment) and Focus (vs. Background). For each of these notions, I outlined their main characteristics in a systematic way so that they can be used to annotate a corpus consistently.

For the notion of Givenness, it is clear that a simple binary distinction between Old and New information is not enough (see Taylor and Pintzuk (2014) for a

systematic evaluation of different annotation schemes). For the present thesis, I annotated the referential status of subjects and objects in the Middle Welsh corpus according to the Pentaset developed by Komen (2013). This type of annotation can help identify effects in word order distributions in combination with annotated syntactic features.

In the section on Topics, I focussed on three different kinds of topics that are found in the Middle Welsh corpus: aboutness, contrastive and familiar topics. The notion of 'Delimitation' as formulated by Krifka plays a crucial role in determining aboutness topics. Like frame or scene setters, they usually occupy the first position in the sentence. Contrastive topics are also found in Middle Welsh. The notion of contrast is thus not necessarily associated with Focus. In final part of this thesis, these kinds of topics are discussed again in their syntactic contexts.

I furthermore presented a detailed overview of different kinds of Focus structures. I illustrated the different types observed in the literature with examples from Welsh and various other languages. I furthermore presented some systematic 'algorithms' to find the focus articulation of copular clauses, based on studies in the history of English by Komen (2013).

Finally, I discussed two further notions that are relevant to information structure: Point of departure and Information Flow. Many so-called 'Points of Departure' of a sentence appear in the form of temporal or circumstantial clauses. In effect, they function as frame setters delimiting the context of the rest of the sentence. The Principle of Natural information flow finally stipulates that old information usually precedes new information. In sentences with the reverse order, the 'flow' of information, or in particular the referential status of the core arguments, is 'marked'.

These three core notions of Givenness, Topic and Focus, in combination with the additional annotation for specific points of departure and information flow are argued to provide a comprehensive insight into the Information Structure of the sentence in its context. The clear definitions and guidelines to find the right labels presented in this chapter facilitate annotation. A consistent analysis of this kind helps to make the study of Information Structure that has suffered from a lot of 'terminological profusion and confusion' more insightful in the language under investigation. But, more importantly, it renders it more useful, because results of such thorough investigation could then be more easily compared between different languages.

CHAPTER 4

Word order patterns in Welsh

4.1 Introduction

> *"The position of words in a sentence depends on the emphasis to be laid on them. In Welsh, as in other languages, the most important word takes precedence. In ordinary discourse, when no particular emphasis is intended to be expressed, or where the verb, as being the main part of the clause, may be regarded as emphatic, the order will stand thus: verb, subject, predicate or object."*
>
> (Rowland, 1876:173)

In his 1876 grammar, Thomas Rowland aimed to give an accurate description of the Welsh language "based on the most approved systems, with copious examples from some of the most correct Welsh writers" (Rowland, 1876:title). As most other nineteenth-century Welsh grammarians, he established VSO as the basic word order in declarative main clauses.

The VSO preference seems to be an innovation of the Insular Celtic languages. Old Irish, the main focus of early research on Celtic by historical linguists, was VSO (cf. Thurneysen (2003 [1946])). According to Vendryes (1912), verb-initial word orders were already a possibility in Indo-European. In Celtic then, this became the only possibility: "L'originalité du celtique est d'avoir généralisé un ordre occasionnel en faisant de cette possibilité une nécessité" (Vendryes, 1912:338). All other branches of Indo-European (e.g. Greek, Indo-Iranian, Balto-Slavic) kept a preferred subject-initial order (SVO or SOV). Syntactic evidence from Continental Celtic languages is scarce, but although verb-initial order was an option, it was certainly not the preferred option in Gaulish (cf. Fife (2010) among others). VSO word order

was therefore one of the main reasons to propose a significant pre-Indo-European substrate in the Insular Celtic branch (cf. Wagner (1959)).

Typologically, preferred verb-initial orders are a minority among the world's languages, though Celtic is far from unique. Other features that are typically found in VSO languages are also found in Celtic languages, e.g. *wh*-words are placed before the verb, they mainly exhibit post-head modification, they are prepositional rather than postpositional, the main verbs follow their auxiliaries, they have sentence-initial particles and, finally, they have SVO as an alternate order (cf. Fife (2010) and Ouhalla (1994)). Concerning this final feature, John Morris-Jones, one of the most famous Oxford Welsh reformers, wrote in his appendix to Rhys & Jones's 1906 *The Welsh People*:

> "(...) there appears in Welsh another form of sentence in which the noun comes first. No distinction is made in any of our Welsh grammars between this and the simple form of sentence in which the verb comes first; and the Welsh translators of the Bible constantly misuse it for the simple form; as *Job a atebodd*, instead of *atebodd Job*, for 'Job answered'. "
>
> (Rhys & Jones, 1902:619)

The 1588 Bible translation had a great influence on Welsh literature for many centuries. From this perspective, as Paul Manning puts it "[i]t was somewhat of a source of chagrin to many to find out that, in effect, biblical figures like Jesus and Job spoke bad Welsh" (Manning, 1997:67). The famous grammarian Rowland notes that "[w]hen the subject of the clause is antithetical, the order of the construction will be subject, verb, predicate or object." Rowland (1876:174). He adds that many Welsh writers "and especially translators" continually express ordinary discourse in this manner for reasons of elegance and "where the same order of words would render the sentences too monotonous". Furthermore, "[i]f the subject is a personal pronoun, it is continually, in affirmative sentences, put before the verb, even when the subject is not antithetic" (Rowland, 1876:175). Nineteenth-century Welsh in the eyes of Rowland thus had 'simple' sentence (VSO), 'somewhat emphatic' sentences (SVO) and 'rhetorical' sentences "for the sake of still greater emphasis and vivacity" (Rowland, 1876:175). In these 'rhetorical' sentences, any constituent could be placed in front of the verb.

The 'somewhat emphatic' sentences listed in Rowland's grammar (without English translation) all exhibit the order Subject - *a* - Verb. The particle *a*, according to Rowland, was a "mere expletive" particle placed immediately before the verb. He quotes Dr Davies who described *a* in the seventeenth century as "adverbium seu particular verbis preposita *nihil significans*".[1]

In Anwyl's 1899 Welsh grammar two patterns are discussed: the 'normal' and the 'inverted' word order. VS + the remainder of the predicate is considered the 'normal' order, whereas the inverted order starts with an emphasised constituent

[1] It should be noted, however, that although this is quoted by various subsequent grammarians, this sentence is actually not found in Dr Davies's grammar of the Welsh language from 1621, where the section on syntax simply states: "Nominativae voces verbis praeponuntur interposito affirmandi adverbio *a* (...). Pro illo tamen *a*, Demetae dicunt *y*" (J. Davies, 1621[1809]:181-182).

followed by a particle *a/y(r)* and the verb with default third-person singular ending. The latter was a complex sentence with a cleft formation and a relative clause. Over the centuries the sentence-initial copula *ys* 'it is' was omitted and thus these disguised complex sentences with inverted order (called the 'Mixed' rather than 'Abnormal' order) were interpreted as 'simple' normal sentences in the Middle Welsh period.

Discussions on the exact origin of the prevalent 'inverted' or 'Abnormal' word orders in Middle Welsh and its development into the Modern Welsh period are continued in the following decades by John Morris-Jones (1931), Henry Lewis (1931, 1942), Melville Richards (1938) and J.J. Evans (1946). With the publication of D. Simon Evans's *Grammar of Middle Welsh* in 1964, the issues are far from solved, but the different word order patterns are now clearly defined:

- (particle)VSO (infrequent in Middle Welsh, but occurs in Old Welsh)
- subject / object / object (or subject) of verbal noun + *a/ry/yr* + verb ('Abnormal Sentence')
- adverb + *y(d)/yt/ry/yr* + verb ('Abnormal Sentence')
- (copula *ys*) + emphasised constituent + relative clause ('Mixed Order')

Formally, the distinctions that were made between the Abnormal and the Mixed orders were based on agreement patterns and negation. The relative verb in the Mixed order usually exhibits default third-person singular endings, but it should be noted that agreement patterns in Welsh vary considerably over time (cf. Koch (1991) and D. S. Evans (1971)). Willis (1998), furthermore notes that the different negative patterns reflect "an entirely unrelated distinction between constituent and clausal negation" (Willis, 1998:6). As soon as the sentence-initial copula *ys* was lost, there was no formal way to distinguish the two patterns. Another crucial question remained: if all these forms were possible which constituent exactly was placed before the verb in which specific contexts? This chapter aims to give a systematic overview of the word order patterns in Welsh. After briefly introducing previous scholarly literature, I list all possible patterns in Welsh and describe their respective word orders in detail with many examples from Old, Middle and Modern Welsh sources.

4.1.1 Functional approaches to word order variation

Proinsias MacCana's 1973 paper on the Welsh Abnormal Sentence initiated a vast body of literature on the variation of word order patterns in various Middle Welsh texts as well. Most of the following contributions were made by T. Arwyn Watkins (1977/78, 1983/84, 1987, 1988, 1991, 1993 and 1997), Erich Poppe (1988, 1990, 1991a, 1991b, 1993, 2000, 2009 and 2014), James Fife (1991, 1993 and, with Gareth King, 1991), Manning (1995), Manning (1997) and, in particular, and Manning (2004) and by MacCana himself (1979, 1990, 1991). Once the synchronic description of the abnormal word order pattern was generally accepted, attention shifted to its usage in various contexts. Why were there various ways of expressing

positive main declarative sentences? When were they subject-, object- or adjunct-initial and why then? Or was there random variation and could all patterns be used in any context?

Since comparing frequencies of different patterns in various texts could not sufficiently answer any of these questions (cf. Poppe (1993)), new researchers took a functional or pragmatic approach to this problem. Erich Poppe discovered that "variation in word order and sentence types is remarkably infrequent in sentences expressing the same or, at least, a very similar information content." (Poppe, 1990:458). Watkins, too, concluded that "we have a small and definable group of exceptions to a near-rigid rule in M[iddle] W[elsh] prose prohibiting the occurrence of the verb as the initial constituent in the positive declarative sentence." (T. A. Watkins, 1993:123). Poppe (1993) suggested a functional analysis for the 'fronting' construction (i.e. the abnormal/verb-second order): "The hypothesis is that frontings can be explained in terms of topic and focus. (...) Topicalization is interpreted to be the basic, unmarked pattern in a positive, main statement in MW prose." (Poppe, 1993:115).

As pointed out in the previous chapter, however, Information Structural terminology like 'topic' and 'focus' remained ambiguous for a long time. Poppe's research initially centered around the idea of "Situationskulisse" or the way in which the sentence can be linked to the situation in the preceding context by placing an adjunct (adverb or prepositional phrases) in initial position. Fife & King (1991) attempt to give clear definitions of various IS categories from a cross-linguistic perspective. But as Poppe notes, there are still instances of functional exceptions and ambiguities (Poppe, 2009:253). According to him, "all attempts to find motivations behind the actual word order patternings of Middle Welsh prose will in the final analysis have to reckon with variation resulting from a text-producer's considerable, but not unrestricted choice of syntactic options available for a specific context." (Poppe, 2014:100)

4.1.2 From Old Welsh to Middle and Modern Welsh

While "the thought of the giants of earlier generations... (Morris-Jones, Sir Ifor Williams and Henry Lewis)... continue to loom large" (Koch, 1991:3), research into the origin and use of the Abnormal order developed into two main directions. MacCana (1991), T. A. Watkins (1977) and Fife (1988) considered it as a mere literary phenomenon:

> "The literati of Middle Welsh took this pre-existing potential [the Abnormal Sentence - MM] and popularized it (among themselves) to the extent of overstepping the bounds of communicative usefulness. At that point fronting was done for fronting's sake alone."
>
> (Fife & King, 1991:144)

Alternatively, D. S. Evans (1968:336-7) and Koch (1991) considered the abnormal order a true feature of (spoken) Middle Welsh. According to Koch (1991), it was an innovation also seen in other Brythonic languages that only entered the literary

language in a later stage (i.e. later than Old Welsh). Willis (1998) argues that "this view is considerably simpler and involves far less 'special pleading', such as references to unverifiable developments and resort to artificial literary languages to explain away contradictory evidence" (Willis, 1998:18). He builds on this account in a generative syntactic framework to explain the subsequent loss of the abnormal order in the Early Modern Welsh period (see Chapter 7 for a detailed diachronic analysis of this construction).

4.2 The question of basic word order

Before moving on to the overview of patterns, we need to address the question of basic word order. Many of the above-mentioned studies of Welsh word order give overviews of the frequency and textual distribution of each pattern. The focus lies on positive declaratives that are main, rather than subordinate clauses. The most frequent pattern is then often called the 'basic' word order. Frequencies of certain patterns can, however, differ in every genre, in which case it would be necessary to specify that pattern X is most frequent in narrative native tales (but maybe not in, for example, historical chronicles). This task, be it somewhat laborious, could be done for each genre, register, style etc. In the frequency tables at the end of this chapter, therefore, all Welsh texts are displayed separately. The question remains: to what extent - if at all - does this say anything about the 'basic' word order in Middle Welsh on the whole (including the spoken language)?

Take for example the following statement from Oliver Currie (where PDMCs means Positive Declarative Main Clauses): "There does not seem to have been any single statistically predominant, basic word order in PDMCs in Middle Welsh prose; (...) In Modern Welsh, in contrast, verb-initial order has been grammaticalized as the basic word order." (Currie, 2000:206). In this context, 'grammaticalized' apparently means 'become statistically predominant in the grammar', which, in turn, means it therefore must be the 'basic' word order. This statement is, however, only meaningful if relative frequency is generally accepted as a decisive indicator for the "basicity" of word order of a language as a whole and if this is the case for all genres, registers etc.

From an information-structural perspective, there are various other ways of determining the 'basic' or 'canonical' word order of a language. Kirk (2012), for example, describes a neutral clause with 'basic' word order as "a clause in which no element has a special topic or focus interpretation"[2] (Kirk, 2012:27) (see also É.Kiss (1998) and Rizzi (1997)). She lists examples of generic and situational sentences, answers to broad focus questions (e.g. 'What happened?') and introductions to parables. These criteria are testable in spoken languages, but it is not always easy to find enough (or any) good examples in historical data.

If we compare the New Testament (NT) examples of situational sentences Kirk (2012:38) finds with VSO and SVO in Greek to their Middle and Modern Welsh

[2]No distinction is made between sentence and discourse topics. 'Topic' is to be interpreted as a constituent that is topicalised for example by ways of fronting.

translations, we see a clear verb-second (abnormal) order in Middle Welsh vs. a verb-initial pattern in Modern Welsh:

(1) New Testament Lk 7:16 'Everyone became afraid.'

a. ***élaben*** *dè phóbos pántas*
seize.PAST.3S PRT fear everyone
Lit. 'Fear seized all (people)' (NT Greek - VS)

b. *Ac ofn a* ***ddaeth*** *ar bawb*
and fear PRT come.PAST.3S on all
Lit. 'And fear came to everyone.' (Middle Welsh - V2)

c. ***Cydiodd*** *ofn ym mhawb*
take-hold.PAST.3S fear in all
Lit. 'Fear rose in everyone.' (Modern Welsh - VS)

(2) New Testament Lk 5:26 'And everyone became amazed.'

a. *kaì éxtasis* ***élaben*** *hápantas*
and amazement seize.PAST.3S everyone
Lit. 'And amazement seized everyone.' (NT Greek - SV)

b. *A syndod a* ***ddaeth*** *ar bawb*
and surprise PRT come.PAST.3S on all
Lit. 'And surprise came to everyone.' (Middle Welsh - V2)

c. ***Daeth*** *syndod dros bawb*
come.PAST.3S surprise through all
Lit. 'Surprise came to everyone.' (Modern Welsh - VS)

Answers to broad focus questions like 'What happened?' have SV(O) order in NT Greek. Their Middle Welsh translations are consistently verb-second and their Modern Welsh equivalents are either translated with VSO patterns or periphrastic constructions in which the finite verb (the auxiliary) is still clause-initial.

(3) New Testament Lk 1:34-35 '(How will this be, since I haven't been with a man?)'

a. *pneûma hágion* ***epeleúsetai*** *epì sé*
spirit holy come.FUT.3S upon you
'The holy ghost will come upon you.' (NT Greek - SV)

b. *Yr Ysbryd Glân a* ***ddaw*** *arnat ti*
the Ghost Holy PRT come.FUT.3S on.2S you
'The Holy Ghost will come upon you.' (Middle Welsh - V2)

c. ***Daw*** *'r Ysbryd Glân arnat*
come.FUT.3S the Ghost Holy on.2S
'The holy ghost will come upon you' (Modern Welsh - VS)

(4) New Testament Lk 1:35 '(The holy ghost will come upon you)'

a. *kaì dúnamis hupsístou* ***episkiásei*** *soi*
and power highest shadow.FUT.3S you
'and the power of the highest will overshadow you.' (NT Greek - SVO)

b. *a nerth y Goruchaf a 'th* ***gysgoda*** *di*
and power the Highest PRT 2S overshadow.FUT.3S you
'and the power of the Highest will overshadow you.' (Middle Welsh - V2)
c. *a* ***bydd*** *nerth y Goruchaf yn dy* ***gysgodi***
and be.FUT.3S power the Highest PROGR 2S overshadow.INF
'and the power of the highest will overshadow you.' (Modern Welsh - AuxSObjClV)

The picture is exactly the same in introductions to parables (although only one example here is cited in NT Greek):

(5) New Testament Lk 14:16 '(And he said to him),'
a. *ánthro:pós tis* ***epoíei*** *deîpnon méga*
man INDEF make.PAST.3S dinner large
'A certain man made a large dinner' (NT Greek - SVO)
b. *Rhyw ŵr a* ***wnaeth*** *swper mawr*
some man PRT do.PAST.3S dinner big
'Some man made a big dinner.' (Middle Welsh - V2)
c. *Yr* ***oedd*** *dyn yn* ***trefnu*** *gwledd fawr.*
PRT be.PAST.3S man PROGR make.INF dinner big
'A certain man made a large dinner' (Modern Welsh - AuxSVO)

The overall pattern in Welsh is very clear: Middle Welsh bible translators chose to use the abnormal sentence or verb-second pattern (SaVO) in each of these contexts. According to Kirk's definition, verb-second would thus be considered the 'basic' or 'neutral, unmarked' word order in Middle Welsh. In Modern Welsh, however, these sentences are consistently translated with verb-initial or auxiliary-initial orders. Modern Welsh could thus be described as having a VSO 'basic' word order in this way.

Since these types of sentence without 'topic' or 'focus' are not always easy to find in historical data, it is useful to consider some more clearly defined notions of information structure. In the previous chapter, ways of finding the focus articulation of a sentence have been described in more detail. According to Lambrecht (1994), Levinsohn (2009) and Van der Wal (2009), basic word order can be observed in sentences with predicate focus (i.e. topic-comment articulations). This is especially the case in narrative literature (Komen, 2013). There are furthermore other factors interacting with the focus articulation: the notions of 'Point of Departure' (or frame setting) and 'the Principle of Natural Information Flow'. Sentences with predicate focus that have no additional Point of Departure or marked information flow could be considered to exhibit 'basic' word order from this point of view.

As will become clear in Chapters 5 and 6, from this perspective the subject-initial or adjunct-initial versions of the abnormal sentence would be the 'basic' word order in Middle Welsh.

4.3 Overview of word order patterns

Word order patterns can be described in various ways. The most basic approach only takes the finite verb and its core arguments (the subject and the direct object) into consideration, resulting in six logical possibilities (SOV, SVO, VSO, VOS, OSV, OVS). This approach is useful when comparing languages on a very large scale. On the very opposite end of the spectrum lie various theoretical frameworks describing the underlying structural configurations and modifications of the different patterns in great detail. The latter can help test predictions and thus verify hypotheses about types of word order variation and change. I leave those types of analyses for the next chapters. In this chapter, I focus on the superficial word order patterns that can be observed in Middle Welsh. Apart from the finite verb and its core arguments, I take adjuncts and other functional elements into consideration as well in order to give an exhaustive overview of all possible patterns.

In this section I present all word order patterns found in positive declarative main clauses in Welsh. The description focusses on the surface order of the verb and its core arguments and how the respective word order patterns are treated in scholarly literature. Copular and non-verbal clauses are discussed as well, though only the syntax of identificatory copular clauses will be analysed in greater detail in Chapter 6. The following types of word order patterns exist in Welsh positive declarative main clauses:

I Verb-initial (VSO)
  (a) VSO (verb absolute clause-initial)
  (b) particle VSO

II Periphrastic constructions with initial auxiliary (AuxSVO)
  (a) with auxiliary *bod*
  (b) with auxiliary *gwneud*
  (c) with auxiliary *ddaru*

III Verb-second after adjuncts ('Abnormal Sentence')
  (a) AdjP *y* VSO
  (b) PredP *y* VSO
  (c) AspP *y* VSO
  (d) AdvP *y* VSO
  (e) PP *y* VSO

IV Verb-second after arguments and VNs ('Abnormal Sentence')
  (a) S *a* V_{agree} O
  (b) O *a* V S
  (c) patient *a* $V_{impersonal}$
  (d) VN *a* DO_{infl} (*gwneuthur*-periphrasis)

V Verb-second after focussed items ('Mixed Sentence')
  (a) *(ys)* focussed noun/pronoun *a* V_{3sg}
  (b) *(ys)* focussed adjunct *y* V_{3sg}

VI Bare verbal nouns
  (a) VN + agent
  (b) VN + *o* + agent
  (c) *a(c)* VN (continuing previous finite clause)

VII Copular clauses
  (a) SCP
  (b) PCS
  (c) CPS
  (d) C S *yn* P
  (e) C S *(ys)sydd* P

VIII Identificational Focus construction
  (a) Sef + DP (+ relative)
  (b) Sef + *yw/oed*
  (c) Sef + *a/y*

IX Non-verbal clauses
  (a) *dyma/dyna/llyma/llyna* + S (truncated copular clause)
  (b) S *(yn)* P
  (c) PS
  (d) Absolutive: Ac S P(P)

4.3.1 Type I: Verb-initial (VSO)

Absolute verb-initial word order is found in all stages of the language, though it is rare and only used in very specific contexts in Middle Welsh. T. A. Watkins (1987) argues that the verb-initial word order is characteristic of Old Welsh prose, but the evidence for this, once embedded and negative clauses are removed from his data, is meagre. There are certainly not enough Old Welsh sources for us to establish what the basic word order was at that time, whichever of the above-mentioned methods (statistical or information-structural) is used.

(6) *prinit hinnoid .iiii. aues*
buy.PRES.3S that four birds
'That buys four birds' (Old Welsh Ox. 234.33 - Willis (1998:10))

In Middle Welsh there are more examples of absolute verb-initial word order, but they seem to be restricted to specific contexts:
(a) Impersonal verbs
(b) Imperatives

(c) verba dicendi ('said he')
(d) answers or direct responses to questions or commands
(e) oaths and other idiomatic sayings

(7) a. *Gorucpwyt hynny.*
do.PAST.IMPERS that
'That was done.' (Impersonal verbs - CO 519)

b. *Aet y porthawr allan*
go.PRES-IPV.3S the gatekeeper out
'Let the gatekeeper go out!' (Imperatives - CO 798)

c. *Amkawd y wrach, Nyd oes plant itaw.*
say.PAST.3S the hag not be.3S children to.3MS
'The hag sad: 'He doesn't have children." (Verba Dicendi - CO 38)

d. *Gwelem arglwyd heb wy mynyd mawr (...)*
see.PAST.1P lord said they mountain big (...)
'We saw, they said, a big mountain (...).' (Answer - Branwen 265)

e. *Henpych gwell. Arglwyd. heb ef*
be.PRES-SUBJ.2S well Lord said he
'Hail Lord, said he' (Idiom - Gereint 32)

In Modern Welsh, VSO order is called *y frawddeg seml* 'the simple sentence' by most grammarians (cf. Richards (1938)). Stephen J. Williams in his 1980 grammar tends to use the term 'normal sentence' alongside 'simple sentence', indicating that this is the most common word order in Modern Welsh. Anwyl (1899) does the same, but Gareth King uses the term 'basic order' (as opposed to what he calls the focussed, i.e. verb-second, order). Examples like (8) are given in most Welsh grammars and also taught in very popular Welsh for Adults courses. Some native speakers, however, seem sceptical about the actual use of these forms. To them, verb-initial orders without either a sentence-initial particle or soft mutation on the initial consonant of the verb like (8) seem highly literary at the very least:

(8) *Gwelodd y plentyn geffyl.*
see.PAST.3S the child horse
'The child saw a horse.' Williams (1980)

Clauses with sentence-initial particles *fe* (in South Wales) or *mi* (in North Wales) like (9c) are commonly found in Modern Spoken Welsh. In Middle Welsh it was also possible to start a sentence with a preverbal particle, but again, examples of those in absolute sentence-initial position are very limited:

(9) a. *Y dywawt Diwrnach (...)*
PRT say.PRET.3S Diwrnach (...)
'Diwrnach says (...)' (CO 1038)

b. *E doeth im heb ef (...)*
PRT come.PAST.3S to.1S said he (...)
'It came to me, said he (...)' (Branwen 148-149)

c. *Fe gyfyd yr afon yn uwch.*
PRT rise.PRES.3S the river PRED higher
'The river will rise higher.' Anwyl (1899)

In contexts of narrative continuity there are many more examples of sentences with preverbal particles in Middle Welsh. According to Willis (1998), however, these examples are only superficially verb-initial. Underlyingly, these sentences exhibit topic-drop and are thus not proper examples of verb-initial order in Middle Welsh. According to Currie (2000), even in the Early Modern Welsh period "we still find several prose texts with either no examples at all of AIV [Absolute Initial Verb - MM] order in the sections analysed." (Currie, 2000:207). In other texts, however, the frequency of verb-initial patterns in positive main declaratives steadily increases. It is furthermore worth noting that verb-initial orders are consistently found after many conjunctions as in (10a), in finite subordinate clauses as in (10b) and in contexts with clausal negation as in (10c) throughout the history of Welsh:

(10) a. *(...) fel y lladdwyf ef*
(...) so that PRT kill.PRES-SBJ.1S him
'so that I could kill him' (b1588 - 1 Sam. 15.19)
b. *O gwnaeth hitheu gam, kymeret (...)*
if do.PAST.3S she wrong take.PRES-IPV.3S (...)
'If she has done wrong, let her take (...)' (PKM 21.17-18)
c. *Ny symudawd Peredur y ar y vedwl (...)*
NEG move.PAST.3S Peredur from 3MS thought (...)
'Peredur did not move from his thoughts (...).' (Peredur 31.2)

Some grammarians call both types (with or without the sentence-initial particle) 'simple' or 'normal' sentences (cf. D. S. Evans (2003 [1964]), Williams (1980) and Richards (1938)), others do not make a distinction between the two (cf. Thorne (1993), King (1993), Morris-Jones (1931) and Anwyl (1899)).

4.3.2 Type II: Periphrastics with initial auxiliary (AuxSVO)

There are different types of periphrastic constructions available in Welsh. These are sentences in which the main verb is a verbal noun and the inflection appears on an auxiliary verb. Three of the main auxiliaries used are inflected forms of *bod* 'to be', *gwneud* 'to do' or *darfod* 'to happen'. The inflected forms of *bod* in Middle Welsh were followed by the subject + an aspectual marker *yn* or *wedi*, resulting in progressive or perfective aspect respectively.

(11) a. *Mae uyg kallon yn tirioni vrthyt.*
be.PRES.3S 1S heart PROGR grow-fond.INF with.2S
'My heart inclines toward you.' (CO 166)
b. *(...) y mae y gwyr hynn yn mynnu an llad*
(...) PRT be.PRES.3S the men these PROGR want.INF 1P kill.INF
'(...) these men want to destroy us' (PKM 54.25)

There are also examples of periphrastic constructions in Middle Welsh in which the auxiliary is not sentence-initial. They can be found in sentences with the abnormal word order or sentences with contrastively focussed elements in sentence-initial position. The examples in (12) with periphrastic constructions are therefore not taken into account here. They are discussed in the sections of their respective word order pattern (types III and V) below.

(12) a. *ac yna yd oyd marchawc y llamysten yn dodi yr*
and then PRT be.PAST.3S knight the sparrow-hawk PROGR place.INF the
ostec
silence
'and then the knight of the sparrow-hawk was ordering silence' (Gereint 277)

b. *mi yd wyt yn y geissaw*
me PRT be.PRED.2S PROGR 3MS search.INF
'It is me you are looking for' (Peredur 28.25-26)

In Modern Welsh these constructions have greatly increased in frequency (cf. Borsley et al. (2007:303)) to the extent that they have taken over the function of the present-tense paradigm to denote present time (causing the present-tense paradigm to shift to function as a modal future). They are abundantly used in the spoken language as well (which auxiliary is preferred is dialectally determined, as shown in examples 13a-c). Even stative verbs are possible, as shown in (13d), indicating that the progressive aspect is not necessary:

(13) a. *Mae Elin wedi/yn prynu torth o fara.*
be.PRES.3S Elin PERF/PROGR buy.INF loaf of bread
'Elin has bought/is buying a loaf of bread.' (Borsley et al., 2007:12)

b. *Gwnaeth Elin brynu torth o fara.*
do.PAST.3S Elin buy.INF loaf of bread
'Elin bought a loaf of bread.' (Borsley et al., 2007:12)

c. *Ddaru Elin brynu torth o fara.*
PAST Elin buy.INF loaf of bread
'Elin bought a loaf of bread.' (Borsley et al., 2007:12)

d. *Dw i 'n gwybod yr ateb.*
be.PRES.1S I PROGR know.INF the answer
'I know the answer.' (Borsley et al., 2007:12n.5)

4.3.3 Type III: Verb-second after adjuncts ('Abnormal')

The third type of word order pattern under investigation is the infamous abnormal sentence discussed abundantly in previous literature as mentioned above. In Welsh grammar, this type of word order is called *y frawddeg annormal* 'the abnormal sentence' (cf. among others Richards (1938)). Anwyl (1899) refers to it as the 'inverted order' and thus does not distinguish this from the other order in which the verb comes in second position following a focussed constituent (see the section on

the 'Mixed Sentence' below). Other names for this construction are 'cleft-fronted' (T. A. Watkins, 1993), 'X1-order' (Poppe, 2009), 'verb-medial' (Currie, 2013) or 'verb-second' (Willis, 1998).

All of these show the finite verb is not the first, but the second core constituent of the clause. The initial position in the sentence could first of all be filled by an adjunct. This first constituent could be an aspectual, adjectival, adverbial (including predicational) or prepositional phrase.

(14) a. *Ac yn ymlad a r pryf hwnnw y colleis i vy llygad*
and PROGR fight.INF with the animal that PRT lose.PAST.1S I 1S eye
'And fighting with that animal I lost my eye.' (Peredur 45.8-9)

b. *Blin a lludedic y th welaf.*
tired and weary PRT 2S see.PRES.1S
'I see you (are) very tired' (WM 168.27-28)

c. *Y trydyd dyd yd ymladawd Arthur e hun ac ef*
the third day PRT fight.PAST.3S Arthur 3MS self with him
'On the third day Arthur himself fought with him.' (CO 1072)

d. *Ac yn diannot y doeth tan o r nef*
and PRED immediate PRT come.PAST.3S fire from the heaven
'And without delay came fire from the sky.' (Dewi 9.10)

e. *Yn yr awr honno y dywedodd yr Iesu wrth y dyrfa*
in the hour that PRT say.PAST.3S the Jesus to the crowd
'In that moment Jesus said to the crowd (...)' (b1588 - Mat. 26.55)

Verb-second sentences with sentence-initial adjuncts are characterised by the form of the preverbal particle *y(d)* (as opposed to the particle *a* found after subjects or objects as in Type IV discussed below). Examples with subordinate clauses preceding the main clause could be considered to be part of this adjunct-initial word order pattern too, since the same particle *y(d)* is used:

(15) *Ban agorer y creu beunyd yd a allan.*
when open.IMPERS the pen each.day PRT go.PRES.3S out
'When the pen is opened every day it goes out.' (PKM 89.3-4)

Sentences of this type are said to bear no particular emphasis on the first constituent. The sentence-initial adjuncts can, however, function as topics (see Poppe (1989) for a description of those constituents as frame setting topics or 'Situationskulisse'). Examples like these are still possible in Modern Welsh as is shown in (16a). Without context, however, it is very difficult to determine whether the initial constituent is focussed or not. Focussed adverbs, like *hwyrach* 'probably' in (16b), are found with the exact same superficial word order pattern (the preverbal particle *y* can be left out):

(16) a. *Yma y gwelsom ef.*
here PRT see.PAST.1P him
'Here we saw him' Williams (1980)

b. *Hwyrach (y) bydd rhaid i chi aros.*
probably (PRT) be.FUT.3S necessity to you wait.INF
'You'll probably have to wait.' (Borsley et al., 2007:124)

Unlike propositional adverbs like *efallai* 'maybe', *braidd* 'hardly' and *hwyrach* 'probably', temporal adverbs in sentence-initial position in Modern Welsh are followed by the preverbal particle *fe*:

(17) *Yfory fe fydd rhaid i chi aros.*
tomorrow PRT be.FUT.3S necessary to you wait.INF
'Tomorrow you will have to wait.'

In Middle Welsh the focussed and topicalised adverbs occupied the same sentence-initial position rendering the same superficial Adjunct-*y(d)*-Verb-Subject. There are, however, also examples with more than one sentence-initial adjunct or with adverbs preceding any of the other word order patterns discussed in this chapter.

4.3.4 Type IV: Verb-second after arguments ('Abnormal')

As mentioned above, core arguments can also appear in sentence-initial position. When subjects or direct objects are preceding the finite verb, the preverbal particle is not *y(d)* (as with adjuncts), but *a*. Subjects in sentence-initial position in Middle Welsh usually agree with the finite verb.[3] Agreement is thus the main feature distinguishing this word order pattern from the other verb-second pattern with focussed sentence-initial constituents (the 'Mixed Sentence') described in the next section (see also chapter 6 for discussion of this issue).

Examples of subject-initial order can already be found in Old Welsh:

(18) *Gur dicones remedaut elbid a-n-guorit*
man create.PAST.3S wonder world PRT-1P-redeem.PRES.3S
'The man who created the wonder of the world redeems us.'(Juv. 5a-b - Willis (1998:10))

In Middle Welsh, this word order pattern can be found with pronouns (as in (19a)), demonstratives (as in (19a)) or full noun phrases in initial position (as in (19c)). Demonstratives and noun phrases in this position could function both as subjects or as direct objects of the finite verb (which can appear in any type of tense, mood or diathesis):

(19) a. *Vynt a gerdassant racdunt.*
they PRT walk.PAST.3P against.3P
'They walked towards them.' (PKM 50.11)
b. *A hwnnw a doeth yma o iwerdon.*
and that PRT come.PAST.3S here from Ireland
'And that one came here from Ireland' (PKM 35.5-6)

[3] But see D. S. Evans (2003 [1964]) for a detailed discussion and some counter-examples.

c. *Duw a ch notho.*
God PRT 2P reward.PRES-SUBJ.3S
'May God reward you.'
d. *a honno a elwir kaer yr Enryfedodeu*
and that PRT call.IMPER castle the wonders
'and that one is called Castle of Wonders' (Peredur 66.9-10)
e. *A deu drws a welynt yn agoret*
and two door PRT see.PAST.3P PRED open
'and they saw two doors that were open' (PKM 46.22)
f. *A hynny a dywetpwyt idi.*
and that PRT say.PAST.IMPERS to.3FS
'And that was said to her.' (PKM 80.11)

Verbal nouns could also occur in sentence-initial position. If this was the case, they were also followed by the preverbal particle *a* because they function as the direct object of the inflected form of the auxiliary *gwneuthur* 'to do'. Transitive verbal nouns could occur with their internal arguments in genitive apposition (20b). As in other genitive constructions in Welsh, pronominal arguments are cliticised and optionally doubled before and after their verbal nouns (20c). Prepositional phrases and other adverbials can also follow the initial verbal noun (20d). This periphrastic VN*a*DO construction can appear with impersonals or passives (20e) as well.

(20) a. *Kynhewi a oruc Pwyll.*
fall-silent.INF PRT do.PAST.3S Pwyll
'Pwyll fell silent.' (PKM 14.12)
b. *a pharattoi y varch a e arueu a oruc.*
and prepare.INF 3MS horse and 3MS weapons PRT do.PAST.3S
'And he prepared his horse and his weapons.' (Owein 231)
c. *A e aros ynteu a wnaeth Manawydan*
and 3MS wait.INF him PRT do.PAST.3S Manawydan
'And Manawydan waited for him' (PKM 56.20)
d. *a y alw attaw a wnaeth*
and 3MS call.INF to.3MS PRT do.PAST.3S
'and he called him to him' (PKM 81.14-15)
e. *Bedydyaw a wnaethpwyt y mab.*
baptise.INF PRT do.PAST-IMPERS the son
'The son was baptised.' (PKM 77.23-24)
f. *A gwybot a wnaeth Arthur (...)*
and know.INF PRT do.PAST.3S Arthur (...)
'And Arthur knew that (...)' (BR 12.16)
g. *A goresgyn y gaer a oruc a e gyuoeth.*
and conquer 3MS castle PRT do.PAST.3S and 3MS wealth
'And he conquered his castle and his wealth' (CO 1241)

Certain verbal nouns like *gwneuthur* 'to do', *bod* 'to be', *geni* 'to be born' or *cael* 'to obtain' never appear in sentence-initial position followed by the inflected form of

gwneuthur 'to do' (cf. T. A. Watkins (1993) who lists other verbs like *gwybod* 'to know' as well, but examples of these do in fact exist, as shown in (20f)).

Prepositional phrases and adverbs can precede or follow the subject or object. The finite verb in these cases appears to be in third or fourth rather than second position. According to Willis (1998), the verb-second analysis can be maintained, however, because it is only possible to add adjuncts before the verb, there are never two core arguments taking up the sentence-initial position. Even 'heavy', i.e. longer and or more complex adjuncts, can appear before the finite verb, as shown in (21e) and (21f). The first constituent (counting for the V2 structure) is shown in parentheses.

(21) a. *a [hwnnw] gwedy hynny a uu escob*
and that after that PRT be.PAST.3S bishop
'and afterwards he was bishop' (Dewi 2.14)

b. *Hir bylgeint [Guydyon] a gyuodes.*
early.morning Gwydion PRT get.up.PAST.3S
'Early next morning, Gwydion got up.' (PKM 82.5-6)

c. *Ac ar hynny [arouun y longeu] a wnaeth ef.*
and on that make-for.INF 3P ships PRT do.PAST.3S he
'And thereupon he made for their ships.' (Branwen 85)

d. *Mi hagen a uydaf gyuarwyd ywch*
I however PRT be.FUT.1S guide to.2P
'But I will be guiding you' (CO 869)

e. *A [chyuarch gwell eissoes y Owein] a oruc ef*
but greet well still to Owein PRT do.PAST.3S he
'But he still welcomed Owein' (BR 14.13-14)

f. *A [gouyn pwy oet] a oruc.*
and ask.INF who be.PAST.3S PRT do.PAST.3S
'And he asked who he was.' (CO 165-166)

According to Fife, "The versions of fronting where the full array of adjuncts is fronted along with the VN seem more natural or unmarked than those where the adjuncts are split up. [...] The reason is that verbs form tighter units with their adjuncts than they do with their subjects." (Fife, 1986:141). Willis (1998) claims that there are four types of adverbs and three possible preverbal positions, before the topic (i.e. fronted constituent), as the topic (word order Type III above) or following the topic. Topic adverbials (*guedy hynny* 'after that') can obviously be in topic position, but they can also precede the topic. Prepositional arguments of verbal nouns (*(trigaw) ar hynny* '(decide) on that') can only appear in topic position. Both consituent adverbials (*hagen* 'however', *heuyt* 'also') and non-topic adverbials (*eiss(y)oes* 'nevertheless') follow the topic, although the latter are also found in pretopical position.

4.3.5 Type V: Verb-second after focussed items ('Mixed')

Superficially, this type of word order pattern is very similar to that of the previous types of the abnormal sentence discussed above. It is also a verb-second pattern, but most Welsh grammarians have kept this type apart because the sentence-initial constituent of the mixed sentence is focussed and the finite verb exhibits default third-person singular inflection most of the time. According to T. A. Watkins (1993), "This sentence reveals an earlier syntactic stage of the cleft sentence, with the copula preceding the fronted constituent." (T. A. Watkins, 1993:126). This fronted constituent is then followed by a relative clause, which would explain the lack of agreement, since agreement is hardly ever found in relative clauses in Welsh.

(22) a. *bydhawt ragot ti gyntaf yd agorawr y porth*
be.FUT.3S to.2S you first PRT open.IMPER the gate
'for you shall the gate be opened first' (WM 456.34)

b. *Oed maelgun a uelun in imuan*
be.PAST.3S Maelgwn PRT see.PAST.1S PROGR fight.INF
'It was Maelgwn that I could see fighting' (YMTh 57.5)

c. *Ys mi a e heirch*
be.PRES.3S I PRT 3FS seek.PRES.3S
'It is I who seek her. (White Book WM 479.29)

Once the copula was lost (through phonological erosion in the Early Middle Welsh period, see Chapter 7), superficially, it was difficult to distinguish these mixed sentences with third-person subjects from their unfocussed abnormal counterparts. There are indeed examples of lack of agreement between verbs and their subjects that should be interpreted as contrastively focussed (e.g. the examples in (23)). But, as shown in example (24), there were also examples in late Middle Welsh at least of contrastively focussed subjects that *do* show agreement.

(23) a. *Mi a e heirch*
I PRT 3FS seek.PRES.3S
'It is I who seek her.' (Red Book CO 566)

b. *neu vinheu a orffei arnaw*
or I PRT overcome.PAST-SUBJ.3S on.3MS
'(he would overcome me) or *I* would overcome him' (Owein 96)

c. *Miui, heb yr Scuthyn, a uyd gwassanaethwr hediw.*
I.strong said Scuthyn PRT be.FUT.3S minister today
'*I*, said Scuthyn, will be minister today.' (Dewi 12.2)

d. *ac euo a welei bawp*
and he.strong PRT see.PAST.3S all
'(no one would see him), but *he* would see everyone' (BR 11.21-22)

(24) *ti a i ddywedaist*
you PRT 3MS say.PAST.2S
'(Are you king of the Jews? Jesus said to him:) It's you who's saying that.' (b1588 - Mat. 11.27)

In Welsh, this type of sentence is called *y frawddeg gymysg* 'the mixed sentence'. It still exists in Modern Welsh and is often referred to as the 'focussed sentence' (King, 1993).

(25) a. *Y plentyn a redodd adref.*
the child PRT run.PAST.3S home
'*The child* ran home.' (Williams, 1980:223)

b. *Dim ond hyn gollais i.*
only that lose.PAST.1S I
'I lost only that.' (Borsley et al., 2007:123)

Further syntactic differences between the abnormal and the mixed word orders are described by, amongst others, Fife and King (1991) and Tallerman (1996) (see chapter 6 for further discussion of this issue).

4.3.6 Type VI: Bare verbal nouns

In Middle Welsh verbal nouns could also be used in declarative main clauses instead of a finite verb. These constructions are called 'historical infinitives' by Tallerman and Wallenberg (2012). The word order in these clauses is Verbal Noun - Subject (or Agent, from a semantic point of view, though other thematic roles are possible as well). It occurs in root and independent clauses in various contexts, some of which are optional, others seem obligatory (Tallerman & Wallenberg, 2012:1). The interpretation is always past tense and the subject can be null as in (26a) or overtly expressed in two ways: in apposition to the verbal noun (26b) or following the verbal noun and a preposition *o* 'of' (26c) or *y* 'to' (26d).

(26) a. *Kymryt gwrogaeth y gwyr a dechreu guereskynn y wlat.*
accept.INF homage the men and begin.INF subdue.INF the land
'He received the homage of the men and began to subdue the land.' (PKM 6.12)

b. *Dyuot Caswallawn am eu penn a llad y chwegwyr*
come.INF Caswallawn about 3P head and kill.INF the six.men
'Caswallon fell upon them and killed the six men.' (PKM 46.2)

c. *A chaffael mab ohonu trwy weti y wlad.*
and get.INF son from.3P through pray.INF the country
'And through the country's prayers they got a son.' (CO 4)

d. *Canu englyn idaw ynteu yna*
sing.INF englyn to.3MS him then
'He sang an englyn then' (PKM 90.9)

Example (26a) furthermore shows that this construction can occur in co-ordinated main clauses as well. Usually, however, the first clause is formally finite and all the following clauses contain just the verbal noun: the subject/agent is very often the same and thus omitted. The abnormal order with a verbal noun + periphrastic form of *gwneuthur* 'to do' frequently occur in the first main clause as in (27), but other types of word order patterns can occur as well as shown in (28).

(27) a. *(...) kyuodi a oruc a dyuot y Lynn Cuch*
(...) rise.INF PRT do.PAST.3S and come.INF to Llyn Cwch
'he got up and came to Llyn Cwch' (PKM 1.8)

b. *Ac yn gyflym diskynnu a oruc Gereint a llidiaw a thynnu cledyf a y gyrchu (...)*
and PRED quick dismount.INF PRT do.PAST.3S Gereint and get-angry.INF and draw.INF sword and 3MS wield.INF (...)
'And quickly Gereint dismounted and he got angry and drew a sword and wielded it (...).' (Gereint 309-310)

(28) a. *Y kyuodes y marchawc enteu a thynnu cledyf arall yn erbyn Gereint.*
PRT rise.PAST.3S the knight however and draw.INF sword other against Gereint
'The knight rose and drew another sword against Gereint.' (Gereint 310-311)

b. *A r llythyr a rwymwyt am uon eskyll yr ederyn a y anuon parth a chymry.*
and the letter PRT bind.PAST.IMPERS on quill the bird and 3MS send.INF towards Wales
'And the letter was bound to the quill of the bird and sent to Wales.' (PKM 38.11-12)

Bare verbal nouns only exist in co-ordinated and subordinate clauses in present-day Welsh. The two tenseless patterns with expressed agents no longer occur on their own.

4.3.7 Type VII: Copular clauses

Copular clauses exhibit various word order patterns in Welsh. In Old Welsh, there is not enough data to be able to establish the context and thus information-structural status of all examples, but it is clear that the copula was always sentence-initial. A cleft construction with *(ys)sydd*, the relative form of the verb *bod* 'to be', could be used to focus the subject.

In Middle Welsh, both copula (C) - predicate complement (P) - subject (S), CPS, and PCS orders existed, though the copula-initial order was on its way out, since *is/ys* phonologically eroded in Early Middle Welsh. It was replaced by other forms of the verb *bod*, like *mae* in initial position. In medial position, the copula took the form *yw/ynt* (present singular/plural) or *oed/oedynt* (past singular/plural).

	Unmarked	**Marked**	
		Focus predicate	Focus subject
Old Welsh	CPS(?)	CPS(?)	C S *(ys)sydd* P
Middle Welsh	CPS & *(y) mae* S *yn* P	CPS & PCS	(C) S *(ys)sydd* P
Modern Welsh	*(y) mae* S *yn* P	PCS	S *(ys)sydd* P

Table 4.1: Copular word orders: C = copula *is/ys*, P = Predicate complement, S = Subject

Not mentioned in the above table are copular sentences with presentational or thetic focus articulation. They can for example be found in Middle Welsh to introduce a narrative tale. Both the subject and the predicate complement represent new information in these cases and the word order is Subject - *(a)* Copula - *(yn)* Predicate complement.

(29) a. *Pwyll Pendeuic Dyuet a oed yn arglwyd ar seith cantref Dyuet.*
Pwyll Pendeuic Dyuet PRT be.PAST.3S PRED lord on seven cantref Dyuet
'Pwyll P.D. was lord of the seven cantrefs of Dyfed.' (PKM 1.1)
b. *Bendigeiduran uab Llyr a oed urenhin coronawc ar yr ynys hon*
Bendigeidfran son Llyr PRT be.PAST.3S king crowned on the island this
'Bendigeidfran son of Llyr was crowned king of this island.' (PKM 29.1)
c. *Math uab Mathonwy oed arglwyd ar Wyned*
Math son Mathonwy be.PAST.3S lord on Gwynedd
'Math son of Mathonwy was lord of Gwynedd.' (PKM 87.7-8)

Unmarked copular clauses in Old and Middle Welsh have topic-comment or 'Predicate focus' articulation. They mainly exhibit CPS word order, but in Middle Welsh, constructions with sentence-initial *mae*, the other inflected form of *bod*, are also found. The subject is in these cases followed by a predicative marker *yn*, as shown in (31b).

(30) a. *is moi hinnoid*
be.PRES.3S more DEM
'this is more' (**CPS**: Old Welsh M&P 23r - Zimmer 1999)
b. *Ys gohilion hwnn*
be.PRES.3S remainder DEM.MS
'He is what remains' (**CPS**: Middle Welsh CO 472)

(31) a. *Ys dyhed a beth gadu dan wynt (...) y kyfryw dyn*
be.PRES.3S bad of thing leave.INF under wind (...) the such man
'Tis a deplorable thing to leave such a man out in the wind (...).' (**CPS**: Middle Welsh CO 133-134)
b. *ac y maent yn barawt*
and PRT be.PRES.3P PRED ready
'and they are ready' (***mae* S *yn* P** - PKM 87.20-21)

Word order patterns that were considered 'marked' by Welsh grammarians are

employed in clauses with constituent focus articulation. The situation in Old and Early Middle Welsh is not exactly clear due to a lack of evidence. CPS, as shown in (32), could be one of the options in Old Welsh. If the predicate complement was focussed, this appeared in sentence-initial position, as in (33). If the subject was focussed, a cleft construction with a relative form of the verb *bod* 'to be' was used, as in (34a).

(32) *Oed gwynnach y chnawd no distrych y donn*
be.PAST.3S whiter 3FS skin than foam the wave
'Her skin was whiter than the foam of the wave.' (**CPS**: (Early) MW CO 491)

(33) a. *A recdouyd ynt y gwraged weithon.*
and chief-giver.P be.PRES.3P the woman.P these.days
'Women are dispensers of gifts these days.'(**PCS**: Middle Welsh CO 17-18)
b. *mab y dynnyon mwyn yw*
son the men gentle be.PRES.3S
'He is the son of gentle folk' (**PCS**: Middle Welsh PKM 23.9-10)

Marked order (Constituent focus subject and (reduced) cleft):

(34) a. *Is aries isid in arcimeir E*
be.PRES.3S Aries be.REL.3S in opposite E
'It's Aries which is opposite E.' (**CS*isid*P** Old Welsh - Comp. 13/4)
b. *Arthur yssyd geuynderw yt*
Arthur be.PRES.3S cousin to.2S
'Arthur is a cousin of yours.' (***Sysydd*P** Middle Welsh CO 57)

In Modern Welsh, predicative copular constructions exhibit the order copula - subject - *yn* predicate. If the Predicate is focussed, it can be fronted, in which case the medial form of the copula *yw/ydy* appears, as in (35d). The subject can also be focussed, resulting in the relative form of the copula *sy(dd)*, as in (35e):

(35) a. *Mae Gwyn yn ddiog.*
be.PRES.3S Gwyn PRED lazy
'Gwyn is lazy.'
b. *Mae Gwyn yn feddyg*
be.PRES.3S Gwyn PRED doctor
'Gwyn is a doctor.' (Borsley et al., 2007:43)
c. *Mae Caerdydd yn ddinas hardd.*
be.PRES.3S Cardiff PRED city beautiful
'Cardiff is a beautiful city.'
d. *Dinas hardd yw Caerdydd.*
city beautiful be.PRES.3S Cardiff
'Cardiff is a beautiful city.' (Borsley et al., 2007:130)
e. *Caerdydd sy 'n ddinas hardd.*
Cardiff be.PRES.REL PRED city beautiful
'It's Cardiff that is a beautiful city. / Cardiff is a beautiful city.' (Borsley et al., 2007:131)

Identity copular constructions are called *brawddeg enwol amhur* 'impure nominal sentence' in Modern Welsh. Presentational interpretations are impossible when the referent of the subject is a member of the set designated by the predicate. The lexical semantics of the subject and predicate are such that the latter cannot be understood as a property predicated of the former. Therefore, example (36a) is infelicitous, but the construction with the medial copular form *yw/ydy* in (36b) with identificational meaning is grammatical:

(36) a. *#Mae 'r ateb yn rhaff.*
is the answer PRED rope
'The answer's a rope.'
b. *Rhaff ydy 'r ateb.*
rope is the answer
'The answer's a rope.' (Zaring, 1996:123)

In the examples in (37), "[t]he more natural interpretation is with *Caerdydd* as topic and *prifddinas Cymru* as new information" (Borsley et al., 2007:130) answering the question in (38a) with a falling intonation on *Cymru* followed by an intonational break. If *Caerdydd* has falling intonation it can be interpreted as new information answering question (38b).

(37) a. *Prifddinas Cymru yw Caerdydd.*
capital Wales be.PRES.3S Cardiff
'Cardiff is the capital of Wales.'
b. *Caerdydd yw prifddinas Cymru.*
Cardiff be.PRES.3S capital Wales
'The capital of Wales is Cardiff.' (Borsley et al., 2007:130)

(38) a. *Beth yw Caerdydd?*
what be.PRES.3S Cardiff
'What is Cardiff?
b. *Pa ddinas yw prifddinas Cymru?*
which city be.PRES.3S capital Wales
'Which city is the capital of Wales?' (Borsley et al., 2007:130)

Similarly with a predicative meaning, example (37b) repeated below as (39) is ungrammatical, but it is perfectly fine with an identificational meaning:

(39) *#Caerdydd yw prifddinas Cymru.*
Cardiff be.PRES.3S capital Wales
('The capital of Wales is Cardiff.') (Borsley et al., 2007:131)

The development of copular constructions is discussed in more detail in chapter 6.

4.3.8 Type VIII: Identificational focus with *sef*

Old and Middle Welsh employed a special construction to focus identity predicates. The definite predicate noun phrase of an identificatory copular clause could be focussed by means of the copula + pronominal anticipatory predicate preceding the subject and focussed predicate. This combination of copula *ys* + pronominal became the petrified *(ys)sef* 'it is it' once the copula was phonologically eroded and the agreement was lost. This subsequently gave rise to further grammaticalisation and the development of different types of '*sef*-constructions' in Middle Welsh (cf. Borsley et al. (2007:318), E. Evans (1958) and T. A. Watkins (1997)).

Old Welsh:

(40) a. *is em hi chet tri uceint torth*
be.PRES.3S it 3FS tribute three twenty loaf
'this is its tribute, sixty loaves' (LL, xlv - Watkins 1997:579)
b. *iss em i anu Genius*
be.PRES.3S it 3MS name Genius
'that's his name, Genius' (gl. *Genius* in Martianus Capella - Watkins 1997:579)

Middle Welsh:

(41) a. *Ys hwy yr rei hynny, Nynhyaw a Pheibyaw*
be.PRES.3S they the ones DEM.P Nynniaw and Peibiaw
'Those are Nynniaw and Peibiaw' (CO 598 - Borsley et al. 2007:318)
b. *Sef seithwyr a dienghis Pryderi Manawydan (...)*
sef seven.men PRT escape.PAST.3S Pryderi Manawydan (...)
'These were the seven men who escaped, Pryderi, Manawydan (...).' (WM 56.34 - Watkins 1997:582)
c. *Sef lle y doethont ygyt y bresseleu*
sef place PRT come.PAST.3P together in Preseleu
'That was the place where they got together, in Preseleu.' (WM 27.28)
d. *Sef kyuryv wr oed Ueuryc guas mavr tec*
sef sort man be.PAST.3P Meurig youth big handsome
'That's the sort of man Meurig was, a big handsome youth.' (BD 72.23)

In Middle Welsh this construction grammaticalised further. The number of clauses with headless relative subjects (see (42)) was increasing giving rise to idiomatic constructions that were no longer focussed, but used in contexts of narrative continuity as well, as shown in (43).

(42) a. *Sef __ a doeth dy nyeint*
sef PRT come.PAST.3S 2S nephews
'That's who came, your nephews.' (WM 89.35)
b. *Sef __ a wystlwys gwrgi*
sef PRT give-as-hostage.PAST.3S Gwrgi
'That's whom he gave as hostage, Gwrgi.' (WM 88.5)
c. *Sef __ y cudyawd y mywn llaw gist*
sef PRT hide.PAST.3S in hand chest
'That's where he hid it, in a small chest.' (WM 93.30)

(43) a. *Sef a gausant yn eu kynghor duunaw ar eu llad*
sef PRT get.PAST.3P in 3P council agree.INF on 3P kil.INF
'This is what they got in their council, they agreed to kill them'(WM 68.8)
b. *Sef a wnaeth y gwraged kyscu*
sef PRT do.PAST.3S the women sleep.INF
'This is what the women did, they slept.' (WM 28.15)

Finally, *sef* grammaticalised further until it was reinterpreted as an element functioning as an adverbial, causing the preverbal particle *a* to change into *y(d)* (which usually followed sentence-initial adjuncts as shown in the description of word order Type III above).

(44) a. *Sef y clywei arueu am ben hwnnw*
sef PRT hear.PAST.3S arms on head that.one
'He could feel armour on that one's head.' (WM 54.28)
b. *Sef y kynhelleis inheu y gyuoeth*
sef PRT withhold.PAST.3S I his dominions
'I withheld his dominions.' (WM 394.42)
c. *Sef y kawssant yn eu kyghor gossot (...)*
sef PRT get.PAST.3P in 3P council release.INF (...)
'They decided to release (...)' (RM 144.17)

Alongside these patterns, the loss of tense (when the copula phonologically eroded) led to the insertion of a medial copula *yw/oed*.

(45) a. *Sef **yw** honno gwreic doget urenhin*
sef be.PRES.3S DEM.FS wife Doged king
'That's who she is, king Doged's wife.' (WM 453.17)
b. *Sef **oed** y rei hynny Gog a Magog (...)*
sef be.PAST.3S the ones DEM.P Gog and Magog
'That's what those were, Gog and Magog (...).' (DB 29.11.12)

In late Middle Welsh *sef* was reanalysed as an NP appositive 'that is':

(46) *(...) llyfr y cofiadur, sef y cronicl*
(...) book the cofiadur, sef the chronicle
'the book of the *cofiadur*, that's to say the chronicle' (b1588 - Esther 6.3)

The development of the *sef*-construction in Welsh is discussed in chapter 6.

4.3.9 Type IX Non-verbal clauses

Sentences with verbal nouns instead of finite verbs were already discussed under type VI above. In co-ordinated sentences, it was also possible to leave out the verb completely. In these elliptical patterns, the finite verb of the previous clause is understood again as the matrix verb. There are also copular sentences in which the copula itself is left out, as in (47a). They usually exhibit the word order Subject - *(yn)* Predicate, though adverbs could interfere as well as shown in (47b).

(47) a. *Gwae uinheu uyn dyuot ar anuab*
woe me 1S come.INF on childless
'Woe me for coming to an childless (man)' (CO 39)
b. *ac angel yn wastat yn getymdeith idaw*
and angel PRED always PRED friend to.3MS
'and an angel will always accompany him' (Dewi 14.1)

These constructions can also appear with prepositional predicates, as in (48):

(48) a. *(Gleif) ennillec yn y law.*
(sword) battle-axe in 3MS hand
'In his hand was a battle-axe' (*gleif* is a gloss on *ennillec* CO 63)
b. *A gwisc ymdan y gwr o pali coch gwedy ry wniaw a sidann*
and garment on the man of satin red after PERF sew.INF with silk
melyn a godreon y llen yn velyn.
yellow and borders 3MS scarf PRED yellow
'And upon the man was a dress of red satin sewn with yellow silk, and yellow were the borders of his scarf.' (BR 5.22-24)
c. *guae ui o m ganedigaeth*
woe mi from 1S birth
'Woe me for my birth/being born' (Branwen 407)

Some of those non-verbal sentences with the order *A(c)* S P(P), functioned as background or had circumstantial readings. Sentences of this type also appear in other languages, for example Biblical Hebrew, where they are called Absolutive Sentences.

(49) a. *ac ynteu yn allmarw y r llawr*
and he PRED stone-dead to the floor
'and he was stone-dead on the floor' (Peredur 14.25)
b. *a thitheu a th lu yn y parth arall*
and you and 2S host yn the part other
'and (meanwhile) you and your host are in the other part' (Branwen 319)

Finally, Welsh employs certain lexical items *dyma/dyna/llyma/llyna/nachaf/wele* 'this is, that is, lo' (cf. French *voici, voilà*) in truncated copular constructions. These clauses still exist with *dyma/dyna* in Modern Welsh.

(50) a. *Llyna Dillus Uarruawc*
behold Dillus Barruawc
'Behold Dillus Barfog/There is Dillus Barfog' (CO 962-963)
b. *Nachaf yr esgidyeu yn ormod.*
lo the shoes PRED plenty
'Behold, the shoes were plenty' (PKM 80.4-5)
c. *ac wele hwynt yn athrist*
and lo they PRED sad
'and behold they were sad' (b1588 - Gen. 40.6)

(51) a. *Dyma gasgliad o feirdd gorau 'r genedl.*
dyma collection of bards best the nation
'Here's a collection of the best bards of the nation' (BBC Cymru - `www.bbc.co.uk/cymru/urdd02/cysylltiadau.shtml`)
b. *Dyna fo!*
dyna he
'There he is!' (Kate Roberts - Te yn y grug)

Another non-verbal clause pattern that is still used in Modern Welsh is illustrated by a sentence like (52a). The inflected form of *bod* can be left out in the present tense. If some other tense is used, the appropriate form of *bod* reappears in sentence-initial position, cf. (52b) and (52c).

(52) a. *Rhaid i mi adael.*
necessity to me leave.INF
'I must leave.' (Lit. 'It is necessary for me to leave')
b. *Bydd rhaid i mi adael.*
be.FUT.3S necessity to me leave.INF
'I will have to leave.'
c. *Roedd rhaid i mi adael.*
be.IMPF.3S necessity to me leave.INF
'I had to leave.' Borsley et al. (2007:66)

In Modern Welsh proverbs it is also possible to leave out the copula, though these sentences really are 'a hallmark of formal rather than casual style' (Borsley et al., 2007:364).

(53) a. *Nid aur popeth melyn.*
NEG gold everything yellow
'All that glitters is not gold.'
b. *Hir pob aros.*
long every wait
'A watched pot never boils.' Borsley et al. (2007:364)

4.4 Frequency of different Types

In this final section, I present an overview of the frequency (both raw counts and percentages per text) of each of the above-mentioned Types in all Middle Welsh texts under investigation. The frequency of verb-second orders of the so-called 'Mixed Sentence' is here only based on 'unambiguous' cases, i.e. cases with plural or pronominal subjects that do not agree with the verb. Verb-initial orders (Type I) include both absolute verb-initial sentences and sentences in which the verb directly follows a conjunction or sentence-initial particle. The total number of main clauses differs from text to text. The Arthurian Romances (*Peredur, Owein* and *Gereint*) are much longer than most of the *Four Branches* or *Llud & Llefelys*. The two manuscript versions of the latter only show small differences in distribution of word order types. The texts presented in the tables below are in rough chronological order starting with the Laws from the beginning of the Middle Welsh period, then *Culhwch* and the *Four Branches*, followed by the Romances and the two Dreams (*Macsen* and *Rhonabwy*). Finally, the two versions of the native tale *Llud* and the Life of St David mark the end of the Middle Welsh period. From 1500 onwards, the language is referred to as (Early) Modern Welsh, exemplified here by the 1588 Bible translation (although the language of this translation is actually not like late Middle or Modern

Welsh as discussed in Chapter 7). Samples of some of the texts were also analysed by Poppe (1989, 1990, 1991a, 1991b, 1993) and Watkins (1977-8, 1983-4, 1988) (and summarised by Willis (1998:54)).

	Laws	CO	Pwyll	Branwen	Manaw.	Math
I Verb-initial	4	74	10	9	9	22
II AuxSVO	0	2	0	1	3	5
III Adjy*VS	66	157	101	88	41	113
IV SaVagr.	112	140	141	78	75	150
IV OaVS	31	36	29	14	23	18
IV VNaDO	3	133	64	22	41	58
V V2 focus	0	3	1	4	0	0
VI VNs	17	142	65	91	73	66
VII Copula	138	171	72	47	30	60
VIII Sef	3	19	22	20	11	37
IX Non-verb.	73	50	36	37	29	29
Total	447	927	541	411	335	558

Table 4.2: Distribution of word order types in positive main declaratives

	Laws	CO	Pwyll	Branwen	Manaw.	Math
I Verb-initial	0.89%	7.98%	1.85%	2.19%	2.69%	3.94%
II AuxSVO	0%	0.22%	0%	0.24%	0.90%	0.90%
III Adjy*VS	14.77%	16.94%	18.67%	21.41%	12.24%	20.25%
IV SaVagr.	25.06%	15.10%	26.06%	18.98%	22.39%	26.88%
IV OaVS	6.94%	3.88%	5.36%	3.41%	6.87%	3.23%
IV VNaDO	0.67%	14.35%	11.83%	5.35%	12.24%	10.39%
V V2 focus	0%	0.32%	0.18%	0.97%	0%	0%
VI VNs	3.80%	15.32%	12.01%	22.14%	21.79%	11.83%
VII Copula	30.87%	18.45%	13.31%	11.44%	8.96%	10.75%
VIII Sef	0.67%	2.05%	4.07%	4.87%	3.28%	6.63%
IX Non-verb.	16.33%	5.39%	6.65%	9.00%	8.66%	5.20%
Total	100%	100%	100%	100%	100%	100%

Table 4.3: Percentages of word order types in positive main declaratives

Text	Adv	Sbj^{Nom}	Sbj^{Pro}	Obj	VN	V1
Branwen	41	17	16	8	14	4
Macsen	43	5	16	20	8	9
Rhonabwy	45	12	6	9	26	2
Culhwch	25	16	12	12	26	9
Llud	39	24	22	4	10	0
Manawydan	24	6	31	12	27	0
Pwyll	38	11	22	10	17	3

Table 4.4: Percentages of word order types from Willis (1998:54) based on Poppe and Watkins

First of all, Poppe and Watkins separate nominal and pronominal subjects.[4] I have analysed this difference systematically in Chapter 5, but I have lumped both together in the tables here. For *Culhwch, Pwyll, Branwen* and *Manawydan* there are some small differences in the frequencies shown here and those presented in the overview by Willis (1998:54) (based on Poppe's and Watkins's earlier papers). The difference in frequencies of subject- and object-initial orders are partly due to a difference in interpretation. For the present corpus, I analysed fronted topics of impersonal verbs as subject-initial. Semantically, they are indeed often interpreted as patients (of passive verbs), but from a syntactic perspective, they could always be argued to function as subjects. To remain consistent throughout the corpuse, I therefore chose the subject-initial analysis, so the numbers for subject-initial sentences are slightly higher. In *Breudwyt Rhonabwy*, the number of object-initial sentences indicated below is again much lower than the number indicated by Poppe (1990) and the same can be observed for *Breudwyt Macsen*. Poppe and Watkins furthermore did not distinguish between auxiliary-initial sentences and verb-initial sentences. This results in some slight differences in frequencies for this category as well. In the present corpus, I furthermore counted sentences with subject or object topic drop for their respective types SaVO and OaVS. These topic drop sentences are not calculated at all in the overviews by Poppe and Watkins. In Type IVd with sentence-initial verbal nouns, Poppe and Watkins sometimes not only include the *gwneuthur*-periphrastics, but also other auxiliaries. Here too, slight differences appear in the counted frequencies. Finally, Poppe and Watkins do not systematicaly present the frequencies of other sentence types (although in some papers, *sef*-sentences and copula-sentences are mentioned). Non-verbal clauses and sentences starting with verbal nouns without any auxiliaries are also not listed. Since they can express positive declarative statements and they function as main clauses as well, I did include them in these overviews.

[4]The percentages in the table are taken from the overview by Willis (1998:54). Note that some of them make up more than 100% per text, most likely due to slight rounding errors. In Willis's overview, one further text was included (*Amlyn ac Amic*) that was not part of the annotated database on which the present corpus study is based. It was therefore not included in the above table. Finally, only the WB version of the tale *Llud* was included.

	Peredur	Owein	Gereint	Rhonabwy	Macsen
I Verb-initial	23	12	36	8	3
II AuxSVO	4	0	2	0	1
III V2 adj.	224	115	204	69	93
IV S a Vagree	420	130	244	39	39
IV O a VS	68	15	60	6	28
IV VN a DO	162	194	196	36	12
V V2 focus	0	4	3	0	0
VI Verbal nouns	134	132	175	36	8
VII Copular	137	96	106	23	12
VIII Sef	19	18	39	18	5
IX Non-verbal	132	88	109	126	32
Total	1323	804	1174	361	233

Table 4.5: Distribution of word order types in positive main declaratives

	Peredur	Owein	Gereint	Rhonabwy	Macsen
I Verb-initial	1.74%	1.49%	3.07%	2.22%	1.29%
II AuxSVO	0.30%	0%	0.17%	0%	0.43%
III V2 adj.	16.93%	14.30%	17.38%	19.11%	39.91%
IV S a Vagree	31.75%	16.17%	20.70%	10.80%	16.74%
IV O a VS	5.14%	1.87%	5.20%	1.66%	12.02%
IV VN a DO	12.24%	24.13%	16.70%	9.97%	5.15%
V V2 focus	0%	0.50%	0.26%	0%	0%
VI Verbal nouns	10.13%	16.42%	14.91%	9.97%	3.43%
VII Copular	10.36%	11.94%	9.03%	6.37%	5.15%
VIII Sef	1.44%	2.24%	3.32%	4.99%	2.15%
IX Non-verbal	9.98%	10.95%	9.28%	34.90%	13.73%
Total	100%	100%	100%	100%	100%

Table 4.6: Percentages of word order types in positive main declaratives

	Llud Mab	Llud Chro	Dewi	b1588
I Verb-initial	0	1	5	87
II AuxSVO	0	0	1	30
III V2 adj.	22	18	88	278
IV S a Vagree	40	50	93	745
IV O a V S	1	2	11	15
IV VN a DO	7	5	21	2
V V2 focus	0	0	0	0
VI Verbal nouns	13	19	56	21
VII Copular	21	20	40	152
VIII Sef	2	1	24	10
IX Non-verbal	3	9	26	67
Total	109	125	365	1407

Table 4.7: Distribution of word order types in positive main declaratives

	Llud Mab	Llud Chro	Dewi	b1588
I Verb-initial	0%	0.80%	1.37%	6.18%
II AuxSVO	0%	0%	0.27%	2.13%
III V2 adjunct	20.18%	14.40%	24.11%	19.76%
IV S a Vagree	36.70%	40.00%	25.48%	52.81%
IV O a V S	0.92%	1.60%	3.01%	1.21%
IV VN a DO	6.42%	4.00%	5.75%	0.14%
V V2 focus	0%	0%	0%	0%
VI Verbal nouns	11.93%	15.20%	15.34%	1.49%
VII Copular	19.27%	16.00%	10.96%	10.80%
VIII Sef	1.83%	0.80%	6.58%	0.71%
IX Non-verbal	2.75%	7.20%	7.12%	4.76%
Total	100%	100%	100%	100%

Table 4.8: Percentage of word order types in positive main declaratives

4.5 Conclusion

We can categorise the large amount of observed word order patterns in positive declarative main clauses in Welsh in nine main Types. First of all, there are verb-initial patterns (Type I). Sentences of this type are rare in Middle Welsh, although variants with sentence-initial conjunctions or declarative particles like *neu(r)* directly followed by the verb are found somewhat more frequently. The second type I described consists of a periphrastic construction with the auxiliary form of the verb *bod* 'to be', rendering the word order AuxSVO. This type is also rarely found, although its frequency increases towards the end of the Middle Welsh period. Word order types I and II (VSO and AuxSVO) are the predominant patterns found in Modern Welsh.

Middle Welsh texts, on the other hand, mainly exhibit the verb-second pattern (the 'Abnormal Sentence') in one of its various forms (Types III, IV or even the focussed Type V, the 'Mixed Sentence'). The adjunct-initial order can appear in many forms: the initial constituent can be an Adverbial Phrase, a Prepositional Phrase or a combination of multiple phrases, as long as the 'topicalised' one functions as an adjunct. The other type of 'Abnormal Sentence', Type IV, on the other hand places a core argument (Subject or Direct Object) in sentence-initial position or a verbal noun followed by the pre-verbal particle *a* and the auxiliary *gwneuthur* 'to do'. In subject-initial sentences, the verb usually agrees with the pre-verbal subject. This is what formally distinguishes the 'Abnormal Sentence' from the 'Mixed Sentence' in which the verb shows default third-person singular inflection (Type V).

Sentences with verbal nouns instead of finite verbs (Type VI) were mainly possible in (Early) Middle Welsh. In early Middle Welsh texts, the verbal noun could appear in non-finite main clauses on their own followed by the subject. These 'verbal noun + agent' almost disappear in independent main clauses. Only sentence-initial verbal nouns in co-ordinated sentences depending on preceding finite clauses continued to exist much longer.

Types VII and VIII are only concerned with copular verbs. There were various ways to express copular predicates in Middle Welsh, with or without overt forms of the verb *bod* 'to be'. These non-verbal sentences were finally labelled as Type IX.

It is clear from the counts in the final table that the language is already changing at the end of the Middle Welsh period. The preferred word order is still the verb-second 'Abnormal' order, but an overwhelming amount of sentences are subject-initial. Verb-initial orders (Type I) and in particular auxiliary-initial periphrastic orders (Type II) are on the rise. In total, over 9000 main clauses were analysed for the present corpus study. In the next chapter, I discuss the potential factors that could influence preferred types of word order and thus explain the distribution found in the Middle Welsh corpus.

CHAPTER 5

Factors influencing word order

> *"The normal word order has become the form of expression suited to the mind in its normal condition of steady activity and easy movement, from which it only departs under the stress of emotion, or for logical reasons, or in conformity to fixed rules."*
>
> (dr. G.O. Curme, Ch. xvii of *A grammar of the English language*)

5.1 Introduction

The previous chapter presented various different types of word order. Why is there more than one way to put words together in a sentence? Do each of these types yield a different meaning? Is the word order changed 'under the stress of emotion, or for logical reasons', as George Curme put it in his description of English grammar? Alternatively, he proposed that deviations of normal word order were 'in conformity to fixed rules' (Curme, 1978). Assuming the latter is a reasonable working hypothesis: what are those 'rules' exactly? Are they based on purely grammatical features, usage, information structure or are there even extra-linguistic features that play a role? This chapter aims to answer all these questions for Middle Welsh.

If we want to describe the true pragmatic nature of Middle Welsh word order it is first of all of crucial importance to have a good description and overview of all the available word order patterns. All possible word order patterns were categorised and described in detail in the previous chapter. After this, all other factors (grammatical (section 5.2), usage-based (section 5.3) and extra-linguistic (section 5.4)) need to be taken into account to check to what extent - if at all - they

interact with these patterns. This then forms our baseline for the main investigation that aims to determine the effect of information-structural notions. First of all, the information-structural notions in themselves need to be analysed in a systematic way. Then we can systematically check their possible effect on the distribution of word order patterns we find. Only when all these considerations (grammatical, usage-based, extra-linguistic *and* information-structural notions) are combined can we find proper generalisations about Middle Welsh word order. The final question that remains then is the following: is it possible to 'predict' the right word order in any specific context in Middle Welsh or is there still (some degree of) random variation? I conclude by addressing this issue of variation with all available evidence presented in this chapter.

5.2 Grammatical factors

In this section I discuss various parts of the grammar and how - if at all - they interact with word order in Middle Welsh. The main focus lies on syntactic features, but some morphological and semantic issues will be taken into account alongside certain lexical items. The underlying assumption is that the different word order patterns described in the previous Chapter reflect different syntactic structures and furthermore that these syntactic structures in turn are the result of differences in various features of the grammar (e.g. tense, mood or transitivity, to mention just a few). Sentences with progressive aspect in Present-day English, for example, differ in syntactic structure from their non-progressive counterparts. This, in turn, can be observed in the different superficial word order patterns, Subject-Aux-Verb*ing*-Object in (1a) vs. Verb-Object (1b):

(1) a. He is kissing Mary.
b. He kisses Mary!

Another example in English that also shows a change of the sequential order of the verb and its core arguments can be observed in different clause types. Interrogative clauses have a different syntactic structure than their declarative counterparts. This is shown by their superficial word order patterns as in (2) (although it could also and/or alternatively be reflected by other linguistic strategies, e.g. differences in prosodic structure).

(2) a. You are at home.
b. Are you at home?

The word order of the verb and its core arguments and the use of different constructions (e.g. auxiliary + *-ing*) in Present-day English can thus be influenced by specific aspects of English grammar. Languages may of course differ with respect to which features in the grammar result in different word order patterns. The main question in this section is therefore to ascertain if - and if so, which ones and to what extent - grammatical features in Middle Welsh result in different superficial

word order patterns.

5.2.1 Clause type

There are four major distinctions in clause type: declaratives, interrogatives, imperatives and exclamatives. The present study is mainly concerned with declarative clauses, which, as I show here, exhibit different word order patterns than imperative or interrogative clauses in Welsh. In this section I also briefly touch on related issues, like the difference between main and subordinate clauses and the role of negation.

Imperative

Welsh, like many other languages employs verb-initial word order in imperative clauses. Even in the Middle Welsh period, when verb-second orders were commonly found, imperative verbs were always found in absolute clause-initial position or directly following a conjunction.

(3) a. *Bydwch lawen a chedwch ych ffyd a ch cret.*
be.PRES-IPV.2P happy and keep.PRES-IPV.2P 2P faith and 2P belief
'Be happy and keep your faith and your belief.' (Dewi 115.4)

b. *Dalet gydymdeithas a mi*
hold.PRES-IPV.3S friendship with me
'Let him be friends with me.' (CO 474)

c. *ac aro ditheu yn kennadwri ninheu*
and wait.PRES-IPV.2S you 1P tidings us
'And wait for our message.' (PKM 41.16)

Interrogative: Questions & Answers

There are different types of interrogatives each reflected by a different superficial word order patterns. Yes/no questions are verb-initial, only preceded by the sentence-initial interrogative particle *a*.

(4) a. *A wydyat llad a chledyf?*
QU-PRT know.PRES.2S kill.INF with sword
'Do you know how to kill with a sword?' (Peredur 7.15-16)

b. *A oes gennwch chwi chwedleu?*
QU-PRT be.PRES.3S with.2P you stories
'Do you have any news?' (PKM 45.24)

Wh-questions have the wh-word in initial position. The word order pattern looks exactly like that of the verb-second order.

(5) a. *Pwy oed hwnnw?*
who be.PAST.3S that
'Who was that?' (PKM 35.4)

b. *Pa dyn a gwyn yn y maendy hwnn?*
which man PRT lament.PRES.3S in the prison this
'Which man laments in this prison?' (CO 914)

c. *Pan doy di*
where come.PRES.2S you
'Where are you from?' (PKM 12.13)

Answers to questions do not necessarily exhibit the same word order as other positive declarative sentences. In Middle Welsh, yes/no questions are frequently answered by repeating the verb in the question as shown in (6). Answers to wh-questions usually start with (or consist solely of) the constituent that solves the variable in question, as shown in (7a), but the verb can be repeated here as well, as shown in (7b).

(6) a. *A wely di y keibedic rud draw? Gwelais.*
QU-PRT see.PRES.2S you the hoed slope yonder see.PRES.1S
'Do you see the hoed slope over there? (Yes) I see (it).' (CO 611-612)

b. *A gaffaf i lety genhyt ti? Keffy.*
QU-PRT get.PRES.1S I stay with.2S you get.PRES.2S
'Can I stay with you? You can.' (Peredur 1251)

c. *A uyd llawn dy got ti uyth? Na uyd.*
QU-PRT be.FUT.3S full 2S coat you ever NEG be.FUT.3S
'Will your coat never be full? It won't.' (PKM 15.8)

(7) a. *Pa ryw aniueileit yw y rei hynny? Aniueileit bychein.*
what sort animals be.PRES.3S the ones those animals small
'What sort of animals are those? Small animals.' (PKM 68.18-19)

b. *Pa du y mae hi? Y mae hi (...) yn Aber Deu Gledyf.*
what side PRT be.3S she PRT be.3S she ... in Aber Deu Gledyf
'Where is she? She's in Aber Deu Gledyf.' (CO 931-932)

Answers to broad focus questions like 'What happened?' are usually assumed to exhibit predicate focus. In translations of the Welsh Bible in 1588, we consistently find subject-initial verb-second patterns here, which could thus be considered to be the 'basic' word order (see previous chapter).

(8) a. *Pa beth a ddigwyddodd, fy mab?*
what thing PRT happen.PAST.3S 1S son
'What happened, my son?' (b1588 - 1 Sam. 4:16)

b. *(...) Israel a ffoawdd o flaen y Philistiaid.*
(...) Israel PRT flee.PAST.3S of front the Philistines
'Israel fled before the Philistines' (b1588 - 1 Sam. 4:17)

(9) a. *Beth yw 'r matter (...)*
what is the matter (...)
'What happened?' (b1588 - 2 Sam. 1:4)

b. *y bobl a ffoawdd o r rhyfel*
the people PRT flee.PAST.3S from the battle
'the men fled from the battle' (b1588 - 2 Sam. 1:5)

Declarative main vs. subordinate clauses

In many languages, main clauses exhibit different word order patterns than subordinate clauses. Certain syntactic phenomena only appear in main clauses or behave differently in subordinate clauses (see Aelbrecht, Haegeman, and Nye (2012) for an overview and discussion). Since the present study is concerned with main clauses only, I will not go into the various word order patterns found in subordinate clauses in Middle Welsh. It suffices to say that they mainly exhibit verb-initial order (cf. D. S. Evans (2003 [1964])). I will, however, briefly discuss relative clauses, since their structure is very similar to the verb-second order observed in Middle Welsh main clauses.

Relatives

Non-restrictive relative clauses in Middle Welsh can be introduced by the demonstrative pronouns *yr hwnn* 'the one (m.)', *yr honn* (f.), *yr hynn* (n.) and *y rei* (pl.). These act as relative pronouns and were introduced in the literary language in imitation of other languages like Latin, English and French (cf. D. S. Evans (2003 [1964]:66) and Willis (1998:80)). Before the introduction of these demonstrative relative pronouns, the word order of relative clauses was Antecedent - *a/y* - Verb. Just like in the verb-second orders in main clauses, the choice of the particle depended on the nature of the preceding constituent. Direct relatives based on subject or objects were followed by the particle *a* (with default third-person singular agreement); indirect relatives with prepositional phrases or adverbial elements were followed by *y*.

(10) a. *Duw a wyr pob peth a wyr bot yn eu hynny*
God PRT know.PRES.3S every thing PRT know.PRES.3S PRED lie that on.1S
arnaf i.
me
'God who knows everything knows that this is a lie about me.'(PKM 21.3-4
)

b. *a r vorwyn a gywirawd yr hyn a adawssei*
and the maiden PRT prepare.PAST.3S the that PRT promise.PLPF.3S
'and the maiden prepared what had been promised.' (Peredur 64.12)

c. *Pwy bynnac a adefo galanas ef a e genedl a e*
who ever PRT confess.PRES-SBJ.3S homocide he and 3MS family PRT 3MS
talant sarhaet y dyn a lader (...)
pay.3P compensation the man PRT kill.IMPER (...)
'Whoever would confess to homocide, he and his family will pay him the compensation of the man who was killed (...)' (Laws 50)

d. *a galw a oruc Arthur ar y gweisson a gadwei y*
and call.INF PRT do.PAST.3S Arthur on the men PRT make.PAST.3S 3MS
wely
bed
'and Arthur called on the servants who made his bed' (Gereint)

e. *y dyd y bei drist y gellyngei y lleill weuyl idaw y*
the day PRT be.PAST.3S sad PRT release.PAST.3S the other lip to.3MS to
waeret hyt y uogel
down till 3MS navel
'the days he would be sad, he would lower his low lip to his navel' (CO 325)

The distinction between direct and indirect relatives made by traditional Welsh grammarians is, however, not always as clear-cut. There is a certain amount of variation in agreement (see Plein and Poppe (2014)), choice of the particle and the use of resumptive pronouns (see Rouveret (1994) and Willis (1998) for a detailed analysis in a generative framework and Chapter 7 of the present thesis on the possibilities of a common analysis for Middle Welsh verb-second and relative clauses).

Negation

A full analysis of negative sentences is beyond the scope of the present study. Negation can be found in many shapes and forms and they each have their own effect on Middle Welsh word order. Diachronically, Welsh seems to have gone through all stages of Jespersen's cycle (cf. Willis (2006)). In Middle Welsh, however, sentence-negation is exhibited by a negative element *ny(t)* in sentence-initial position, directly followed by the finite verb:

(11) a. *Ny daw ef o e uod genhyt ti*
NEG come.PRES.3S from 3MS will with.2S you
'He will not come with you out of his own will' (CO 580)

b. *Ny wnn i dim y wrth honno.*
NEG know.PRES.1S I anything from that
'I don't know anything about that.' (PKM 54.9)

There are also some examples of noun phrases preceding the negation as in (12a), but this is far less common. Although this type of word order superficially resembles the abnormal verb-second order in positive declaratives in Middle Welsh, it cannot be exactly the same in all of these cases. (12b) and (12c) for example show resumptive pronouns, either attached to the negative *ny-* or preceding the verbal noun. In positive declaratives, such resumptives are never found.

(12) a. *Afles ny wnaf inheu.*
harm NEG do.PRES.1S I
'I will do no harm.' (Peredur 29.23)

b. *a merch inheu nys keffy*
and daughter my NEG.3FS get.PRES.2S
'and my daughter you won't get' (CO 711)
c. *Vyg kywilyd ny ellwch y dalu y mi*
1S shame NEG can.PRES.2P 3MS pay to me
'My shame you cannot compensate to me' (PKM 74.26-27)

Negative counterparts of mixed word orders with focussed initial constituents always have the negative element *nid* directly preceding the focussed constituent, yielding Neg - Foc - *a/y* - V order, as shown in (13):

(13) a. *Nyt o hynny y goruydir*
NEG from that PRT prevail.IMPER
'It's not because of that one is successful.' (PKM 68.11)
b. *Na marchawc na phedestyr y del itaw*
NEG knight NEG soldier PRT come.PRES-SBJ.3S to.3MS
'Nor a knight, nor a soldier would come to him.' (Gereint 57-58)

Negation was thus possible in different types of word order patterns, but the most common way to negate an entire proposition was by placing the negative particle *ny* in front of the verb in sentence-initial position.

5.2.2 Tense & Aspect

"The past is always tense, the future perfect."
(Zadie Smith)

When tense is expressed by inflectional morphology on the finite verb, it is not immediately associated with variation in word order. A complete lack of tense, however, or a lack of overtly expressed tense at least, can result in different word order patterns. In tenseless main declaratives in Middle Welsh, verbal nouns occupy the first position in the sentence (Type VI), followed by their agents (see section 5.2.4 below). Loss of tense over time, for example because of phonological erosion as seen in the copula *ys* can in turn trigger the creation of new types of word order as well. This can be observed in one type of the *sef*-construction, *sef + yw/oed* (see detailed discussion of the diachronic development of this construction in Chapter 7). If tense is expressed, another question arises: do different tenses yield different word order patterns? Or, vice versa, do certain word order patterns occur typically or only in present or preterite tense, for example? According to Poppe (1993), the latter is the case for periphrastic constructions with the verbal noun + the inflected form of *gwneuthur* 'to do' (Type IVc). Out of almost 1000 instances of this type in the Middle Welsh corpus under investigation, there are indeed only 40 examples in which *gwneuthur* exhibits non-preterite (i.e. imperfect, perfect or pluperfect) inflection.

(14) a. *Ac yna clymu a wnaethant.*
and then halt.INF PRT do.PAST.3P
'and then they halted' (PKM 72.18-19)

b. *a phaup ual y delynt kyuarch guell a wneynt idaw*
and all as PRT come.IMPF-SBJ.3P greet.INF well PRT do.IMPF.3P to.3MS
'and all greeted him as they came in' (PKM 4.9-10)

c. *a gwedy hynny kyscu a wna*
and after that sleep.INF PRT do.FUT.3S
'And after that he will sleep.' (CO 968-969)

	Imperfect	Perfect	Pluperfect	Present	Preterite
Type I Verb-initial	16	1	4	139	143
Type II Periphrastic	7	0	0	41	1
Type III Adj y VS	276	4	8	317	1072
Type IVa SaVO	311	3	12	789	1382
Type IVb OaVS	74	0	5	114	166
Type IVc VN a DO	14	0	1	25	916
Type V Focus	2	0	0	5	8
Type VII Cop	319	0	4	702	100
Total Frequency	1019	8	35	3101	2129

Table 5.1: Tense & Aspect in Middle Welsh relevant word order types

	Imperfect	Perfect	Pluperfect	Present	Preterite
Type I Verb-initial	1.57%	12.50%	11.76%	6.53%	3.78%
Type II Periphrastic	0.69%	0%	0%	1.93%	0.03%
Type III Adj y VS	27.09%	50.00%	23.53%	14.89%	28.30%
Type IVa SaVO	30.52%	37.50%	35.29%	36.92%	36.46%
Type IVb OaVS	7.26%	0%	14.71%	5.35%	4.41%
Type IVc VN a DO	1.37%	0%	2.94%	1.17%	24.18%
Type V Focus	0.20%	0%	0%	0.23%	0.21%
Type VII Cop	31.31%	0%	11.76%	32.97%	2.64%
Total Frequency	100%	100%	100%	100%	100%

Table 5.2: Tense & Aspect in Middle Welsh relevant word order types in percentages

These results indicate that there is a significant relation between word order type IVc VNaDO and tense (comparing Preterite to Present in argument-initial sentences (VN-initial vs. Sbj/Obj-initial order), $\chi^2 = 397.21$, df $= 1$, $p < 0.0001$, Fisher's exact $p < 0.0001$,). The question is whether this is inherent to the syntax of this particular construction. Since there are also examples, however few, of verbal-

noun constructions with present or imperfect auxiliaries, it cannot be a syntactic constraint. The context in which verbal-noun constructions tend to appear - mainly in continuous narrative - could be related to the preference for preterite tense. Section 5.5 below will shed more light on this particular matter.

According to T. A. Watkins (1993), some verbs 'resist inflection', in which case they were exclusively found as verbal nouns and in these cases inflected forms of *gwneuthur* 'to do' had to be inserted. He lists, among others, *kyuarch gwell* 'to greet', *kyuedach* 'to carouse', *kynhewi* 'to become silent', *meithryn* 'to nurse' and *ymchwelut* 'to return'. Although it might have been a contributing factor, these verbs resisting inflection alone could not have caused the rise of the periphrastic order with 'to do'. First of all at least some of the verbs he lists actually do exhibit inflected forms in Middle Welsh already (e.g. *ymchoeles* 'returned' (Laws) or *kyuarchaf* 'I will greet' (Pwyll 30) and *kyuarchawd* 'greeted' (PKM 16.9), others are attested from the 16th century at least (e.g. *meithrinesit* 'he was brought up' (Testament Newydd gan Salesbury 1567) or *faethrinodd* 'brought up' (E. James Homily 1606)). There was furthermore another type of word order available in Welsh in which the verb could stay uninflected: Type VI with verbal nouns + agents. The frequency of this type rapidly declined in the Middle Welsh period, however, which might be inversely correlated to the increase in use of the periphrastic 'to do' construction that could express tense overtly.

Another periphrastic construction in Middle Welsh was used to render progressive or perfective aspect. Only certain verbs exhibit perfect inflectional endings and, as shown in the table above, these occurred very infrequently. The periphrastic construction with inflected forms of *bod* 'to be' + the aspectual particles *yn/wedi* could yield progressive or perfective interpretation as well. Although in Middle Welsh there are not very many examples of this yet, it was increasingly used with an even wider aspectual range from the late Middle Welsh period onwards. The verbal noun could precede or follow the auxiliary, but in information-structurally neutral contexts (see section 5.5 below), the clause would start with a preverbal particle *y* followed by the auxiliary.

(15) a. *Yn hela yd oedwn yn iwerdon dydgueith*
PROGR hunt.INF PRT be.PAST.1S in Ireland one.day
'One day I was hunting in Ireland' (PKM 35.11-12)

b. *y mae gvedy mynet gyd a Gwenhwyuar y hystauell.*
PRT be.PRES.3S PERF go.INF with Gwenhwyfar to.3FS chamber
'She has gone with Gwenhwyfar to her chamber.' (WM 408.7)

c. *ac y maent yn symudaw enweu*
and PRT be.PRES.3P PROGR change.INF names
'and they change names' (PKM 68.20-21)

d. *yny doeth rybudyeu idaw, a menegi uot y crydyon*
until come.PAST.3S warnings to.3SM and indicate.INF be.INF the shoemakers
wedy duunaw ar y lad.
PERF conspire.INF on 3SM.GEN kill.INF
'until he was warned the shoemakers conspired to kill him.' (PKM 58.17)

5.2.3 Mood

Apart from indicative mood, Welsh also has a separate set of verbal endings for present and imperfect subjunctive mood. In Modern Welsh, the use of the subjunctive sounds quite archaic, but in Middle Welsh texts various examples can be found, as in (16). These examples are distributed over many different word order types, although the frequency of subject-initial Abnormal orders is higher, due to the great number of idiomatic greetings and blessings in dialogues (16b). If we just look at the distribution of the different types of abnormal order (argument- or adjunct-initial), there is no significant difference in the use of indicative or subjunctive. The frequencies of subjunctive verbs in other types (verb-initial or auxiliary-initial) are too low to achieve any reliable statistical results here. In conclusion, mood does not seem to have an effect on choice of initial constituent within the preferred abnormal order in Middle Welsh.

(16) a. *Amaeth a amaetho y tir hwnnw*
farmer PRT plough.PRES-SBJ.3S the land that
'A farmer who would plough that land.' (CO 578)

b. *Duw a rodo da ywch*
God PRT give.PRES-SBJ.3S good to.2P
'May God give you good (things).' (PKM 30.11)

c. *Henpych gwell, Yspadaden Penkawr, o Duw ac o dyn.*
be.PRES-SBJ.2S well Yspadaden Penkawr, from God and from man
'May you be well, Y.P., from God and from man.' (CO 513-514)

	Present Subj.	Imperfect Subj.	Present Ind.	Imperfect Ind.
Type I Verb-initial	6	4	133	12
Type II AuxSVO	1	0	40	7
Type III Adj *y*VS	19	32	298	244
Type IVa SaVO	81	32	705	280
Type IVb OaVS	8	9	106	65
Type IVc VNaDO	3	2	22	12
Type V Focus	2	0	3	2
Type VII Copula	13	12	689	308
Total	133	91	3788	928

Table 5.3: Frequency of Subjunctive & Indicative Mood in Middle Welsh relevant word order types

	Present Subj.	Imperfect Subj.	Present Ind.	Imperfect Ind.
Type I Verb-initial	4.51%	4.40%	6.66%	1.29%
Type II AuxSVO	0.75%	0%	2.00%	0.75%
Type III Adj *y*VS	14.29%	35.16%	14.93%	26.29%
Type IVa SaVO	60.90%	35.16%	35.32%	30.06%
Type IVb OaVS	6.02%	9.89%	5.31%	7.00%
Type IVc VNaDO	2.26%	1.29%	1.10%	1.29%
Type V Focus	1.50%	0%	0.15%	0.22%
Type VII Copula	9.77%	13.19%	34.52%	33.08%
Total	100%	100%	100%	100%

Table 5.4: Percentages of Subjunctive & Indicative Mood in Middle Welsh relevant word order types

5.2.4 Transitivity

The word order of clauses with transitive verbs can be different from intransitive clauses because of the position of the additional direct object. It is therefore strictly speaking impossible to fairly compare the word order of transitive clauses with that of intransitives. If there is no direct object, its position in the clause is irrelevant. Many studies of word order therefore focus on sentences with transitive verbs only (e.g. Kirk (2012)). This, of course, limits the amount of data we can work with.

Furthermore, if 'overtness' of arguments of the verb is a criterion, the question is what we define by overt. If the subject was pronominal, in Middle Welsh, it could be expressed by the verbal inflection only (Middle Welsh, in other words, was a 'pro-drop' language). In some sentences, overt pronominal subjects did appear in post-verbal position, but this was by no means obligatory, as shown in (17). Unless inflectional endings count as overt subject arguments, our data set would be limited even further if we take these sentences out as well.

(17) a. *a hynny a elly yn haut*
and that PRT can.PRES.2S PRED easy
b. *a hynny a elly di yn haut*
and that PRT can.PRES.2S you PRED easy
'And that you can (do) easily.' (White Book vs. Red Book PKM 3.3-4)

There are also sentences that contain transitive verbs with elided objects. Many of these elliptical constructions are found in answers to questions, for example, or in other contexts in which the direct object can be easily understood. In addition to that, some verbs take prepositional arguments (that are not optional) as shown in (18a) and (18b). Other transitive verbs can also appear as intransitives as in (18c).

(18) a. *Keffy. myn vyg cret.*
get.PRES.2S by 1S belief
'You will get (it), on my word' (Peredur 11.51)

b. *Nyt ymedewis ef a hwnnw.*
NEG depart.PAST.3S he with that
'He did not depart from that.' (PKM 43.3)
c. *Gwn. heb ynteu.*
know.PRES.1S said he
'I know, said he' (Gereint 438)

Finally, transitivity, or how many compulsory arguments the verb takes, can be subject to change. A good example in Middle Welsh is the verb *kyrchu* 'to make for, go to', which can occur with a nominal direct object or with a prepositional phrase, even in one and the same text:

(19) a. *Ac y r neuad y gyrchwys y diarchenu.*
and to the hall PRT go.PAST.3S to disrobe
'And he went to the hall to disrobe.' (Intransitive - PKM 4.7)
b. *Yr orssed a gyrchyssant.*
the mount PRT go.PAST.3P
'They went to the mount.' (Transitive - PKM 10.19)

Note that the interpretation of *yr* in (19b) as *y* 'to' + *r* 'the' is unlikely, because the preverbal particle is *a* (only used with preceding arguments) rather than *y* (used with preceding prepositional phrases and other adjuncts). Degrees of transitivity are also relevant in certain types of intransitive verbs, as shown in the section on Intransitives in Welsh below.

Transitive

Transitive verbs occur in clauses with different types of word order in Middle Welsh, as shown in 5.5 (since copular clauses are intransitive by definition, Types VII, VIII and IX are omitted):

	Transitive	Intransitive	
Type I Verb-initial	105 (34.65%)	198 (65.35%)	303
Type II Periphrastic	18 (36.73%)	31 (63.37%)	49
Type III Adj y VS	899 (53.64%)	777 (46.36%)	1676
Type IV OaVS	357 (100%)	0 (0%)	357
Type IV SaVO	948 (37.95%)	1550 (62.05%)	2498
Type IV VN a DO	487 (50.94%)	469 (49.06%)	956
Type V Focus	9 (60%)	6 (40%)	15
Total	2823	3031	5854

Table 5.5: Transitive and intransitive clause in Middle Welsh positive main declaratives

The relative order verb-subject (or verb-agent) occurs in Types I, III, IVb and VI. Verb-object (VO) order occurs in all types, apart from the abnormal sentence with

direct objects in initial position. In Types I and III, the subject occurs in between the verb and the object, but the relative VO order remains. Object-verb (OV) word order is only observed in sentences with pronominal objects, but these always appear in the form of preverbal clitics in this case (as in (20a)) and another pronominal element optionally follows the verb, as shown in (20b).

(20) a. *a gwidonot Kaer Loyw a e lladassei*
and witches Gloucester PRT 3MS kill.PLQPF.3S
'And the witches of Gloucester killed him' (Peredur 33.28)
b. *ac Arthur a y lladawd ynteu*
and Arthur PRT 3MS kill.PAST.3S him
'And Arthur killed him.' (CO 284)

From the above table, it is clear that subjects occupy the preverbal position in the abnormal sentence more often than objects. According to T. A. Watkins (1993) the lower frequency of fronted objects is the result of their higher degree of 'markedness'. The subject is the least 'marked' constituent of the sentence and will thus appear in first position (unless any of the other 'constraints' on word order apply, like the 'Imperative constraint', which places verbs in sentence-initial position). According to Poppe (1993), the choice of subjects as topics of their sentences (and thus fronted elements of verb-second clauses) is 'natural'. In section 5.5 and Chapter 6 I go into this issue in further detail.

Subjects could also appear in sentence-final position. According to Borsley et al. (2007), this occurs when a noun phrase is either heavy, as in (21a), or when "the clause presents some new element in the discourse" (Borsley et al., 2007:316) (see also section 5.5 below). Late subjects can also occur with unaccusative verbs. According to Borsley et al. (2007), this is only possible if the subject is a pronoun following the complement of the verb like *yma* in (21c). The complement could also consist of a prepositional phrase, as in (21b). Direct objects and noun phrases could also function as the patient of verbs with impersonal inflection, as shown in (21d) and (21e) respectively.

(21) a. *kanys ny wisgawd arueu eiryoet uarchawc urdawl well noc ef*
because NEG wear.PAST.3S arms ever knight honourable better than him
'since a better knight than he never bore arms.'(YSG 3972-3 - Borsley et al. (2007))
b. *Dypi iti hynny.*
come.FUT.3S to.2S that
'You shall have that' (Lit. 'That shall come to you') (CO 535)
c. *Pa neges y dodyvch yma chwi?*
which mission PRT come.PERF.2P here you
'On what mission have you come here?' (CO 476-7)
d. *Gellwng y mywn wy*
let.IPV.2S in them
'Let them in' (PKM 81.27)

e. *ac y lladwyt yna Twrch Llawin.*
and PRT kill.PAST.IMPERS there Twrch Llawin
'And Twrch Llawin was killed there.' (CO 1147)

Indirect objects are expressed by a prepositional phrase introduced by *y* 'to'. The English-type 'dative-alternation' ('He gave Mary a book' vs. 'He gave a book to Mary') is not found in Middle Welsh. The order of direct and indirect object varies, as shown in (22). Indirect objects could also be passivised ('Mary was given the book'), although with the verb *dywedyt* 'to say, tell about' there are examples of raising of arguments that were not the patient of the verb, as in (22c). :

(22) a. *Y rodet y march y r fab.*
PRT give.PAST.IMPERS the horse to the boy
'The horse was given to the boy.' (PKM 24.4-5)
b. *Mi a dangossaf ytti dyn bychan.*
I PRT show.PRES.1S to.2S man small
'I will show you a small man.' (Owein 130)
c. *Kei a dywedit y uot yn uab itaw.*
Kei PRT say.IMPF.IMPERS 3MS be.INF PRED son to.3MS
'Kei was said to be his son.' (CO 265)

Intransitive

Intransitive verbs can be further categorised as unergative (with an external argument) or unaccusative (with an internal argument).[1] According to Tallerman and Wallenberg (2012), the arguments of verbal nouns in Middle Welsh (Type VI above) exhibit an ergative case-marking pattern: the subject and object are grouped together vs. the agent. There are two possible word order patterns available: the first pattern is VN + $\text{Sbj}_{\text{unacc.}}$ (with preverbal clitics in case of pronominal subjects), the second pattern is VN + preposition *o/y* + $\text{Sbj}_{\text{unerg.}}$. The verbal nouns themselves display either split or fluid intransitivity, i.e. some can use both patterns depending on additional factors like animacy.

Examples of Pattern 1: VN + $\text{Sbj}_{\text{unacc.}}$

(23) a. *Marw y urenhines.*
die.INF the queen
'The queen died.' (CO 22)
b. *Kyuodi yna Kei.*
rise.INF then Kei
'Then Kei got up.' (CO 384)

[1] Intransitive verbs can be split into unergatives and unaccusatives. Unergative verbs have an external argument, usually an agent, e.g. to dance. Unaccusative verbs have an internal argument (the argument that is usually the complement of a transitive verb), e.g. to arrive. In languages like Dutch, the difference between unergatives and unaccusatives is clear from the choice of auxiliary in the perfect (have or be), e.g. *Ik heb gedanst* 'I have danced' vs. *Ik ben aangekomen* 'I have arrived'.

c. *A e dyuot ynteu y r llys*
and 3FS come.INF he to the court
'And he came to the court.' (CO 46)

Examples of Pattern 2: VN + preposition *i/o* + $\text{Sbj}_{\text{unerg.}}$

(24) a. *Emystynnu idaw ynteu yn y peir*
stretch.INF to.3MS him in the cauldron
'He stretched himself out in the cauldron' (PKM 44.19)
b. *Canu englyn idaw ynteu yna*
sing.INF englyn to.3MS him then
'He sang an englyn then' (PKM 90.9)
c. *Yna agori y safyn y 'r llew*
then open.INF 3MS mouth to the lion
'Then the lion opened its mouth' (YBH 31.1296-7)

The direct object or prepositional phrases following the verb could precede (as in (25a) and (25b)) or follow the agent, as shown in (25c-g):

(25) a. *A chaffael mab ohonu trwy weti y wlad.*
and get.INF son from.3P through pray.INF the country
'And through the country's prayers they got a son.' (CO 4)
b. *Ymrodi y gerdet ohonaw ynteu.*
undertake.INF 3MS walk.INF of.3MS him
'He started to walk.' (CO 1145)
c. *Galw o Arthur ar Gyndylic Kyuarwyd.*
call.INF of Arthur on Cyndylic Kyuarwyd
'Arthur called on Cyndylic Kyfarwyd' (CO 399)
d. *Marchogaeth o Galaath*
ride.INF of Galaath
'Galaath rode' Tallerman and Wallenberg (2012:4)
e. *Kerdet ohonu y dyt hwnnw.*
walk.INF of.3P the day that
'That day they walked.' (CO 413)
f. *Ryuedu o Owain.*
marvel.INF of Owain
'Owain marvelled.' Tallerman and Wallenberg (2012:4)
g. *a goruot o Wyn a dala Greit mab Eri*
and overcome.INF of Gwyn and take.INF Greit son Eri
'And Gwyn won and took Greid son of Eri' (CO 992)

In the sentence following (25g), however, the verb *dala* appears again with a prepositional phrase introduced by *o* that can clearly not be interpreted as the agent:

(26) *A dala o Penn uab Nethawc (...)*
and take.INF of Penn son Nethawc (...)
'And he took Penn son of Nethawc (...)' (CO 993)

Penn uab Nethawc and the following names are the ones who were taken prisoner by Gwyn (the agent of the previous sentence). If we want to maintain Tallerman and Wallenberg's ergative distinction, we have to assume that either the *o* in this one particular example is a mistake or that the actual agent *Gwyn* was accidentally omitted between *o* and *Penn*. Transitive verbal nouns with both agents and patients expressed usually exhibit VN - *o* Agent - Patient order as in (27), but the order of the arguments could also be reversed, as in (28):

(27) a. *galw o Uendigeiduran y mab attaw*
call.INF of Bendigeidfran the boy to.3MS
'Bendigeidfran called the boy to him.' (PKM 43.13-14)

b. *Clybot oheni hitheu eu trwst yn dyuot.*
hear.INF of.3FS her 3P noise PROGR come.INF
'She heard the noise of their coming.' (CO 459)

(28) a. *Kymryt crip eur o Arthur*
take.INF comb gold of Arthur
'Arthur took a gold comb.' (CO 164)

b. *Keissaw gwiscaw y uodrwy ohonaw ac nyd aei*
seek.INF wear.INF the ring of.3MS and NEG go.PAST.3S
'He sought to put on the ring, but it would not go' (CO 442)

There are two types of *bod* 'to be' in Middle Welsh. One form was used as the copula (Type VII) and could occur in sentence-initial position in the form *ys* in Old and Early Middle Welsh, as shown in (29a). Though in the Bible translation from 1588, examples with the preterite form of *bod* in sentence-initial position still occur, as shown in (29b), the copula could also occur in medial position in the form *yw/ydy*, as shown in (29c) or (29d) and (29e) with other tenses. Finally, with focus on the subject, the relative form of the copula could be used (in any tense) immediately following the subject, as in (29f).

(29) a. *is moi hinnoid*
be.PRES.3S more this
'this is more' (Old Welsh M&P 23r)

b. *A bu Ddafydd gall yn ei holl ffyrdd.*
and be.PAST.3S David smart in 3MS all ways
'And David was smart in all ways.' (b1588 - 1 Sam. 18.14)

c. *Trydyd yw kamarver o e wreic.*
third be.PRES.3S abuse of 3MS wife
'Third is the abuse of his wife.' (Laws 14)

d. *Budugawl oed Kei.*
victorious be.PAST.3S Kei
'Kei was victorious' (CO 387)

e. *Dilesteir uyd dy hynt.*
unhindered be.FUT.3S 2S road
'Your path will be unimpeded.' (PKM 3.26)

f. *Mi a uydaf porthawr y Arthur pob dyw kalan Ionawr*
I PRT be.FUT.1S gatekeeper to Arthur every day first.January
'I will be gatekeeper to Arthur on every first of January.' (CO 83-84)

The verb *bod* was furthermore used as the substantive verb 'to be, to exist'. This substantive form behaved like any other verb and thus occurred with various word order types as shown in (30). The verb-initial order preceded by the preverbal particle *y* was very common.

(30) a. *y buant ulwydyn gyt a mi*
PRT be.PAST.3P year with me
'They were with me for a year.' (PKM 35.24)

b. *a llawen uuwyd vrthunt yno*
and joy be.PAST.IMPERS to.3P there
'And there they were made welcome.' (Gereint 1337)

c. *Mae yna carw. ac ewic. ac elein gyt ac wynt.*
be.PRES.3S there stag and doe and fawn with them
'There was a stag and a doe and a fawn with them.' (PKM 75.12-13)

Finally, *bod* functioned as an auxiliary in periphrastic constructions. This construction was used more and more after the Middle Welsh period, as we see in the 1588 Bible translation, shown in (31a). But examples of this construction can occasionally also be found in earlier Middle Welsh texts, as shown in (31b).

(31) a. *yr wyf fi yn cofio fy meiau heddyw*
PRT be.PRES.1S I PROGR think.INF 1S sins today
'I am thinking about my sins today' (b1588 - Gen. 41.9)

b. *Ac y mae matholwch yn rodi brenhinaeth iwerdon y*
and PRT be.PRES.3S Matholwch PROGR give.INF kingdom Ireland to
wern
Gwern
'And Matholwch is giving the kingdom of Ireland to Gwern' (PKM 41.9)

5.2.5 Diathesis

One way to distinguish active from passive voice in Welsh is the use of a special set of verbal endings (in all tenses and moods) called 'the impersonal inflection'. The distribution of impersonal verb forms over the different types of word order is shown in table 5.6 below.

	Active	Impersonal	
Type I Verb-initial	290 (93.94%)	12 (6.06%)	198
Type II Periphrastic	30 (96.78%)	1 (3.22%)	31
Type III Adj y VS	610 (78.53%)	167 (21.47%)	778
Type IVa SaVO	1416 (91.16%)	137 (8.86%)	1546
Type IVb OaVS	354 (100%)	0 (0%)	357
Type IVc VN a DO	446 (95.10%)	23 (4.90%)	469
Type V Focus	6 (100%)	0 (0%)	6
Total	3048	340	3388

Table 5.6: Diathesis of transitive verbs

When we again compare adjunct- and argument-initial abnormal orders (Type III vs. Type IVab), we find a significant difference in diathesis ($\chi^2 = 111.12$, df $= 1$, $p < 0.0001$, Fisher's exact $p < 0.0001$). This can already be observed from the above frequency table: impersonal inflections appears much more often with adjunct-initial word order. Some examples are given in (32a), (32b) and (32c). There are also some other word order types in which impersonals occur, as shown in (32d), (32e) and (32f).

(32) a. *a Fferis brenhin Freinc, ac am hynny y gelwir Kaer Paris*
and Peris king France and for that PRT call.PRES.IMPERS Paris
'And Peris, king of France, and because of that it was called Paris'(CO 278)

b. *A bydydaw y mab a orucpwyt.*
and baptise.INF the boy PRT do.PAST.IMPERS
'And the boy was baptised.' (CO 9-10)

c. *kam y m byrywyt i doe*
wrong PRT 1S hit.PAST.IMPERS I yesterday
'It was wrong that I was hit yesterday' (Owein 192)

d. *Gorucpwyt hynny.*
do.PAST.IMPERS that
'That was done.' (CO 519)

e. *ac y gwnaethpwyt y ffyrd*
and PRT make.PAST.IMPERS the roads
'and the roads were made' (BM 9.16)

f. *A mynegwyd i Saul gan ddywedyd*
and tell.PAST.IMPERS to Saul by say.INF
'And it was told to Saul, saying' (b1588 - 1 Sam 19.19)

Impersonal inflection was very often interpreted as a passive as in (33d), but true

impersonal examples existed as well in Middle Welsh, as shown in (33a), (33b) and (33c).

(33) a. *Pa gyueir heb y Gereint yd eir yma.*
what reason said Gereint PRT go.PRES.IMPERS here
'What's the reason one goes here? said Gereint' (Gereint 1404)
b. *kyweirher i minheu vy march*
prepare.IPV.IMPERS to me 1S horse
'Let my horse be prepared for me.' (Peredur 26.8)
c. *a m rodi y wr o m hanwod yd ydys.*
and 1S give.INF to man from 1S unwill PRT be.PRES.IMPERS
'And they were giving me to a husband against my will.' (PKM 12.23-24)
d. *Ac yna gyntaf y guarywyt broch yg got.*
and then first PRT play.PAST.IMPERS badger in bag
'And then 'Badger in the Bag' was played for the first time.'(PKM 17.13-14)

5.2.6 Agreement

'Agreement' can refer to various aspects of the grammar of a language. For this study, I am only concerned with agreement between the subject and the verb and, to some extent, topic agreement reflected as the particles *a* or *y*, depending on the type of fronted constituent in verb-second clauses. As mentioned in the previous chapter, lack of subject-verb agreement was one of the main features to distinguish the 'Abnormal' from the 'Mixed' (or 'focussed') verb-second patterns in Middle Welsh. The mixed order was a reduced cleft sentence and as such it featured a relative clause after the focussed constituent. Since agreement did not (usually) occur in relative clauses in Middle Welsh (cf. D. S. Evans (2003 [1964]) and Borsley et al. (2007:334) among others), it also did not occur in the mixed sentence.

The abnormal pattern superficially looked exactly like the mixed sentence. They both had similar types of fronted constituents and both featured the (relative) particle *a/y* with the same distribution (*a* following arguments, *y* following adjuncts). The fact that most of these abnormal clauses *do* exhibit subject-verb agreement, even with plural noun phrases, requires an explanation. Agreement with full noun phrases did not occur in any other word order pattern, e.g. in patterns with subjects following the verbs like Type I VS or Type III Adj*y*VS. In these cases, the verb very often exhibited default third-person singular endings.

(34) a. *A r bore ym bronn y dyd drannoeth yd ymordiwedawd rei*
and the morning in edge the day next.morning PRT overtake.PAST.3S some
o r gwyr ac ef.
of the men with him
'And on the early morning the next day some of the men caught up with him.' (CO 1119)
b. *Ac yna y dechreuawd y seint bregethu bop eilwers.*
and then PRT begin.PAST.3S the saints preach.INF every moment
'And then the one by one the saints started to preach.' (Dewi 13.7)

c. *Yna y doeth kennadeu.*
then PRT come.PAST.3S messengers
'Then messengers came.' (PKM 79.27)

	Plural agreement	Plural + default 3S
Laws	2	
Culhwch		6
Pwyll	2	2
Branwen	7	2
Math	2	1
Owein		1
Peredur	2	3
Gereint	6	2
Lludd WB	1	
Lludd CH	2	2
Rhonabwy	1	
Macsen		4
Dewi	8	1
Beibl 1588	43	

Table 5.7: Agreement with plural noun phrases in Middle Welsh subject-initial clauses

The overall numbers of plural full DPs are very low. In most texts, we only find fewer than ten examples like the ones in (35). There is no clear pattern in terms of agreement vs. default third-person singular, apart from the large amount of agreement examples in the Bible translation. Combined, the excerpts of the bible are longer than most other texts, so chances of finding plural DP subjects are higher to begin with. The complete lack of third-person singular patterns in such a large text suggests the preferred standard for the Bible translation was plural agreement.

(35) a. *uy aeleu ry syrthwys ar aualeu uy llygeit*
1S eyebrows PRT fall.PAST.3S on balls 1S eyes
'My eyebrows have fallen on my eyeballs.' (CO 547-548)

b. *Y gwyr a dywawt wrth Arthur.*
the men PRT say.PAST.3S to Arthur
'The men said to Arthur.' (CO 839)

c. *Y gwyr a wiscawd amdanunt ac a nessayssant attunt y wayret.*
the men PRT arm.PAST.3S on.3P and PRT approach.PAST.3P to.3P down
'The men armed themselves and went down towards them.'(PKM 29.22-23)

d. *Deu uarchauc a doeth i waret*
two knight PRT come.PAST.3S down
'Two knights came down.' (PKM 32.18-19)

In a number of cases, the facts are further complicated because it is actually unclear what the 'expected' agreement pattern should be. This is mainly the case in noun phrases that contain numerals and/or quantifiers (those difficult cases are therefore excluded in the above table). Numerals preceded the noun, which was mostly found in the singular, rather than the plural in that case. Nonetheless, the entire phrase was more often found with verbs with plural endings than other plural phrases (cf. Nurmio and Willis (2016)). Number itself was a complex feature of Middle Welsh grammar: there were singulars and plurals, but also duals, collectives and, from those, new singulatives were derived (cf. Nurmio (2015)). It was possibly as a result of all this as well that 'mixed' agreement patterns like the ones shown in example (36) were found.

(36) *A phan yttoedynt y deu amherawdyr ar eu bwyt y doeth y*
and when be.PAST.3P the two emperor on 3P food PRT come.PAST.3S the
Brytanyeit wrth y gaer
Britons at the town
'and while the two emperors were at their meat, the Britons came to the town' (BM 11.11)

Poppe (2009) concludes after reinvestigating several Middle Welsh texts that "the rules of concord were not systematically exploited, at least in the case of fronted plural subjects, in order to distinguish between the pragmatic functions of topic an focus" (Poppe, 2009:258). Instead of the more rigid distinction between the abnormal (topicalised) order and the mixed (focussed) order he in his earlier studies claimed to exist, he now proposed 'a pragmatic cline' from topic to focus reserved for constituents that are fronted as the centre of attention. After presenting more examples of 'unexpected (lack of) concord', he goes even further saying that "[t]hese examples are embarrassing for any attempt to relate the formal differences to pragmatic differences." (Poppe, 2009:257)

According to T. A. Watkins (1988), agreement between subjects and verbs in abnormal sentences must have been an innovation. More than that, he called it a "solely literary development" (T. A. Watkins, 1988:11). D. S. Evans (1971) suggests that this happened under the influence of Latin grammar: "It was always there, but naturally its influence was doubly exerted on the translators who had a Latin text at their elbow." (D. S. Evans, 1971:56). This argument does not always hold when comparing Welsh translation to their Latin originals (cf. Plein and Poppe (2014)). Plein and Poppe (2014) note a methodological flaw in his study: since he only collects instances of 'unexpected (lack of) agreement', there is no way to contrast this with the number of instances that do exhibit the expected pattern. They conclude that Latin influence is likely, but "the amount of variation attested in the *Historia* shows that the syntactic system of Middle Welsh permitted and tolerated such variation" (Plein & Poppe, 2014:13).

There is one other reason why it is difficult to examine the exact agreement rules in Middle Welsh. As Koch (1991) points out, it is not altogether clear that the third-person conjunct plural ending *-nt* has actually survived apocope. If it did not survive, it would strictly speaking have been very difficult - if at all possible - to distinguish the singular from the plural verbal endings. Plural *-nt* could have been analogically restored later in some paradigms, but the proper inherited forms of the singular and plural would have been the same. This could also account for (or at least contribute to) the puzzling variation in agreement patterns. Not all historical phonologists believe Koch (1991) to be right here, but it is impossible to test his hypothesis. It seems reasonable since all final consonants in Proto-British were lost because of apocope (apart from word-final *-r*, but there are no other cases of word-final *-nt* to compare this to (cf. Peter Schrijver p.c.).

I examine this variation and the limits thereof further in section 5.6 below. In chapter 6 I furthermore present a case study of the interaction between syntax and information structure about this exact problem with the traditional distinction between the abnormal and the mixed word order patterns in Middle Welsh.

5.2.7 Types of argument phrases

In the previous section I have shown that different types of subjects yield different agreement patterns. Pronouns exhibit agreement in other parts of the grammar as well (e.g. inflected prepositions), whereas full noun phrases never do. This is called 'the complementarity principle' (cf. Anderson (1982), Sproat (1983) and Borsley (1989) among others). Since agreement was already discussed above, in this section I only focus on the remaining issues concerning different types of arguments.

Subject vs. object pronouns

In preverbal subject position, three types of pronouns could appear in Middle Welsh: simple, conjunctive and reduplicated pronouns, as shown in Table 5.8:

	Simple	Conjunctive	Reduplicated
I	*mi*	*minneu*	*miui*
you (sg.)	*ti*	*titheu*	*tidi*
he	*ef*	*ynteu*	*efo*
she	*hi*	*hitheu*	*hihi*
we	*ni*	*ninneu*	*nini*
you (pl.)	*chwi*	*chwitheu*	*chwichwi*
they	*wy*	*wynteu*	*wyntwy*

Table 5.8: Middle Welsh Preverbal subject pronouns, cf. Willis (1998:134)

Conjunctive pronouns were used in close connection with the preceding context (mainly to switch the topic, but see section 5.5). Reduplicated pronouns were

always focussed.

(37) a. *hyt nas gwelei neb vynt ac vyntvy a*
so.that NEG.3S see.IMPF-SBJ.3S no.one them but they.REDUP PRT
welynt pawb
see.IMPF-SBJ.3P all
'so that no one could see them, but THEY could see everyone' (CO 4.358)

b. *a e uenegi idi a wnaeth. Hitheu a gymerth diruawr*
and 3MS tell.INF to.3FS PRT do.PAST.3S she.CONJ PRT take.PAST.3S great
lywenyd yndi.
pleasure in.3FS
'And he told it to her. She, then, took great pleasure in (hearing) it.'(Math 1.561)

As became clear from the frequency tables in the previous chapter, there are far fewer examples of verb-second orders with initial objects than there are with initial subjects. One of the reasons for this is grammatical restriction of the Welsh language: subject pronouns can appear independently (and are thus possible in sentence-initial position) as in (38a), but object or genitive pronouns cannot. Genitive pronouns are used as possessives. They appear in two forms, depending on the preceding word (originally a phonological distinction between words ending in vowels or consonants).

	Object	Possessive
I	*'m*	*vy/'m*
you (sg.)	*'th*	*dy/'th*
he	*'e/s*	*y/'e*
she	*'e/s*	*y/'e*
we	*'n*	*yn/'n*
you (pl.)	*'ch*	*ych/'ch*
they	*'e/s*	*eu/'e*

Table 5.9: Middle Welsh dependent pronouns, cf. Willis (2011b)

Pronominal direct objects always appear as clitics between the preverbal particle and the inflected verb, as shown in (38). They can optionally be 'doubled', i.e. apart from the clitic a further pronominal form known as the 'echo pronoun' could follow the inflected verb, as shown in (38c). The infixed object clitic is compulsory. Note that (38b) without an infixed clitic for this reason cannot mean 'Llewelis loved *him* most'. Pronominal direct objects of verbal nouns take their possessive form, treating the verbal noun as any other noun.

(38) a. *Ac ef a welei neuad.*
and he PRT see.PAST.3S hall
'And he saw a hall' (Peredur 3.976)

b. *Llewelys hagen a karey ef en wuyhaf o y vrodyr.*
Llewelis however PRT love.PAST.3S he PRED most of 3MS brothers
'Llewelis, however, he loved most of all his brothers.' (Llan 267.12)

c. *Yr Arglwyd a m anuones i attat ti*
the Lord PRT 1S send.PAST.3S me to.2S you
'The Lord sent me to you' (Dewi 2.9)

d. *Auory mi a th ganhadaf di e ymdeith*
tomorrow I PRT 2S allow.PRES.1S you to go.INF
'Tomorrow, I allow you to go.' (PKM 85.28)

e. *Ac eu gorchymyn y enyt a wnaeth.*
and 3P entrust.INF to Enid PRT do.PAST.3S
'and he entrusted them to Enid' (Gereint 857)

Pronominal subject 'echo pronouns' could also be left unexpressed in Middle Welsh if the verbal inflection was sufficient to disambiguate the potential subjects. Although there is no overt subject in these pro-drop cases, the verbal inflection is counted as the subject, thus yielding Verb-Subject order. In sentences with clause-initial pronominal subjects, the verb is still inflected, but the word order is analysed as Subject-Verb (instead of Subject-Verb-Subject, with the final subject reflecting the inflection on the verb only).

Expletives

Expletives form a very specific kind of pronominal subject. In Middle Welsh, they can be found in the same sentence-initial position as other subject pronouns.

(39) *Ef a doeth makuyueit a guesson ieueinc y diarchenu*
it PRT come.PAST.3S squires and lads young to.3MS disrobe.INF
'There came squires and young lads to disrobe him.' (PKM 4.8-9)

Expletive subjects are found in three contexts in Middle Welsh: before unaccusatives (mostly verbs of motion) as in (40a), with impersonal verbs as in (40b) and, finally, "in the topic position of some main clauses containing postposed clausal arguments" (Willis, 1998:151), as shown in (40c).

(40) a. *Ef a gyuodes Pwyll y uynyd*
it PRT rise.PAST.3S Pwyll up
'Pwyll got up.' (PKM 18.27)

b. *Ef a dywetpwyt idaw.*
it PRT say.PAST.IMPERS to.3MS
'It was said to him.' (PKM 80.9-10)

c. *Ac ef a tebygei Owein bot yr awyr yn edrinaw*
and it PRT suppose.PAST.3S Owein be.INF the air PROGR reverberate.INF
rac meint y gweidi
against amount the shouting
'And Owain supposed that the air was reverberating with the noise of the shouting' (Owein 346-7)

From the sixteenth century onwards, expletives could also be found with transitive verbs. According to Willis (1998), "One major cause of the spread of verb-initial word order at lower stylistic levels is the spread of the expletive construction beyond the environment to which it is restricted in the Middle Welsh tales." (Willis, 1998:149). I turn to these diachronic implications in chapter 7.

Nominal subjects

Currie (2000) notes that "in contrast to the pattern with a fronted verbal-noun object (...), an expressed nominal subject is frequently used in sentences with a fronted adverbial expression, i.e. in 57%." (Currie, 2000:223). The frequencies for Middle Welsh are listed in table 5.10 below.

	Nominal subject	Pronominal subject
Type III Adj *y* VS	622 (30.37%)	902 (28.76%)
Type IVab SaVO-OaVS	1061 (51.81%)	1666 (53.13%)
Type IVc VN*a*DO	365 (17.82%)	568 (18.11%)
Total	2048 (100%)	3136 (100%)

Table 5.10: Nominal subjects in Middle Welsh relevant sentence types

If we run a chi-square test, we see that there is actually no significant difference between nominal and pronominal subjects in relation to different kinds of verb-second word orders (Types III, IVab and IVc). Both nominal and pronominal subjects are possible in all word order types. This is thus not a grammatical constraint. In section 5.5 below, I discuss this difference again and try to seek an explanation related to the degree of Givenness of these subjects.

'Heavy' constituents

As noted above, subjects do not usually appear in clause-final position, as in (41). If the subject noun phrase is a complex or 'heavy' constituent, however, clause-final position was an option in Middle Welsh.

(41) *kanys ny wisgawd arueu eiryoet uarchawc urdawl well noc ef*
because NEG wear.PAST.3S arms ever knight honourable better than him
'since a better knight than he never bore arms.' (YSG 3972-3)

Since there are very few examples of these late subjects in the Middle Welsh corpus under investigation, it is very difficult at this stage to determine if this was more than just an option. Fronting of 'heavy' constituents was in itself not problematic in Middle Welsh. In word order type IVc, with fronted verbal nouns, there are many examples in which not just the verbal noun, but its entire complement and even the rest of the sentence is fronted as well. An analysis of 'optional' late subjects when they are 'heavy' thus seems more likely.

(42) a. *Galw y hathro attei a oruc hitheu.*
call.INF 3FS teacher to.3FS PRT do.PAST.3S she
'She called her teacher.' (CO 20-21)

b. *Bwrw badeu allan a wnaethont wynteu*
throw.INF boats out PRT do.PAST.3P they
'They threw the boats out.' (PKM 30.8-9)

c. *a chwynaw yn luttaf yn y byt rac Aranrot a*
and complain.INF PRED stubborn in the world against Arianrhod PRT
wnaethant
do.PAST.3P
'and they complained to Arianrhod in the most stubborn way in the world' (PKM 83.17-18)

5.2.8 Grammatical words and phrases

Some lexical items have a grammatical function in addition to (or instead of) their semantic content. In this section I discuss the most common functional elements and also some fixed expressions that are associated with specific word order types.

Fixed expressions

S. Davies (1995) lists various types of idiomatic phrases, formulae and frequently-used expressions in Middle Welsh narrative tales. For greetings, for example, one of the following expressions is used:

— *Dyd da itt* 'good day to you'
— *Kyuarch gwell* 'greetings'
— *Duw a rodo da itt* 'May God give you good (things)'
— *Craesaw Duw wrthyt* 'God's welcome to you'

There are furthermore certain recurring patterns in opening and closing statements (called *fformiwlâu* 'formulae' by S. Davies (1995)). The proper name of the main protagonist is in sentence-initial position, followed by his title or status, which is in turn followed by the extent of their kingdom (or their location). Examples of this can be found in various tales of the *Mabinogion*, as shown in (43). Closing statements of narrative tales, on the other hand, frequently exhibit the pattern *felly* 'thus' or *fel hyn* 'like this' + preverbal particle *y* + a verb that sums up or literally finishes the tale, as shown in (44).

(43) a. *Pwyll Pendeuic Dyuet a oed yn arglwyd ar seith cantref Dyuet.*
Pwyll Prince Dyfed PRT be.PAST.3S PRED lord on seven cantref Dyuet
'Pwyll Prince of Dyfed was lord of the seven cantrefs of Dyfed.' (PKM 1)

b. *Bendigeiduran uab Llyr a oed urenhin coronawc ar yr ynys hon,*
Bendigeidfran son Llyr PRT be.PAST.3S king crowned on the island this
ac ardyrchawc o goron Lundein.
and invested with crown London
'Bendigeidfran son of Llyr was crowned king of this island and invested with the crown of London.' (Branwen 1)

(44) a. *Ac yuelly y teruyna r geing hon yma o r Mabinogyon.*
and thus PRT end.PRES.3S the branch this here of the Mabinogion
'Thus ends this branch here of the Mabinogion.' (PKM 27.27-28)
b. *Ac uelly y kauas Kulhwch Olwen merch Ys. P.*
and thus PRT get.PAST.3S Culhwch Olwen daughter Y. P.
'And thus Culhwch obtained Olwen daughter of Y.P.' (CO 1245)

The main protagonists of the *Mabinogion* in particular often found themselves in need of counsel. There was a very specific set of phrases used for this procedure. Getting counsel was expressed with the phrase *kymryt kynghor* 'taking counsel'. The result of this was usually presented in a *sef*-construction:

(45) a. *Sef a gahat yn y kynghor rodi branwen y uatholwch.*
sef PRT get.PAST.3S in 3P council give Branwen to Matholwch
'This is what they got in their council: giving Branwen to Matholowch' (PKM 30.28-29)
b. *Sef y kawssant yn eu kyghor, gossot kanwr ym pop tri chymwt ym Powys o e geissaw.*
sef PRT get.PAST.3P in 3P council place.INF 100.men in every three Commot in Powys of 3MS seek.INF
'They determined to place a hundred men in each of the three Commots of Powys to seek for him.' (BR 1.14)

Similarly, the *sef*-construction was very often used to describe the table settings of big feasts.

(46) *Sef ual yd eistydassant o r neilltu y Ereint yd eistedawd y iarll ieuanc*
sef how PRT sit.PAST.3P from the one.side to Gereint PRT sit.PAST.3S the earl young
'This is how they were sitting: the young earl sat on the one side of Gereint.' (Gereint 366-367)

There is very little variation in word order when one of these formulae or expressions were used. One type of the *sef*-construction that gained particular high frequency was the variant with periphrastic *gwneuthur* 'to do', as shown in (47).

(47) a. *Sef a wnaeth ynteu y deimlaw ef yny gauas y benn.*
sef PRT do.PAST.3S he 3MS feel.INF him until get.PAST.3S 3MS head
'And he felt about it until he came to the man's head.' (PKM 42.27)
b. *Sef a wnaeth ynteu maglu y llinin am uynwgyl y llygoden*
sef PRT do.PAST.3S he noose.INF the string on neck the mouse
'Then he put the noose around the mouse's neck' (PKM 63.5)

Initially, the *sef*-construction was employed to focus the predicate of an identificatory copular clause, but this interpretation was lost in the Middle Welsh period. In chapter 6 I discuss the exact diachronic development of all *sef*-constructions in greater detail.

Finally, there are some examples of figurae etymologicae. These examples where the internal argument is repeated by the verb, are often found in the Hebrew Bible. There is, however, also one example of this in the early Middle Welsh tale *Culhwch ac Olwen* with verb-initial word order:

(48) *Tyghaf tyghet it na latho dy ystlys vrth wreic*
swear.PRES.1S oath to.2S NEG strike.PRES-SBJ.3S 2S area with wife
'I declare to thee, that it is thy destiny not to be suited with a wife' (CO 50)

Focus particles

As in most languages, certain lexical items in Middle Welsh were used to focus preceding or following constituents. The most common particles preceding the focussed constituent are *hyd yn oet* 'even' and *dim ond* 'only'. Others follow the focussed constituent, like *hagen* 'however' and *eyssioes* 'nevertheless, still', or could either follow or precede, like *heuyd* 'also, too'. Words that mean 'the same' or 'the other (one)' also denote one specific item in a set of alternatives and are therefore related to constituent focus as well. In Middle Welsh, *un* 'one', could also mean 'the same' and occurred just like the numeral in front of the modified constituent. The adjective *arall/ereill* 'other (sg/pl)', like most other adjectives, followed it.

The word order of the whole clause did not necessarily change when one constituent was focussed. The mixed sentence could be used, but constituent focus also appeared with other word order types (see also section 5.5 below).

(49) a. *Velly hagen y gorfuost ar lawer onadunt wy*
thus however PRT prevail.PAST.2S on many of.3P them
'Thus, however, you triumphed over many of them.' (Peredur 32.23-24)
b. *Ti a geffy hynny heuyt.*
you PRT get.PRES.2S that too
'You will get that too.' (PKM 64.1)

There was also a fixed set of originally demonstrative pronouns (or contraction of demonstratives and certain adverbs 'see here/there', cf. French *voilà*) that was used to introduce a character or item in the story with an element of surprise (i.e. a mirative reading). *Llyna, dyma, nachaf* and *wele* 'lo, behold' were the most common. The word order pattern was that of a truncated copular clause (Type IXa). The interpretation was not always mirative, according to Sturzer (2001), because "'finding people in and around a fort or castle going about their business is an expected and ordinary circumstance" (Sturzer, 2001:41). This word order pattern could therefore also be used simply to draw attention to a character or situation.

(50) a. *llyna y marchawc yd aeth Gereint yn y ol*
behold the knight PRT go.PAST.3S Gereint in 3MS back
'Behold the knight Gereint went after him.' (Gereint 430)
b. *Dymma ei ddeongliad ef*
behold 3MS interpretation him
'Behold his interpretation/Here is his interpretation.'(b1588 - Gen. 40.12)

Dyma/llyna/nachaf could be used as adverbials as well. In this case, they were followed by the preverbal particle *y* and then the inflected verb, resulting in the adjunct-initial Type III, as shown in (51).

(51) *Nachaf y gwelynt o pebbyll gwynn penngoch*
behold PRT see.PAST.3P of tent white top.red
'Behold they saw a white tent with a red canopy.' (BR 11.31)

Welsh has special particles for focussed questions as well, but these are beyond the scope of the present study.

Conjunctions & complementizers

Conjunctions and complementizers always introduce the main or subordinate clause. Some conjunctions that introduce main clauses, like *a(c)* 'and' could appear before any word order pattern. Others, mainly subordinate conjunctions and complementizers are directly followed by the inflected verb in all stages of Welsh. Since this study is concerned with main clauses, I will not discuss the subordinate conjunction and their verb-initial word orders here.

There is one conjunction that deserves further attention: *canys* 'because'. This is a contraction of earlier < *can* 'since, for' + *ys* 'it is'. The copula in sentence-initial position resulted in a following cleft sentence pattern in an earlier stage of the language. The constituent following the copula was originally the predicate, followed be a relative clause to modify it. Since relative clauses usually did not exhibit agreement, even if the antecedent was a plural noun or pronoun, we would not expect plural inflection on the relative verb, as shown in (52a). However, as D. S. Evans (1971) and Borsley et al. (2007) point out, there are also some examples with agreement, as shown in (52b) and (52c). In example (52d), with a following preverbal particle and auxiliary *mae* 'is', it is clear that *canys* was completely grammaticalised as the conjunction meaning 'because'.

(52) a. *Canys Arabyeit yssyd yn chwerwdic yn y ymlil*
because Arabs be.REL-PRES.3S PRED angry PROGR 3MS pursue.INF
'Because the Arabs are pursuing him angrily.' (YBH 3958-3962)

b. *Canys Israel a r Philistiaid a fyddinasent fyddin yn erbyn byddin.*
because Israel and the Philistines PRT marshal.PAST.3P army against army
'Because Israel and the Philistines prepared army against army for battle.' (b1588 - 1 Sam. 17.21)

c. *canys eu llygaid hwy oeddynt drymmion.*
because 3P eyes them be.PAST.3P heavy.P
'Because their eyes were heavy.' (b1588 - Mat. 26.43)

d. *Canys y mae cariad Crist yn ein cymhell ni*
because PRT be.PRES.3S love Christ PROGR 1P spur.on.INF us
'Because the love of Christ spurs us on.' (b1588 - 2 Cor. 5.14)

Since *canys* can introduce main clauses as well, these examples are analysed and categorised according to their word order types in this study; conjunctions introducing subordinate clauses, like *pan* 'when', *ual* 'as, like' or *hyt* 'until' are not.

5.2.9 Semantics

Certain features of the grammar that have not been discussed so far are usually categorised as being semantic in nature. These include scope effects, animacy, but also certain lexical constraints. Since issues of scope in Middle Welsh that can influence word order are all related to negation, they are not relevant for the present study of positive main clauses. In this section I therefore focus on animacy, accessibility and lexical constraints only.

Animacy

Harlos, Poppe, and Widmer (2014) claim that animacy and accessibility of sentence-initial constituents play a role in Middle Welsh word order. They rate the level of animacy of constituents on a scale ranging from 'self' and 'human' to 'location' and 'abstract' on the lower end. Accessibility for them is the relationship between cognitive accessibility of a referent in the memory store of a participant in communication and the morphosyntactic encoding of the referent (Harlos et al., 2014:134n.31). I discuss this latter feature further in section 5.5 below on 'givenness'.

In the very small sample they investigate, they find a higher frequency of animate than inanimate subjects and, unsurprisingly, the reverse is true for direct objects. They furthermore claim that in clauses with indirect objects "animacy has an effect on the distribution of possible word order patterns" Harlos et al. (2014:145). If we test the statistical significance of the animacy (divided into two categories here, rather than a scalar notion) related to word order patterns, we indeed find there a significant result for indirect objects (χ^2 test with Yates's continuity correction: $\chi^2 \approx 6.55$, $df = 1$, $p \approx 0.0105$; Fisher exact test: $p \approx 0.0079$).

The animacy of subjects, however, does not give any statistically significant results in relation to choice of word order ($\chi^2 \approx 0.78$, $df = 1$, $p \approx 0.7768$; Fisher exact test: $p \approx 0.5578$). Nor are there any significant effects if we collapse subjects and indirect objects to look at animacy of arguments in general. The tables below are based on the counts presented by Harlos et al. (2014:140) for indirect objects only, but we observe a similar pattern for direct objects (i.e. animacy of objects is significant, but animacy of subjects or animacy of both subjects and objects in general has no significant effect).

	Animate Ind.Obj.	Inanimate Ind.Obj.
S-V-Ind.Obj.	27 (93.10%)	24 (63.18%)
Ind.Obj.-V-S	2 (6.90%)	14 (36.84%)
Total	29	38

Table 5.11: Animate & inanimate indirect objects in *Pwyll* from Harlos et al. (2014)

	Animate Subj.	Inanimate Subj.
S-V-Ind.Obj.	47 (77.05%)	2 (66.67%)
Ind.Obj.-V-S	14 (22.95%)	1 (33.33%)
Total	61	3

Table 5.12: Animate & inanimate subjects in *Pwyll* from Harlos et al. (2014)

It is, however, very difficult to draw any conclusion based on such a small sample. The word order pattern with sentence-initial indirect objects is in fact always a pattern with a sentence-initial prepositional phrase (since indirect objects always require a preposition in Welsh). These would be categorised as word order type IIIe, or adjunct-initial (including PP-initial) verb-second (see previous chapter). The different word order patterns Harlos et al. (2014) mention are, however, not *all* possible patterns. There are of course also sentences with both direct and indirect objects and it is unclear what word order pattern would be preferred in those cases (Type III with initial indirect object or Type IVa or IVb with initial subject or direct object respectively).

	AniSbj-AniObj	AniSbj-InObj	InSbj-AniObj	InSbj-InObj
I Verb-initial	15 (2.16%)	126 (5.55%)	1	0
II AuxSVO	13 (1.87%)	17 (0.75%)	2	0
III V2 Adj.	164 (23.63%)	443 (19.52%)	2	3
IVa SaVO	359 (51.73%)	1044 (46.01%)	8	2
IVb OaVS	41 (5.9%)	301 (13.27%)	1	1
IVc VN*a*DO	98 (14.12%)	336 (14.81%)	0	0
V Focus	4 (0.58%)	2 (0.09%)	0	0
Total	694 (100%)	2269 (100%)	14	6

Table 5.13: Animacy Subject-Object in entire corpus (A = animate, In = inanimate)

If we look at the animacy level of subjects and (indirect) objects in the entire Middle Welsh corpus under investigation, distributed over all these word order types (see table 5.14), we see a clear and expected pattern: subjects are mostly animate

and objects are inanimate. There is a significant difference between Subject- and Object-initial word orders in terms of animacy ($\chi^2 = 28.0198$, df $= 1$, p-value $<$ 0.00001). The relation between animacy of objects and word order Types III vs Type IV (both subjects & objects combined) is also significant, though the p-value is much higher ($\chi^2 = 3.9221$, df $= 1$, p-value $= 0.04766$).

It is more difficult to analyse the animacy of 'indirect objects' in the same way Harlos et al. (2014) did it for the tale of *Pwyll*.

	AniSbj-AniIObj	AniSbj-InIObj	InSbj-AniIObj	InSbj-InIObj
I Verb-initial	34 (2.92%)	32 (4.44%)	0	6
II Aux-initial	4 (0.34%)	1 (0.14%)	0	0
III Adj*y*VS	307 (26.37%)	258 (35.78%)	15	28
IVa SaVO	513 (44.07%)	209 (28.99%)	25	52
IVb OaVS	26 (2.23%)	9 (1.25%)	0	0
IVc VN*a*DO	276 (23.71%)	212 (29.40%)	0	0
V Focus	4	0	0	1
Total	1164 (100%)	721 (100%)	40	87

Table 5.14: Animacy Subject-Ind. Object in entire corpus (A = animate, In = Inanimate)

For active verbs, animacy of the indirect objects seems to be significant for word order Type III (argument-initial) vs type IV (adjunct-initial) (split in Type III vs Type IVa vs Type IVb: $\chi^2 = 38.2175$, df $= 2$, $p < 0.0001$)(Type III vs. Type IVa & b combined: $\chi^2 = 38.2735$, df $= 1$, $p < 0.0001$). There seems to be no significant difference between subject- and object-initial orders ($\chi^2 = 0.0703$, df $= 1$, $p = 0.7908$). When it comes to the animacy of direct objects, however, there is a difference between subject- and object-initial orders ($\chi^2 = 27.0993$ df $= 1$, $p < 0.0001$) and also (though only slightly) significant for word order Type III vs IV (combined a & b) ($\chi^2 = 4.4672$ df $= 1$, $p = 0.03455$). Animacy of the subject does not make any difference in preferred word order type.

In Middle Welsh texts, however, more distinct categories of animacy are not always easy to determine. There are many examples of magic changing people into animals (in Math) or creating people out of non-organic material (Blodeuwedd) or little boys that grow out of lumps of flesh (in Math). Even in religious texts this distinction between human and other animate beings is sometimes difficult to maintain, as in example (53). For the present study, therefore, only a basic animate vs. inanimate distinction was made.

(53) *a daeth yspryd yr Arglwydd ar Ddafydd o r dydd hwnnw*
and come.PAST.3S spirit the Lord on David from the day that
allan.
onwards
'And from that day onwards the spirit of the Lord came to David' (b 2.488)

Lexical constraints

As T. A. Watkins (1993) notes, there seem to be some lexical constraints as well interacting with word order types. Although the list of verbs that 'resist inflection' is not completely accurate, in some cases his generalisation does hold. Even in the large sample of Middle Welsh texts under investigation, there are for example no cases of the verbs *gwneuthur* 'to do' or *bod* 'to be' in the periphrastic verb-second construction. Sentences like **Gwneuthur a wnaeth*, literally 'doing he did' or **Bod a wnaeth* 'being he did' never occur. This is quite likely a simple semantic restriction.

Other verbs like *darfod* 'to happen' (a combination of a preverb + *bod* 'to be') occur more often in sentence-initial position in texts like *Breudwyt Rhonabwy* and *Peredur* (cf. Poppe (1993:96)). It should be noted though, that 1 of the total amount of 2 examples in *Breudwyt Rhonabwy* is the imperative *derffit*, which as an imperative would occur sentence-initially anyway. The other examples (also in *Peredur*) are actually preceded by a preverbal particle like *neur* most of the time, as shown in (54b). This tendency to appear in sentence-initial position (or, not in verb-second position) probably has to do with the meaning of the verb again. Especially in historical narratives, many sentences start with 'It happened that...'. Since sentence-initial forms of *bod* 'to be' were increasingly found in the late Middle Welsh period, it is not surprising a similar sentence-initial position was preferred for compounds with *bod*. Even in the 1588 Bible translation, however, this verb could also still appear in verb-second position, as shown in (54a):

(54) a. *AC fe a ddarfu wedi i r Iesu orphen y geiriau hyn oll.*
and it PRT happen.PAST.3S afterward to the Jesus finish.INF the words that all
'And it happened afterwards that Jesus ended all these words.' (b1588 - Mat. 26.1)

b. *Neur deryw y r maccwy llad llawer o th lu.*
PRT happen.PRES.3S to the lad kill.INF many of 2S host
'The lad happens to kill many of your men.' (Peredur 38.19)

5.2.10 Interim summary

In the above sections, various grammatical features were discussed in relation to the different types of word order patterns. There seem to be some absolute restrictions, in particular related to clause type (e.g. imperatives always occur in sentence-initial position). But most of the observations exhibit strong or weak tendencies, e.g. Type IIIc VN*a*DO is almost exclusively found in the preterite tense. This does not mean, however, that the reverse is automatically the case. It also does not tell us why this is the case. In types of phrase, we can also find some patterns in the distribution over the different word order types. Object-initial pronouns are clearly impossible in Middle Welsh grammar, but the reason why pronominal subjects exhibit a different distribution than nominal subjects cannot be explained by this grammatical difference alone. In the next section, I therefore explore various information-structural factors and their relation to word order patterns in Middle Welsh.

5.3 Usage-based factors

Languages are mainly studied by observing the data in use. In order to accurately compare different texts, speakers and/or stages of a language, it is of crucial importance to be aware of the *type* of data we are dealing with. Even a single speaker can use one and the same language in different ways, for example in different contexts with different interlocutors. If there are differences in genre, register or style within a language, it is strictly speaking impossible to fairly compare different types of word order in each of the texts under investigation. If we had an unlimited amount of data, it would be easy to just select texts of the exact same genre, register, etc. But the available data for Middle Welsh are limited. In this section, I briefly touch upon some issues related to how language is used and how this complicates the research question.

5.3.1 Spoken vs. written language

In any historical linguistic study, the difference between spoken and written language should be emphasised. Both spoken and written language is subject to change over time, but not necessarily in the same way or at the same rate. It may take years and years before a specific linguistic construction that is already widely used in spoken language, enters the written form of the language as well. Formality and standardisation of written language play a big role in this respect.

When the data are limited to written sources, like in the current study of Medieval Welsh, we have to take various extra-linguistic factors into account as well (see section 5.4). Some written data may be closer to the spoken language at the time than others, some genres might even render spoken language almost verbatim, e.g. witness or defensive statements in certain documented court cases. But, if anything, the conclusions drawn in this study say something about the written form of Middle Welsh as we find it in available manuscripts today. This certainly does not represent the Middle Welsh language as a whole. But even this written form was part of the language and an accurate description of this particular part of it thus helps us to understand this stage of the language better.

5.3.2 Direct vs. indirect speech

Written narratives often contain both direct and indirect speech. Direct speech in turn can be used for both monologues and dialogues or other forms of conversation. Monologues can be very similar to any other narrative sequence, but there are also examples of monologues centered around the experiences of one particular speaker, starting every sentence with *mi* 'I':

(55) a. *Mi a uum gynt y Ghaer Se ac Asse (...)*
I PRT be.PAST.1S before in Caer Se and Asse
'In the past, I have been in Caer Se and Asse'

b. *Mi a uum gynt yn yr India Uawr a r India Uechan*
I PRT be.PAST.1S before in the India Big and the India Small
'In the past, I have been in Greater and Lesser India' (with more examples of *mi a uum*... CO 117-118)

Because of their interactive nature, dialogues frequently employ very specific word order types to render questions, answers or commands (see section 5.2). In this type of direct speech, there are hardly any examples of the word order types that are typically used in continuous narratives (see section 5.5 above): adjunct-initial or verbal noun-initial orders. In many of these examples, the sentence-initial constituent is focussed, as shown in (56).

(56) a. *Mynet a wnaf i a th wyneb di a dygaf i genhyf.*
go.INF PRT do.1S I and 2S honour you PRT take.1S I with.1S
'(Everyone has received his boon, and I yet lack mine,) I will go and take your honour with me.' (CO 328-329)

b. *ac attat titheu y mae y neges ef.*
and to.2S you PRT be.3S 3MS message he
'And for you was his message.' (BR 12.20)

Most sentences with direct speech that are not questions, answers or commands exhibit argument- and in particular subject-initial word order (Type IVa). Because of the nature of the dialogue, the subjects are usually personal pronouns (cf. T. A. Watkins (1977:390-391)).

(57) a. *A thitheu, heb ef, mi a th gymeraf yn wreic im.*
and you.CONJ said he, I PRT 2S take.1S PRED wife to.1S
'And you, he said, I'll take as my wife.' (PKM 74.16-17)

b. *Mi a e dywedaf itt yr ystyr.*
I PRT 3MS tell.1S to.2S the meaning
'I will tell you the meaning of it.' (BR 4.29)

5.3.3 Poetry vs. Prose

Syntactic analyses tend to keep apart prose and poetry, because the word order in poetry can be subject to specific patterns like rhyme and metre that are not found in prose. For Middle Welsh, this is particularly relevant when looking at word order. According to Willis (1998), the frequency of absolute verb-initial sentences "is close to nil in Middle Welsh texts" (Willis, 1998:102). The texts he refers to are only prose texts; (Early) Middle Welsh poetry is not taken into account in most Welsh word order studies, because the syntax is indeed very different.

Verb-initial orders are often found in poetry from the Early Middle Welsh period onwards, but these are not taken into account in the present study. The excerpts of the Bible translation chosen for the present corpus are therefore also only narrative

prose (Joseph's and David's stories (Genesis 37-42 and 1 Samuel 16-19), the gospel of Matthew and Paul's letter to the Corinthians).

In his 2013 study, Currie also includes various excerpts of the Bible translations. He concludes that absolute verb-initial order was found in Middle Welsh after all, because he finds frequencies up to 41% in his corpus. If we look more closely at his data, however, these high frequencies are found in the translations of the Psalms (Salesbury 41% and Morgan 24.8%) and the Book of Isaiah (24.8%), whereas the other Biblical excerpts (the gospel of Mark and the book of Esther) do not even reach 10%. This is not surprising, considering the fact that the Psalms and the Prophets were written in a very different form of Hebrew that certainly did not look like the regular narrative prose found in the rest of the Bible.

Without an in-depth analysis of the original Hebrew of the Psalms and Prophets compared to the narrative prose, it is thus impossible to draw any conclusions on the resulting word order frequencies in the Welsh translations. According to Currie (2013), the high frequency of verb-initial orders in the psalms might be due to their highly elevated style. Style is thus another factor we need to control for when comparing word order types.

5.3.4 Genre, register and style

Style can vary between different genres and registers, but also within one and the same text itself. Most texts in the Middle Welsh corpus are narrative prose, but one of the native tales of the *Mabinogion*, *Llud and Llefelis*, is also found in a manuscript of a completely different genre: chronicle literature.

Another example of chronicle literature in the corpus is *Buched Dewi* 'The Life of Dewi'. Table 5.16 compares the frequencies between an excerpt of the Laws, two narratives tales of the Mabinogi and two chronicles: the chronicle version of *Llud* and the Life of David.

	Laws	Math	Llud (nar.)	Llud (chr.)	Dewi
Type I Verb-initial	4	14		1	2
Type II Periphrastic		3			1
Type III Adj y VS	52	45	15	8	58
Type IV SaVO	96	98	24	31	65
Type IV OaVS	31	18	1	2	11
Type IV VN a DO	2	25	5	4	7
Type VIII Sef		22	2	1	21
Total	185	225	47	47	165

Table 5.15: Word order types of transitive sentences in different genres

	Laws	Math	Llud (nar.)	Llud (chr.)	Dewi
Type I Verb-initial	2.16%	6.22%		2.13%	1.21%
Type II Periphrastic		1.33%			0.61%
Type III Adj y VS	28.11%	20.00%	31.91%	17.02%	35.15%
Type IV SaVO	51.89%	43.56%	51.06%	65.96%	39.39%
Type IV OaVS	16.76%	8.00%	2.13%	4.26%	6.67%
Type IV VN a DO	1.08%	11.11%	10.64%	8.51%	4.24%
Type VIII Sef		9.78%	4.26%	2.13%	12.73%

Table 5.16: Percentage of word order type of transitive sentences in different genres

As it turns out, there is not a big difference between the two genres. Chronicles like *Buched Dewi* tend to employ the adjunct-initial word order (Type III) more often, because they often relate sequential events that are linked to a specific time or location, but this is not observed in the chronicle version of *Llud*. Verb-initial orders are hardly ever found overall. Subject-initial sentences are most frequently found in all the genres, though significantly less in the chronicle of Dewi.

For the Middle Welsh period, it is very difficult to take into account various registers of the language since the extant corpus is very limited. Stylistic differences can be found when we compare native tales to translations and retellings of stories from Latin and/or French origin. Differences in agreement patterns were the subject of investigation in Welsh translated literature in particular, because agreement with plural noun phrases in Welsh was claimed to have come from Latin (cf. D. S. Evans (1971)). Plein and Poppe (2014) conclude, however, after closely comparing the Welsh *Historia Gruffudd vab Kenan* to its Latin original, that this is not necessarily the case: "We are currently unable to identify potential triggers in the Latin text for the realization in the Welsh text of expected default third-singular and unexpected verbal agreement respectively." (Plein & Poppe, 2014:155).

	Culhwch	Branwen	Peredur	Macsen	B1588
Type I Verb-initial	64	5	16		49
Type II Periphrastic		1	1	1	21
Type III Adj y VS	67	44	81	51	136
Type IV OaVS	36	104	68	28	17
Type IV SaVO	76	42	245	26	476
Type IV VN a DO	70	16	79	5	
Type V Focus	3	1			
Type VIII Sef	10	12	12	3	
Total	326	225	502	114	699

Table 5.17: Word order type of transitive sentences in different genres

	Culhwch	Branwen	Peredur	Macsen	B1588
Type I Verb-initial	19.63%	3.70%	3.19%		7.01%
Type II Periphrastic		0.74%	0.20%	0.88%	3.00%
Type III Adj y VS	20.55%	32.59%	16.14%	44.74%	19.46%
Type IV OaVS	11.04%	10.37%	13.55%	24.56%	2.43%
Type IV SaVO	23.31%	31.11%	48.80%	22.81%	68.10%
Type IV VN a DO	21.47%	11.85%	15.74%	4.39%	
Type V Focus	0.92%	0.74%			
Type VIII Sef	3.07%	8.89%	2.39%	2.63%	

Table 5.18: Percentage of word order type of transitive sentences in different genres

When we put another set of texts with a different genre or background together, the most striking frequencies are found in *Macsen*. It is the only text in which the object-initial type constitutes almost a quarter of all sentences and most other sentences are adjunct-initial. If we look closer at the object-initial examples, however, we find they are primarily used with one particular verb *gwelet* 'to see', which is no doubt due to the nature of the text. It is a narration of what someone saw in a dream at a particular time and place. The observed objects are usually new in the narrative and could thus occur in initial position (see section 5.5 above).

Compared to the other (later) Middle Welsh texts, *Culhwch* is different in that it employs a great number of verbal noun constructions (21.47%) and it has more verb-initial sentences than any other text in the corpus. The latter are mainly verbs of saying, however, and there are some fixed expressions amongst those as well (see section 5.2). In later Middle Welsh texts, like the Arthurian Romance of *Peredur*, we see subject-initial orders are gaining more and more majority. In the 1588 Bible translation, this subject-initial word order type represents the overwhelming majority of sentences. In this case, we are dealing with a translated text as well. The choice of word order of the translator, however, is not at all influenced by the verb-initial word order that was dominant in the Hebrew original.

Overall, it is important to be aware of stylistic differences, within authors/texts, but also those that are due to different genres or registers. The number of texts for various genres and registers in Middle Welsh is, however, rather limited, so it is difficult to draw any definite conclusions. There are furthermore various extra-linguistic factors that play a role here. I turn to these in the next section.

5.4 Extra-linguistic factors

Working with historical linguistic data is not just challenging because of the limited amount of data. In this section, I discuss some further issues that should be taken into account when we interpret the results of any diachronic investigation of Welsh.

5.4.1 Philology: the scribes and their manuscripts

One particular problem historical linguists are faced with is the lack of necessary philological background for their data. Even in close collaboration with philologists, it is not always possible to establish for example, the exact date of a certain text. A related problem is the question of the text itself: to what extent does the version we have represent the 'original'? If there are more manuscripts with the same text: do we choose one or the other or do we work with the diplomatically edited version?

Even when we can make these decisions and justify them, we are still dealing with the problem of the origin of the text. Even if we know when and where it was written down in a certain manuscript, this hardly ever gives us any information on when and where the text was originally composed. If there are several centuries between the date of composition and the written down version we have now, it severely complicates any accurate dating of linguistic phenomena. When scribes and copyists were set to work, what exactly did they do? Did they blindly copy any 'mistakes' they found or did they 'update' the language in such a way they thought it would be easier for their contemporary audience to understand it.

According to T. Charles-Edwards (2001), there were 'fluid' and 'fixed' textual traditions. In his eyes, the Four Branches of the Mabinogion were more or less fixed, i.e. the extant versions found in different manuscripts do not exhibit a great amount of variety when closely compared. The Romances like *Peredur*, however are part of a fluid tradition that exhibit a degree of variation that is not due to normal copying errors "but introduced by the scribe for some other editorial purpose" (Vitt, 2011). According to Russell (2003:65-66), therefore, "Fluid texts are the bane of the classical textual critic". For historical linguists, of course, the problem of a 'fluid' text is even worse, because not only should we be able to account for linguistic phenomena in one version of the text we find, our description of the language ideally encompasses all other possible versions as well. We crucially do not know which of the versions was correct or if both versions were for different people or in different periods of time. Copyists made mistakes, but explaining away all unexpected variation as 'scribal errors' is too easy.

The texts analysed in the present study represent one manuscript version. Other manuscripts have, however, been systematically compared to these versions for any differences in word order. The most common differences between manuscripts are found in verbal noun constructions. The auxiliary 'to do' is either omitted in one of the versions or a different form is used (*gwnaeth* vs. *goruc* 'did').

(58) a. *Dyuot y porthawr ac agori y porth*
come.INF the porter and open.INF the gate
'The porter came and opened the gate.' (Culhwch *White Book* 786)

b. *Dyuot a oruc y porthawr ac agori y porth*
come.INF PRT do.PAST.3S the porter and open.INF the gate
'The porter came and opened the gate.' (Culhwch *Red Book* 786)

These types of variation between manuscripts do not have any significant effect on the hypotheses concerning word order distribution. If anything, it indicates that

the verbal noun construction with the auxiliary was an innovation used by later scribes. Main clauses with bare verbal nouns in initial position were already rare in the earliest Middle Welsh tale in the corpus, *Culhwch ac Olwen*. It is not surprising that the Red Book scribe, known to 'modernise' his text while copying (cf. Rodway (2004)), added the auxiliary of the verb 'to do' resulting in the commonly-used verbal noun construction with verb-second word order. In later Middle Welsh, however, this construction became used less, taken over by adjunct- and subject-initial word orders. The Middle Welsh Bible from 1588 only has very few examples of sentence-initial verbal nouns.

There are, however, also other types of variation found in different manuscript versions. For example, between subject- and object-initial word order as shown in (59):

***Owein* l. 652**
So they returned, and Owain pressed forward until he met the Earl. And Owain drew him completely out of his saddle, and turned his horse's head towards the Castle, and, though it was with difficulty, he brought the Earl to the portal, where the pages awaited him. And in they came.

(59) a. *a 'r iarll a rodes Owein yn anrec y 'r iarlles*
and the earl PCL gave.3S Owein PRED gift to the countess
'And Owein presented the earl as a gift to the countess.' (OaVS - White Book)
b. *ac Owein a rodes y iarll yn anrec y 'r iarless*
and Owein PCL gave.3S the earl PRED gift to the countess
'And Owein presented the earl as a gift to the countess.'(SaVO - Red Book)

(60) a. *A 'e dyuot hitheu*
and 3FS come.INF her
'and she came' (CO 487: VN + agent - White Book)
b. *Dyuot a oruc hitheu*
come.INF PCL did.3S she
'and she came' (CO 487: VN + do - Red Book)

In this example, the older White Book manuscript has object-initial word order, where the later Red Book prefers the subject in initial position. Since there is a clear focus on the object in this context, the object-initial order is not unexpected (see section 5.5). The Red Book scribe is generally known to 'update' and 'correct' his work. If this was indeed what he did, this could be an example of the object-initial order becoming less prominent towards the end of the Middle Welsh period. Object-initial orders were hardly ever used in the 1588 Bible translation and even if they were, they were always contrastively focussed. In a more 'fluid' text, like *Owain* or any of the other Romances, the Red Book scribe might have felt free to 'update' the syntax of this particular sentence to the subject-initial order that sounded far more familiar in his ears. It remains difficult though, to speculate on the basis of one single example. It is striking, however, that this is one of the very few examples in

the present corpus with a variation in word order in different manuscripts. Another one found in Peredur exhibits the same difference:

(61) a. *A hynny a wnaeth y makwyf. Yr orssed a gyrchyssant*
and that PRT do.PAST.3S the lad the mound PRT make.for.PAST.3P
'And the young lad did that. They made for the mound' (White Book)
b. *Y gwas a wnaeth hynny. Dyuot yr orssed a orugant*
the lad PRT do.PAST.3S that come.INF the mound PRT do.PAST.3P
'The groom did that. They came to the mound' (Red Book)

Again, there seems to be a preference for subject-initial order in the later Red Book version. Although this type of knowledge about the scribe and manuscript can help to establish the relative chronology of linguistic phenomena, it remains difficult to get a detailed diachronic description because of the lacunae in our data and metadata. Even if we can establish where a particular text was written down, this does not always tell us more about the origin of the text and how much the language was modified before it was put into writing. For this reason, it is impossible to be more precise about the exact dates than 'early' and 'late' Middle Welsh. With the present corpus, it is furthermore impossible to tie the results to any particular region in Wales. We know that there were different dialects of Welsh in the medieval period, but a lot more data is needed to say anything about preferences or patterns in different types of word order in different regions of Wales.

5.5 Information-structural factors

In this section I discuss how information structure relates to Middle Welsh word order. First I investigate the focus domain of Middle Welsh sentences in the corpus. Two of the three core notions of information structure, topic-comment and focus-background, are discussed in this section. The third notion of information structure, givenness, sheds light on the distribution of different types of argument phrases in Middle Welsh. The Principle of Natural Information flow, i.e. old information is followed by new information, is discussed in this context as well. The final section concerns text cohesion: how is a particular sentence linked to the preceding context. Framesetters and points of departure are crucial in establishing whether there is textual continuity or a deliberate break or change of scene for example signalled by a shift of topics.

5.5.1 Focus Articulation

As described in detail in Chapter 3, there are three focus articulations or domains. The most common in narrative texts is the domain that focusses the verb and the rest of the predicate, also known as 'topic-comment' domain or 'predicate focus'. If one particular constituent is in focus, i.e. if there is a relevant alternative, then the focus domain is 'constituent focus'. If all the information in the sentence, i.e. the subject

as well as the predicate, convey new information and no constituent is focussed in particular, the focus domain of the sentence is THETIC or PRESENTATIONAL. THETIC or PRESENTATIONAL FOCUS is almost exclusively found in sentences with copular or existential forms of the verb 'to be'.

THETIC and PRESENTATIONAL Focus

Opening statements of narratives often present new protagonists in the context of where they live or rule. These types of sentences exhibit presentational focus, because all the information is new and the leading character is introduced to the storyline as in (62):

(62) a. *Pwyll Pendeuic Dyuet a oed yn arglwyd ar seith cantref Dyuet.*
Pwyll Prince Dyfed PRT be.PAST.3S PRED lord on seven cantref Dyuet
'Pwyll Prince of Dyfed was lord of the seven cantrefs of Dyfed.' (PKM 1)
b. *Bendigeiduran uab Llyr a oed urenhin coronawc ar yr ynys hon,*
Bendigeidfran son Llyr PRT be.PAST.3S king crowned on the island this
ac ardyrchawc o goron Lundein.
and invested with crown London
'Bendigeidfran son of Llyr was crowned king of this island and invested with the crown of London.' (PKM 29.1)

New characters can also be introduced sentence-finally or sentence-initially, as in example (63):

(63) a. *mae yna carw*
be.PRES.3S there stag
'there was a stag' (PKM 75.12-13)
b. *Trychanhwr teulu yssyd idi*
300.man host be.PRES.3S to.3FS
'She has a host of 300 men.' (Peredur 45.22)

A final way to introduce new characters to the discourse is by using the contracted form *dyma* or *llyma, llyna* 'here, there is' in non-verbal word order Type IX:

(64) a. *Llyna Dillus Uarruawc.*
there.is Dillus Barfawg
'There is Dillus Barfawg.' (CO 1013)
b. *Llyma pump morwyn yn dyfot o ystafell y r neuad.*
here.is five maiden PROGR come.INF from room to the hall
'There came five maidens from the room to the hall.' (Peredur 23.14)

PREDICATE Focus

Topic-comment sentences can be found in Middle Welsh in various word order types. The topic, in this case, is the topic of the sentence. This is not necessarily the same as the topic of the entire discourse. Topics in Middle Welsh are frequently found in

sentence-initial position, resulting in the verb-second order with the verb following a topical adjunct, subject or object. Frame-setting topics are usually adjuncts in sentence-initial position, they set the scene and/or delimit the space or time in which the event described in the following comment takes place. Aboutness topics are not further defined here than that which the sentence or discourse is about. They frequently show up as subjects, but can also be found as (indirect) objects of the sentence, as shown in (65):

(65) a. *Kyuodi a oruc* ***yr heusawr*** *y uynyd.*
rise.INF PRT do.PAST.3S the giant to up
'**The giant** got up.'
b. *Mal y kyuyt, rodi modrwy eur a oruc Culhwch itaw.*
as PRT rise.PRES.3S give.INF ring gold PRT do.PAST.3S Culhwch to.3MS
'As **he** got up, Culhwch gave **him** a golden ring.'
c. *Keissaw gwiscaw y uodrwy ohonaw.*
try.INF put.on.INF the ring of.3MS
'**He** tried to put on the ring.' (CO 440-442)

In (65), the discourse is about the giant: he gets up and is given a ring, which he then tries to put on. This topic is first the subject with a periphrastic VN + do construction. In the sentence directly following, it is the subject of the subordinate clause and the indirect object in the inflected preposition *itaw* 'to him' of the main clause. Finally, the giant is the agent of the main verb again, but this time, there is no conjugated verb and the agent is rendered by the inflected preposition *ohonaw*.

Sentences with PREDICATE FOCUS usually have topics that contain old information. The new information is then rendered by the following comment. In some sentences, the referential status of the topicalised constituent is not completely old, but linked to the preceding context in some other way, e.g. by an identity anchor as in (66):

(66) *Os ynteu a 'm llad ynheu, vy angklot a gerda ar draws*
if he PRT 1S kill.PRES.3S me my infamy PRT walk.PRES.3S on surface
y byt yn dragywyd.
the earth PRED always
'If it is him who kills me, my infamy will spread over the world forever.' (CO 402-404)

Contrastive topics are also found in Middle Welsh (see Chapter 4). The few examples we find have subject-initial order and belong to the PREDICATE FOCUS (topic-comment) articulation.

CONSTITUENT FOCUS

CONSTITUENT FOCUS can occur with designated focus particles like *hefyd* 'also, too' or *hyd yn oed* 'even', but there are also other cases of constituent focus that are more difficult to detect. In these cases the constituent in focus has to have a possible alternative, which is not mentioned. The constituent in focus thus reflects one of

all relevant alternatives. Sometimes the alternatives are overtly contrasted, as in (67) and a reduplicated pronoun can be used as well in this case.

(67) a. *hyt nas gwelei neb vynt ac vyntvy a*
so.that NEG.3S see.IMPF-SBJ.3S no.one them but they.REDUP PRT
welynt pawb
see.IMPF-SBJ.3P all
'so that no one could see them, but THEY could see everyone' (CO 410)
b. *Vn o r ffyrd hyn a a y m llys i*
one of the ways those PRT go.PRES.3S to 1S court my
'One of those ways goes to my court.' (Peredur 48.19)

Although constituent focus is often found with special constructions like the *sef*-construction of Type VIII, sentences with constituent focus can also exhibit other types of word order. The focussed constituent is most frequently found in sentence-initial position as in (68a) and (68b), but this is not necessarily the case, as shown in (68c).

(68) a. *Yr eil fford a a y r dinas yssyd yna yn agos.*
the second way PRT go.PRES.3S to the town be.REL.3S there PRED close
'The second road goes to the town that is close to there.' (Peredur 48.30)
b. *ti a gereis*
you PRT love.PAST.1S
'I loved YOU.' (CO 501)
c. *ac a i rhoddes i Ddafydd. a i wiscoedd. ie hyd yn oed*
and PRT 3FS give.PAST.3S to David with 3MS clothes yes even
ei gleddyf
3MS sword
'and he gave it to David with his clothes, yes even his sword' (b1588 - 1 Sam. 18.14)

Constituent focus is also found in answers to questions. The focussed constituent always appears in sentence-initial position in that case.

In what manner didst thou receive them?

(69) *Eu rannu ym pob lle yn y kyuoeth.*
3P divide.INF in every place in the kingdom
'I dispersed them through every part of my dominions' (Branwen 64)

And what are you doing, Lord?

(70) *Crogi lleidyr a geueis yn lledratta arnaf.*
hang.INF thief PRT get.PAST.1S PROGR steal.INF on.1S
'Hanging a thief I caught stealing from me.' (PKM 62.2-3)

What are you asking?

(71) *Vyg kymryt yn wr itt.*
1S take.INF PRED husband to.2S
'To take me as your husband.' (Peredur 49.28)

5.5.2 Givenness

The referential status of constituents, in particular subjects and objects, is one of the most-studied aspects of information structure in various languages. For Middle Welsh, Erich Poppe, among others, has studied the relation between information status and agreement. He concludes that a preference for concord or non-concord is not related to information status (Poppe, 2009:257). Earlier I argued that a simple distinction between 'old' and 'new' information cannot always capture fine-grained differences in pragmatic usage. I therefore annotated all subjects and objects in the database according to the Pentaset, that captures the difference between linked and unlinked information (to the previous context or to something known by the hearer). Some constituents convey information that is technically new, but can be inferred from the previous context in some way, e.g. a set relation. The results presented in this section are based on this more fine-grained annotation.

Principle of Natural Information Flow

According to the Principle of Natural Information Flow, old information precedes new information in unmarked contexts. In verb-second sentences with either the subject or the object in initial position, the null-hypothesis would thus be that the information status of the initial arguments is old (or older at least) than that of the rest of the sentence. If this is not the case, i.e. if the initial argument conveys new(er) information, then the sentence does not comply with this Principle of Natural Information Flow and is thus somehow 'marked'.

For Middle Welsh this means that we could check this from the point of view of referential status of the core arguments. If both subject- and object-initial word orders are unmarked, the referential status of these initial subjects and objects should be older than the information in the rest of the sentence. Table 5.19 below, however, shows that this is not always the case with sentence-initial DPs (pronouns are not taken into account here because sentence-initial object pronouns are grammatically impossible in Middle Welsh).

	Unmarked: Old - New	Marked: New - Old
Type IVa SaVO	152 (99.35%)	13 (13.68%)
Type IVb OaVS	1 (0.65%)	82 (86.32%)
Total	153	95

Table 5.19: Information Flow in Subject- and Object-initial sentences

Subject-initial sentences with marked information flow, i.e. with new subjects preceding old(er) direct objects do not occur very often. The 13 instances in the database contain subjects that can all be interpreted as (contrastively) focussed: the subjects represent one person/item of a set of relevant alternatives, as shown by the examples in (72). In example (72a) for example, there are many things/people that were threatening the land in those times (foreign invaders, plagues, etc.), so the famine is chosen as the significant item from this set of relevant alternative things that could have destroyed the land.

(72) a. *a newyn a ddifetha y wlad.*
and famine PRT destroy.3S the country
'and a famine shall destroy the country' (b1588 - Gen. 41.30)

b. *Y gwr yssyd tat inni bieu y llys hon.*
the man be.REL.3S father to.1P own.3S the court this
'The man who is our father owns this court.' (Peredur 43.9)

c. *Yna Michol merch Saul a garodd Ddafydd.*
then Michol daughter Saul PRT love.PAST.3S David
'Then Michol daughter of Saul loved David.' (b1588 - 1 Sam. 18.20)

By far the most sentences with marked information flow are object-initial. The one example with old information preceding new information in sentences with object-initial order clearly contains a contrastive focus of the sentence-initial constituent:

(73) *a r hanner arall a dal y neb a losco ac ef.*
and the half other PRT pay.3S the one PRT burn.SUBJ.3S with him
'and the one who would burn (it) with him pays the other half' (Laws 85)

Sentences with object-initial word order in Middle Welsh are thus marked, if only from the perspective of the Principle of Natural Information Flow.

(74) **Generalisation**
Object-initial sentences in Middle Welsh are always marked, unless the object is a familiar topic.

Subject- vs. Object-initial sentences

The question is what this generalisation tells us about Middle Welsh word order and the information-structural notion of givenness. What is the distribution of old(er) and new(er) subjects and objects in subject- and object-initial sentences? Table 5.20 gives an overview of the referential status of the arguments and their respective word order types (ID = Identical to what is already in the hearer's short-term memory because it was mentioned in the immediately preceding context).

	Sbj ID - Obj ID	Sbj ID - Obj New
Type IVa SaVO	38 (76%)	107 (62.57%)
Type IVb OaVS	12 (24%)	64 (37.43%)
Total	50	171

Table 5.20: Referential status of DPs in argument-initial word order types

	Sbj ID - Obj ID	Sbj ID - Obj New
Type IVa SaVO	27 (90%)	106 (63.10%)
Type IVb OaVS	3 (10%)	62 (36.90%)
Total	30	168

Table 5.21: Referential status of DPs (excl. demonstratives) in argument-initial word order types

There is a significant difference between subject- and object-initial orders with identical (old) objects and new objects when demonstrative pronouns are not taken into account ($\chi^2 = 7.1803$, df $= 1$, p-value $= 0.007378$, Fisher's p=0.002957). The strong generalisation that all sentence-initial objects have to contain new information does not seem to hold, because there are 12 examples of objects with referential status Identity (= Old). If we look closer at those examples, however, we find that 9 of those objects are demonstrative pronouns continuing the topic of the immediately preceding sentence, as shown in (75).

(75) a. *A hynny a oruc y gwyr oll.*
and that PRT do.PAST.3S the men all
'And all the men did that.' (Gereint 386)

b. *A hynny a wnaeth y makwyf.*
and that PRT do.PAST.3S the lad
'And the lad did that.' (PKM 10.18-19)

c. *Hynny a dywot y guas idi hitheu.*
that PRT say.PAST.3S the lad to.3FS her
'The lad said that to her.' (PKM 84.20)

There are also examples of topic continuity that repeat the topic phrase entirely, instead of referring to it with a demonstrative:

(76) a. *A r pymp arueu a rodes yn y pymp kyfrwy*
and the five armours PRT give.PAST.3S in the five saddle
'and he place the five suits of armour on the five saddles.' (Gereint 838)

b. *A nawd a rodes Gereint itaw.*
and mercy PRT give.PAST.3S Geraint to.3MS
'And Geraint gave him mercy.' (Gereint 1051)

Most other examples with initial objects that are not new are (contrastively) focussed, often by overt focus particles like *hefyd* 'also, too' or *hagen* 'however'. This is also often seen with pronominal subjects.

(77) a. *Yr vn peth hefyd a edliwiodd y lladron.*
the same thing also PRT taunt.PAST.3S the thieves
'The thieves taunted the same thing.' (b1588 - Mat. 27.44)

b. *Y mab hagen a gymeraf i.*
the boy however PRT take.1S I
'The boy, however, I will take.' (PKM 75.20)

c. *ei ferch hefyd a rydd efe iddo ef.*
3MS daughter too PRT free.3S he to.3MS him
'His daughter, too, he released for him.' (b1588 - 1 Sam. 17.25)

d. *E gedymdeithas oreu a allwyf i.*
the friendship best PRT cause.SUBJ.1S I
'I would show the best friendship.' (PKM 50.3-4)

Argument-initial sentences with nominal arguments in Middle Welsh are mostly subject-initial (165 out of 248 examples). Object-initial orders are also possible, but they are always marked somehow. They are either (contrastively) focussed or their referential status is New (new information focus).

Other examples of object-initial orders all exhibit direct topic continuity, either by repeating the topic noun phrase mentioned in the previous sentence or by referring back to it with a demonstrative pronoun. To conclude, givenness or the referential status of the core arguments, in particular the direct objects does influence the type of word order in Middle Welsh.

Givenness and other word order types

In the previous section I showed that object-initial sentences only appear under certain conditions: the object has to be focussed (either because it is new information or contrastively) or it continues the immediately preceding topic. What about the notion of givenness in relation to other verb-second structures in Middle Welsh?

Table 5.22 shows that the referential status of the core argument of impersonal verbs (the patient) is usually 'Identity' (= old). There are more examples of impersonal verbs with adjunct-initial word order (Type III), but there are more sentences with adjunct-initial order overall (see Chapter 4). Whenever the patient contains new information, however, it is far more often placed in sentence-initial position. This difference is significant (chi-square = 18.5707 df = 1 $p < 0.0001$).

	Patient = ID	Patient = New
Type III Adjunct-initial	126 (63.96%)	32 (36.36%)
Type IVab Argument-initial	71 (36.04%)	56 (63.64%)
Total	197	88

Table 5.22: Referential status of patients of impersonal verbs

In all other verb-second word order types, direct objects also more often convey new information than subjects, but there is no significant relation between referential status of the object and adjunct-, argument- or verbal noun-initial orders. In conclusion, within argument-initial orders there is a strong preference to place the subject in first position. Objects and patient phrases of impersonals can also be found in initial position, but only if their referential status is New. Other word order types do not contain enough tokens to compare.

Late subjects and objects

Givenness finally seems to interact with word order in the case of delayed subjects and objects. These postposed constituents are only possible if they convey new information, as shown in the following examples:

(78) a. *kanys ny wisgawd arueu eiryoet uarchawc urdawl well noc ef*
because NEG wear.PAST.3S arms ever knight honourable better than him
'since a better knight than he never bore arms.' (YSG 3972-3)

b. *ac y lladwyt yna Twrch Llawin.*
and PRT kill.PAST.IMPERS there Twrch Llawin
'And Twrch Llawin was killed there.' (CO 1147)

5.5.3 Text Cohesion

In the previous section, one particular form of textual cohesion was already mentioned: topic continuity. So far, I have mainly looked at the information structure at sentence-level. In this section, I focus on information-structural features that play a role on the level of the paragraph and/or bigger sections of the discourse. There are various ways to link a sentence to the preceding context, but it is also possible to change the topic and/or scene. Points of departure or framesetters are frequently-used devices to render textual continuity or change. I discuss the most important examples of these in Middle Welsh in the section below.

Points of departure

Points of departure come in different shapes and forms. In Middle Welsh, various adverbial expressions in sentence-initial position determine the point of departure or the frame in which the predication of the rest of the sentence holds. These

adverbials are mostly temporal, spatial (i.e. referring to a specific location) or referential. Examples of these in Middle Welsh are:

(79) a. *O hynny allan y gelwit Goreu mab Custenhin.*
from that onwards PRT call.IMPERS Goreu son Custennin
'And from then on he was called Goreu son of Custennin.' (CO 811)
b. *Ac y r dref y doyth y uorwyn*
and to the town PRT come.PAST.3S the maiden
'And the maiden came to the town.' (Gereint 213)
c. *Y Beli Uawr vab Manogan y bu tri meib.*
to Beli Mawr son Manogan PRT be.PAST.3S three son
'And Beli Mawr son of Manogan had three sons.' (Llud WB 1)

The adverbial is almost exclusively followed by the particle *y* + the inflected verb, resulting in word order Type III (adjunct verb-second). There are some examples of points of departure followed by other types of word order, but these are the exception rather than the rule:

(80) a. *a chynn kyscu genthi dyuot Gwynn uab Nud*
and before sleep.INF with.3FS come.INF Gwynn son Nud
'And before sleeping with her, Gwynn son of Nud came' (CO 989-990)
b. *Ac ar hynny eu taraw a r hutlath*
and on that 3P hit.INF with the magic wand
'And after that he struck them with the magic wand.' (PKM 75.19)
c. *A gwedy eu heisted gofyn a orugant y r wrach ...*
and after 3P sit.INF ask.INF PRT do.PAST.3P to the hag
'And after they sat down they asked the hag ...' (BR 2.27)

Some sentence-initial adverbials have a less specific semantic content. They are mainly used as connectives (cf. Poppe (1993:112)) indicating a sequential course of events. The most common examples of these in Middle Welsh are *yna, yno, gwedy hynny* 'then, there, after that'. These connectives can also be found in sentence-initial position, followed by any various word order types. In the Middle Welsh biblical narratives, these connectives occur more often than any other sentence-initial adverbial. They are either followed by a preverbal particle *y* and the inflected verb or by the subject:

(81) a. *Yna r eisteddasant i fwytta bwyd.*
then PRT sit.PAST.3P to eat.INF food
'Then they sat down to eat food.' (b1588 - Gen. 37.25)
b. *Yna efe a ddywedodd wrthynt*
then he PRT say.PAST.3S to.3P
'Then he said to them' (b1588 - Mat. 26.10)

Continuity

Narrative cohesion in Middle Welsh is most frequently established by the use of *a(c)* 'and' or any of the above-mentioned other connectives. These could be followed by any word order type. Sentences with initial verbal nouns, either followed directly by the agent or by the auxiliary 'to do' (Type VIc), signal topic continuity as shown in (82). Verbal nouns could also continue inflected verbs (Type VIc) as shown in (83), but can subsequently be continued by an inflected verb again.

(82) a. *Kychwynnu a oruc Arthur (...)*
start.INF PRT do.PAST.3S Arthur
'Arthur set out (...)'
b. *a mynet ym Prytwen y long*
and go.INF in Prydwen 3MS ship
'and went in Prydwen his ship'
c. *a dyuot y Ywerdon*
and come.INF to Ireland
'and came to Ireland' (CO 1040-1043)

(83) a. *Yna y kyuodes ynteu o r ennein*
then PRT rise.PAST.3S he from the bath
'Then he rose from the bath'
b. *a guiscaw y lawdyr amdanaw*
and wear.INF 3MS trousers on.3MS
'and put his trousers on'
c. *ac y dodes y neilltroet ar emyl y gerwyn*
and PRT put.PAST.3S 3MS one.foot on edge the tub
'and he put his one foot on the edge of the tub.' (PKM 87.27-88.2)

In the 1588 Bible translation, narrative continuity is more and more found with subject-initial word order as well. In these cases the topic is mentioned in the beginning of the sentences, but dropped in the following clauses, until there is a topic switch or some intervening noun phrase that could be the new topic.

(84) a. *Hefyd efe a freuddwydiodd etto freuddwyd arall*
also he PRT dream.PAST.3S still dream other
'He also dreamt another dream.'
b. *ac a i mynegodd i w frodyr*
and PRT 3FS tell.PAST.3S to 3MS brothers
'and told it to his brothers'
c. *ac a ddywedodd (...)*
and PRT say.PAST.3S
'and said: (...)' (b1588 - Gen. 37.9)

Topic continuity can also occur with points of departure or framesetters. In this case, the adjunct-initial word order type III is used. The continued topic, the third person plural pronoun 'they' is in this case merely rendered by the inflectional ending of the verb. This type of continuous prodrop is always found when topics

remain the subjects of the immediately following sentences.

(85) a. *A thrannoeth y kymeryssant eu hynt*
and next.day PRT take.PAST.3P 3P way
'And the next day they went on their way'
b. *dros Elenit y doethant*
through Elenit PRT come.PAST.3P
'(and) they came through Elenit'
c. *A r nos honno y buant y rwng Keri ac Arwystly (...)*
and the night that PRT be.PAST.3P to between Keri and Arwystly
'and that night they were between Keri and Arwystly'
d. *Ac odyna y kerdyssant racdunt*
and from.there PRT walk.PAST.3P against.3P
'and from there they walked on' (PKM 71.4-7)

A specific form of continuity of a certain theme from one sentence to the other is the use of lead sentences (cf. T. A. Watkins (1993:126)). As already pointed out above, it was possible in Middle Welsh narratives to continue the topic of the immediately preceding sentence by repeating it in sentence-initial position in the following sentence.

(86) a. *A nawd a rodes Gereint itaw.*
and mercy PRT give.PAST.3S Geraint to.3MS
'('Mercy, Lord!') And Geraint granted him mercy.' (Gereint 1051)

b. *Amser a doeth udunt e uynet e gyscu, ac y*
time PRT come.PAST.3S to go.INF to sleep.INF and to sleep.INF
gyscu yd aethant.
PRT go.PAST.3P
'Time came for them to go to sleep, and to sleep they went.'(PKM 4.26-27)

Change

'Change' in context take various shapes and forms. There can be a change of scene in the narrative, like a significant change of time or a change of location (see also Poppe (2014:99) for a discussion of the idiom *mynet ymdeith* 'go away' in the context of sudden changes in the narrative). Sentence-initial subordinate clauses and adverbials like the different kinds of points of departure and framesetters discussed above, can indicate discontinuity, in this case, a change of time:

(87) a. *Dyuot a oruc Arthur hyt yn Esgeir Oeruel (...)*
come.INF PRT do.PAST.3S Arthur until in Esgeir Oerfel
'Arthur came to Esgeir Oeruel (...)'
b. *Gellwng kwn arnaw o bop parth.*
release.INF dogs on.3MS from every side
'Dogs were let loose at him from all sides.'

c. *Y dyd hwnnw educher yd ymladawd y Gwydyl ac ef.*
the day that dawn PRT fight.PAST.3S the Irish with him
'The next day at dawn the Irish fought with him.' (CO 1122-1124)

Stories often display many changes in referential points of view as well: subjects and topics can vary from sentence to sentence. Argument-initial verb-second word orders in Middle Welsh (subject-initial Type IVa and object-initial Type IVb) were specifically used to introduce new topics into the discourse or to change the discourse-topic from that of the preceding context. There is a specific set of 'conjunctive' pronouns in Middle Welsh (see Table 5.23 repeated below) used in contrastive contexts like topic shift, but noun phrases could also be used.

(88) a. *Ac yna yd aeth Llwytawc hyt yn Ystrat Yw*
and thence PRT go.PAST.3S Llwydawg until in Ystrat Yw
'And from there Llwydawg went to Ystrat Yw'
b. *ac yno y kyuaruu gwyr Llydaw ac ef*
and there PRT meet.PAST.3S men Brittany with him
'and the men from Brittany met him there'
c. *ac yna y lladawd ef Hir Peissawc brenhin Llydaw (...)*
and there PRT kill.PAST.3S he Hir Peissawg king Brittany
'and there he slew Hirpeissawg the king of Brittany (...)'
d. *Ac yna y llas* ***ynteu.***
and there PRT kill.IMPERS he.CONJ
'and there was he himself slain.'
e. ***Twrch Trwyth*** *a aeth yna y rwng Tawy ac Euyas*
Twrch Trwyth PRT go.PAST.3S there to between Tawy and Euyas
'T.T. went from there to between Tawy and Euyas.' (CO 1217-1221)

	Simple	Conjunctive	Reduplicated
I	*mi*	*minneu*	*miui*
you (sg.)	*ti*	*titheu*	*tidi*
he	*ef*	*ynteu*	*efo*
she	*hi*	*hitheu*	*hihi*
we	*ni*	*ninneu*	*nini*
you (pl.)	*chwi*	*chwitheu*	*chwichwi*
they	*wy*	*wynteu*	*wyntwy*

Table 5.23: Middle Welsh Preverbal subject pronouns, cf. (Willis, 1998:134)

The conjunctive pronoun can be used in apposition to a noun phrase to emphasise the contrast meaning 'however, meanwhile, on the other hand'. But they are also used to repeat or pick up the discourse topic again in which there is an intervening noun phrase that could otherwise be interpreted as the topic. This is shown in (89) where the topic *wynteu* 'they' has to be overtly mentioned, since there is a plural

noun phrase *merchet* 'daughters' intervening, but the men are the ones who deserve to get all the drinks and love, according to this passage.

(89) a. *A r gwyr racko a gaffant med a bragawt yn enrydedus*
and the men there PRT get.PAST.3P mead and bragget PRED honourably
'And these men get lots of mead and bragget'
b. *ac a gaffant gorderchu merchet teyrned Ynys Prydein yn*
and PRT get.PAST.3P woo.INF daughters kings Isle Britain PRED
diwaravun
freely
'and they get to woo the daughters of the kings of the Island of Britain'
c. *ac wynteu a dylyant hynny*
and they.CONJ PRT merit.3P that
'And this they (i.e. the men) deserve' (BR 7.12-15)

5.5.4 Interim Summary

In this section I have presented the results of the investigation of the most important notions of information structure in Middle Welsh. A particular Focus Articulation or Domain (PRESENTATIONAL, PREDICATE or CONSTITUENT FOCUS) does not automatically yield one word order type in particular. Presentational focus can be found in subject-initial sentences (often with copular verbs), but new protagonists can also be introduced by non-verbal sentences (Type IX) with presentational idioms like *llyma, dyna* 'here is, there is'. PREDICATE FOCUS can be found in most word order types, though verb-second orders are always preferred and thus most frequently found in narrative contexts. CONSTITUENT FOCUS, finally, puts the focussed constituent in sentence-initial position or uses a very specific construction altogether to identify a constituent (the *sef*-construction of Type VIII).

Givenness and in particular the referential state of subjects and objects turns out to play an important role in making more fine-grained distinctions between different types of argument-initial word order. Direct objects can only be in initial position under certain conditions: they are either focussed (contrastively or conveying new information) or they continue/repeat a highy familiar topic from the immediately preceding context.

Different types of word order are finally employed in textual cohesion. Devices like points of departure or framesetters can be used to continue or change the scene. Continuous narratives without change in topic or scene are rendered by verbal noun-initial orders (Type IVc or Type VI), but as soon as there is a break, the new scene, time, location or protagonist is introduced in sentence-initial position by word order type III or IVab.

Overall, Information Structure played a significant role in the 'choice' between the various word order types in Middle Welsh.

5.6 Variation in word order

A study of the variation in word order of a particular (period of a) language is only meaningful if it is possible to control for any variables that could potentially influence the type of word order. Variation in this sense can then be:

1. 'all other things being equal' sometimes we find word order Type X and sometimes Type Y
2. if we change 1 variable from the 'standard, base', we find Type Y rather than Type X

The first scenario entails true optionality, but before we can draw that conclusion, we have to be 100% sure that 'all other things' are 'equal' indeed. It requires a very systematic analysis of all possible factors that could influence word order. The second scenario presents a very different approach, but this can only be employed when there is general agreement on what the 'standard' or 'base' is.

'True optionality' can give room for authorial choice: variation in word order could in this case be due to a preference for one type of word order or the other. According to Currie, in Early Modern Welsh this authorial choice appears "to be a decisive factor in determining the frequency of use of AIV (absolute verb-initial - MM) order" (Currie, 2000:211). For Middle Welsh, Poppe in particular has studied the variation in word order and agrees with Cappelle that "free choice in making grammatical choices [which] is not an illusion in some cases" (Cappelle, 2009:197) (cf. Poppe (2014) among others).

In order to systematically control for 'all other things being equal', this chapter presents the role of various grammatical, pragmatic (or information-structural), usage-based and extra-linguistic factors. In many of these cases, it turns out there is in fact no random variation at all. For some factors, clear rules and/or constraints can be formulated because there are no examples of a particular word order type in the database. For others, the distribution of the different types of word order over the possible variables reveals significant patterns. But only when all these factors are systematically and thoroughly investigated and combined can we accurately describe the variation and possible limits thereof.

Middle Welsh grammar indeed had many 'options' in terms of word order: for positive main declarative sentences alone, we can identify 9 different types. But not all of those could be used for transitive sentences, with past indicative inflection, subjects that conveyed new information in a constituent focus domain - to mention just one possible combination of variables. As was shown in the previous section, when all these factors are combined, variation in Middle Welsh word order was, in fact, rather limited.

5.6.1 The 'choice' of a particular word order type

The question is whether we can take this 'rather limited' statement one step further: is it possible to predict the type of word order if we take into account all these grammatical, pragmatic and other factors? To a certain extent, this indeed

seems to be the case. Figure 5.1 is a schematic representation of a 'decision-making' tree yielding the word order found in each of the possible (grammatical) contexts in Middle Welsh. It starts with the Numeration, the collection of things the speaker/writer wants to get across next. Which words and functional items appear in the Numeration depends on the language. In Middle Welsh, for example, aspect and tense played a role in the grammar, but evidentiality - an important linguistic feature in Amerindian and Tibetan languages - did not. Tense and aspect are thus expected to be part of the Numeration in Middle Welsh, but evidentiality is not. With the intended message ready in the Numeration, the syntax can build the sentence that will ultimately yield one of the word order types. In transitive statements in narrative contexts, a possible 'algorithm' determining the word order of each sentence taking all factors in the sentence and the context investigated in this chapter into account looks like Figure 5.1. Needless to say this algorithm is a very basic representation based on the tendencies found in the present corpus. If more texts are added and more variables are taken into account, this will probably have to be extended to cover all the data. I present this now in the form of a decision-making algorithm, however, because it forces us to be extremely explicit and precise in our analyses of word order variation. It furthermore provides a good starting point for future studies in Middle Welsh word order.

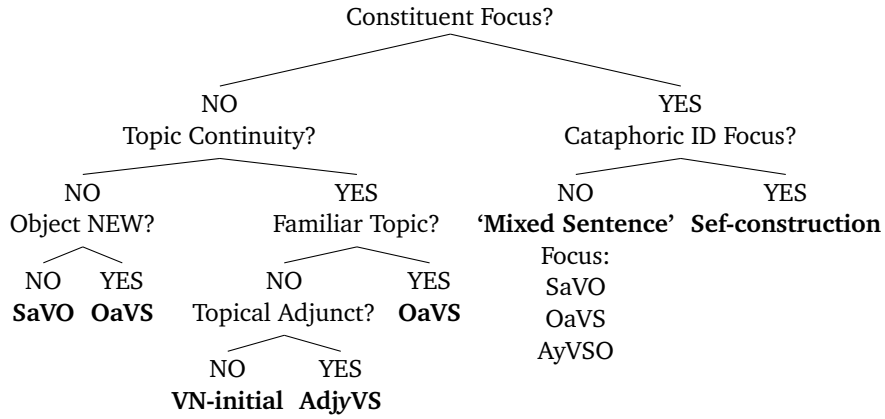


Figure 5.1: Decision algorithm 'predicting' the word order pattern in Middle Welsh

If there is an adjunct, for example a connective adverb like *yna* 'then' or *ar hynny* 'upon that', it can be added in front of any of these word order patterns, rendering Adjunct-OaVS or Adjunct-SaVO, for example. Note furthermore that in the course of the Middle Welsh period, the *sef*-construction developed in various ways, some of which were no longer marked for focus (see Chapters 6 and 7).

If the above was indeed correct for transitive sentences in Middle Welsh narratives, why can we still observe variation in word order in parallel passages or different manuscript versions of one and the same text. The context and grammar should be the same in these cases, so variation here requires further explanation.

One possibility already hinted at in the previous section is diachronic development of the language. In other words, the above-sketched decision-making scheme may have looked differently in different stages of the Middle Welsh language. Verbal-noun constructions were less frequently found towards the end of the Middle Welsh period, as were object-initial sentences. Manuscripts written by different scribes in different periods could give us more insight in the diachronic development.

Absolute verb-initial word order was for example only found under very restricted circumstances in Middle Welsh (oaths, idioms and quotative constructions as well as imperative and negative contexts). This changed in the Early Modern Welsh period: as Willis (1998) and others show, century after century, verb-initial order was increasingly found. But in the late Middle Welsh period, verbal-noun and object-initial orders were lost and at the same time the frequency of adjunct-initial order as well as periphrastic orders with the auxiliary *bod* 'to be' was already increasing. The implications of these diachronic developments in Late Middle Welsh are discussed in detail in Chapter 7.

5.7 Conclusion

As has become clear from this chapter, there are indeed various factors that could influence the word order of a sentence. They could work independently from each other, but many of those are likely to interact when used in different combinations. As Fried (2009:297) points out, even in modern languages speakers may have multiple options when it comes to choosing one particular word order pattern. Which patterns are available may be guided by discernible grammatical or pragmatic rules and cognitive principles, but it is not always all that clear "how the potential conflicts are resolved and whether or not they form coherent networks of combinations, both within individual languages and cross-linguistically." (Fried, 2009:297).

In Middle Welsh, there are nine main word order types (see Chapter 4). Some of those, for example, the argument-initial verb-second pattern contain different subtypes as well (i.e. subject-, object- or verbal-noun-initial orders). The main question I tried to answer in this chapter was which factors have an effect on the observed distribution of word order patterns. I systematically went through all language-internal and -external factors to determine if and how they exert any influence.

Starting with possible grammatical factors, verb-second sentences with verbal nouns in initial position (Type IVc) almost exclusively occur with verbs in the preterite tense. The significance of (preterite) tense as a factor is likely to be related to the fact that these verbal-noun patterns are the basic word order in indirect speech passages of narrative tales. In direct speech, on the other hand, subject-initial orders are most frequently attested. Another interesting finding concerns active vs. impersonal inflection. Impersonal verbs are most frequently found in verb-second sentences with initial adjuncts (Type III). This can be explained if the sentence-initial position is a topic position. The agent in impersonal and passive

constructions is unlikely to be the preferred topic because it is demoted. If there are other candidates to fill the topic position, for example adjunct frame- or scene-setters, these will be preferred in sentence-initial position. A final grammatical factor that plays a role in the preferred types of verb-second order is animacy of objects and indirect objects. Inanimate objects tend to appear in object-initial orders more frequently than expected. This might have something to do with information structure, to which I turned in the final section of this chapter.

The first information-structural notion under investigation was Givenness. After determining the referential status of the core constituents in the corpus, I found that direct objects in initial position almost exclusively convey New information. In this way, the 'Natural information flow' of the sentence (going from old to new) is disturbed and these object-initial sentences are thus marked. The only exceptions to this generalisation are familiar topics, mainly in the form of demonstrative pronouns referring back to the the last-mentioned item/person/concept in the immediately preceding context.

In terms of text cohesion we can make two further observations. First of all 'points of departure' or frame-setters clearly occur most often in verb-second sentences with adjunct-initial order in which they function as the topic. They can also be found with other types of verb-second order, for example in combination with subject-initial word order, but this is not the preferred pattern. A second observation in this context concerns textual continuity achieved by sentences starting with verbal nouns. To achieve close cohesion, these initial verbal nouns can be placed in sentence-initial position. They are either relying on an inflected verb in the previous sentence (Type VI) or are continued with an inflected form of the auxiliary 'to do' (Type IVc). Again this is part of the preferred narrative style.

Focus can finally be observed in the dedicated (reduced) cleft order called the 'Mixed Sentence' (Type V). Focus of the identificatory predicate can furthermore be found in the special *sef*-construction (Type VIII). Not all sentences with *sef* are focussed, however (see Chapter 7 for an overview of the diachronic development). In Chapter 6 I will examine four different case studies concerning the most important notions in information structure and how they are manifested in Middle Welsh syntax.

CHAPTER 6

Information structure and word order in syntax

6.1 Introduction

"MM: What are you teaching this term?
YT: One session on vampires (found some really nice old texts) and one on rules concerning archery ceremonies.
MM: That sounds great!
YT: One of them is about a primordial ladyvamp who descends to earth!
MM: Then what happens?"

Previous chapters focussed on the core notions of information structure and Middle Welsh word order. If we look at the above conversation between two academics, we clearly see that information-structural primitives like givenness and focus appear in sentences with 'abnormal' word-order patterns: even wh-elements that usually appear sentence-initially can be preceded by other elements. In this chapter the main question therefore is: how do information structure and word order relate to the syntax of Welsh?

To answer this question it is first of all important to define syntax itself in relation to word order. Early syntactic research often merely concentrated on the word order of the verb and its core arguments. Languages that did not seem to have a preference for one particular basic word order were called 'non-configurational' (cf. K. Hale (1983) on the Austronesian language Warlpiri). This as opposed to configurational languages in which the 'grammar' determined the order of words in the sentence. But what part of the 'grammar' is this? In functional traditions like the Prague School, discourse-semantic notions could also play a role in structural

relations. This was formalised in syntactic accounts by, amongst others, Jackendoff (1972) and Horvath (1981). Around the same time, Li and Thompson (1976) distinguish subject-prominent languages from topic-prominent languages in which the morphology and syntax highlight topic-comment distinctions, rather than grammatical functions like subject or object. This then led to a third type of language: discourse-configurational. According to É.Kiss (2001), languages are discourse-configurational if they link either or both of the discourse-semantic functions topic and focus to particular structural positions.

This leaves some interesting questions open. First of all, are these discourse-semantic functions an overall property of the language or do they, for example, only play a role in a certain domain? If there is a 'particular structural position', where in the sentence can we find this? And, finally, is this the same cross-linguistically and if not, how do we account for language variation?

This chapter aims to address some of these issues that are relevant for Middle Welsh. It discusses how the information-structural notions introduced in Chapter 3 can be integrated into syntax. The corresponding Middle Welsh word order patterns discussed in Chapters 4 and 5 are then analysed syntactically. Each of the core notions of information structure are finally considered in greater detail in case studies on focus, topic, givenness and text cohesion.

6.2 Integrating IS and word order in syntax

According to Lambrecht (1994:6-13), language is a tripartite system consisting of syntax, semantics and information structure. Semantics is concerned with the *meaning* of words and utterances. Information structure is a pragmatic notion signalling how a certain message is conveyed or, following Lambrecht, 'why there are so many sentence structures' (Lambrecht, 1994:9). Syntax, finally, is the form or formal structure. It is often broadly described as 'sentence construction': the way words group together in phrases and sentences (Tallerman, 2011:1). The questions and answers in the introductory conversation above show various linguistic strategies (e.g. wh-movement, but also if we read it out loud, special intonation on the word *then*, for example). These strategies can be paired with certain interpretations (e.g. aboutness topics, contrastive focus, etc.). As **?**:1 points out, however, this pairing "does NOT mean that the interpretation is there BECAUSE of the linguistic strategy $\Rightarrow$ correlation $\neq$ causation."

This section gives a brief overview of formal ways to integrate information structure into syntax and marks the basic assumptions for the present study of historical Welsh.[1]

[1] Dependency grammars are not included in the present overview, since they are traditionally less concerned with linear word order than, for example, phrase structure grammar. There are, however, attempts to implement information-structural notions in lexicalised dependency grammar formalisms, like Topological Dependency Grammar (TDG) (cf. Kruijff and Duchier (2003)).

6.2.1 Formal combination of IS and syntax

There are various ways to formalise this 'grouping of words'. In theory, this could be done by a dedicated set of rules predefined for a certain language. Starting from grammatical functions, for example, a language like English could have the very basic rule to group the core arguments of the verb together in the order 'subject-verb-object'. To account for all possible variation, both within one language, but also cross-linguistically, we would have to define a vast amount of rules for each specific context or sentence type. This is undesirable for many reasons, not in the least because it cannot *explain* why the 'grouping of words' is the way it is and why it differs from other types of sentences or other languages and, crucially, why that is not always the case.

Syntacticians have therefore tried to formalise this system, abstracting away from a predefined set of rules. Language, and in particular grammatical knowledge was since the work of Noam Chomsky in the 1950s viewed as a modular cognitive system in the generative approach. This system is considered to be a computational system (C_{HL}) interfacing with other cognitive modules like the conceptual-intentional system concerned with meaning and the sensory-motoric system producing and processing sounds.

The constructivist or usage-based view denies this modularity of the grammatical system. Linguistic representations are instead grounded in experiences of language use (cf. Langacker (1988)). In construction grammar (cf. Fillmore, Kay, and O'Connor (1988), Goldberg (1995)) this means that both grammatical rules as well as words consist of pairings of form and meaning: sounds and meaning are linked according to conventions of the speech community leading to an inventory of constructions: a Constructicon. Constructions in the Constructicon are assumed to bear different kinds of relationships to each other (cf. Beekhuizen (2015:14-16)). Both lexical and grammatical constructions can be combined like building blocks creating larger and more complex linguistic units. In such a system, information-structural phenomena (like topic or focus) must be coded as properties of constructions. Features are used to indicate these 'rhetorical relations' (cf. Östman and Virtanen (1999:92-93)) in the construction matrix, just like grammatical relations (Subject, Object, etc.), semantic roles (Patient, Agent, etc.) and situational frame-roles (like 'buyer' or 'seller' in a commercial transaction).

In Lexical Functional Grammar (cf. Bresnan (2001)), on the other hand, information structure is considered to be one of the possible structures that are hypothesised in the LFG framework. Language consists of multiple dimensions of structure, e.g. the representation of grammatical functions (f(eature)-structure), syntactic constituents (c(onstituent)-structure), but also semantic, morphological and phonological structures. Information-structural notions are thus combined (and constrained) like any other part of language.

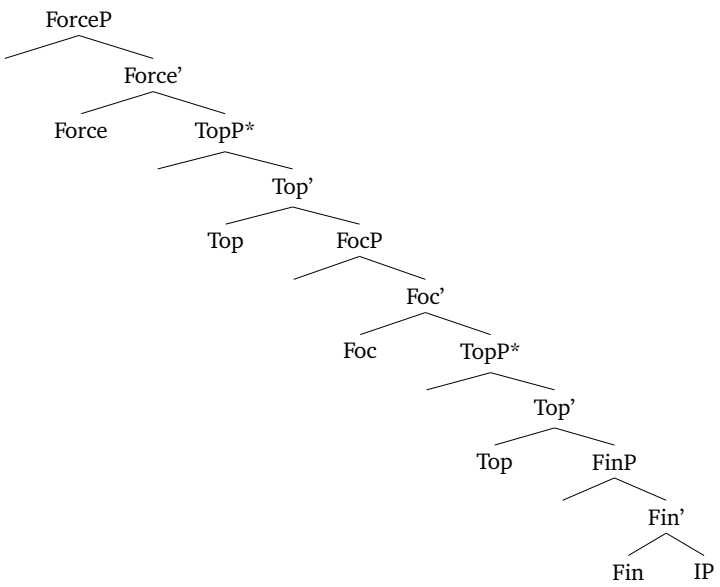
Extra levels have also been proposed in a generative framework. López (2009) takes discourse to be "a computational module that assembles sentences (and possibly other units) into Discourse Representation Structures" (López, 2009:22). He further posits a pragmatics module that "assigns features relevant for the

insertion of a syntactic object into a discourse structure to constituents in certain positions." (López, 2009:22). These 'positions', according to López, are the edges of syntactic phases (in the sense of Chomsky (2000)). The relevant features for him are binary +/- Anaphoric and +/- Contrast (rather than Topic or Focus).

A featural approach to information-structural notions is crucial in other frameworks as well. In Head-driven phrase structure grammar (HPSG) the relevant units of linguistic information are signs (cf. Pollard and Sag (1987) and Pollard and Sag (1994)). These signs explicitly express phonological, syntactic, semantic and pragmatic information, formalised as typed feature structures. Engdahl and Vallduví (1996) implemented information structure in this framework as a set of features in the CONTEXT (the part representing pragmatic information) of the feature matrix.

A different way of implementing information-structural features is to sequence them in a universal hierarchy of functional heads. Cartography was the first proposal 'mapping' the information-structural features in such a way in the left periphery of the clause (Rizzi, 1997). His work is based on various types of topic and focus phrases found in clause-initial position in Italian (and other Romance languages). Cinque (1999) subsequently added a similar detailed structure for adverb positions. A central hypothesis in this framework is that this fine hierarchy (see example (1) based on Rizzi (1997)) and order of functional projections is universal, i.e. it can be found in all languages.

(1)



A major test case for the Cartographic framework is thus presented by other languages than Italian (or Romance) on the basis of which this articulated structure was originally proposed. The main question is whether it is necessary to assume this rigid hierarchy for languages that do not overtly show these types of topic

and focus constituents. A further question is whether languages that do exhibit multiple topic and focus phrases in the left periphery always order those in the same way. In light of the latter, various scholars working on for example (Old) Germanic (Frascarelli, 2007), Chinese (Badan & Del Gobbo, 2011) or Hungarian (Lipták, 2011) have suggested refinements or additions to Rizzi's original proposal.

Cartography is not the only way to integrate information structure in the syntax. What could be argued to be the opposite view of cartographic syntax 'full' of information structure is 'Clean Syntax' (cf. **?**:2). In this other extreme point of view advocated by, amongst others, Fanselow and Lenertová (2011), information structure and syntax are completely independent (see also experimental work by Onea and Beaver (2011) and Destruel and Velleman (2014)). Both these extremes - a syntax full (Cartography) or completely devoid (Clean) of information structure face empirical challenges (for examples from Bantu languages, see Cheng and Downing (2012) and **?**).

Another solution is presented by interface approaches developed by, amongst others Neeleman and Van de Koot (2008) and Kučerová and Neeleman (2012). In their framework, syntax is mapped to information structure at the interface, with movements being driven by the necessity of the complements of topics and foci to be constituents at the interface. This line of research is based on the frequently-found interaction between 'marked' prosodic patterns and information structure. Conditions or rules at the interface between syntax and phonology restrict the possible derivations and interpretations. From this point of view, information structure and syntax interact only indirectly, mediated by prosodic manifestations (see also Szendrői (2001), Zubizarreta (1998) and Horvath (2010)). To account for syntactic focus movement, Horvath (2010) introduces an Exhaustive Identification Operator requiring stress-based (information) focus within its c-command domain. Topic-comment structures, on the other hand are dealt with via the Comment Mapping Rule posited by Neeleman and Van de Koot (2008):

(2) **Comment Mapping Rule**
If XP in (3) is interpreted as topic, then interpret N2 as comment.

(3) $[_{N1}$ XP $[_{N2}$ ... t ... $]]$

According to Aboh (2010), however, information-structural features such as topic and focus must have their origin in the Numeration just like Case and φ-features. He emphasises that in a minimalist approach to the study of language, syntax is the computational system C_{HL} that maps some array of lexical choices (the Numeration) to the sound-meaning pairs (π, λ).[2] Sentences are built from the items in the Numeration only and features can thus not be added during the derivation (i.e. during the structure-building). This is called the Inclusiveness Condition:

> "Given the numeration N, C_{HL} computes until it forms a derivation that converges at PF and LF [...] A "perfect language" should meet the

[2]'Sound' could also be a sign in sign languages.

> condition of inclusiveness: any structure formed by the computation [...] is constituted of elements already present in the lexical items selected for N; no new objects are added in the course of computation apart from rearrangements of lexical properties." (Chomsky, 1995:228)

From this point of view topic and focus, for example, but also interrogative force or the concept of contrast, are part of the numeration and project in syntax. This could result in a Cartographic hierarchy of information-structural heads and phrases in the left periphery of the clause. Alternatively, topic and focus features could be clustered on a single C (or Force/Fin) head, at least in languages without multiple phrases in the left periphery of the sentence. The status of the C-domain in itself (articulated or not) is a topic of various recent studies. Since constituents in the left periphery of the C-domain often interact with other linguistic domains such as prosody, they can be argued to exist in a dimension that differs from the core argumental syntax. Constituents that are information-structurally marked, for example, exist on a different plane and can therefore be targeted by prosody. Examples of interface studies suggesting such an approach are Cheng and Downing (2012) (for focus in Zulu) and D'Alessandro and Van Oostendorp (2016) (based on truncated vocatives in various languages).

6.2.2 Assumptions for the present study

Despite the lack of spoken data, phonological interface approaches as the ones mentioned above have been developed for older stages of Germanic languages (cf. Hinterhölzl (2009)). These studies have to make certain assumptions about the phonological phrases and their relation to syntactic structure. Hinterhölzl (2009:56) suggests for example that a "right-headed phonological phrase (in a verb cluster) must sit on a right branch with respect to the syntactic head that is to become its prosodic sister". Word order preferences are due to violable interface conditions defining ideal mappings between syntactic and prosodic structures (cf. Hinterhölzl (2009)).

There is, to my knowledge, no systematic study of prosodic structure in Middle Welsh in relation to syntactic phrases. This severely complicates drawing any conclusions using any of the above-mentioned phonological interface approaches. For the present study, I therefore adopt Aboh's (2010) view with information-structural features starting out in the Numeration with other linguistic items the speaker chooses to express. In the course of the (narrow) syntactic derivation, these features can then enter into an Agree relation with a probing head in the C-domain.

6.2.3 Middle Welsh syntax

Traditional Welsh grammarians were intrigued by Middle Welsh because of its 'abnormal' i.e. 'non-verb-initial' word order as discussed in the previous chapters. Information structure was considered to have played an important role as 'a pragmatic constraint' on the syntax (cf. Poppe (1991), Fife (1991)). From such

a functionalist view, the word order or syntax was determined by information-structural notions like topic or focus. Studies along this line of research mainly focussed on the description and distribution of various possible word orders (e.g. subject-initial, object-initial or adjunct-initial). This left the questions of how and why these information-structural notions interacted with the syntax unanswered.

The puzzle of Middle Welsh word orders

At a glance, the puzzle of Middle Welsh word order patterns is the following. From a synchronic, Middle Welsh, point of view, there seem to be two main strategies. Traditional Welsh grammarians have distinguished those based on functional (topic vs. focus) and grammatical (subject-verb agreement vs. default third-person singular agreement) characteristics. 'Topicalised' sentences exhibit subject-verb agreement and are traditionally called 'Abnormal Sentences' or, in Welsh *brawddeg annormal* (see Chapter 4). 'Focalised' sentences do not exhibit agreement and are called 'Mixed Sentences' (*brawddeg gymysg*).[3] Typical examples of abnormal and mixed sentences are shown in (4) and (5):

(4) *A 'r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
'And the nobles came together' (Abnormal Sentence - PKM 90.27)

(5) *Mi a 'e heirch.*
I PRT 3FS seek.3S
'(it is) I who seek her' (Mixed Sentence - WM 479.24)

Abnormal Sentences like (4) typically show agreement in number between the preverbal subject and the finite verb.[4] This sentence is thus not only 'abnormal' because it is not verb-initial (like Modern Welsh), but also because it shows agreement with plural full DP subjects. This is unique in both Middle and Modern Welsh, since usually only pronouns show agreement in any part of the language (the verbal system, but also as 'inflected' prepositions). Full DPs never cause agreement and the language thus exhibits a similar type of Complementarity Principle as was observed by, amongst others, Borsley and Stephens (1989a) and Stump (1989) for Breton. From a functional perspective "no special emphasis is intended for the word or phrase which comes at the beginning" (D. S. Evans, 2003 [1964]:180).

The Mixed Sentence exemplified in (5) on the other hand is used "[w]hen a part of the sentence other than the verb is to be emphasized" (D. S. Evans, 2003 [1964]:140). It was originally preceded by a form of the copula and followed by a relative clause. Since relative clauses usually do not exhibit agreement (D. S. Evans, 2003 [1964]:60-61), the verb is found in the default third-person singular form even when the subject/antecedent is a pronoun as seen in (5) with a first-person singular pronoun *mi* 'I'. Sometimes, the original copula is still found, shown in (6):

[3] Names in Modern Welsh are given because much of the secondary literature on this topic was written in Modern Welsh. I will keep using the traditional Abnormal and Mixed labels for the sake of convenience.
[4] Agreement in Gender is never found on inflected verbs in Welsh.

(6) *Ys mi a 'e heirch.*
is.3S I PRT 3FS seek.3S
'It is I who seek her.' (WM 479.29)

If we disregard any notions of information structure, it is impossible to make a formal distinction between the abnormal and the mixed sentence if the subject is a singular noun or a third-person singular pronoun. The verb in these cases would exhibit third-person singular inflection anyway. According to D. Simon Evans, "[f]ormal divergence is found only when the sentence is negative" (D. S. Evans, 2003 [1964]:180), as shown in (7):

(7) a. *Y dyn ny doeth.*
the man NEG come.3S
'the man didn't come' (Abnormal Sentence)
b. *Nyt y dyn a doeth.*
NEG the man PRT come.3S
'it was not the man who came' (Mixed Sentence)

Willis (1998:6) notes, however, that this difference might simply reflect the distinction between negation of the entire proposition or of a single constituent. In the examples he gives with the abnormal order (7a), the negation follows the subject, whereas in the mixed order (7b), it precedes the emphasised/fronted phrase in sentence-initial position. Negation in Abnormal Sentences as found in (7a) is not often found, however. The preferred word order for sentence negation is NegVSO as shown in (8):

(8) a. *Ny chymerwn ninheu y gan y tayogeu hynny.*
NEG take.1P we from with the churls these
'We will not take that from these churls.' (PKM 53.28)
b. *Ny welei ef y twrwf rac tywyllet y nos.*
NEG see.PAST.3S he the commotion as darkness the night
'He could not see the commotion as the night was so black.' (PKM 22.23)

Apart from subjects, direct objects or adjuncts (adverbs or prepositional phrases) could also appear in sentence-initial position, either with or without 'emphasis'. Just like the antecedents of relative clauses, subjects and direct objects were obligatorily followed by the preverbal particle *a* (as seen in (4), (5) and (6) above). This particle caused lenition or soft mutation of the immediately following verb. Whenever an adjunct appeared in sentence-initial position, the preverbal particle *y(r)* was used (without any form of consonant mutation), as in (9):

(9) *Yna y doeth y kennadeu.*
then PRT come.PAST3S the messengers
'Then the messengers came.' (PKM 79.27)

From a synchronic syntactic point of view, the most important question is how the Abnormal and Mixed Sentences are derived? Furthermore, apart from their

agreement patterns, do these patterns differ in any way? If that is the case: how are they different? And, furthermore, do these differences arise from differences in their information-structural status?

Although the observed generalisation of topic (agreement) vs. focus (no agreement) seems to hold most of the time, there are many exceptions. There are examples of sentences with agreement that clearly contain contrastively focussed subjects (see (10a)). But there are also cases without expected agreement where no focus can be detected either (see (10b)). To make matters worse, as Poppe (2009) points out, there are cases in which differences in agreement appear in the exact same (con)text, but in different manuscript versions, as shown in (11).

(10) a. *Miui hagen a uydaf gyfarwyd ywch*
I.EMPH however PRT be.1S familiar to.2P
'I, however, will be familiar to you.' (Culhwch 899)

b. *Kennadeu a aeth at uranwen.*
messengers PRT go.PAST-3S to Branwen
'Messengers went to Branwen.' (PKM 40.1-2)

(11) a. *Ti a y gwelho*
you PRT 3FS see.SBJ-3S
'You will see it' (White Book CO 451)

b. *Ti a y gwelhy*
you PRT 3FS see.SBJ-2S
'You will see it' (Red Book equivalent)

The Middle Welsh word order situation is further complicated by the fact that other types of word order appear alongside the above-mentioned Abnormal and Mixed sentences. There are verbal noun constructions with and without auxiliary verbs appearing in contexts of narrative continuity (see Chapter 5). But there was also a special type of copular clause with sentence-initial *sef* marking the focussed identificational predicate (see section 6.3 below). In the course of the Middle Welsh period, however, this *sef*-construction further developed and the original identificational focus of the predicate was lost, resulting in yet another option to express propositions in a narrative context.

Syntactic studies of Middle Welsh

According to various Welsh scholars (MacCana (1973), Fife (1988), T. A. Watkins (1977)), the Abnormal Sentence was never part of the spoken language in Middle Welsh. Verb-initial order according to them had always been the norm and these 'fronting' constructions with sentence-initial subject or objects were a purely literary device (Fife, 1991:89-90).

Willis (1998), however, convincingly showed based on cross-linguistic as well as language-internal evidence that this cannot be the case. The abnormal and mixed orders cannot be a literary device, but must be a case of a verb-second constraint on the grammar of Middle Welsh. From a cross-linguistic point of view, it is unlikely that a highly literary rule as proposed from Middle Welsh would have developed

in related languages independently. Breton and Cornish also exhibit subject- and object-initial word orders, so it is more likely that these were present already in the parent language Brythonic. From a language-internal point of view, it is difficult to explain how such a syntactically complex rule as topicalisation could be learnt for purposes of writing only (see also Borsley et al. (2007:292-293)). According to Willis, this requires "an awareness that constituents other than the subject could be fronted and a *conscious* awareness of the notion of 'topic'." (Willis, 1998:13).

Tallerman (1996) proposed to explain the difference between the abnormal sentence and the mixed sentence by positing different derivations for each of them. Abnormal Sentences involve adjunction of the topic XP to CP and the syntactic subject is realised as *pro*, triggering subject-verb agreement. According to Borsley et al. (2007:293), however, this is problematic, because it predicts multiple topics should be possible. Topicalisation in Middle Welsh was not recursive, according to Willis (1998): only one of the preverbal constituents could be an argument (hanging topics and left-dislocations aside): "[a]ll other fronted elements are adverbial" (Borsley et al., 2007:293).

Alternatively, Willis (1998) proposes the difference between agreement and the lack thereof in subject-initial sentences in Middle Welsh is based on a difference in movement. Topicalised Abnormal Sentence involve A-movement of the subject via SpecAgrSP, whereas focalised mixed sentences are derived by A'-movement. The focalised subject skips the higher agreement projection and goes straight from SpecTP (where it receives Nominative Case) to SpecCP. One possible objection to this approach is that additional assumptions have to be made about the trace or copy of full DP subjects. This is unexpected according to the Complementarity Principle that seems to hold in all other parts of the grammar: full DPs never seem to cause agreement. An additional assumption that the trace or copy of the full DP *can* result in number inflection on the verb thus has to be made.

Interim summary

Studies of Middle Welsh word order patterns have initially focussed on functional descriptions of the various verb-second orders that deviated from the Modern Welsh verb-initial norm. Though much progress was made describing various information-structural patterns, these 'purely pragmatic' approaches (like for example Poppe (1991) or Fife (1991)) ran into problems accounting for the variation in agreement and, crucially, the lack thereof (as pointed out in detail by Poppe (2009)). These difficulties arose not in the least because there was no consensus on what the basic notions of information structure were and how they could be defined and implemented in systematic analyses of the language. This I have tried to remedy by clearly outlining information-structural methodology and terminology in Chapter 3. In Chapter 4 I furthermore concluded that the distribution of word order patterns in Middle Welsh could be the result of multiple factors interacting with each other. Information-structural features do play a role, but they cannot be taken into account in complete isolation. In the remaining part of this chapter I therefore examine examples from each of the core notions of information structure discussed

in Chapter 2 focussing on how they are integrated into (or part of) the syntactic system of Middle Welsh.

6.3 Case Study I: Focus-background

As has become clear from Chapters 4 and 5, there are various ways to exhibit focus in Middle Welsh. In this section I propose a syntactic analysis of one particularly frequently found focus construction in Middle Welsh: identity predicate focus by means of the lexical item *sef*. There are various so-called '*sef*-constructions' in Middle Welsh, all of which derived from the identity copular clause with anticipatory predicates. A diachronic analysis of the various stages of the grammaticalisation process is presented in Chapter 7. This section focusses on the syntactic derivation of the *sef*-construction, starting from the derivation of the two types of unmarked copular clauses.

6.3.1 Identity predicate focus: the data

As shown in Chapter 4, copular matrix clauses in Middle Welsh exhibit two possible word order patterns as shown in the schemas in (12a) and (12b):

(12) a. *ys* - Predicate Complement - Subject (CPS)
b. *mae* - Subject - *yn* Predicate Complement (CS*yn*P)

In the present tense each of these constructions yields a different form of the copula: *ys* or *mae*. In (12b) there is a special predicate marker *yn* introducing the predicate complement. This predicate marker *yn* is never found in examples with CPS word order with the schema presented in *excop*. This difference goes back to the traditional Celtic distinction of true copulas and substantive verbs (cf. for example Lash (2011) on Old Irish). Examples reflecting this distinction are presented in (13a) and (13b):

(13) a. *Ys gohilion hwnn*
be.PRES.3S remainder DEM.MS
'That one is remaining.' (CO 472)
b. *Ac y mae y enw yn parawt.*
and PRT be.PRES.3S 3MS name PRED ready
'and his name is ready' (PKM 76.19)

Willis (2015) notes that a third type of word order is found in non-finite subordinate copular clauses with the infinitival copula *bod* 'to be', as shown in schema (14):

(14) *bod yn* Predicate Complement - Subject (C*yn*PS)

This schema of subordinate copular clauses *does* exhibit the predicate marker *yn*, but the Subject and Predicate complement are in the same order as the matrix

copular clauses *without* the marker *yn*. An example of this Predicate-Subject order in subordinate clauses is shown in (15):

(15) *Duw, a wyr pob peth, a wyr bot yn eu*
God REL know.PRES.3SG every thing PRT know.PRES.3SG be.INF PRED false
***hynny** arnaf i.*
that on.1SG me
'God, who knows everything, knows that that is a lie about me.' (PKM 21.3)

Finally, a special form of the copular clause with focus on the identificational predicate puts a petrified form of the copula and the anticipatory predicate in initial position (*(y)s + ef > sef*), followed by the subject and the predicate in that order (*sef* Subject - Predicate:

(16) *Sef gwreic a uynnawd gwreic ieuank*
sef woman PRT want.PAST.3S woman young
'That was the woman he wanted, a young woman.' (YBH 6)

This *sef*-construction took up many shapes and forms during the Middle Welsh period. In Chapter 7, I argue that these forms represent different stages in a process of grammaticalisation. In the following section I zoom in on the synchronic syntactic analyses of the above copular clauses and the *sef*-construction in particular.

6.3.2 Identity predicate focus: syntactic analysis

There are various possible ways to derive the above sentences that explain the superficial difference in Subject-Predicate vs. Predicate-Subject word order. Assuming that the subject starts out in the specifier of the Predicate Phrase, some form of predicate raising is necessary to arrive at copula-initial word orders. Adger and Ramchand (2003) propose such raising analyses for Scots Gaelic (to SpecTP). In the following sections I show how their approaches can be extended to account for the various word orders found Middle Welsh copular clauses, including the identificational predicate focus clauses or so-called '*sef*-constructions'.

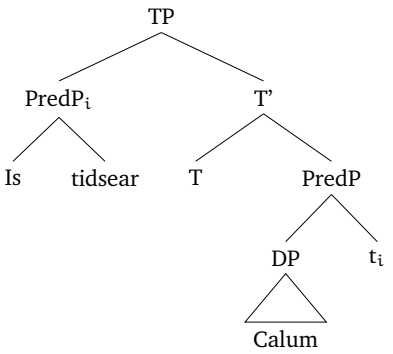
Adger and Ramchand (2003) propose an analysis raising the copula and the predicate together for what they call 'Inverted Copular Clauses' (ICCs) in Scots Gaelic with the same Predicate-Subject word order. Consider the following example in Scots Gaelic (SG) and the derivation in (18) (cf. Adger (2011:4)):

(17) *Is tidsear Calum.*
COP-PRES teacher Calum
'Calum is a teacher.' (SG ICC - Adger and Ramchand (2003:335))

The raising of the predicate is motivated to satisfy the EPP property of T: the copula raises and pied-pipes its complement. The copula could not raise on its own due to its 'extreme phonological weakness, so head movement to adjoin to T does not occur' (Adger and Ramchand (2003:336)). The EPP triggers the movement of Pred',

in the notation of Adger and Ramchand (2003). Under Minimalist assumptions of Bare Phrase Structure, this would be considered 'PredP' and as such it could be moved as a phrase (see also Adger and Ramchand (2003:336n6)).[5]

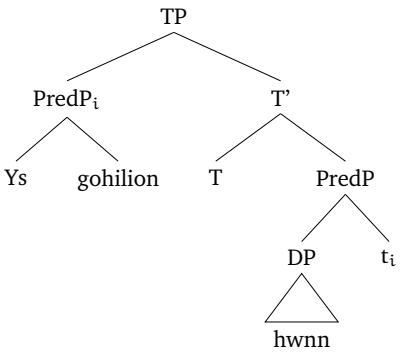
(18)



This predicate raising analysis would yield the copula-predicate-subject (CPS) order for Early Middle Welsh sentences like (19) as shown in (20).

(19) *Ys* *gohilion* *hwnn*
be.PRES.3S remainder DEM.MS
'That one is remaining.' (CPS - CO 472)

(20)



One way of explaining the difference between this CPS order and an example with the predicate marker *yn* like (21) is to leave the Predicate Phrase *in situ* and satisfy the EPP of T by (first) merging the copula *mae* there.[6]

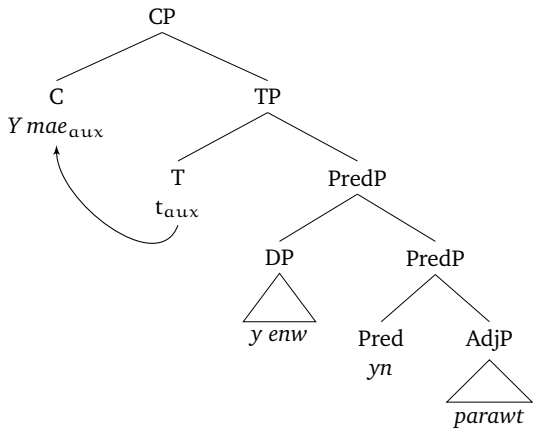
(21) *Ac* *y* *mae* *y* *enw* *yn* *parawt.*
and PRT be.PRES.3S 3MS name PRED ready
'And his name is ready' (CS*yn*P - PKM 76.19)

[5]Technically, we are in fact dealing with 'optional' pied-piping of the predicate complement in this case. If the Pred-head is probed and therefore moved to SpecTP, it can pied-pipe its complement.

[6]Note that movement of the subject DP to SpecTP would be possible, but entirely string-vacuous in this derivation.

The difference thus lies in the presence of the lexical predicate marker *yn* in the Numeration. This external merger of *mae* further creates the option to move it up to the C-domain as suggested by, amongst other, Roberts (2005) for all inflected forms of *bod* 'to be' in Welsh (which would also allow the subject to move to SpecTP to agree with the inflected verb). Agreement with the subject could be established by the auxiliary form of *bod* 'to be' in the T-head probing the subject in SpecPredP and subsequently moving up adjoining the sentence-initial particle in the (higher[7]) C-head. The Predicate *yn* and the Adjectival Phrase *parawt* can remain *in situ* lower down in the clause in this configuration.

(22)



Adger & Ramchand's (2003) analysis of the copular constructions has a solid semantic background involving a **holds** predicate that predicates a property of an individual as follows (cf. Adger (2011:4)):

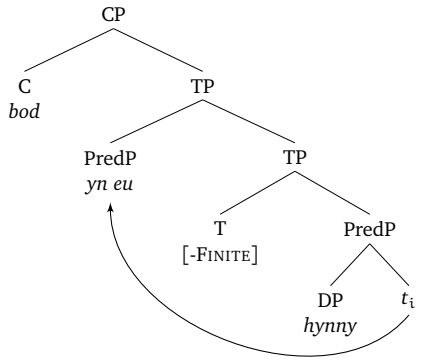
(23) a. [[Pred']] = λx.**holds(teacher,*x*)**
b. [[Calum]] = **Calum**
c. [[PredP]] = **holds(teacher, Calum)**

Recall that in non-finite subordinate copular clauses introduced by *bod*, the word order was copula-*yn*-Predicate Complement-Subject. The copula in this case consists of the infinitival form *bod* 'to be'. When introducing a subordinate clause, however, *bod* can be analysed as the complementiser in the C-head of the clause. In this case, the infinitival T-head is empty and can probe the Predicate head that again moves to SpecTP pied-piping its complement just as in the matrix CPS orders. A derivation of the subordinate clause in (24) is shown in (25):

[7] Roberts (2005) argues that auxiliary forms of *bod* 'to be' end up in the higher C-head of an articulate CP he labels ForceP, but I leave out the details of the C-domain here, because they are not relevant to the present discussion. In Chapter 7, however, I will return to this issue.

(24) ... *bot yn eu* ***hynny*** *arnaf i.*
bod PRED false that on.1SG me
'... knows that that is a lie about me.' (PKM 21.3)

(25)



The characteristics of the T-head, rather than the phonological strength of the copula in the Pred-head (as Adger and Ramchand (2003) argue) might thus be the reason why movement to SpecTP is triggered or not.[8] (26) shows the three possibilities and characteristics of T and the Numeration in greater detail:

(26) a. *ys* - Predicate Complement - Subject (CPS)
Numeration: { $T_{[+FIN]}$, DP_{Sbj}, Copula *ys*, $DP_{PredComp}$ }
⇒ empty finite T-head bears EPP attracting PredP

b. *mae* - Subject - *yn* Pred. Complement (CS*yn*P)
Numeration: { $T_{[+FIN]}$, Aux. *mae*, Pred. marker *yn*, DP_{Sbj}, $AdjP_{PredComp}$ }
⇒ Aux first-merged in finite T: EPP may attract subject

c. (Matrix) ... *bod yn* Pred. Complement - Subject (...C*yn*PS)
Numeration: { (Matrix), $T_{[-FIN]}$, complementiser *bod*, Pred. marker *yn*, DP_{Sbj}, $AdjP_{PredComp}$ }
⇒ *bod* first-merged in C: empty non-finite T attracts PredP

In both (26a) and (26c) the T-head is empty and therefore able to attract the PredP to its specifier. In (26b), on the other hand, the auxiliary must be first-merged in the T-head (to receive tense inflection), therefore movement of PredP does not take place. In the non-finite subordinate clauses finally, *bod* has no tense inflection and can be directly merged as the complementiser in the C-head.

[8] Willis (2015) also presents a predicate-raising proposal based on featural differences in the T-head. His analysis involves raising to the outer specifier of an extra VPredP and further remnant movement of the predicate complement, which results in the same possible range of word order patterns. I do not adopt Willis's proposal here, however, since it presents further complications when it comes to explaining the (historical) developments in the various different kinds of *sef*-constructions. As I argue in the next sections and in Chapter 7, Adger & Ramchand's (2003) approach *can* be extended to account for those as well, which is why I adopt and extend their approach for Scots Gaelic here.

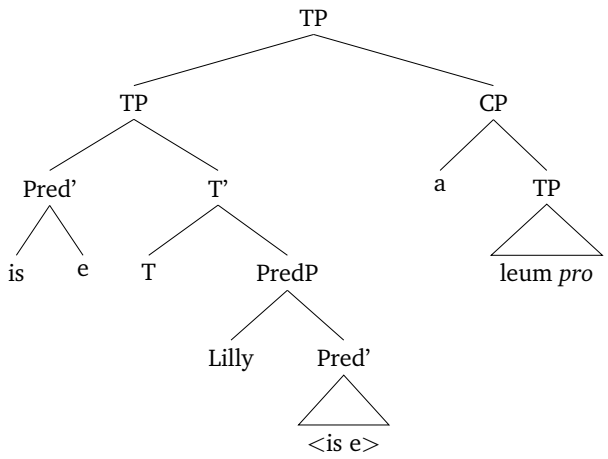
Now the basic structures of the copular clauses are clear, let us turn to the information-structurally marked options with cleft structures. In Gaelic, the Inverted Copular Clause (ICC) with predicate raising is now somewhat archaic, but it was used to build many other constructions in the language such as clefts. These clefts were eventually preferred over the ICC orders as shown in (27a) and augmented copular clauses as in (27c).

(27) a. *'S e tidsear a tha ann an Calum.*
COP-PRES it teacher REL be.PRES in Calum
'Calum is a teacher.' (Preferred cleft structure - Adger (2011:3))
b. *Is e Lilly a leum.*
COP it Lilly that jumped
'It's Lilly that jumped.' (Cleft - Adger (2011:5))
c. *'S e Calum an tidsear.*
COP-PRES AUG Calum the teacher
'Calum is the teacher.' (ACC - Adger and Ramchand (2003:339))

Adger's (2011) derivation of a cleft sentence like (27b) is shown in (28) (semantically) and (29) (syntactically):

(28) [[Cleft]] = **holds**
($\lambda x \exists e$.jump(e) $\wedge$ agent(x, e) $\wedge$ past(e), Lilly)

(29)



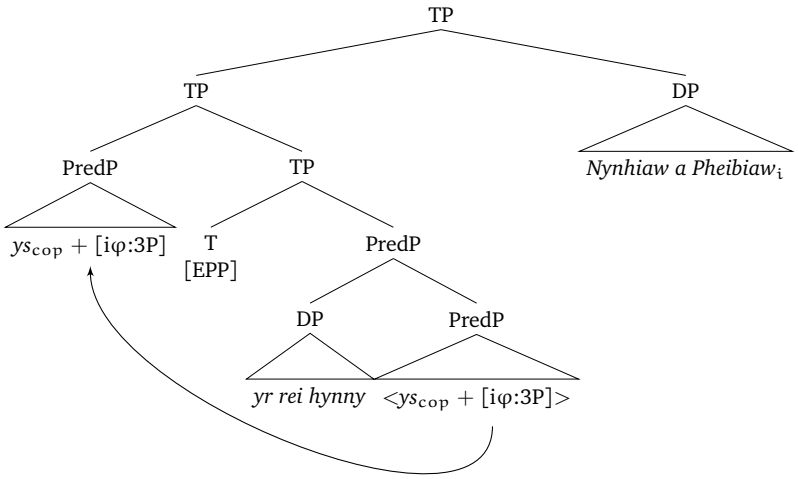
Middle Welsh also used a cleft structure containing a sentence-initial copula *ys* with a directly following anticipatory predicate like the *e* in Scots Gaelic in example (29). From an information-structural point of view, these constructions can be analysed as a clear focus of the (identificational) predicate. Considering the common background of the languages and further similarities in the copular system (like the distinction between substantive verbs and true copulas), it is tempting to extend Adger's (2011) analysis to these Middle Welsh constructions as well. The word order schema of these sentences is given in (30). It resembles that of the first

type of copular clauses with the order CPS. In these constructions with focussed identificational predicate complements, an extra 'anticipatory' predicate appears in the form of an agreeing pronoun, just like the *e* in the above example in Scots Gaelic. In (31), the anticipatory pronoun *hwy* agrees with the plural predicate identifying the names of the two oxen 'Nynnyaw and Peibiaw'. This predicate complement is focussed and adjoined to TP. The subject *yr rei hynny* 'those ones' remains *in situ* in the specifier position of the PredP. The derivation of example (31) would look like (32):

(30) Copula *ys* - anticipatory predicate - Subject - Focussed Pred. Complement

(31) *Ys hwy yr rei hynny, Nynhyaw a Pheibyaw*
be.PRES.3S they the ones DEM.P Nynniaw and Peibiaw
'Those are Nynniaw and Peibiaw'
(Lit. 'This is who those are NYNNIAW AND PEIBIAW.') (CO 598)

(32)



This particular sentence in the only example in Middle (or Old) Welsh showing agreement between the anticipatory predicate *hwy* 'third-plural pronoun' and the coindexed predicate *Nynhiaw a Pheibiaw* adjoined to TP. All other examples exhibit the third-person singular pronoun *ef*, which later merged with the predicate yielding the petrified focus marker *sef* (from copula *ys* + *ef*, see Chapter 7 for a diachronic analysis of the subsequent changes). It is difficult to draw any conclusions from one single example, but if agreement was indeed an (earlier?) option, then the focussed predicate complement *Nynhiaw a Pheibiaw* is likely to be extraposed (right-dislocated) to TP from its base-generated position as the complement of the predicate *ys*.[9] Agreement can then be achieved via two possible strategies:

[9]Alternatively, in a framework that does not permit rightward movement, *Nynhiaw a Pheibiaw* could be

1. The extraposed predicate complement leaves its φ-features behind, which are subsequently spelled out as the third person pronoun *hwy* that surfaces as the anticipatory predicate (cf. Trace Conversion as proposed by Fox (2002)).
2. In the Colon Phrase approach (cf. Koster (2000) and others) the nominal predicate actually contains a co-existing third-person plural pronoun *hwy* AND the nominal predicate *Nynhiaw a Pheibiaw*: in this case the pronoun simply moves up with the copula to the Specifier of TP

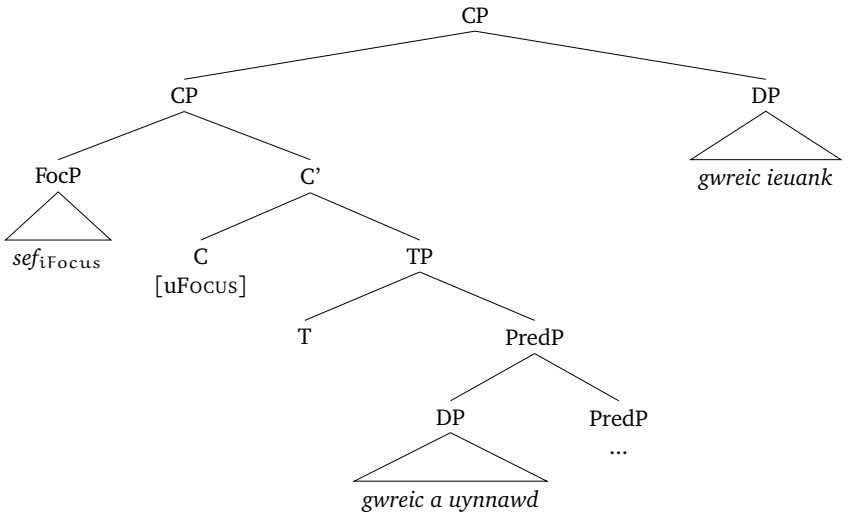
Either way, the semantic representation of the identificatory copular clause in example (31) remains the following:

(33) a. [[Pred']] = λx.**holds(*Nynhiaw a Pheibiaw*, *x*)**
b. [[those ones]] = **those ones**
c. [[PredP]] = **holds(*Nynhiaw a Pheibiaw*, those ones)**

As soon as the copula and anticipatory predicate pronoun merged to *sef*, it became a mere marker of focus merged in the C-domain to satisfy the [*u*FOCUS] on the C-head. An example of this is given in (34). The coindexed predicate will then be adjoined to CP, to receive the focussed interpretation, yielding a derivation like (35):

(34) *Sef gwreic a uynnawd gwreic ieuank*
sef woman PRT want.PAST.3S woman young
'That was the woman he wanted, a YOUNG woman.' (YBH 6)

(35)



moved to the Specifier of some higher phrase and then everything else could be moved leftward across *Nynhiaw a Pheibiaw*.

Extraposition of the predicate complement is string-vacuous in these configurations. This in turn, gave rise to possible reanalyses and other types of *sef*-constructions. In Chapter 7, I present a detailed account of the entire process of grammaticalisation including the reanalyses and extensions leading to possible new forms of *sef*-constructions in which the focussed interpretation and the association with identificatory predicates was lost. These innovated forms of the *sef*-construction included headless relative subjects, medial copular forms and adjunct phrases.

6.3.3 Conclusion Case Study I: Focus-Background

To conclude, in this section I presented a case study related to the information-structural notion of Focus, in particular a special case of focussed predicates. I argued that Adger & Ramchand's (2003) predicate-raising analysis of Scottish Inverted Copular Clauses can be extended to both the two word order patterns found in matrix copular structures and the inverted order in subordinate clauses in Middle Welsh. It can also explain the difference between 'true copulas' and substantive constructions with predicate marker *yn*.

In addition to this, Adger's (2011) analysis of clefts could be used as a starting point for the analysis of Middle Welsh identificatory copular clauses with focussed predicate complements: the *sef*-constructions. Raising of the entire predicate phrase to SpecTP (and possibly higher up to SpecCP in the end) can account for all types of *sef*-constructions, two of which were discussed in this chapter.

6.4 Case Study II: Topic-Comment

As presented in Chapters 3 and 4, there are different types of 'topics' in Middle Welsh. This section is dedicated to the puzzling agreement data of the Abnormal and Mixed sentences shown in section 6.2.3. It focusses on the synchronic derivation of sentences with subject-verb agreement. These sentences are argued to contain a base-generated aboutness topic in the left-periphery of the clause. Agreement is established with the coindexed subject in the form of a minimal pronoun (similar to referential *pro*). The derivation of these subject-agreement sentences called Abnormal Sentence crucially differs from their 'Mixed' counterparts without agreement. The lack of agreement in Mixed Sentences is, however, expected if these sentences involve reduced clefts with relative clauses, since Welsh relatives never exhibit agreement. I discuss the synchronic derivation of the Abnormal and Mixed sentences here and turn to their diachronic origin in Chapter 7.

Section 6.4.1 first presents the relevant data and introduces the crucial concept of the Complementarity Principle in Brythonic languages. Section 6.4.2 then continues to work out the details of the syntactic derivation. Finally, in section 6.4.3 I develop a comprehensive account of both agreeing and non-agreeing positive declarative sentences in Middle Welsh.

6.4.1 Topics: the data

Welsh, just like Breton, exhibits the Complementarity Principle according to which full DPs (usually[10]) do not show trigger agreement morphology, where pronouns (either an overt pronominal form or *pro* as in (37b)) do. This distinction can be observed in the verbal domain, but also with inflected prepositions, as shown in (36). Many prepositions can be combined with seven different possible person-number (and gender in 3SG) endings.

(36) a. *at Uatholwch*
to Matholwch
'to Matholwch' (PKM 32.7)

b. *attat titheu*
to.2S you.CONJ
'to you' (BR 12.20)

c. *y 'r llys*
to the court
'to the court' (PKM 11.13)

d. *idaw*
to.3MS
'to him' (PKM 1.3)

e. *y Arthur*
to Arthur
'to Arthur' (BR 19.4)

f. *wrthyf i*
to.1S me
'to me' (PKM 7.14)

g. *wrth y wreic*
towards the woman
'towards the woman' (PKM 7.24)

(37) a. *Y th law di nu y rodaf i.*
in 2S hand your now PRT give.1S I
'I now place in your hand.' (Gereint 644)

b. *ac y r neuad y kyrchyssant (pro).*
and to the hall PRT go.PAST.3P
'and they went to the hall' (PKM 59.22)

c. *Y hela y moch yd aeth y kynnydyon yna oll.*
to hunt.INF the pig PRT go.PAST.3S the huntsmen there all
'All the huntsmen went there to hunt the pig.' (CO 731)

d. *Yna y doeth y kennadeu.*
then PRT come.PAST.3S the messengers
'Then the messengers came.' (PKM 27.12)

Middle Welsh had an elaborate pronominal system consisting of dependent and independent pronouns with different forms according to their function.[11] In contexts with agreement, as shown by the examples above, the dependent affixed form of the pronoun can optionally be spelled out as the 'echo' pronoun as in (36f). Tables 6.1 and 6.2 show the range of forms (based on Borsley et al. (2007)):[12]

[10] In some cases, collective DPs do trigger plural agreement.

[11] Many of these forms trigger different types of consonant mutation, like soft, nasal or aspirate mutation.

[12] At first glance there seem to be many ambiguous forms consisting of a single letter, e.g. *e* for third-person Accusative & Genitive singular and plural. However, each of these trigger different kinds of initial consonant mutation of the verbs and nouns directly following them. I have left out these details in the present table, because the mutation effects complicate the representation and are not always found in Middle Welsh orthography anyway. For the present study of agreement in Abnormal and Mixed Sentence, the distinction is not relevant.

	Affixed conjunctive	Accusative	Genitive	Affixed (echo)
I	*inheu*	*'m*	*vy / 'm*	*(u)i*
you (sg.)	*ditheu*	*'th*	*dy /'th*	*di/ti*
he	*ynteu*	*'e/s*	*y /'e*	*ef*
she	*hitheu*	*'e/s*	*y /'e*	*hi*
we	*ninheu*	*'n*	*yn/'n*	*ni*
you (pl.)	*chwitheu*	*'ch*	*ych/'ch*	*chwi*
they	*wynteu*	*'e/s*	*eu/'e*	*wy(nt)*

Table 6.1: Dependent pronouns: conjunctive, accusative, genitive and affixed

	Independent	Conjunctive	Reduplicated
I	*mi*	*minheu*	*miui*
you (sg.)	*ti*	*ditheu*	*tidi*
he	*ef*	*ynteu*	*efo*
she	*hi*	*hitheu*	*hihi*
we	*ni*	*ninheu*	*nini*
you (pl.)	*chwi*	*chwitheu*	*chwichwi*
they	*wy(nt)*	*wynteu*	*wyntwy*

Table 6.2: Independent pronouns: 'normal', conjunctive and reduplicated forms

Recall the aberrant plural inflection of the verb in the so-called Abnormal Sentences in Middle Welsh with preverbal full DP subjects. The Mixed Sentences, on the other hand, also feature preverbal subjects, but in these constructions even pronouns do not trigger agreement of the verb.

Abnormal Sentences:

(38) a. *A 'r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
'And the nobles came together' (PKM 90.27)

b. *Ac ef a welei lannerch yn y coet.*
and he PRT see.PAST.3S clearing in the forest
'And he saw a clearing in the forest.' (PKM 1.13-14)

c. *Ac ni a gredwn iddo.*
and we PRT believe.1P in.3MS
'and we believe him' (b1588 - Mat. 27.42)

d. *a mi a fyddaf eu Duw hwynt.*
and I PRT be.FUT.1S 3P God 3P
'And I will be their God.' (b1588 - 2 Cor. 6.16)

Mixed Sentences:

(39) a. *Mi a 'e heirch.*
I PRT 3FS seek.3S
'(it is) I who seek her' (WM 479.24)

b. *y guyr hynny a y godiwawd*
the men those PRT 3FS overtake.PAST.3S
'Those men overtook her.' (PKM 32.20-21)

c. *Kimri a oruit*
Welshmen PRT prevail.FUT.3S
'(is shall be) the Welsh that shall conquer' (BBC 59.4)

d. *os tydi yw Crist Mab Duw.*
if you.REDUP be.PRES.3S Christ son God
'...if you are Christ, son of God' (b1588 - Mat. 26.63)

The formal difference between the two can only be observed in sentences with preverbal plural DP or pronominal subjects. As pointed out in section 6.2.3 above, a 'purely' pragmatic distinction between the two as topicalisation with agreement vs. focalisation without agreement is difficult to maintain. There are examples with focussed reduplicated pronouns in agreement contexts, as shown in (40a), but there are also examples of preverbal plural DPs without focus or agreement, as in (40b) (see also, amongst others, Poppe (2009)).

(40) a. *Miui hagen a uydaf gyfarwyd ywch*
I.EMPH however PRT be.1S familiar to.2P
'I, however, will be familiar to you.' (Culhwch 899)

b. *Kennadeu a aeth at uranwen.*
messengers PRT go.PAST-3S to Branwen
'Messengers went to Branwen.' (PKM 40.1-2)

These examples give rise to a number of questions. What is the difference between the Abnormal and Mixed sentences rendering these superficial agreement patterns (and the lack thereof). Do topic or focus or any other information-structural features play a role if both options (with and without agreement) are grammatical? If so, how do they influence the respective syntactic derivations of these sentences?

Some of the above questions are addressed in Chapter 7 when their diachronic syntax is taken into account. In this section, I focus on the question concerning the underlying syntax of the 'Abnormal' sentences with full DP subjects and verbs with third-person plural inflection (as in (38a)). How can these be derived in a language that usually adheres to the Complementarity Principle?

6.4.2 Topics: the analysis

There are - to my knowledge - two relevant analyses of sentences with plural agreement as in (38b): Willis's (1998) A-movement approach and Tallerman's (1996) CP-adjunction approach. As pointed out in section 6.2.3 above, both of these meet with difficulties. In this section, I first discuss the details of each of their

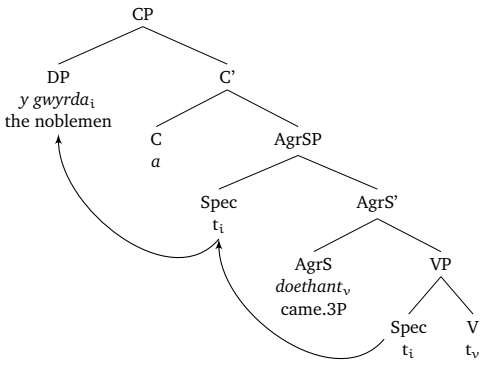
analyses with their respective advantages and disadvantages. Then I proceed to propose an alternative way of deriving Abnormal Sentences like (41):

(41) *A ’r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
‘And the nobles came together’ (PKM 90.27)

Willis’s (1998) approach: A-movement through AgrSP

Willis (1998) proposes a movement analysis for both Mixed and Abnormal sentences. The crucial difference in agreement arises because in Abnormal sentences (with agreement), the subjects moves through SpecAgrSP where it triggers agreement inflection of the verb. Although this approach makes the right prediction for Abnormal sentences with pronominal subjects, it fails to account for the third-person plural agreement in sentences with A-moved full DP subjects, since full DP subjects normally do not trigger agreement (cf. the Complementarity Principle). Willis’s (1998:93) derivation of the Abnormal Sentence is as follows:

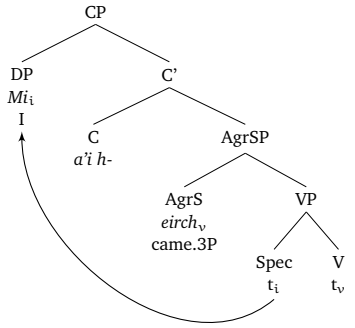
(42) **Abnormal Sentence**



If the operation Agree operates in the same way as it would if the subject were preverbal, plural inflection is still unexpected because full DPs never trigger agreement under the Complementarity Principle. The difference must then lie in the nature of the trace or copy of the full DP left in SpecAgrSP. Willis (1998) has to assume (though this is not made explicit in his proposal) that this copy *can* somehow trigger plural inflection. If the copy of the full DP is ‘reduced’ (cf. the Reduced Copy Theory, van Koppen (2007)) or ‘converted’ (cf. Trace Conversion, Fox (2002)) to a pronoun, this could perhaps indeed account for the plural inflection on the verb.

In Mixed Sentences without subject-verb agreement, Willis (1998:92) assumes the subject is fronted via A’-movement, skipping the A-position in AgrSP as shown in (43):

(43) **Mixed Sentence**

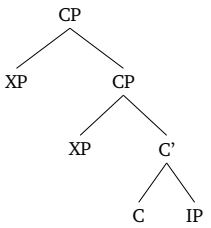


In addition to agreement with plural DPs in Abnormal Sentences, the lack of agreement in Mixed sentences has to be accounted for. Willis (1998) stipulates that Mixed sentences do not exhibit subject-verb agreement, because the subject does not move through SpecAgrSP. Fronting of the subject in Mixed sentences is then A'-movement, skipping the A-position in AgrSP. Some extra mechanism is thus required to prevent A'-movement through a position where it can trigger subject-verb agreement.

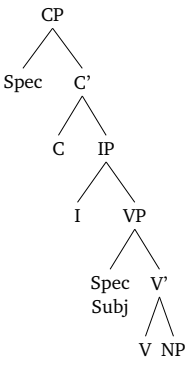
Tallerman's (1996) approach: adjunction to CP

Tallerman (1996:111), on the other hand, proposes a derivation for Abnormal Sentences where the topic occupies a position adjoined to CP:

(44) **Abnormal Sentence**



(45) **Mixed Sentence**



This approach correctly predicts the impossibility of Abnormal Sentences in embedded clauses, because embedded clauses are s-selected by lexical heads (following the Adjunction Prohibition as formulated by McCloskey (1992:11)). Agreement in Abnormal Sentences is not with the topic adjoined to CP, but with the null pronominal subject *pro* (residing in SpecTP or SpecAgrP presumably, although this is not specified).

Mixed Sentences are clefts, according to Tallerman (1996:107) exactly parallel to that found in wh-questions and relative clauses. As such they involve A'-movement to the specifier of the lower CP (CP_2 in a recursive CP configuration) and do not exhibit agreement, because the empty NP in the canonical subject position is a wh-trace. This analysis is parallel to that proposed by Borsley and Stephens (1989b) for Breton topicalisation structures that also do not show subject-verb agreement. An example of the basic structure for Mixed Sentences is given in (45):

One major difficulty with deriving abnormal sentences via adjunction of the topic to CP is that it wrongly predicts multiple topicalisation for Middle Welsh. As Borsley et al. (2007) point out, this is in fact not what we find. Although it is possible to find sentences with multiple constituents preceding the inflected verb, only one of those can be an argument. All other preverbal elements must be non-argument adverbials (Borsley et al., 2007:293). The single (topical) argument determines the form of the preverbal particle: *a* for subjects or objects or *y* for adjuncts (prepositional phrases or adverbs). Subject and objects can never occur in preverbal position in the same sentence, unless one of them is clearly left-dislocated (in which case a resumptive can be found as well).

Further observations in the data

My proposal for this agreement puzzle is based on an additional observation in the Middle Welsh data concerning the pronominal system. In Modern Welsh there is a clear distinction both in form and distribution between so-called 'strong' (independent) and 'weak' (dependent) pronouns. The data presented in (46) is from Modern Welsh (Borsley et al., 2007:213-214) and it shows the clear difference in grammaticality. There is no reason to assume the distribution was any different in Middle Welsh, because the ungrammatical forms in (46) and (47) are never found (while there are plenty of examples of the grammatical ones, i.e. plenty of 'missed opportunities').

(46) a. *Fi welodd y ceffyl.*
I see.PAST.3S the horse
'It was I that saw the horse.'
b.**Gwelais fi 'r ceffyl.*
see.PAST.1S. I the horse
('I saw the horse.')

(47) a. *Gwelais i 'r ceffyl.*
see.PAST.1S I the horse
'I saw the horse.'
b.**I welodd y ceffyl.*
I see.PAST.3S the horse
('It was I that saw the horse.')

The weak or 'echo' pronoun *i* 'I' in (47) can only be found in the context of agreement inflection, like the first-person singular inflection of the verb. This is why they are characterised as 'dependent' affixes in table 6.1 above. These

echo pronouns (both conjunctive and affixed) are found in Middle Welsh in many agreement contexts (represented in bold in the following examples):

(48) a. *genhyt* ***ti***
with.2S you
'with you' (WM 121.20)

b. *A phaham y gouynhy* ***di****, Arglwyd?*
and why PRT ask.2S you Lord
'And why do you ask, Lord?' (PKM 61.24)

c. *amdanaf* ***i***
about.1S me
'of me' (RM 87.27)

d. *E dodeis* ***inheu*** *ar gynghor uy gwlat ...*
PRT put.PAST.1S I.conj on council 1S country
'I referred to the council of my country...' (PKM 36.4)

Crucially, however, these optional echo pronouns are never found in abnormal sentences with preverbal subjects (see also Willis (2007a)). In the following examples, we see the person-number inflection, but *not* the echo pronoun:

(49) a. *Mi a af y ymwelet a 'r pryf.*
I.ind PRT go.1S to visit.INF with the worm
'I will go to encounter the Worm.' (WM 161.13)

b. *Mi a gredwn ac a dywedwn y taw ti oed Bown.*
I PRT believe.1S and PRT say.1S PRT FOC you be.PAST.3S Bown
'I would believe and say that thou wert Bown' (YBH 24.1541)

c. *Ti a welaist hyn*
you PRT see.PAST.2S that
'You saw that.' (b1588 - 1 Sam. 19.5)

d. *Ti a e keffy yn llawen.*
you PRT 3FS get.2S PRED glad
'You will get it gladly' (BM 11.27)

e. *A chwi a uydwch ar y ford yn hir*
and you PRT be.2P on the way PRED long
'And a long time will you be upon the road.' (PKM 45.2)

Although this remains an argument *ex silentio*, there is no reason to assume that these optional pronouns are *incidentally* always absent in abnormal sentences with preverbal subjects. If this indeed no coincidence, we can distinguish four different surface forms of φ-features in Middle Welsh:

1. φ-inflection on verbs and prepositions ("*pro*")
2. dependent 'weak' or 'echo' pronouns
3. independent 'strong' pronouns
4. full lexical DPs (carrying interpretable φ-features)

Strong pronouns and full lexical DPs exhibit similar distributional patterns according to the Complementarity Principle. They differ only in that strong pronouns never occur in immediate post-verbal position. Weak pronouns do occur in this dependent position, but only in the context of overt inflection on preceding verbs or prepositions (and even then the dependent pronouns are optional). The φ-inflection itself in turn looks like configurations with empty *pro* often found in null-subject languages (NSLs). Middle Welsh is also a null-subject language, but it does allow the optional spell-out of the echo pronoun in (dependent) agreement contexts. Examples of each of the above-mentioned instantiations of φ-features are given in (50):

(50) a. *Kythreulyeit llawer a ’m* ***kylchynassant****.*
demons many PRT me surround.PAST.3P
‘Many demons have surrounded me.’ (‘*pro*’ - B x 54.9)

b. *Eissoes negessawl wyf* ***i*** *y gan Arthur attat.*
yet messenger be.1S I from Arthur to.2S
‘Yet I am a messenger to thee from Arthur.’ (‘weak’ - WM 143.11)

c. *A phoet* ***euo*** *a ’th danuono drachevyn*
and be.SBJ.3S he.RED PRT you send.SBJ.3S back
‘and may it be he who shall send thee back.’ (‘strong’ - SG 15.15-16)

d. *Yna yd aeth* ***kennadeu*** *yn y erbyn.*
then PRT go.PAST.3S messengers to 3MS against
‘Then messengers went to meet him ’ (full DP - PKM 85.2)

Assuming φ-features are the underlying cause for the observed agreement, let us now turn once more to the patterns usually found in Middle Welsh.[13] For the explanatory purposes, I for now use the traditional denotation of Topic and Focus for pre-verbal subjects of Abnormal and Mixed sentences respectively:

Full φ-agreement between subject and verb:
- Nominal Topic - Verb + agreement
- Pronominal Topic - Verb + agreement
- XP - Verb + agreement - Weak subject pronoun

Default third-person singular agreement:
- XP - Verb 3sg - Full DP subject
- Nominal Focus - Verb 3sg
- Pronominal Focus - Verb 3sg

The question is now which of the four above-mentioned φ-feature patterns is involved in each of these observed agreement patterns. All other things being equal, the crucial variables for the sentences with Topic or Focus seem to be the type

[13]There are some ‘occasional’ exceptions to these observations. Many of these have to do with singular or plural nouns that can have a collective interpretation as well. I turn to some of those examples below. See also Nurmio and Willis (2016) for details about the problematic number category in Middle and Early Modern Welsh noun phrases.

of DP (pronoun or full DP) and the syntactic derivation of the sentence-initial subject (internal or external merge). If nouns and pronouns are probed in the same way, there are still four logical possibilities: both topic and focus are derived by internal merge, both by external merge, one by internal and the other by external merge or vice versa. If both the focussed and topicalised sentences are derived by internal merge, we have to assume (like Willis (1998)) the lack of agreement is due to the focussed constituent 'skipping' the position where it can agree with the verb (SpecAgrSP for Willis (1998)). In addition to that, we have to assume some form of Trace Conversion (cf. Fox (2002)) allowing us to treat the trace/copy of the moved full DP differently from the original DP somehow so that it *can* cause number agreement. As pointed out above, this might be possible, but the amount of extra assumptions in this approach make it worth exploring other options.

6.4.3 Topics: a comprehensive account

If we take the Complementarity Principle as a starting point, agreement with nominal topics (as in (51a)) and the lack of agreement with focussed pronouns (as in (51b)) is unexpected. In this section, I zoom in on two alternative approaches to this conundrum: derivation by external merge for both topics and foci and a combined approach of internal merge for foci and external merge for topics.

(51) a. *A 'r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
'And the nobles came together' ('Topicalised' - PKM 90.27)

b. *Mi a 'e heirch.*
I PRT 3FS seek.3S
'(it is) I who seek her' ('Focalised' - WM 479.24)

Following Willis (1998), I assume the dedicated pre-verbal position for both topical and focussed constituents is the specifier of C. From a cross-linguistic perspective, this is not an odd assumption. As pointed out in section 6.2.1 above, most topicalisation and focalisation structures involve constituents in the C-domain (in SpecCP or in the specifier of, for example, a topic or focus projection in a proliferated C-domain). Further evidence from Welsh comes from agreement with the complementiser or pre-verbal particle. The element in SpecCP can agree with the complementiser to yield its correct surface form: *a* following arguments, *y* following adjuncts. Assuming SpecCP as the dedicated pre-verbal position furthermore makes the correct prediction that multiple topics are impossible.[14]

If we want to avoid the complications of moving the subject through an agreeing position to SpecCP, we could assume these topical subjects are base-generated instead. If the topic is base-generated in the C-domain, however, agreement still needs to be explained. I propose agreement can be realised with a minimal pronoun that does not possess any φ-features (and therefore cannot be spelled out overtly, see below), but is co-indexed with the base-generated topic (via predication with a

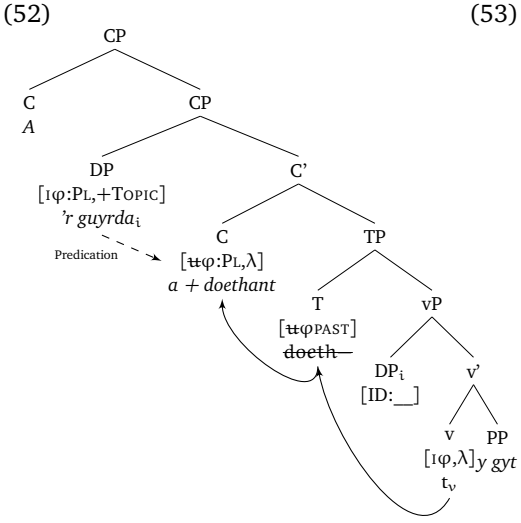
[14]Unless the C-head projects multiple specifiers, which I assume not to be the case in Middle Welsh.

λ-feature as I will explain in detail in the next section). This minimal pronoun is the equivalent of the referential *pro* as postulated by Frascarelli (2007) for sentences with base-generated aboutness topics. This explains the subject-verb agreement, even if it is co-indexed with a full DP topic, which under the Complementarity Principle would not trigger subject-verb agreement moving through the T-domain. This type of analysis is in fact similar to the one advocated by Tallerman (1996) for topics in Abnormal Sentences. The main difference is that she suggested base-adjunction to CP resulting in the incorrect prediction that there could be multiple topics in Middle Welsh. If the topic is base-generated in SpecCP instead, this poses no problems, because multiple topics or focussed constituents are not predicted to coexist in SpecCP. A derivation of this kind is presented in (52) below.

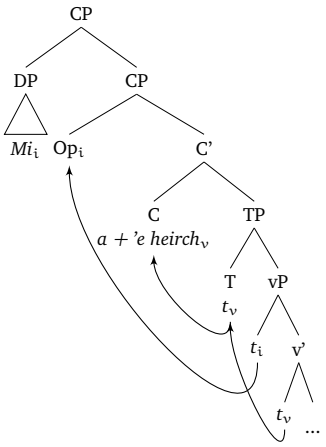
Middle Welsh sentences with pre-verbal subjects *without* agreement (as in (53)), the so-called 'focalised Mixed Sentences', are then simply analysed in exactly the same way as relative clauses (from which they originate, see Chapter 7 for a diachronic analysis). Sentences like (51b) are reduced clefts with an externally headed relative. The lack of agreement is expected in the same way it is expected in relative clauses: empty operators can bind the variable in subject-position, but do not license agreement.

Abnormal Sentence **Mixed Sentence**

(52)



(53)



Deriving the Abnormal Sentence with agreement

As noted above, on the basis of prosodic evidence from Aboutness Topics in Italian, Frascarelli (2007) advocates an analysis with a base-generated topic in the C-domain that is coindexed with a referential *pro* lower down in argument position in

the clause. For Middle Welsh, of course it is impossible to provide similar evidence based on differences in prosodical patterns for Aboutness and Familiar topics. From the context, however, these subjects in sentence-initial position do seem to (re)introduce the topic that the sentence is about. Furthermore, if the topic/subject stays the same, a silent or null topic (that could be analysed as *pro*) is found:

(54) *Peredur*$_i$ *a ordinawd y varch ac (pro*$_i$*) a 'e kyrchawd yn*
Peredur PRT spur.PAST.3S 3MS horse and (pro$_i$) PRT 3MS attacked PRED
llityawcdrut (...) ac (pro$_i$*) a 'e gwant dyrnawt gwenwyniclym ...*
angry (...) and (pro$_i$) PRT 3MS hit.PAST.3S blow incisive ...
'Peredur spurred his horse and attacked him angrily (...) and struck him an incisive blow...' (Peredur 41.27-33)

Willis (1998) argues that the silent topic in Middle Welsh is not a *pro*, however, but an empty topic operator. Evidence for the operator analysis comes from coordinated sentences with null objects, rather than null subjects. Null objects in Welsh can only be found in the context of agreeing object clitics (cf. the first type of φ-features above that are only found in the context of inflected verbs or prepositions). Examples of null objects in the second conjunct are exceedingly rare, but they do exist, as Willis (1998:126) points out:

(55) *Ac yna y kanhatwyt y Chyarlys bot yn Ager* ***gawr***$_i$ *Ffarracut y*
and then PRT reported.IMPERS to Charles be.INF in Ager giant Fferracud his
enw o genedyl Goliath ac (pro$_i$*) a dathoed o eithauoed Sirya*
name from race Goliath and (pro$_i$) PRT come.PLQPF.3S from extremes Syria
ac (pro$_i$*) a anuonassei Amilald vrenhin Babilon (t*$_i$*) y ryuelu (...)*
and (pro$_i$) PRT send.PLQPF.3S Amilald king Babilon to make.war (...)
'And then it was reported to Charles that there was in Ager a giant named Fferracud from the race of Goliath, and (he) had come from the ends of Syria and Amilald King of Babylon had sent (him) to make war (...)' (YCM 25.12-15)

If the null object is a topic operator, rather than a *pro* (as indicated in the above example), it can bind a variable in object position. In this type of configuration with an empty operator, according to Willis (1998:127) there is no need for agreeing object clitics. Lack of agreement with operators is indeed expected from a cross-linguistic perspective. Problems arise, however, in sentences with silent topics that *do* exhibit agreement. From a theoretical perspective it would be inelegant to say the least to postulate an empty operator for a silent topical object alongside a referential *pro* for a silent topical subject. Postulating referential *pro* in null-subject languages is in fact undesirable from a Minimalist point of view as well. Most recent analyses of NSLs involve either deletion of the subject after it satisfies EPP (cf. Holmberg (2005), Sheehan (2007) and Roberts (2009)) or a hybrid approach in which either the verb or the subject can satisfy the EPP (after which the subject is deleted as well, see Sheehan (2015) or, in a somewhat different version, Biberauer and Richards (2006)). In both these types of analyses of NSLs, referential *pro* is

removed from the system. Postulating it here for Middle Welsh silent topics would thus be undesirable from this perspective as well.

The type of empty element we need is a pronoun with a defective feature set. It should be able to be bound, coindexed or 'identified' in the derivation by the sentence-initial Aboutness topic, but it does not have a separate set of φ-features of its own. In a semantic account as propagated by Kratzer (2009), these kinds of bound variables are 'minimal pronouns'. According to Kratzer (2009:187), these include local fake indexicals, relative pronouns, reflexives and PRO. The features they are missing can be acquired in the course of the derivation from verbal functional heads that carry λ-operators to bind them in a 'predication' configuration.

This 'minimal pronoun' is very similar to the pronoun without φ-features but with Identity features that they mark as '[ID:_]'. Adger and Ramchand (2005) propose to explain the difference between structure with internal and external merge in Scots Gaelic (and beyond). The (non)identity effects on which their account is based are difficult to test in Middle Welsh, because Welsh no longer has case morphology and there is no definiteness agreement between prepositions and their complements. They furthermore predict (correctly for Scots Gaelic) that multiple wh-questions are impossible, because of the clefted nature of the copular and wh-constructions. Although Welsh questions look superficially similar to the Scots Gaelic clefts and are originally based on clefts, multiple wh-questions are possible in Modern Welsh[15]. In the same way, evidence from non-identity effects in parasitic gaps cannot be readily found in Middle Welsh or points towards the opposite direction (cf. Sproat (1985) for the possibility of parasitic gaps in Welsh). In short, despite their superficial similarities and their common background, Adger & Ramchand's (2005) analysis cannot be readily transposed to Middle Welsh (see also Willis (2011b) for an analysis of Modern Welsh relative structures that faces the same difficulty): Welsh and Scots Gaelic diverge too much.

This does not mean, however, that their basic intuition about the featural differences in 'minimal pronouns' and resumptives is wrong or that the analysis of those pronominal elements being bound by a λ-operator on a functional head cannot be implemented in Middle Welsh at all (they actually implement it in Modern Welsh relatives in the same paper). Their approach furthermore allows for cross-linguistic variation: there are different types of Merge (i.e. base-generation with co-indexation), but in addition Move (i.e. internal merge) is still an option (this is in fact the predicted strategy for languages like English with over relative pronouns). For now, I leave this as an option that is worth exploring in future work. I only take their notion of 'Identity' and the featural representation '[ID:_]' for bound variables that are 'minimal pronouns' in Kratzer's sense, because this is the exact type of variation in φ-features that *could* account for the variation in Welsh agreement patterns.

[15] Since these types of questions are rare in general, it is difficult to find examples in the medieval data. It is possible that they did exist in Middle Welsh, however, which would make a similar analysis of Scots Gaelic and Middle Welsh impossible.

The implementation of this type of empty category, the 'minimal pronoun' with the identity category, but without inherent φ-features, proceeds as follows. I first assume the verb can enter the derivation with interpretable φ-features. Recall from the above-mentioned discussion on the hybrid approach for the EPP on T, that postulating φ-features or, in other words a D-feature on V was already necessary to account for probing of the verb in null-subject languages (see, amongst others, Biberauer and Richards (2006) and Sheehan (2015)). The topic is a DP with a full set of φ-features, base-generated in SpecCP (as postulated above). The empty category it binds, however, enters the derivation without φ-features in SpecvP. Because it has no φ-features, it cannot be probed by T. It is, however, bound by the λ-operator on the verb to realise coindexation with the topic in SpecCP.

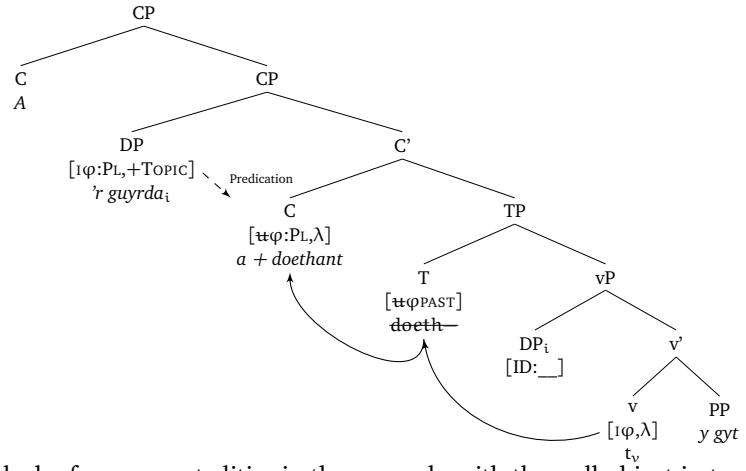
The derivation then proceeds in the same way as described for passives or unaccusatives in null-subject languages (or any other configuration in which SpecvP is not occupied by a phrase bearing φ-features). The verb thus moves to T and is subsequently probed by C. Following Adger and Ramchand (2005) and Kratzer (2009), I assume C can carry a λ-operator. Under the principle of predication - as formulated by Kratzer (2009) - the φ-features of the DP in the specifier of CP are then united with those of the C-head.

(56) PREDICATION (Specifier-Head Agreement under Binding)
When a DP occupies the specifier position of a head that carries a λ-operator, their φ-feature sets unify. (Kratzer, 2009:196)

These features can now be spelled out as the inflection on the verb agreeing with those of the topic in SpecCP. This pronoun, like any regular referential pronoun, does enter the derivation with dedicated φ-features '[ID:φ]'. If it had entered the derivation in SpecvP as a true subject without a topic feature (as in the adjunct-initial examples above), its set of features would be the same. As we saw in the null-subject derivations above, these φ-features would be incorporated into the verb and be optionally spelled out as a weak or echo pronoun. Topical pronouns, however, look different because they are spelled out as 'strong pronouns'. This is not due to a featural difference, however, but solely to their position preceding the complementiser *a* and the verb: they cannot be incorporated and spelled out as (weak) clitic pronouns. Since the difference between weak and strong pronouns is only related to their surface position, postulating the exact same feature set for both is an elegant solution allowing us to treat them uniformly, strictly in accordance with their exact same semantic properties (i.e. unlike their 'minimal pronoun' counterparts, they are referential). A sample derivation of this kind based on example (57) is again given in (58).

(57) *A 'r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
'And the nobles came together' ('Topicalised' - PKM 90.27)

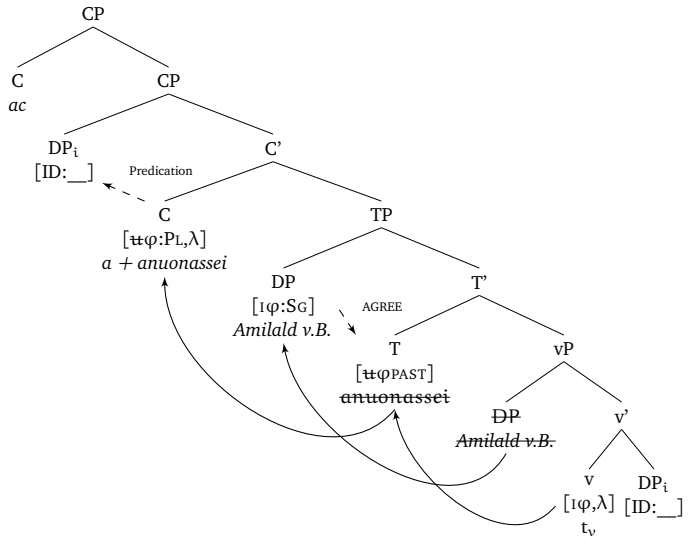
(58)



The lack of agreement clitics in the example with the null object in topic position can now also be straightforwardly explained. The φ-features in that configuration are also not there, because just like the minimal subject pronoun in the above example, the minimal object pronoun does not carry φ-features when it enters the derivation. It carries an Identity feature so that it can be bound by the topic via the λ-feature on the transitive verb: [ID:_]. The derivation of the last part of the long coordinated sentence presented above is now shown in (60):

(59) *ac (pro$_i$) a anuonassei Amilald vrenhin Babilon (t$_i$) y ryuelu*
and (pro$_i$) PRT send.PLQPF.3S Amilald king Babilon to make.war
'and Amilald King of B. had sent (him) to make war' (YCM 25.12-15)

(60)



6.4.4 Conclusion Case Study II: Topics

In this section I presented a base-generated approach to derive sentences with initial (aboutness) topics. These topics appear to trigger subject-verb agreement, yielding the 'Abnormal Sentence' in Middle Welsh. These sentences are not just 'abnormal' because of their verb-second word order, but mainly because sentence-initial full DP subjects often agree with the verb, which is unexpected in a language that is commonly assumed to abide by the Complementarity Principle. According to this principle, also observed in Breton, only pronouns cause agreement on verbs or inflected prepositions, while full DPs always co-occur with default third-person singular inflection or uninflected prepositions.

Given the Copy Theory of Movement, it is difficult to explain this subject-verb agreement with plural full DP subjects in sentence-initial topic position. Extra assumptions have to be made to convert the copy of the DP into a pronoun-like element that *can* trigger agreement when it moves through a canonical subject-position. In focussed constructions - which never show agreement, not even with pronouns - further assumptions are necessary to 'prevent' subject-verb agreement with pronominal subjects. Therefore, alternatives with base-generated topics and the coindexed 'minimal pronoun' were explored. Based on this data and the further observation that topicalised subjects never co-occur with the spell-out of 'weak' pronouns, I presented a four-way overview of the occurrance of φ-features in Middle Welsh:

1. agreement inflection only (on verbs or prepositions)
2. weak or echo pronouns (only in positions following agreement inflection)
3. strong pronouns (NOT following agreement inflection)
4. full DPs (NOT causing agreement on verbs or prepositions)

I argued that agreement in the first context is in fact just the spell-out of the φ-features of the verb. The verb enters the derivation with interpretable φ-features and λ-binders on functional heads in the derivation can establish the link with the bound variable. This bound variable is a minimal pronoun in the sense that it enters the derivation without φ-features. It only carries an identity feature [ID:__] that allows it to be bound by a λ-operator on a functional head, e.g. the verb or the C-head. This allows us to not only explain the observed agreement patterns in topicalised sentences with full DPs, it also offers a solution to the lack of agreement clitics with topicalised objects. If the null object, just like the null subject, enters the derivation without φ-features, those features cannot appear as clitics on the verb. In these configurations, the verb agrees with the subject DP rendering the usual agreement pattern. Subject-verb agreement inflection on the verb with topicalised subjects is a reflection of the interpretable φ-features the C-head receives from the DP in its specifier via predication. Finally, this way of looking at the pronominal system solves the awkward distinction between strong and weak pronouns in Welsh. These pronouns can now be considered the same, both carrying φ-features, when they enter the derivation. They only differ in terms of their position at spell-out: weak pronouns are incorporated φ-features and strong pronouns are independent.

Mixed Sentences *without* subject-verb agreement were analysed involving an

operator just as in Welsh relative clauses. These operators always trigger default third-person singular agreement. In Chapter 7 I present a diachronic analysis of the Mixed Sentence arguing it originates in clefts with relative clauses.

An afterthought on examples with 'messy agreement'

Although the above-sketched agreement patterns occur with such regularity, the Complementarity Principle is not a full-proof generalisation in Middle Welsh. There are exceptional cases of plural noun phrases triggering agreement even when they follow the verb.[16] One of these exceptional cases is shown in (61):

(61) *e uelly e dianghassant e gelynyon wedy caffael eu golwc*
thus PRT escape.PAST.3P the enemies after get.INF 3P sight
'thus the enemies escaped having received their sight' (B ix 337.20-21)

Apart from the fact that these sentences involve subjects that could be considered collectives or that involve numeral phrases, I have no ready solution for these now. In Chapter 7 I discuss these cases again in the light of their diachronic background.

There are some other cases of abnormal sentences that show challenging agreement patterns. Examples of those can occasionally be found in coordinated structures. See (62) for coordinated DPs. The first-person inflection on (62) is not immediately expected under the currently-adopted base-generation approach:

(62) *Miui a 'm bydin a ruthraf udunt hwy.*
I.RED and 1S host PRT hurry.1S to.3P them
'I and my host will attack them.' (HGK 15)

In the base-generation analysis described above, for a sentence like (62) we would have to assume that it is the first-person singular φ-feature that transfers to the C-head under predication as soon as the coordinated topic phrase is merged in SpecCP. It is not so clear why these features would be preferred over those of the second conjunct (or those of the conjoined phrase combined). Welsh always exhibits first-conjunct agreement with conjoined noun phrases and this is usually analysed as such because in a VSO order, the first conjunct is the closest, but this is clearly not the case here.

Would a movement analysis of these constructions not be better? If there were movement, the trace/copy of the conjoined phrase would have to be converted to something that can cause φ-agreement, but, crucially, cannot be spelled out as the weak pronoun. After Agree takes place with the subject, the copy of the dislocated subject phrase thus has to be converted to the minimal pronoun, a DP with [ID:_] features, postulated above. This would not explain plural agreement in sentences with dislocated plural noun phrases, however, because if Agree takes place first, plural inflection is unexpected. For these sentences, we would first again have to

[16] To my knowledge, there are no examples in Middle Welsh of (plural) inflection on prepositions preceding full noun phrases, however.

assume the copy is converted to some empty category that behaves like third-person plural pronoun, but without the optional (and quite unexpected) spell-out of the echo pronoun. Then only after this conversion, the verb Agrees with this empty category. This order of events seems undesirable: movement (or re-merge) and Agree should go together on current minimalist assumptions.

The only way to 'save' this movement derivation would be to postulate a spell-out rule stating full DP noun phrases always agree with and thus transfer their φ-features to their probing functional heads, but plural agreement is simply not spelled out if the DP immediately follows the inflected verb (or preposition), yielding the Complementarity Principle. If we do not want to resort to such a spell-out rule, a movement analysis cannot readily explain the facts and thus the base-generated analysis should be adopted. In this case, agreement with the first-conjunct is not a linear, but a structural requirement: only the φ-features of the first conjunct are transferred to C. Since the Welsh phrase *a'm bydin* does not necessarily mean 'and my host', but can also be a prepositional phrase 'with my host', this preference for the head noun is not unexpected.

A final category of difficult cases of agreement in Middle Welsh are presented by coordinated CPs[17] as shown in (63) and (64). These are also discussed by Poppe (2009:257), but he does not provide any syntactic analysis. The default third-person singular inflection following the plural noun phrases is unexpected if these phrases are abnormal sentences in which agreement usually occurs. They could be analysed as collectives or simply as mixed sentences without agreement. But then the third-person plural agreement in the second conjunct following the dropped topic is unexpected again.

(63) *Y gwyr a wiscawd amdanunt ac a nessayssant attunt.*
the men PRT dress.PAST.3S on.3P and PRT go.PAST.3P to.3P
'The men armed themselves and went towards them.' (PKM 29.22-23)

(64) *Y guyr hynny a 'y godiwawd ac a ouynyssant idaw*
the men these PRT 3MS overtake.PAST.3S and PRT ask.PAST.3P to.3MS
'These men overtook him and asked him...' (PKM 32.20-21)

These sentences seem highly problematic for any approach that attempts to give a uniform analysis of agreement patterns. Equivalent sentences in English can (optionally[18]) be pronounced with an overt (unstressed) pronoun *they* in the equivalent: 'The men armed themselves and (they) went towards them.'. The dropped topic in the second conjunct (the optional 'they' in English) has to carry plural φ-features. But if it gets those φ-features from the topic of the first conjunct, why is there no agreement in the first conjunct?

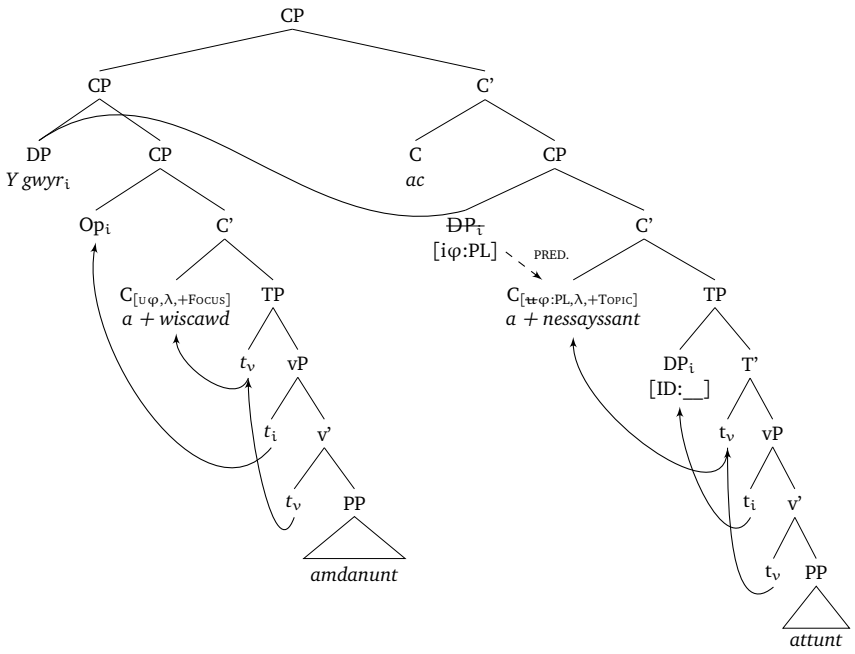
[17] The presence of the complementiser *a* provides evidence for a coordinated CP analysis. VP coordination is thus excluded.

[18] This is not necessarily a case of true optionality in the sense of Biberauer and Richards (2006). A British English informant tells me that adding the overt pronoun indeed has the same meaning, but it could make you wonder for a short while if it is perhaps *not* coreferenced with 'the men'. True optionality would then only be found in contexts in which this is made explicit somehow.

Recall that Mixed Sentences without agreement were analysed with an empty operator as relative clauses, because the relative operator always yields third-person singular agreement, but also because they are ultimately derived from relative clauses (see Chapter 7 for a full diachronic analysis). How can we combine this with agreement in the second conjunct?

At first glance, it seems difficult to maintain this type of analysis for coordinated sentences with mixed agreement like (63) or (64). We would have to assume an operator for the first conjunct and, at the same time, a dropped topic in the second conjunct that carries the same φ-features as the original topic in the first conjunct. This would be a combination of a Mixed and an Abnormal Sentence as in (65):

(65)



Although this derivation yields the right agreement pattern, many questions remain. What exactly is the deleted DP in the specifier of the second conjunct, for example? If it is a multidominance structure in which both conjuncts are 'dominated' by the the subject/topic 'the men', why are they not derived in exactly the same way and why does this not appear at the edge of the sentence (as usual for multidominance structures)? 'The men' are not focussed in the first conjunct, so why is there no agreement?

Can these questions be answered by adopting a movement approach? First of all, this seems undesirable if we take the coordinate structure constraint seriously. Furthermore, for the first conjunct, the lack of agreement with a full DP subject moving through SpecTP would be expected. For the second conjunct, we cannot assume movement of the full DP (unless we 'convert' the trace/copy and assume

Move and Agree are not tied together, as outlined above). We have to postulate a 'topic operator' that carries the same φ-features as the plural DP. But at the same time, we cannot readily assume this topic operator was 'born' with φ-features. If that were the case, we would first of all expect the possibility of spelling out the weak pronoun (which in theory is possible, but never seen in this configuration). Furthermore, for null objects, as explained above, we have to postulate a topic operator without φ-features, because we do not see agreement clitics on the verb. It seems undesirable to postulate two different kinds of topic operators: one for subjects with φ-features and one for objects without. Adopting a movement approach does not fare much better than the outlined base-generated approach.

Let us now get back to the earlier question about the lack of focus in the first conjunct ('the men' is actually an aboutness-shift topic in the context). From an information-structural perspective, the notions [+TOPIC] and [+FOCUS] seem to have been rendered meaningless here. They are only mentioned in the derivation as an indication for an agreeing and non-agreeing structure respectively. If we recall some examples with 'unexpected' agreement patterns from the introduction of this section, however, we see the same 'pattern' (or 'lack of association between Topic/Focus and Agree/No Agree'):

(66) a. *Miui hagen a uydaf gyfarwyd ywch*
I.EMPH however PRT be.1S familiar to.2P
'I, however, will be familiar to you.' (Focus, but Agree - CO 899)
b. *Kennadeu a aeth at uranwen.*
messengers PRT go.PAST-3S to Branwen
'Messengers went to Branwen.' (Topic, but no Agree - PKM 40.1-2)

In a movement analysis, the above agreement pattern (Agree with pronoun and no Agree with plural DP) is exactly what we would expect. If the information-structural features were not strictly associated with a particular derivation anymore, could it be the case that the language we observe was actually representing a grammar in transition from a base-generated to a movement analysis of verb-second clauses? If we look at the other puzzling example from the introduction from two different manuscripts - the older White Book and the later Red Book - this might actually hint at this transition.

(67) a. *Ti a 'y gwelho*
you PRT 3FS see.SBJ-3S
'You will see it' (White Book CO 451)
b. *Ti a 'y gwelhy*
you PRT 3FS see.SBJ-2S
'You will see it' (Red Book equivalent)

In Chapter 7, I return to this issue putting these difficult agreement patterns in a diachronic context.

6.5 Case Study III: Givenness

The present corpus of Middle Welsh was annotated for Givenness with the Pentaset (Komen, 2013). Recall that according to this annotation scheme, constituents are first divided by whether they are somehow 'linked' or not. If they are not linked, a further distinction is made between constituents that are not active as possible antecedents in the following context (labeled INERT) and those that can be referred to (labeled NEW).

If a constituent is linked, the first question is whether it is linked to something that previously occurred in the text or not. If that is not the case, it can still be linked to something that is considered common knowledge by the speaker and listener (or writer and reader) in that particular situation. A divine figure like 'God' for example, is assumed to be commonly known even if 'God' was not introduced in the immediately preceding context. In the Pentaset, these constituents are labeled ASSUMED.

Constituents can be identical to an item or person still in the working memory of the listener, because it appeared in the preceding context. A clear example is a pronoun 'he' referring to a full DP 'the man' in the previous sentence. Since these constituents both refer to one and the same man, they receive the IDENTITY label. Finally, there are constituents that are not identical to something or someone previously mentioned, but they are related to them in another way, for example a set or part/whole relation. These constituents are labeled INFERRED. The Pentaset allows us to make meaningful distinctions on a scale of Givenness, rather than a black-and-white old vs. new distinction.

The Case Study related to Givenness I present in this section is concerned with the referential status of object. As pointed out in Chapter 4, direct objects are hardly ever found in sentence-initial position, but if they are, they either convey New information or information that is 'newer' on the scale than that of the sentence subject. In the exceptional cases their status is not new(er), they are always familiar topics (so different from the aboutness-shift topics presented above). In section 6.5.1 I first outline the data and in section 6.5.2 I present a syntactic analysis along the lines of the approach for topic sentences in the previous section.

6.5.1 Givenness: the data

Most sentence-initial objects convey New information, as shown in the examples in (68a) and (68b). The subjects in these sentences often convey Old information: their referential status is IDENTITY. These object-initial examples are thus marked according to the Principle of Natural Information flow, because New information precedes Old information instead of the more common 'Old-before-New' pattern.

Subject IDENTITY + Object NEW

(68) a. *Ac val y deuth y mywn. gwydbwyll a welei yn y*
and as PRT come.PAST.3S to in Gwyddbwyll PRT see.PAST.3S in the

neuad.
hall
'And as he came in he saw a *gwyddbwyll*[19] in the hall.' (Peredur 66.23-24)

b. *Kymmeu a welei, (a diffwys, a cherric uchel ...)*
valleys PRT see.PAST.3S and steep place and rock high a ...
'And he saw valleys (and a steep place and a high rock ...).' (BM 2.19-20)

In some sentences, the sentence-initial direct objects contain constituents that are not literally mentioned before, but they are somehow linked to the preceding context. These objects are INFERRED. In these cases, the sentence is still marked, because Newer information precedes Old information. The direct objects in these sentences are often focussed as well: they pick out one or a part of a possible set of alternatives.

Subject IDENTITY + Object INFERRED

Finally, there are object-initial sentences in which the referential status of the subject and the object is both IDENTITY: they both convey 'old' information. These objects are very closely linked to the preceding context. They mostly repeat either the exact same constituent that was mentioned last or they refer to the same context with a demonstrative pronoun. As such, they are annotated as 'Familiar topics'. In section 6.6 about textual cohesion I discuss these further.

So Peredur took half of the meat and of the liquor himself,

(69) *a r llall a adawd yghyfeir y vorwyn.*
and the other PRT leave.PAST.3S for the maiden
'and the rest he left for the maiden.' (Peredur 10.28)

If there are gifts for the husband via the wife, it belongs to the husband until the end of seven years; and if she gets to the third night of the seventh year,

(70) *haner y da oll a geiff y wreic pan yscaront.*
half the goods all PRT get.3S the wife when divorce.SBJ.3P
'the wife gets half of all the goods when they divorce.' (Laws 520)

Subject IDENTITY + Object IDENTITY

And without further parlance, they encountered one another, and immediately Peredur overthrew the knight, and he besought mercy of Peredur.

(71) *Nawd a gehy gan gymryt y wreic hon yn briawt.*
mercy PRT get.2S by take.INF the woman that.FS in marriage
"Mercy shalt thou have by taking this woman in marriage" (Peredur 22.5)

[19] Some kind of chessboard.

And on the chair sat a lovely auburn-haired maiden, with a golden frontlet on her forehead, and sparkling stones in the frontlet, and with a large gold ring on her hand. (...) "My mother," said he, "told me, wheresoever I saw a fair jewel, to take it." "Do so, my soul," said she.

(72) *Y vodrwy a gymerth Peredur.*
the ring PRT take.PAST.3S Peredur
'Peredur took the ring.' (Peredur 11.4)

"When first I met the mother of this maiden, nine bushels of flax were sown therein, and none has yet sprung up, neither white nor black; and I have the measure by me still."

(73) *Hwnnw a vynnaf inheu y gaffel yn y tir newyd draw*
that PRT want.1S I 3MS get.INF in the land new over.there
'I require to have the flax in the new land under.' (CO 606-607)

The main question for this section is: how are these object-initial sentences derived syntactically? In the following section, I propose an analysis in line with the base-generated approach for topics and foci outlined above.

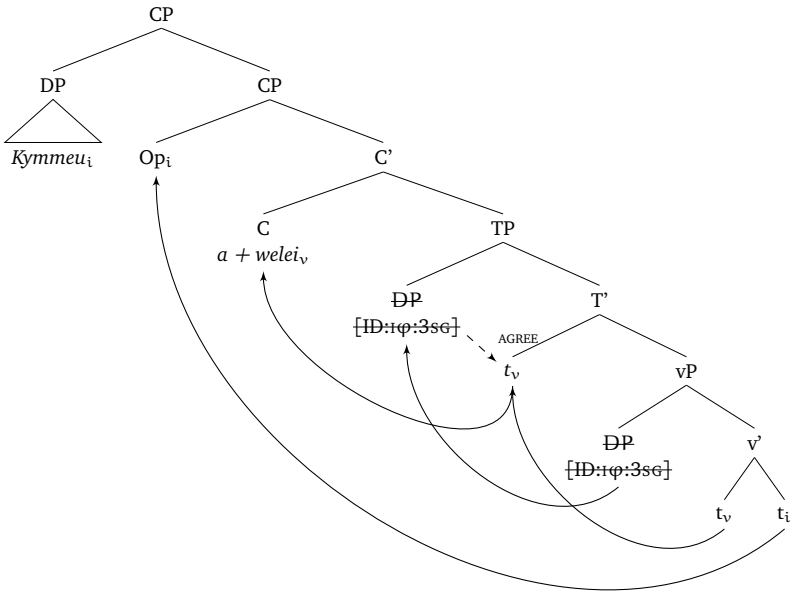
6.5.2 Givenness: the analysis

In object-initial sentences, the tricky agreement problem discussed in the second Case Study is not observed. Sentence-initial objects can furthermore only be full DPs in Welsh, since pronominal objects are always cliticised to the verb, so the Complementarity Principle cannot be observed either. The subject is always post-verbal when the object is in sentence-initial position and in these contexts we find agreement according to the Complementarity Principle as expected. So how are object-initial sentences derived?

If sentence-initial objects contain New information (and are thus marked in terms of the Principle of Natural Information Flow), they could be analysed as sentences containing New Information Focus. If New Information Focus structures are derived in the same way as contrastively focussed structures, we expect the focussed constituent to be base-generated and co-indexed with an Operator in the specifier of CP. The derivation of (68b) repeated here as (74) would then look like (75).

(74) *Kymmeu a welei*
valleys PRT see.PAST.3S
'And he saw valleys' (BM 2.19-20)

(75)

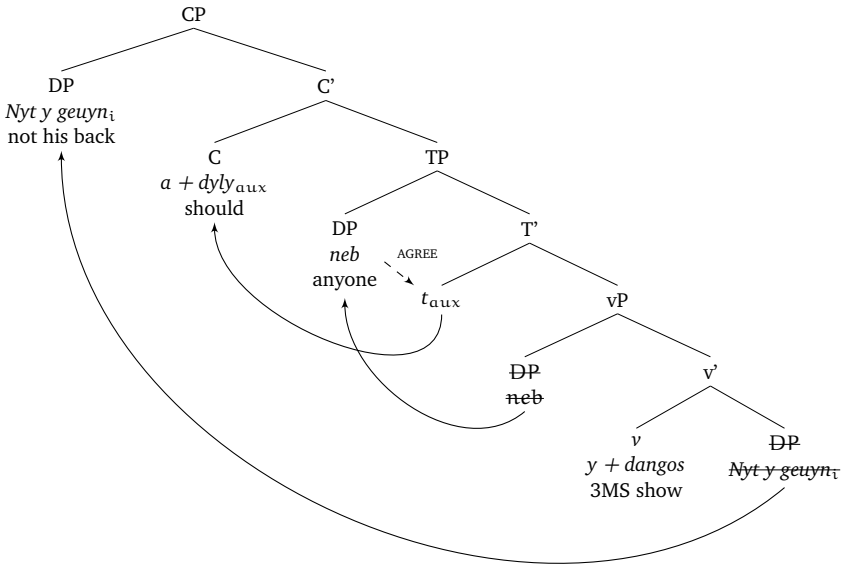


An analysis involving movement of the object would yield the same result. There is no agreement with full DPs, so we do not expect agreement clitics on the verb, which is exactly what we find. A uniform analysis of all focus structures, both for initial subjects as well as objects, would be preferred, so adopting a movement approach for these object-initial sentences needs to be well motivated. One reason for adopting a movement analysis could come from structures involving (local) Binding. An example is given by Borsley et al. (2007:293):

(76) *Nyt y geuyn a dyly neb y dangos y*
NEG 3MS.GEN back PRT should.PRES.3S anyone 3SM.GEN show.INF to.3MS.GEN
elynnyon.
enemies
'It is not his back that anyone should show to his enemies' (i.e. 'No one should show his back to his enemies.') (YCM 140.26-7)

We expect the possessive pronoun *y* in the sentence-initial object constituent *y geuyn* 'his back' to be bound by the quantifier *neb* 'anyone' in its base position where it can be c-commanded by the quantifier. The derivation (of the first part of the sentence) with the moved direct object would then look like (77):

(77)



There are, however, very few examples that can provide such evidence for a preferred movement approach. In addition to that, similar problems with local binding are observed in relative clauses. Recall that relative clauses are also claimed to involve a null-operator (by, amongst others, Borsley et al. (2007) and Willis (2011b)). As Willis (2011b) points out, an example like (78) from Modern Welsh needs 'some mechanism' to "ensure that the operator is in some sense linked to anaphor *ei hun* 'himself' in the antecedent of the relative clause" (Willis, 2011b:213 n.16).

(78) *Dyma 'r llun o 'i hun mae Ifan yn ei leicio fwyaf.*
this-is the picture of 3MS REFL be.PRES.3S Ifan PROG 3FS like.INF most
'This is the picture of himself that Ifan likes most.'(Willis, 2011b:213)

Again, such examples are not frequently found in the limited historical corpus. It is therefore difficult to assess first of all whether they existed in Middle Welsh. Secondly, it is not clear that whatever this 'mechanism' entails, would also 'solve' the quantifier-binding example presented above. Adopting a movement analysis for these relative clauses as well, however, is not self-evident for various reasons (e.g. lack of agreement with subjects). For now, I assume the base-generated approach for any focussed constituents, to give a uniform account of the data as it was in one particular stage of Middle Welsh. As in the previous section, I do not exclude the movement approach as an option to derive these sentences. Again, perhaps this option became available in the course of the Middle Welsh period and example (77) represents that option.

6.5.3 Conclusion Case Study III: Givenness

In this section one particularly interesting Case Study related to the information-structural concept of Givenness was presented: object-initial word orders that are marked because they present New information before Old information in the sentence (by subjects with the referential label IDENTITY). These object-initial sentences can be derived in many different ways. If they are considered to be focus structures equivalent to the 'mixed' sentences with contrastive focus presented in section 6.4 above, we can postulate the exact same derivation with an operator in SpecCP.

Some sentences with initial objects, however, seem to fare better with a movement approach to ensure the fronted object can be locally bound by, for example, a quantifier. If these sentences contain a null-operator, just like relative clauses, some mechanism is needed to ensure binding is possible with the object in a base-generated sentence-initial position. To conclude, object-initial sentences seem to provide some evidence for a movement-based analysis (in the form of one example with quantifier binding). As pointed out in the previous section, however, a movement analysis presents some serious difficulties for sentence-initial subjects. The only way to solve this puzzle is to assume both analyses were possible, perhaps because the grammar of Middle Welsh was in transition. This option will be explored further in Chapter 7.

6.6 Case Study IV: Text Cohesion

As pointed out in detail in Chapter 3, apart from topic, focus and givenness, there is a fourth information-structural notion that plays an important role in Middle Welsh syntax: text cohesion. This notion is concerned with how sentences are linked together within a paragraph or text as a whole. In Middle Welsh, as in many other languages, a frequently-found strategy to achieve textual cohesion is by means of sentence-initial 'Points-of-Departure'. These Points of Departure can set the scene (like scene setting topics) and introduce a new section. Very often, however, they are linked to the situation in the previous sentence by a prepositional or adverbial phrase referring to a specific time or place. These constructions with sentence-initial adjuncts are the most frequently found word order patterns in the Middle Welsh corpus. Some adverbial or prepositional phrases in initial position appear to occur before the topic or the focussed constituent, yielding superficial V3 patterns.

In addition to that, there are other ways to achieve a high degree of cohesion within a paragraph. In passages with direct speech, for example, a familiar topic in the form of a sentence-initial object is often used to provide a close link to the immediately-following narrative. The syntactic analysis of both these options will be discussed in section 6.6.2. In section 6.6.1, I first present the data.

6.6.1 Text Cohesion: the data

Some points of departure occur before subject or objects in topic/focus position. They usually consist of a prepositional or adverbial phrase, but can also be complete subordinate clauses setting the scene for the matrix clause. In some sentences, exclamatives like *nachaf* 'lo, behold' or temporal adverbs like *yna* 'then' are used to introduce the matrix clause following the subordinate clause that sets the scene:

(79) *Yr awr y kymerth hi y bara yn y gyluin. hi a syrthawd o*
the hour PRT take.PAST.3S she the bread in 3FS beak she PRT fall.PAST.3S from
r prenn yn varw y r llawr.
the branch PRED dead to the ground
'The moment she took the bread in her beak, she fell from the branch dead to the ground.' (Dewi 12.12)

(80) *A chynn y dyuot y r gynnulleittua honno.* ***nachaf*** *y gwelynt*
and before 3P come.INF to the assembly that.3F lo PRT see.PAST.3P
yn dyuot yn y herbyn gwreic wedw gwedy marw y hun mab.
PROGR come.INF yn 3P back woman widow after die.INF 3FS own son
'And before coming to that assembly, lo, they saw a widow coming towards them whose own son died.' (Dewi 16.1)

(81) *A gwedy eu diflannu hyt nas gwelei.* ***yna*** *y kyfaruu ac*
and after 3P disappear.INF until NEG-3P see.PAST.3S then PRT meet.PAST.3S with
ef. yn eisted ar ben cruc. y wreic teccaf o r a
him PROGR sit.INF on top mound the woman most.beautiful of those PRT
welsei eiroet.
see.PLQPF.3S ever
'And after they disappeared so he couldn't see them, then met with him sitting on the mound the most beautiful woman he had ever seen.'(Peredur 47.9-11)

Most adverbial or prepositional phrases in sentence-initial position, however, are directly followed by the preverbal particle *y* and the inflected verb. As such, they occupy the preverbal position in which argument topics and foci can also reside.

Temporal adverbials

(82) a. *A* ***thranoeth*** *y kyfodes y maccwyeid racdunt.*
and next.day PRT rise.PAST.3S the youths towards.3P
'And the next day the youths rose towards them.' (Peredur 18.29)
b. *Ac* ***yna gyntaf*** *y dywetpwyt y geir hwnnw.*
and then first PRT say.IMPERS the word that.3MS
'And then first that word was said.' (PKM 41.1-2)
c. *A* ***r nos honno*** *y buant yno yn diwall ... ganthunt.*
and the night that.3FS PRT be.PAST.3P there PRED safe ... with.3P
'And that night they safely ... stayed there.' (PKM 46.25-26)

Temporal prepositional phrases

(83) a. *Ac* ***erbyn*** *hanner dyd drannoeth. yd oed yn y uedyant y*
and by half day next.day PRT be.PAST.3S in 3MS possession the
dwy dyrnas.
two kingdom
'By noon next day the 2 kingdoms were his.' (PKM 6.13-14)

b. *Ac* ***yn*** *yr amser hwnnw yd oed yn arglwyd ar Wynt Ys Coet*
and in the time that.3MS PRT be.PAST.3S PRED lord on Gwynt Is Coed
Teirnon Twryf Uliant.
Teirnon Twryf Uliant
'Now at that time Teirnyon T. V. was Lord of Gwent Is Coed' (PKM 22.1-2)

Spatial adverbial and prepositional phrases

(84) a. *Ac* ***yno*** *y bum seith mlyned yn penydyaw.*
and there PRT be.PAST.1S seven years PROGR do.penance.INF
'And there I was seven years in penance.' (BR 5.15)

b. *Ac* ***y*** *r neuad y gyrchwys y diarchenu.*
and to the hall PRT go.PAST.3S to undress.INF
'And he went to the hall to take off his boots' (PKM 4.7-8)

c. *Ac* ***yn*** *y ty yd oed cassec.*
and in the house PRT be.PAST.3S mare
'And in the house was a mare.' (PKM 22.3)

Familiar topics form another category of elements that can realise textual cohesion. Examples of these often contain the exact same lexical items or demonstrative pronouns that refer back to a situation or a thing/person just described in direct speech or the immediately-preceding narrative context.

(85) a. *A hynny a gawssant ual y notteynt.*
and these PRT get.PAST.3P as PRT note.PAST.3P
'Those they got as they named it.' (BM 8.7)
b. *Y rei hynny a rithassei ef o r madalch.*
the those those PRT form.PAST.3S he from the fungus
'Now these he had formed of fungus.' (PKM 70.22-23)

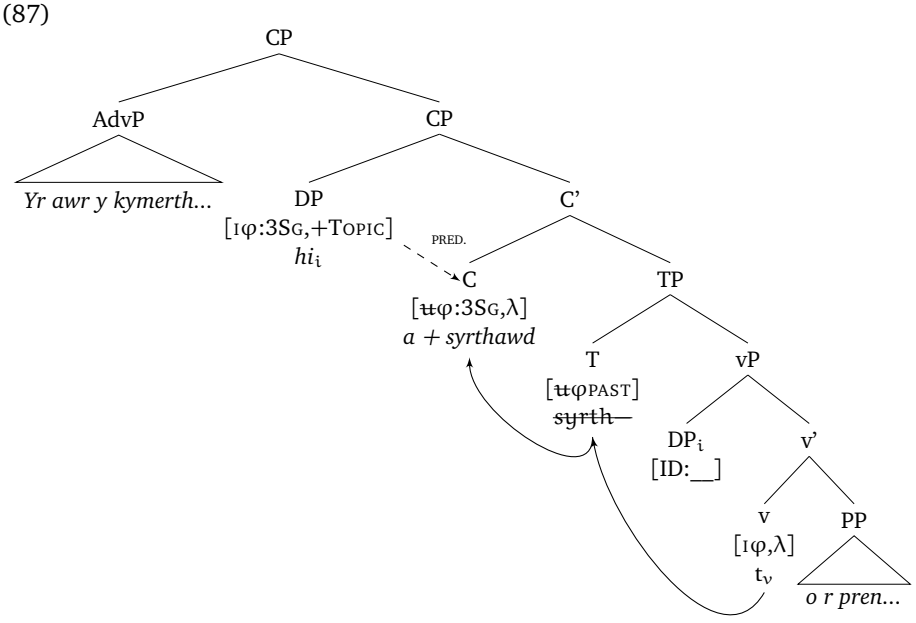
How are these adjunct-initial sentences derived? Are the object-initial examples with familiar topics the same as the objects conveying New information we saw in the previous section? In the next section I outline the syntactic derivations for these sentences with constituents in initial position for reasons of textual cohesion.

6.6.2 Text Cohesion: the analysis

Sentences with adverbial or prepositional phrases preceding a topical element can be analysed as containing scene-setting topics. These scene-setting topics are always in the highest projection of a proliferated left-periphery, in a ForcePhrase or a dedicated 'Frame' or 'Scene-setting' Phrase just below that. Since in Middle Welsh topic and focus never co-occur in sentence-initial position, I have so far limited the C-domain to one Specifier-position of CP. As pointed out in section 6.4 above, hanging topics and left-dislocated topics were also possible in Middle Welsh. Again, in a 'rich' left-periphery, these each receive their own 'Hanging Topic' and 'Left Dislocated' phrase. Whatever the name of the phrase, it is clear that there should be an extra position to the left of the CP for any of these elements. For now, I remain agnostic about the name of this position and use an extra CP layer for all of these elements.The derivation of a sentence like (86) is shown in (87).

(86) *Yr awr y kymerth hi y bara yn y gyluin. hi a syrthawd o*
the hour PRT take.PAST.3S she the bread in 3FS beak she PRT fall.PAST.3S from
r prenn yn varw y r llawr.
the branch PRED dead to the ground
'The moment she took the bread in her beak, she fell from the branch dead to the ground.' (Dewi 12.12)

(87)

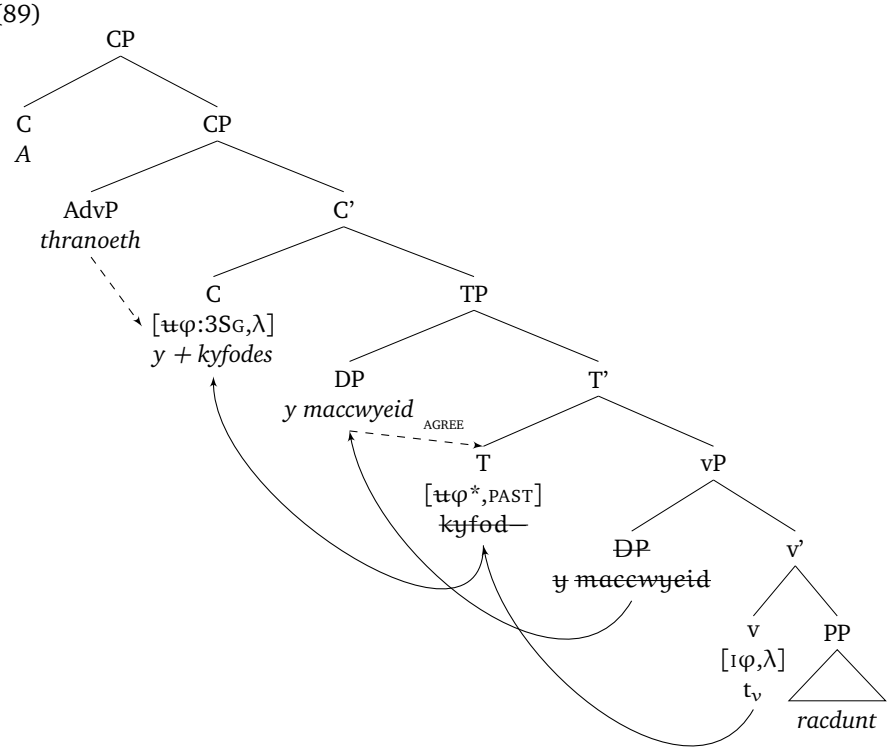


Whenever there is no argumental topic or focus in SpecCP, the adverbial or prepositional phrase can be base-generated in that position and control the form of the

complementiser in the C-head (*y* rather than *a* following subjects or objects). The derivation of (88) is presented in (89):

(88) *A* ***thranoeth*** *y kyfodes y maccwyeid racdunt.*
and next.day PRT rise.PAST.3S the youths towards.3P
'And the next day the youths rose towards them.' (Peredur 18.29)

(89)



For an adverb like *tranoeth* 'the next day' it is no problem to be base-generated in SpecCP. Are there any prepositional arguments of verbs in this position as well? Examples with an argumental PP dependent on the verb as in (90) are more likely derived via movement, however.

(90) *Ac ar y kynghor hwnnw y trigyssant.*
and on the advice that PRT settle.PAST.3P
'And on that advice they settled.' (PKM 25.5)

If argumental PPs are derived via movement, however, there is no reason to assume non-argumental PPs and adverbial phrases should be base-generated in SpecCP. They could all be derived via movement at this stage, although to prove this, we would need more examples with possible Principle C-effects.

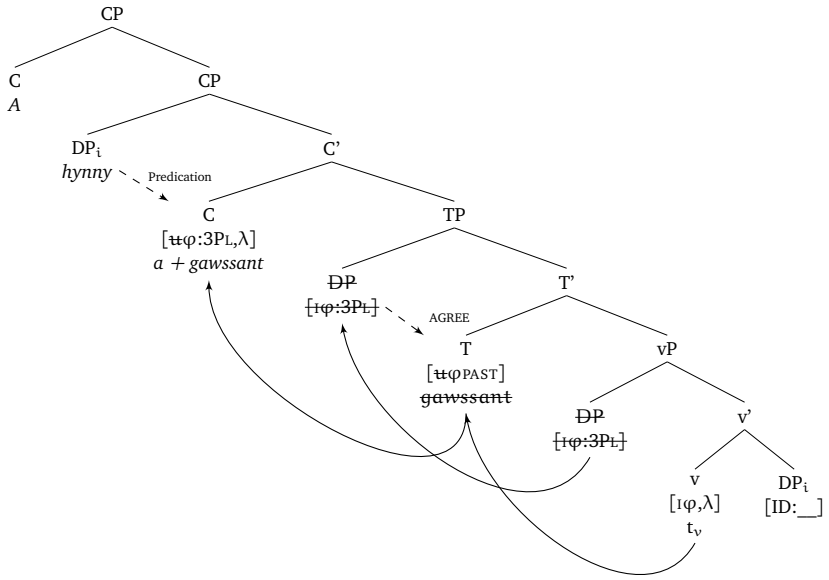
Finally, let us consider the 'Familiar topics' in sentences like (92). Just like in the case of the objects conveying New information in the previous section, it is very difficult to decide whether these involve a movement or a base-generated strategy. As for their landing site, Frascarelli and Hinterhölzl (2007) propose a hierarchy of phrases in the left periphery, as shown in (91).

(91) [ForceP [ShiftP [GroundP [ContrP [FocP [FamP [FinP]]]]]]]

Familiar topics occupy a position much lower in the clause, directly above Fin, whereas aboutness/shift topics occupy the highest projection. For Welsh, however, it seems better to assume one dedicated position in the C-domain, since the constituent in the specifier of the CP determines the form of the complementiser. If we were to postulate various phrases in the left periphery, we would have to assume the heads of these phrases can all contain that same complementiser. Alternatively, we would have to find a mechanism via which a phrase in the specifier of a higher position in the CP can still agree with / determine the form of the complementiser in the C-head. This requires extra assumptions and is thus a less desirable solution. I therefore keep analysing the preverbal phrases in the specifier of CP, assuming this CP can host foci and both aboutness as well as familiar topics. A possible derivation of the main clause in (92) with base-generation is presented in (93).

(92) *A hynny a gawssant ual y notteynt.*
and these PRT get.PAST.3P as PRT note.PAST.3P
'Those they got as they named it.' (BM 8.7)

(93)



6.6.3 Conclusion Case Study IV: Text Cohesion

In this section I presented different ways of achieving textual cohesion with constituents that are linked to the preceding context in the initial position of the sentence. This can result in V3 word orders with, for example, frame or scene setters or hanging topics (or even V4 sentences if extra adverbials are added to the C-domain).

If there are no topicalised subjects or objects, however, adverbials and prepositional phrases denoting a dedicated time or location (or any other type of Point of Departure discussed in Chapter 3) appear in SpecCP modifying the form of the complementiser (yielding *y* rather than *a*). In principle, these structures could be derived through the base-generation approach outlined for abnormal and mixed sentences above. If the initial prepositional phrase is an obligatory argument of the verb, however, a movement strategy seems better suited.

Sentence-initial objects that are 'Familiar topics' can be analysed as aboutness topics. They do determine the form of the complementiser in C (yielding *a*), but they do not transfer their φ-features to C. The C-head's uninterpretable φ-features are already matched by those of the verb moving (with inflection and thus interpretable φ-features) to incorporate into the C-head. A movement analysis for these familiar topics is possible if they are objects; for subjects, a base-generation approach is still preferred for reasons of agreement discussed in section 6.4 above.

6.7 Conclusion

In this chapter I discussed four different case studies related to the most important information-structural features in Middle Welsh. The aim of this chapter was to provide a syntactic analysis for those information-structural phenomena and to see how notions like topic, focus and givenness are implemented in the syntax of the language. As generally assumed in current minimalist approaches, many of these IS features are ultimately postulated to reside in the left periphery of the clause. Although in many recent approaches, this left periphery is argued to consist of various phrases with dedicated heads for all kinds of topics and foci, it is difficult to prove this is also a necessary assumption for Middle Welsh.

Middle Welsh allowed only one topic position. Although V3 and even V4 structures are attested, those can only involve either hanging topics or scene setters or other adverbial elements preceding or following the topic.

Two different types of analyses were presented and discussed in greater detail: a base-generation approach for topical and, with a null-operator, for focussed constituents and a movement approach. A movement approach creates problems for sentence-initial subjects, because Middle Welsh seems to adhere to the Complementarity Principle in general, but Abnormal Aentences do exhibit agreement with plural full DP subjects. At the same time, focalised pronouns do not exhibit agreement. Both of these facts are unexpected and difficult to account for under a movement analysis.

Under a base-generation approach, these different agreement patterns can be explained. There are, however, also examples that present a greater challenge for a base-generation analysis, such as sentence-initial constituents that must be (locally) bound by a quantifier (see section 6.5) and argumental PPs (see section 6.6).

Finally, there are some very challenging examples with coordinated clause exhibiting mixed agreement patterns. All in all, we seem to be forced to conclude both a base-generation approach and a movement approach are necessary to account for all the Middle Welsh data. In the next Chapter, I will sketch a diachronic analysis in which both of these options play an important role in the development of Middle Welsh grammar.

CHAPTER 7

Diachronic syntactic change

> *"As there are sophisticated methods for its reconstruction, the common ancestor language of Welsh, Cornish, and Breton is so accessible that with a bit of practice we would be able to strike up a conversation with a second-century British Celt in his native language and explain to him how his language had changed - quite dramatically as a matter of fact - by the end of the sixth century."*
>
> (Schrijver, 2014:30-31)

7.1 Introduction

Schrijver's above-sketched optimistic scenario is based on the success of the Comparative Method reconstructing the sounds and words of older stages of languages we no longer have direct access to. Following in the footsteps of Sir William Jones (1746-1794), a philologist and judge of Welsh descent (see Silk (2014)), this method led to a number of late nineteenth-century breakthroughs by Neogrammarians like Hermann Osthoff and Karl Brugmann (see their famous manifesto in the preface to volume 1 of their *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, Osthoff and Brugmann (1878)). The field of historical linguistics was so influential that it inspired Darwin with the early discoveries leading up to his theory of evolution (see Alter (1999)). Schrijver therefore argues that "every educated human being should be aware of the method", however, "hardly anyone actually is" (Schrijver, 2014:6).

Schrijver then goes on to explain this comparative method, first by providing a step-by-step example, then by reconstructing much of the phonology, morphology and lexicon of Proto-British, the language of his second-century British Celt. Syntax,

however, is not discussed '[f]or practical reasons' (Schrijver, 2014:1). These reasons actually go beyond the 'time frame' and 'nature of the available source material' (or the lack of material altogether). They even go beyond linguistic expertise or preference, as we can read in a study entirely dedicated to diachronic syntax: 'one can no more reconstruct the syntax of a proto-language than one can reconstruct last week's weather[1]' (Lightfoot, 2002:135). It seems then, that we can only converse with our British Celt in a language that, by necessary assumption, has the same syntax as that of his descendants, who did write things down in their vernacular, be it Breton, Welsh or Cornish. However, although these three were closely related, it matters which one we use as an exemplar for our Proto-British grammar.

To illustrate this let us examine one example again that was frequently used in Middle Welsh, the Abnormal Sentence. The agreement patterns observed in Middle Welsh are hardly ever found in Middle Breton: both pronouns and plural nouns yield default third-person singular inflection on the verb in Breton. The question is therefore when speaking to our British Celt, should we Agree or not Agree? If we want to make sure we are not making any syntactic 'errors', we might be better off greeting him in Late Latin, a language conveniently found in various sources around that period. Depending on our Celt's social status and, arguably, his place of birth, there is a good chance he was perfectly capable of speaking or at least understanding this language of the Roman invaders.

Once we are done with formalities and chit-chat about the reconstructed weather, we would like to get down to business and tell this British Celt all about the drastic changes his language will undergo in the next few centuries. Not just the sounds, but also the order of "the magical letters S, V and O" (C. Watkins, 1976:305) will be changed. We could answer the puzzled look on his face reassuring him that those magical new word orders will not last much more than a thousand years. This might be an adequate answer to his first question ('When?'), but can we give any satisfactory explanation as to *how* and *why* it changed so dramatically?

This chapter aims to shed more light on the 'how' question regarding some major syntactic changes in the Middle Welsh period. In section 7.2 I first discuss the main mechanisms of diachronic syntax and which specific challenges it presents from an empirical, theoretical, and - depending on our definition of syntax - ontological perspective. A reconstruction of Proto-British syntax goes beyond the scope of the present study, but I will emphasise the importance of comparative studies illustrated with some examples from Middle Breton. If we want to gain a better understanding of the syntactic history of the Welsh language, a solid methodology for both historical syntax and syntactic reconstruction is indispensable. The challenging examples presented in the previous chapter are addressed again in the context of mechanisms of grammaticalisation and reanalysis in section 7.3. Finally, in section 7.4 I take a closer look at the role of information structure in the study of diachronic syntax.

[1] This frequently-cited metaphor has its origin in early work by Jerzy Kuryłowicz on the laws of analogy (Kuryłowicz, 1949), who observed that historical linguists cannot predict when it would rain, even if the presence of gutters predict that it would.

7.2 Approaches to diachronic syntax

If the object of study in syntactic research is limited to the competence of the individual speaker-hearer, their I-language (see Chapter 1 and Chomsky (1965) and Chomsky (1986)), we need to ask ourselves whether 'diachronic syntax' exists at all. Children are extremely successful in acquiring grammar, but as soon as they grow up speaking a language with, for example, SVO word order, they are unlikely to change to VSO at any given stage during their lifetime. Change over time in the internalised grammar of individuals is generally very restricted[2] (see, amongst others, Clahsen (1991)).

But syntactic innovations have been found in historical documents in various periods of time. Modern French, Romanian and Italian word order, for example, differ from Latin, just as Modern Greek differs from the language spoken by Plato and Socrates. So how do we account for that? Arguably the easiest way out of this apparent paradox is to challenge the premise: maybe I-language is not the right object of study in diachronic syntax? After all, we can only ever study I-language through spoken or written sources of E-language and even then it remains difficult to be sure that we are in fact dealing with the 'real' I-language.

When comparing Middle Welsh texts to Modern Welsh texts, we can indeed observe that the basic word order has changed. If we want to account for either the Middle Welsh or Modern Welsh sentence structures we observe to reach descriptive adequacy, we need to go beyond mere observations. Adequate generalisations can, however, only be made abstracting away from the observed examples in a systematic way. Therefore even to answer the question of how a sentence/construction/word order pattern is derived synchronically, we need a certain level of abstraction and thus a syntactic framework that gives us tools and methods of analysis. Explanatory adequacy then goes even further in addressing the 'why' question: why do we find pattern X (and not pattern Y) or - in our diachronic scenario - why does pattern Y replace pattern X?

In the first part of this section I briefly discuss some approaches to diachronic syntax that have been used to explain various phenomena in historical Welsh syntax. Examples include (Cognitive) Construction Grammar and the loss of V2 in Early Modern Welsh, contact-induced change by language shift in Early Brythonic and generative acquisition-based models of change. Some of these overlap and/or share specific mechanisms proposed to account for syntactic change. Before moving on to the diachronic analysis of the Welsh data presented in the previous chapter in a generative framework, I discuss these mechanisms and the most important challenges in the study of historical syntax.

[2] 'Unlikely' in the previous sentence refers to that fact that such rigorous word order changes are not observed by researchers studying language change. This lack of evidence does not exclude the possibility of such changes occurring in individual grammars. Although Crisma and Longobardi (2009) assert that 'within an I-language, there seems to be no such a thing as change' (Crisma & Longobardi, 2009:4), there are certain subtle changes, mainly in frequencies rather than in the emergence of new structures (see also Walkden (2014:35n20)).

7.2.1 Diachronic Construction Grammar

Within Construction Grammar (CxG) the main focus on diachronic syntax is centred around how constructions (form-meaning configurations larger than morphemes and words) change over time. The interaction of frequency and constructionalisation has played an important role, as well as how constructions develop to become more lexicalised or more schematised (Barðdal, Smirnova, Sommerer, & Gildea, 2015:20). Explanatory adequacy within Cognitive Construction Grammar is achieved through the concept of *motivation*: each construction must be motivated by principles of grammaticalisation, discourse demands, iconic or general principles or appeal to constraints on acquisition (Goldberg, 2006:17). Goldberg (1995) formulates the Principle of Maximized Motivation as follows: "If construction A is related to construction B syntactically, then the system of construction A is motivated to the degree that it is related to construction B semantically ... Such motivation is maximized." (Goldberg, 1995:67).

Currie (2013) employs this principle of motivation in his study on the loss of V2 in Early Modern Welsh. According to Currie, adverb-initial word orders 'motivated' verb-initial word orders, because of the perceived parallelism between sentences with clause-initial adverbs and those without: "[t]he basis for this motivational relationship is the formal similarity between the respective pairs of constructions and the fact that clause-initial adverbial phrase could be analysed as a clause connector, separate from the verbal phrase, so that the following construction - XP + verb or verb - could be perceived as clause-initial" (Currie, 2013:67).

The concept of 'motivation' is criticised in other corners of the field of Construction Grammar. Within Unification Construction Grammar (a non-usage-based version of CxG focussing on unification-based formalism, see Kay and Fillmore (1999)) the concept of 'motivation' is discarded, because it fails to make any testable predictions. According to Goldberg (2006), however, this is a misinterpretation of the concept of 'motivation'. She argues that "[w]hile motivation is distinct from prediction insofar as a motivated construction *could have been otherwise*, it typically could not have had the opposite values of the properties claimed to provide motivation" (Goldberg, 2006:219).

At first glance, however, it seems unclear what this means in the case of the loss of V2 in Welsh, because the two available structures (adverb-initial and verb-initial) are claimed to motivate *each other* (Currie, 2013:67). According to Currie, the lack of verb-initial orders in Middle Welsh is due to the lack of Adverb + Verb orders in that same period. The 'prediction' in this sense must therefore be that because of this correlation of mutually motivating word order patters, verb-initial orders would not develop if Adverb + Verb orders had not increased in frequency. To the extent this makes any predictions concerning the change of word order from V2 to verb-initial in Early Modern Welsh, we are still left with what Roberts (2007) calls a 'Chicken-and-Egg' problem of syntactic reanalysis: which is the cause and which the effect of change?

According to Willis (1998), the loss of preverbal particles *a* and *y* was a crucial factor in the loss of V2. It not only led to an environment in which sentence-initial

subject-pronouns could be reanalysed as clitics (see also Willis (2007a)), but it furthermore led to an increase in Adverb + verb orders (since, as we have seen in Chapters 4 and 5, Adverb + *y* + Verb was by far the most frequently-found word order pattern towards the end of the Middle Welsh period). Currie, however, states that "we cannot say the decline in the use of *y* necessarily caused the increase in use of AIV [absolute verb-initial - MM] order" (Currie, 2013:67). Other factors, such as synchronic variation in word order patterns in Early Modern Welsh and the importance of the Welsh Bible translations were just as much part of the 'motivation' for the change from V2 to VSO (Currie, 2013:71).

This variation, according to Currie, does not correlate with any socio-linguistic factors (e.g. class, dialect, register or genre): "the main parameter of variation appears to be stylistic choice by individual writers" (Currie, 2013:69). This then explains the 'gradual' pattern of the loss of V2 and should thus serve as an argument against Willis's parametric approach since the change took centuries to fully complete (see Willis (1998), but also the discussion on 'discrete' and 'gradual' change from a generative point of view below). The theoretical 'mechanism' behind this pattern of individual variation is borrowed from the Cognitive Sociolinguistic framework. Within this framework, Coupland defines the concept of *styling* where speakers "can frame the linguistic resources available to them in creative ways, making new meanings from old meanings" (Coupland, 2007:84) (as cited by Currie (2013:69-70)). Some Early Modern Welsh authors chose to use more verb-initial sentences in prose, because these verb-initial orders already frequently occurred in poetry and in the first Welsh Bible translations (and they wanted to imitate this elevated poetic style); others did not.

It should be noted, however, that Currie's (2013) conclusions regarding the high frequency of verb-initial orders in various excerpts of the Bible translation are slightly misleading, because he is conflating different types of Biblical genres. Crucially, this high(er) number of verb-initial orders is found in the *Book of Isaiah* (41.0% verb-initial order according to his Table 1) and the *Psalms* (24.8%), neither of which contain the narrative prose found in, for example, the *Book of Esther* (with only 9.4% verb-initial orders) or the *Gospel of Mark* (6.5%). According to both the Christian as well as the Judaic tradition, the *Psalms* belong to the Poetic texts of the Bible along with, for example, *Job* and *Proverbs* (see, amongst others, Vriezen and Van der Woude (2000:96)).[3]

In other diachronic studies within Construction Grammar, usage-based motivation is often specified from a structural, referential, semantic, discourse-pragmatic and/or contextual point of view. This then, in combination with the relative frequencies of various constructions, aims to give a comprehensive explanation of the particular syntactic change under investigation (see Fried (2009) on the development of the subjective epistemic particle *jestli* '[in-my-opinion-]maybe' in conversational Czech and the rise of the dative substitution in Icelandic by Barðdal

[3] See furthermore Watson (1973:2) and Green (2005:60) for the poetic nature of the language of the *Book of Isaiah*. Since verb-initial orders were already (more) frequently found in poetry, this distribution is not at all surprising.

(2011)). In Currie's study of the loss of V2 in Welsh, however, many questions remain. For example, to what extent did the authors' choice to imitate poetic style reflect their daily speech, if at all? Why did they choose to imitate Biblical poetry, rather than Biblical prose (which was still subject-initial V2)? Furthermore, if using verb-initial orders was indeed a stylistic literary choice of some authors, how and why did VSO become the prevalent word order in Modern Welsh?

Overall, it is not only intuitively attractive, but arguably also necessary to look for 'motivations' of syntactic change beyond the structural domain. To the extent it is possible working with limited historical data, evidence from semantic, information-structural and sociolinguistic variation should definitely be taken into account. These factors are built into usage-based Construction Grammar. In theory then this seems a reasonable approach to problems in diachronic syntax. In practice, however, looking at Currie's (2013) account of the loss of V2 in Welsh, many questions remain unanswered and it is not clear why - if at all - this approach achieves more 'explanatory adequacy' than, for example, the arguments originally put forward by Willis (1998) in a generative framework (and Willis (2007a) or, in 'flexible syntax' by Bury (2002).[4]

7.2.2 Sociolinguistic variation and language contact

One of the important factors in diachronic syntax also touched upon in the previous section is 'variation'. The source of variation can lie in sociolinguistic factors, but also in (combination with) situations of language contact. There are several approaches to language change that focus on characterising the exact nature of variation. After all, "[i]t is speakers and not languages that innovate" (Milroy, 1992:169). In what is arguably the most influential study of sociolinguistic variation (Weinreich et al., 1968), language is a form of 'orderly heterogeneity' (see also Nevalainen and Raumolin-Brunberg (2003:12)). Rissanen (2008:56) groups the most important extralinguistic factors that affect the choices of variants in the following way:

1. Sociolinguistic ⇒speaker's/writer's social status, education and the relationship between discourse participants
2. Textual ⇒genre, topic or purpose of text, discourse situation and medium
3. Regional ⇒language contact

He notes that many of these extralinguistic factors overlap. Research into variation and change thus necessarily needs to take a combination of these factors into account as well as "internal processes of change" (like, for instance, grammaticalisation or analogical levelling discussed below) (Rissanen, 2008:57). A balanced corpus with extensive metadata on the origin and philological background of the texts is indispensable in this type of approach.

[4]For more on Construction Grammar and explanatory adequacy, see the series of papers discussing this problem by Adele Goldberg and David Adger in Goldberg (2006) and Adger (2013a) *et seq.*.

Language contact in a historical context is a particularly difficult field of study. Labov presents the 'Principle of Contingency' according to which specific instances of change require specific (rather than universal) explanations (Labov (2001:503) and also Walkden (2014:46) for discussion). Contact can lead to change, but - surprisingly - also to continuity in grammars. Bilingualism and the ability of children to acquire more than one language perfectly if they learn both from a young age, can play a role in this. This is shown by studies of a corpus of Welsh-English bilingual speech in which only one possible instance of convergence (i.e. contact-induced transfer) was found (modifier-head order within noun phrases) (P. Davies & Deuchar, 2010). Although there is a large amount of bilingualism in North Wales (and there has been for a long time), P. Davies and Deuchar (2010) conclude that Welsh grammar - in particular the noun phrase under investigation - exhibits continuity rather than change.

Whenever there *is* contact-induced change, it appears to come in different forms. Thomason and Kaufman (1988:50) present a 'scale of interference' according to which the extent and type of contact determine the type of change from lexical borrowing with minimal contact to structural changes with intensive long-term contact. Winford (2005) characterises this distinction as recipient or source language agentivity. In the case of recipient language agentivity transfer of linguistic material typically includes the borrowing of open class vocabulary items and it is likely to lead to complexification of the recipient language. Cases of source language agentivity, on the other hand, are called 'imposition'. Here the transfer mainly consists of phonological and syntactic features.

In the following section, I describe two cases of language contact and syntactic change in the history of Welsh. First I discuss the proposal of language shift (resulting in 'imposition') in British Celtic put forward by Schrijver (2002; 2007; 2014). Then I briefly discuss proposed cases of Latin influence on Welsh grammar in a later stage (due to literary translations and/or adaptations from Latin originals).

Language shift in early Britain

Schrijver (2002) (and also Schrijver (2007) and Schrijver (2014)) sketches a scenario of language contact, in particular language shift in the history of the Brythonic languages to account for various morpho-syntactic phenomena found in the British Celtic languages (but, crucially, not in Irish). According to Bede's description of Britain (written in the first half of the eighth century), there were five languages present at the time: English, British, Irish, Pictish and Latin. In the centuries after the collapse of the Roman empire, there is evidence (in the form of inscriptions) for three of these in Wales: British, Irish and Latin (see, amongst others, Sims-Williams (2003), Falileyev (2003) and Russell (2012)). The extent to which each of these three was spoken and in daily use is a matter of ongoing debate (cf. Adams (2007), T. M. Charles-Edwards (2013) and Schrijver (2014)), but it is clear that what distinguishes Brythonic languages from Irish is the loss of final syllables and the case system. After 'the departure of Rome', both Latin and Brythonic were spoken and there was probably a high degree of bilingualism

(Russell, 2012:216-218).

The scenario outlined by Schrijver (2002) involves a split between speakers of Celtic in the lowlands and the highlands. Highland British Celtic is argued to be the predecessor of Welsh, Breton and Cornish in the west, whereas Lowland British Celtic (with a more Irish-like phonological system) and Late British Latin influenced the sound system of the Anglo-Saxon invaders in the southeast. During the Roman period, Latin was a superstrate language and as such it donated many lexical items to Brythonic. After the collapse of the empire, however, the situation was reversed rendering Brythonic a superstrate language as opposed to speakers of Latin who then became of lower status. Based on the language contact theory of Thomason and Kaufman (1988), Schrijver (2002) proposes that the observed Latinised morpho-syntactic features in Brythonic are the result of language shift. Speakers of the then substrate-language Late Latin moved to the 'Highland Zone' and rapidly shifted to speaking Brythonic, keeping a Latin accent (and Latin-like morpho-syntax), but avoiding Latin vocabulary (Schrijver, 2014:32).

According to Russell (2012:220-221), there are various geographical and sociolinguistic problems with this scenario. Here I focus on the proposed morpho-syntactic influence from Latin transferred by language shift. The mentioned features include the loss of neuter gender and the case system and the development of the pluperfect in Brythonic languages. The first two are equally problematic, according to Russell (2012). Loss of neuter gender, first of all, also happened in Irish, so this is not necessarily a feature of the grammar of Brythonic languages only (it might have been on its way out in Celtic in general) (Russell, 2012:222). As for the loss of the case system, the nominative and the genitive arguably survived the longest in Brythonic. In Old French, however, the nominative and accusative are both still attested. If British Latin "shared north-western Romance features with the Latin of northern Gaul" (Russell, 2012:222) as Schrijver (2002) suggests, this is a problem. The reconstructed paradigms of Late Spoken British Latin in Schrijver (2014:46-47), however, show that for all five declensions, the genitive survived alongside the collapsed/combined nominative-accusative (or, in the fifth declension type *homō*, the nominative-vocative **omō* was distinguished from the accusative-genitive **om(ɪ)nɪ* and dative **om(ɪ)ni*). If Schrijver's (2014) reconstructions of the Late British Latin nominal paradigms are correct, the loss of the case system in British Celtic indeed followed a parallel development with Late British Latin. This pattern was unlike that found in Old Irish, in which five distinctive cases survived (Thurneysen, 2003 [1946]).

This distinction between Irish and Brythonic languages also exists in the development of the pluperfect in the latter, but not in the former branch of Celtic. MacCana (1976) first proposed that this new paradigm observed in Welsh, Breton and Cornish was influenced by Latin. Russell (2012:223) argues, however, that it is unlikely that the periphrastic origin of the form *amauerat* 'had loved' was still discernible in British Latin, since its pronunciation had developed to /a'ma:rat/. If the periphrastic form *amauerat* still existed on a high literary level, it probably had little impact on spoken British Celtic. Even if it had existed, it could hardly serve as

a model for Brythonic *carassei* 'had loved', because this cannot be decomposed as a form of the preterite + the imperfect of the verb 'to be' (Russell, 2012:223).

Overall, the presented scenario involving language shift with speakers retaining elements of their native Late British Latin grammar is certainly possible. The evidence of syntactic similarities put forward by Schrijver (2002) is in the present state, however, still inconclusive. Out of the three suggested syntactic innovations in Brythonic, Russell (2012) argues only the one about the pluperfect is potentially convincing.

Later Latin influence on Welsh

According to D. S. Evans (2003 [1964]) and D. S. Evans (1971), influence from Latin can also be found in later stages of Brythonic languages, in particular in Welsh translations of Latin texts in the (Early) Middle Welsh period. This type of contact is not language shift by speakers of the substrate Late British Latin, but rather textual influence on a literary level. Examples of these literary Latinised features are third-person plural agreement with plural nouns (going against the 'Complementarity Principle' discussed in the previous chapter) and the use of the definite article + demonstrative as relative pronouns (e.g. *yr hwnn, yr hynn* 'that, which'). With respect to the plural subject-verb agreement, Schumacher (2011) points out that this is the only possible pattern in Old Welsh prose, regardless of whether the subject preceded (as in (1a)) or followed the verb (as in (1b)):

(1) a. *enuein di sibellae int hinn*
names of Sibyllae be.PRES.3P these
'These are the names of the Sibylls' (*MC*)

b. *imguodant ir degion*
beseech.PAST.3P the nobles
'the nobles besought one another' (*Chad* LL xliii)

Strachan (1909:61) already mentioned that agreement in Old and Middle Welsh shows 'certain peculiarities'. Just like in Middle Welsh prose, he argues "[i]n the earlier poetry the plural is quite common, and in corresponding constructions in Old Irish the plural is regular. In Welsh there has been an encroachment of the singular upon the plural, as there has been in later Irish." (Strachan, 1909:62). Koch (1991) notes that default third-person singular agreement must have been well established in Old Welsh, giving examples from, among others, the same Old Welsh marginalia in the Lichfield Gospels Schumacher mentioned above (*Chad* LL xliiff):

(2) *imaliti duch cimarguithejt*
lead.3S you story-tellers
'as the story-tellers would lead you' (*Chad* 3)

For neuter plural subjects, default third-person singular agreement is not unexpected from an Indo-European point of view. Examples of this are found in Hittite,

Greek and Old Avestan that are argued to go back to old collective nouns (see Beekes (1995:173) and Fortson (2010:132)). For masculine and feminine plurals, plural agreement was found in most Indo-European languages, including Celtic. The question is thus exactly when and how the Complementarity Principle came into being in the history of British Celtic. Koch notes that the third-person plural verbal ending *-nt* (in the old conjunct paradigm of the verb[5]) could have been lost by regular sound change (i.e. apocope in Proto-British) in which case the singular and plural ending of the verb coalesced completely.[6] The formal similarity of the singular and plural conjunct forms could be the base for analogical levelling in the rest of the verbal system. This then would explain the lack of agreement with plural nouns following verbs in Middle Welsh and the lack of agreement altogether in Breton and Cornish (although there too, are exceptions). It still does not explain the agreement with preverbal plural nominal subjects in the Middle Welsh Abnormal Sentence. In section 6.4 below I put these cases in a diachronic and cross-linguistic perspective.

Interim Summary variation and contact

Variation no doubt plays a significant role in language change. Language contact and in particular situations in which speakers of one (substrate) language shift to another (superstrate) language can result in more variation and change in the morpho-syntactic domain as well. There is, however, very little data, both linguistic and socio-historical, from the time of intensive contact between speakers of Brythonic and British Latin in the crucial period after the collapse of the Roman empire. The extent of variation and change caused by language contact is therefore difficult to ascertain. We need a comprehensive description of the syntax of Late British Latin and a sound methodology for reconstructing the syntax of Proto-British. Neither of these are provided by the above-mentioned approaches.

The second example of language contact in a later period (from translating Latin) is of a very different kind. Especially if we have the original text in Latin, grammatical similarities between the two languages are easier to expose. Since contemporary native tales are also available in that period, it would be possible to distinguish phenomena that are typically inherited from Proto-British (or even Celtic) from those borrowed from Latin, like the relative pronoun *yr hwnn*. If those constructions are fully incorporated in the language, we still need tools to adequately describe their formal function within Middle Welsh grammar and how (and why) they changed (again) in Early Modern Welsh.

[5] Insular Celtic had two separate paradigms of verbal endings that can still be found in Old Irish. Traces of the old absolute forms can also be found in Old Welsh, so presumably this system was still found in Brythonic.

[6] According to the 'standard doctrine' (VKG §152 and L&P §88), Proto-British word-final *-nt* survived apocope, but it is not altogether clear why this would be the case, since all final consonants except **-r* disappear in Proto-British. Koch's suggestion is, however, impossible to verify - for now, at least - since, according to Peter Schrijver (p.c.) the third-person plural conjunct verb form is the only (reliably) attested example of word-final **-nt*.

In the following section I present both tools to handle syntactic change as well as a proposed methodology for the reconstruction of syntax based on a generative acquisition-based approach.

7.2.3 Syntactic change in generative grammar

The most foundational study on diachronic syntax within generative grammar is Lightfoot's *Principles of Diachronic Syntax* (Lightfoot, 1979). It transfers mechanisms and insights of the first decades of generative grammar to diachronic syntax and, most importantly, identifies the 'source' or 'starting point' for any syntactic change as language acquisition (see also Paul (1920 [1880]) and Harris and Campbell (1995) for a non-generative approach with the same starting point). Subsequent work in the field (in particular Lightfoot (1991), Lightfoot (1999), Roberts and Roussou (2003), Van Gelderen (2004), Roberts (2007) and Van Gelderen (2011)) is built on the same assumption connecting syntactic change to learnability and acquisition.

In section 7.2 I questioned the usefulness of I-language in the study of diachronic syntax, because the research question in the field typically concerns observations in E-language. Within the Minimalist Program (MP), the syntactic component itself is considered to be invariant, therefore 'syntactic change' as such cannot exist (see the introduction of Biberauer and Walkden (2015) and Walkden (2014:31n14) for discussion and M. Hale (1998) who made the original point). A pure I-language approach to diachronic syntax might not exist (Walkden, 2014:31), but the progress and various breakthroughs in the field (see in particular the annual conferences on *Diachronic Generative Syntax* (DiGS) and the volumes resulting from the conferences, e.g. Biberauer and Walkden (2015)) show that it is worthwhile to keep a notion of I-language and thus a generative approach to syntactic change. This allows us to share the tools and mechanisms of the Minimalist Program analysing how language works and it furthermore gives access to related research in language acquisition.

Not all generative syntactic tools and insights can be straightforwardly applied to diachronic syntactic problems, however. In this section, I discuss some of these challenges and the solutions that have been proposed within the field of diachronic generative syntax. Continuing from the previous section, I start with the notion of variation as a possible source for language change. I then move on to various types of syntactic change such as Reanalysis and Grammaticalisation and how they can be accounted for in an acquisition-based model. Finally, I explore the dynamics of change and the possibilities and limitations of syntactic reconstruction.

Variation in generative grammar

What is syntactic change or language change in general? An instance of 'change' can be defined as a case in which the grammar of a language ('Grammar 2' or G2) that is derived from another language (Grammar 1) differs from this G1. We are thus dealing with variation between two grammars (or two languages or dialects)

over time (a historical relation or 'H-relation', as Crisma and Longobardi (2009:5) call it). One great advantage of the early generative Principles & Parameters approach was the reconciliation of the universal principles solving the Poverty-of-Stimulus problem with the parameters attempting to account for cross-linguistic variation. It specified the relation between the language experience (Primary Linguistic Data or PLD, the input for the language learner) and the innate language faculty of Universal Grammar (UG). To illustrate this: a very crude example of a universal principle could be 'combine the verb with a direct object'. An example of a parameter for a particular language could then be 'let the direct object precede the verb' resulting in languages with linear OV order. An example of syntactic change could be the resetting of that parameter, e.g. OV order changed into VO order (the so-called 'Head Parameter', cf. Travis (1984), Koopman (1984) and Pintzuk (1991), Pintzuk (2002) and Lightfoot (1991) for the diachronic example). Kroch (1989) described this as a situation of grammar competition: a language with parameter-setting 'OV order' (Grammar 1) competes with a later stage of that language in which the parameter switched to 'VO order', resulting in Grammar 2.

Upon closer investigation of the data of these and other proposed parameters, this view of a binary setting that must be switched in a catastrophic fashion turned out to be too simplistic (see also the section on *The dynamics of syntactic change* below). Examples found in the history of English OV and VO word orders suggest for example that this change consisted of various different stages. OV order with quantified and negative objects was lost at a later stage, for example, and, most importantly, the major catastrophic switch from OV to VO seems to have taken centuries to complete (see Pintzuk (2002) for evidence from Old English and Van der Wurff and Foster (1997) for surface OV up until the sixteenth-century). Questions arose on whether certain syntactic changes (always) clustered together and, if so, how and why those changes in particular and not others? Were there non-parametric changes as well and, if so, how can they be characterised and formalised within the system?

Various empirical problems with the traditional parametric approach have been put forward by Newmeyer (2005). In addition, there are specific problems of implementation. It is for example first of all controversial what triggers a certain parameter setting (cf. Dresher (1999) and Lightfoot and Westergaard (2007)): what counts as a cue? The parametric approach furthermore suffers from the Linking Problem (cf. Pinker (1984), but also Beekhuizen et al. (2014) on why this particular problem is so far not solved by any linguistic theory and therefore not just a challenge for parametric theory as described in Chapter 1 of this thesis). Finally, from the point of view of acquisition, parameters need to be learned in the right sequence and there seems to be a growing number of parameters that have to be acquired (cf. Gibson and Wexler (1994), J. D. Fodor (1998) and Evers and Van Kampen (2008)).

According to Newmeyer (2004), the parametric approach of syntactic variation has a further major disadvantage: it lacks what Longobardi (2003) termed 'evolutionary adequacy' (see also Gianollo, Guardiano, and Longobardi (2008)).

This is a new level of empirical adequacy added to the well-known three advocated by Chomsky (1964). Beyond the observational, descriptive and explanatory level, a theory of linguistics should also aim to reach 'evolutionary adequacy', i.e. why did we evolve to have precisely the type of language faculty we have today and why do we have the attested variety of languages (and not others)? Newmeyer (2004) proposes a rule-based grammar instead: variation or "language-particular differences can be captured by difference in language-particular rules" (Newmeyer, 2004:183). A major disadvantage of any rule-based system, however, as Holmberg and Roberts (2005) point out, is that it is unrestrictive in the sense that in principle 'anything goes'. This is typically not what we find in human languages, however (see also Biberauer, Holmberg, and Roberts (2014)).

More recent studies on (parametric) variation within the Minimalist Program have therefore moved the source of variation from 'switchboard-style' parameters in UG to functional features in the lexicon. This was first suggested by Borer (1984) and picked up by Chomsky in early Minimalist work:

(3) **The Borer-Chomsky Conjecture (BCC)** (M. Baker, 2008:353)
All parameters of variation are attributable to the features of particular items (e.g. the functional heads) in the lexicon.

From the point of view of first-language acquisition, this is a real advantage, because it puts the burden of learning (back) to acquiring vocabulary with idiosyncratic properties.[7] According to Walkden (2014:22-23) it furthermore makes more (and clearer) predictions about possible languages than, for example, the rule-based alternative put forward by Newmeyer (2004). According to Roberts and Holmberg (2005), parameters represent points of underspecification and as such are not really primitives of UG. The grammatical system becomes operative once these underspecifications are filled. According to Chomsky (2005) there are three factors in language design. Biberauer, Holmberg, Roberts, and Sheehan (2014) argue that parameters arise as a result of the interaction of the three factors:

1. Factor 1 ⇒ innate endowment (UG): basic operations Merge and Agree (plus a formal feature template [iF]/[uF], and a very small subset of [F]s not derivable from the input)
2. Factor 2 ⇒ primary input experience (PLD) giving evidence for movement, doubling, systematic silence and multifunctionality
3. Factor 3 ⇒ non-language-specific innate capacities: general computational conservatism of the learning device, e.g. Feature Economy (FE) and Input Generalisation (IG)

The third factor is perceived as generally applicable learning biases (Biberauer,

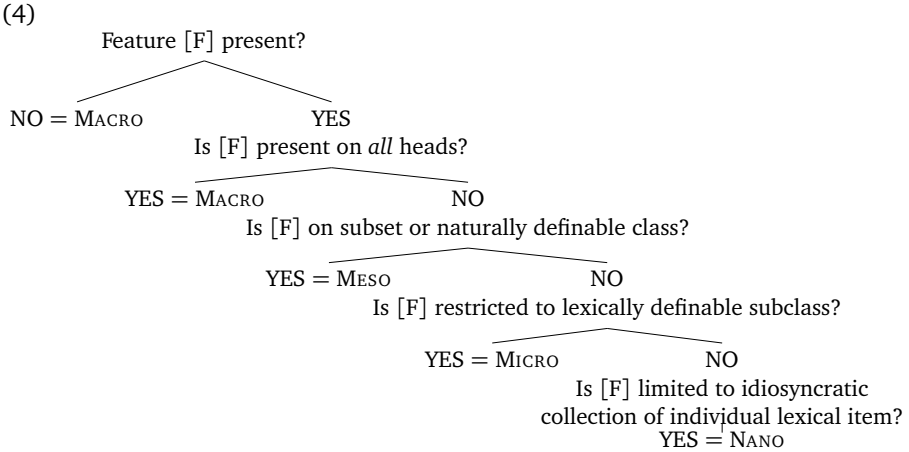
[7] As Walkden (2014) points out, such a lexical approach to variation is somewhat similar to the 'Constructicon' in Construction Grammar discussed in the previous chapter (see also Barðdal and Eythórsson (2012)). The tools and mechanisms in the Minimalist Program, however, differ considerably from those available in Construction Grammar, like the concept of 'motivation' discussed above.

Holmberg, Roberts, & Sheehan, 2014). It consists of a 'minimax' search or optimisation algorithm and is thus wholly in line with Minimalist assumptions making maximal use of minimal means. Feature Economy (FE) is generalised from Roberts and Roussou (2003) as the bias to postulate as few features as possible (i.e. possible to account for the input). Input Generalisation (IG) stipulates that learners maximise the use of the available features (see also Roberts (2007)). This kind of 'emergent parameter' approach is used as a counterargument against Newmeyer's comment on the lack of 'evolutionary adequacy' in a parametric approach to language variation. The underspecification of formal features can appear in three forms (cf. Biberauer, Holmberg, Roberts, and Sheehan (2014:108)):

1. association of formal features with (functional) heads
2. values of formal features, triggering Agree
3. purely diacritic features triggering movement

Clustered syntactic changes can now be thought of in terms of 'cascading parameters' (Biberauer & Roberts, 2008) and networks of parametric changes (Roberts, 2007), or, in line with the third-factor learning biases and the latest output of the project on *Rethinking Comparative Syntax* ('ReCoS') at the University of Cambridge: parameter hierarchies (see Biberauer et al. (2014) and much other work available via the ReCoS project website). The hierarchy consists of different levels, ranging from macroparameters (all (functional) heads share the value v_i of feature [F]), to mesoparameters (all functional heads of a given naturally definable class, e.g. [+V], share v_i), to microparameters (a small subclass of functional heads shows v_i, e.g. pronouns or modal auxiliaries), and finally, nanoparameters (one or more individual lexical items is/are specified for v_i) (Biberauer et al. 2014) and (Ledgeway, 2016):

(4)



Examples of hierarchies and parameters on different levels are given by Roberts

(2012) and Biberauer et al. (2014) (see also the contribution to the volume *Parameter Theory and Linguistic Change* (Galves, Cyrino, Lopes, Sandalo, & Avelar, 2012)). From a diachronic perspective, macroparameters are expected to be highly stable (e.g. rigid head-finality). Null-subjects in the history of Romance are much less stable and would count as a mesoparameter. Microparameters are even more likely to change over time (e.g. English modals). Nanoparameters, finally, could literally come and go with one lexical item (e.g. the relics of the Conditional Inversion in English).

This framework thus gives us very concrete tools to describe and explain variation, either synchronic or diachronic. It leaves the nature of the formal features unspecified, however. This issue is related to a final question concerning variation: does 'free' variation or 'true optionality' in one single grammar exist?

According to Biberauer and Richards (2006), there are indeed cases in which 'the grammar does not mind'. In their study on the EPP feature, they show the *option* of pied-piping of the whole phrase bearing the interpretable φ-feature with examples from auxiliaries in Afrikaans. Since we often have very little or no information about the sociolinguistic situation in earlier stages of languages, it is difficult to make any such claims in diachronic syntactic studies. If two grammars are 'in competition' (as advocated by Kroch (1989) and Pintzuk (1991) among others), we cannot be certain whether this variation is a genuine case of 'true optionality' or whether the variants were distinct on some (sociolinguistic) level, with evidence for this having been obscured over time (Roberts, 2007:331). Walkden (2014) finds some further issues with Biberauer & Richards's necessary rejection of derivational determinism asking why there would be no difference between, in their pied-piping example, moving a small or a big category and what determines which of the two options will be taken. In the end, speakers/writers *do* make a decision, but if an algorithm is non-deterministic it is unimplementable. He therefore concludes that "[f]or a given selection of lexical items, there is only one possible derivational outcome" (Walkden, 2014:23). This means there can be no 'true optionality' or 'free variation' within a single grammar. He furthermore adds that speakers have access to multiple varieties of their language and that there is a 'user's manual' regulating the choice between them (cf. Culy (1996:114)). This variation can be subtly conditioned, not semantically (in the strict truth-conditional sense), but functionally or contextually. Walkden argues that these sociolinguistic factors or 'social knowledge' should be treated as part of the lexicon. As such they can enter the derivation like any other type of formal feature (Walkden, 2014:28-31).

Certain types of formal features are (relatively) uncontroversial, such as referential features (φ-features), negative (polarity) features or features related to questions, such as wh-features. The exact nature of the EPP feature is still an issue of debate, but the fact that there must be some sort of movement-triggering feature (as an 'Edge feature' or simply in the form of a diacritic caret ˆ) is not. As discussed in the previous chapter, a wide variety of information-structural features has been proposed, such as TOPIC, FOCUS or ANAPHOR. Whether there should be more or fewer of those and whether that might be language-specific is still a matter of

ongoing cross-linguistic research. As I have argued in the previous chapter, for Middle Welsh, we need Topic and Focus features at the very least (with possibly, an added distinction between different subtypes of topics, such as Aboutness, Familiar and Contrastive). A further set of 'social' features might exist, as Walkden (2014) suggests to resolve the issue of 'free variation', but the exact nature and effect of those in Middle Welsh is difficult to ascertain on the basis of the corpus under investigation.

In conclusion, within the Minimalist Program, variation can still be thought of as parametric variation with the locus of parameters in the formal (functional) features of the lexicon. Clusters or cascading changes in the grammar of a language can be captured by a hierarchical structure of parameters. With these tools in mind, I first discuss two major mechanisms of syntactic change before moving on to the complex issues concerning actuation and diffusion of syntactic innovations.

Types of syntactic change

In principle, any element of the grammar that can exhibit variation (within a language or between different languages/dialects) could be subject to change. Diachronic changes have been studied in the core domains of argument structure (thematic roles and grammatical functions, e.g. English psych verbs (Allen, 1995)) or passives (Dreschler, 2015) and complementation (e.g. in Latin *ut*-clauses (Vincent, 1988)). The earlier diachronic syntactic descriptions furthermore focussed on major changes in word order. The change from OV to VO in English already discussed above, could for example be seen as a change in head-directionality. But a simple parametric switch from head-final to head-first does not adequately account for the complex data in the history of English. However, parametric change in the much more fine-grained sense of change of functional features in a parameter hierarchy within a Minimalist framework could be the right approach to all these types of change.

Syntactic innovation can also change the underlying structure of a certain pattern without necessarily modifying the surface manifestation. This is called syntactic reanalysis (see, among others, Harris and Campbell (1995)). The preconditions for diachronic reanalysis are structural ambiguity and a preference for simplicity. The hearer assigns a specific parse to the input that is different from the structure assigned by the speaker (Walkden, 2014:39). An often-cited example is the reanalysis of *for...to* in Middle English creating a complementiser marking Case on subjects in nonfinite clauses as presented by Fischer (1992:330-334) and Fischer, Van Kemenade, Koopman, and Van der Wurff (2000:214-200):

(5) a. PREDICATE [$_{PP}$ [$_{P}$ *for* NP] [$_{TP}$ *to* VP] ⇒
PREDICATE [$_{CP}$ [$_{C}$ *for*] [$_{TP}$ NP *to* VP]]
b. It is bad [$_{PP}$ for you] [$_{TP}$ to smoke] ⇒
It is bad [$_{CP}$ for [$_{TP}$ you to [$_{VP}$ smoke]]]

Willis (2016) cites this example under 'spontaneous syntactic innovation' and notes that this standard account is sharply criticised by, among others, Garrett (2012:55-66). If reanalysis becomes a possibility at any time (but is never required) it fails to explain why it actually happens (see also the discussion on triggers of change and acquisition in the next section). Just as the above-mentioned types of syntactic change, diachronic reanalysis of this kind might be reduced to a parametric change. In this particular case the category of the lexical item *for* changed from preposition to complementiser. Since the preposition *for* still exists in other constructions, it looks like a second lexical item *for* was created in the lexicon with a different featural and categorial makeup so that it can function as a complementiser selecting a TP (instead of a preposition selecting an NP or DP).

Another very well-studied area of syntactic change is grammaticalisation. Grammaticalisation is a specific type of reanalysis in which 'less grammatical items', for example simple open class lexical (content) items, become 'more grammatical'. In other words grammaticalisation is "the dynamic, unidirectional historical process whereby lexical items in the course of time acquire a new status as grammatical, morphosyntactic forms, and in the process come to code relations that either were not coded before or were coded differently" (Traugott & König, 1991). The term was first coined by Meillet (1958 [1912]) and presented in comprehensive discussion in, among others, Heine and Kuteva (2002). Apart from being defined as a historical process, the term is also used to describe a research framework (Hopper & Traugott, 2003). According to Campbell and Janda (2000), there are different processes involved in grammaticalisation, such as phonological reduction (e.g. English 'let us' > 'let's'), loss of 'syntactic freedom' (e.g. French *pas* 'step' > *pas* as a negative marker), pragmatic inferencing (e.g. English 'since' from temporal sequence to inferred causation) and semantic bleaching (e.g. German *Mann* 'man' > *man* 'one, some human being').

Campbell (2000:141) argues that grammaticalisation is in itself not a mechanism of change. It relies primarily on the above-mentioned mechanism of reanalysis and also on the extension of the construction in question. As such it could also be viewed as a parametric change or a change in the featural makeup of a lexical item. As Roberts and Roussou (2003) describe it in a formal (generative) account grammaticalisation is a categorial reanalysis driven by change in properties of functional heads. When a new exponent of a functional head F is created, it may also involve creating new parametric properties (triggering Agree or internal Merge) associated with that head. A good example of this type of reanalysis in the history of Welsh is the specifier-to-head reanalysis of personal pronouns becoming complementisers (Willis, 2007a). Another example that I will describe in greater detail in the second part of the chapter is the grammaticalisation of the so-called *sef*-construction in Middle Welsh.

The logical problem of language change

As noted above, within the generative model language change is defined as two distinct grammars in a historical relation. This leaves two logical possibilities for

the locus of change. First, this could be first language acquisition whereby syntactic change is driven by 'abductive' reanalysis (i.e. the result of a transmission failure) (cf. Andersen (1973), Lightfoot (1999) and Van Kemenade (2007) among many others). A second option could be a change in the internalised grammar of an adult speaker. Such change, however, is not considered to constitute a case of 'real' diachronic change "until a future generation of speakers have adopted the mixed system as their own." (Faarlund, 1990:10). Although some cases of this type of change have been argued to be fully completed during one generation, I focus on first-language acquisition here.

If children are generally successfully acquiring the syntactic system of the language of their parents for generations, how can they suddenly be unsuccessful in doing so? This 'logical problem of language change' has received much attention in diachronic syntactic literature (cf. Clark and Roberts (1993:12), Kroch (2000:699-700), Lightfoot (2006:15) and Willis (2016)). Even non-generative approaches must address this question if they want to speak of causation in syntactic change.

Figure 7.1 shows the traditional Z-model of abductive change presented by Andersen (1973:767). The main idea behind this model is that the child may make an error of abduction and mistake a similar case of a structural analysis for the actual case uttered by the speaker of Grammar 1. This can happen because there is no direct link between Grammar 1 and Grammar 2: contact between the two I-languages is mediated by the E-language output. The mismatches that can arise in such situations are in fact the reanalyses we find in syntactic innovation.

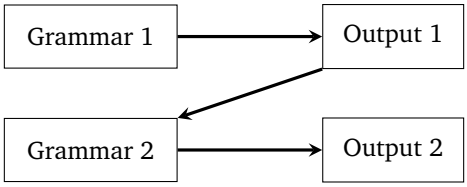


Figure 7.1: Andersen's (1973) Z-model

The difficulty mostly lies in the assumption that first-language acquisition is fully deterministic: children always succeed in acquiring the language perfectly. This Z-model is highly idealised, however, since the primary corpus ('Output 1') is never generated by the grammar of a single individual. It consists necessarily of indeterminate evidence: a finite set of sentences uttered by different individuals, each of whom have a grammar that is not 100% the same as 'Grammar 1' in the model (see Niyogi and Berwick (2009) and Walkden (2012) for further discussion).[8] In an attempt to solve this paradoxical issue in acquisition, Roberts and Roussou (2003) propose a form of 'weak determinism' saying that "the goal of language acquisition is to fix parameter values on the basis of experience; all parameter values must be fixed, but there is no requirement for convergence with the adult grammar" (Roberts & Roussou, 2003:13). Some models of L1 acquisition depart from the

[8] 'Abduction' in itself might furthermore not be the right description of the phenomenon in diachronic syntax (see Lass (1997), Deutscher (2002) and Walkden (2011)).

deterministic assumption altogether. The models proposed by Gibson and Wexler (1994) and C. D. Yang (2002) each contain probabilistic components. In such a scenario, the child can posit more than one grammar (i.e. parameter setting) on the basis of the input she receives. The PLD, in other words, is ambiguous and/or leaves certain options unspecified. A probabilistic model (based on frequency of syntactic structures that count as cues or triggers for a certain grammar, for example) helps the child to determine which grammar to choose.

The type of syntactic innovation or reanalysis based on the ambiguous structure of a sentence in the input could be categorised as 'spontaneous innovation'. Such a purely endogenous solution to the actuation problem lacks explanatory power. Why, for example, does the reanalysis take place at a given time and place, as Weinreich et al. (1968) already pointed out. Willis (2016:3) notes, however, that "if misparsing by children (leading to reanalysis) is distributed randomly in the population (perhaps with some social contexts, such as population mixing, favouring it), it would be pointless to expect more" (of an explanation). Our task then is first of all to accurately describe the conditions and pathways of the reanalysis in a plausible way (i.e. not violating any principles of grammar that are well-established from research into synchronic variation and L1 acquisition). Furthermore, we need to investigate whether the reanalysis is indeed randomly distributed and, to the extent this is possible in our historical context, what the possible social contexts are favouring one pattern rather than the other.

Typological approaches form another kind of endogenous solution proposed already in the earliest stage of historical linguistic research (from the Universals listed by Greenberg (1963) to Indo-Europeanists like Lehmann (1973) and Vennemann (1974)). The core argument consists of applying synchronic restrictions on the ways in which languages combine features to diachronic syntactic changes. These approaches "make system-based predictions about possible and impossible changes" (Willis, 2016:3). In this manner possible pathways for changes are predicted, but the changes themselves do not *have to* occur. If they do, these pathways can still not predict when this will happen (cf. Hawkins (1990:99) and Willis (2016:§3)). Within a generative approach Biberauer, Sheehan, and Newton (2010) argue that the 'Final-over-Final constraint' (FOFC) restricts possible diachronies. For diachronic syntax, FOFC predicts that a change from head-final to head-initial word order must follow a particular order to avoid head-final over head-final structures. A possible explanation for this could lie within the cognitive domain as a processing preference. These types of cognitive preferences may in fact lie behind more (or all?) observed typological universals. For this highly deterministic approach it is first of all important to confirm the cognitive claims with data from thorough psycholinguistic experiments. It is furthermore of crucial importance to have a comprehensive description and adequate analysis of all cross-linguistic data of the phenomenon under investigation.

Another solution to the actuation problem is based on language usage. An increased frequency of use can, for example, explain cases of grammaticalisation. If a particular sequence is often used, but rarely varied, children could acquire it as

a single unit. Within Construction Grammar, this is called 'constructionalization'; in other frameworks, it is simply referred to as 'lexicalisation' (Willis, 2016:6). This moves the 'why'-question to language usage, i.e. why was this particular pattern used more frequently? The answer to this might be irretrievable as long as we do not have access to an accurate description of the sociolinguistic history of the period under investigation.[9] Again we can nonetheless aim to identify the factors that might have aided the increased frequency of a particular pattern.

A further important question arises here (which is also relevant in the context of change in general): how frequent does the pattern have to be to be 'grammaticalised', 'lexicalised' or 'reanalysed'? In other words: what is the so-called 'tipping point' for Grammar 1 to change to Grammar 2 and can this be described in terms of (relative) frequency alone? In order to answer this question, we first need to be clear on the exact cue or trigger for the change. In the case of lexicalisation or grammaticalisation, this is often very straightforward: the frequency of the pattern in the target context (vs. other contexts) could be retrieved from a historical corpus. Assuming we are dealing with a well-balanced corpus that accurately reflects different stages of the language,[10] we can define the frequency required for the change with relative ease. In studies of the acquisition of a particular type of word order, e.g. V2 word order, however, the situation is more complex. Sentences with initial subjects, for example, cannot count as 'triggers' for the child to postulate a V2 grammar (even though subject-initial sentences are V2 in many Germanic languages). The evidence would not be sufficient, however, because an English-like SVO grammar is also possible on the basis of that input. To convince the child to opt for a positive 'V2 setting' of the parameter in question (see section 6.4 below), she needs a significant input of non-subject-initial word orders (followed directly by a finite verb). Lightfoot (1999:154) estimated that roughly speaking, an average of 30% of the sentences should have this type of $XP_{Non-Subject}$-V_{Fin} order to convince the child that her language has a V2 grammar. With syntactically annotated corpora, this estimated number could be compared to a sample of real data. In a corpus of Modern Dutch, C. D. Yang (2000:114) found that 23% of the sentences had $XP_{Non-Subject}$-V_{Fin} order. Since Dutch children successfully acquire V2, he concluded that Lightfoot's estimation of 30% might be too high. On the basis of the Dutch corpus study it seems that 23% should be sufficient. Westergaard (2009:67) conducted a similar study of Norwegian corpora. She finds only 13.6% in her child-directed corpus. These numbers found in spoken corpora might differ in historical written corpora, because it is not always clear to what extent the written data reflect the spoken language at the time. This type of research in first-language acquisition is nonetheless extremely useful in attempting to accurately describe situations of historical change.

The actuation problem in historical syntax can also be 'solved' by considering

[9] See also Lass (1980:101-103) and Walkden (2012:897-898) on Popper's methodological version of the principle of causality (Popper, 1968:67) and why it might not be appropriate to ask ourselves this particular kind of even further-removed or deeper 'why'-questions in the study of historical syntax.

[10] This is a somewhat idealised situation, because there are various practical limitations building a well-balanced historical corpus, as discussed at length in Chapter 2.

changes in other parts of the language. Apocope or the loss of final syllables discussed in section 7.2.2 can, for example, lead to the loss of a case system, because the morphological endings no longer function as distinctive features. 'Phonological erosion' (whatever causes it) is mentioned by Willis (1998) as the crucial trigger for the loss of V2 in Early Modern Welsh. When the preverbal particles *a* and *y* disappeared, the acquisition of the V2 system became obscured at first and then completely impossible. As shown in the previous chapters, Middle Welsh allowed V3, V4 and V5 orders with adjuncts in the preverbal domain alongside the standard V2 'abnormal' and 'mixed' sentences. With the loss of the preverbal particle *y*, adjunct-initial sentences could easily be reanalysed as Adjunct + VSO orders (see also section 7.2.1 above). Pronominal subjects in initial position were reanalysed as main-clause complementisers after the loss of the preverbal particle *a* (see section 7.3.2). Object-initial orders were very infrequent already towards the end of the Middle Welsh period. According to Willis (2016:9) (building on Willis (1998) and Willis (2007a)), this phonological source of change led to other parallel changes as well, such as the reanalysis of the expletive pronouns as affirmative particles. One phonological change can also lead to another, e.g. a change in the stress pattern can lead to the reduction of vowel quality or even syncope or apocope. What ultimately triggers the initial change in this case is difficult to ascertain. Again psycholinguistic experiments on language production could prove revealing, although the question remains why certain changes were not 'triggered' in the same way centuries earlier, for example.

This leads us to language-external approaches to the logical problem of language change. In principle, external sources in the form of language contact do not necessarily lead to language change. Children are perfectly capable of acquiring more than one language if they get the right input in the earliest stages of their lives. They grow up to be bilingual, fluent in two (or even more) languages or dialects and they can distinguish and use the two grammars without any problems (see also the study on Welsh-English bilingual code-switching and the conclusion of grammatical continuity rather than change by P. Davies and Deuchar (2010) cited above). Syntactic change, however, also occurs in contact situations. According to Meisel, Elsig, and Rinke (2013), a change in the core grammar can in fact *only* occur when non-native speakers form a large part of the speech community (see Meisel et al. (2013:171-182) and Willis (2016)). As discussed above in the section about language shift, syntactic changes are often considered to require a specific type of contact. One possible situation would be the shift of speakers of the substrate to the superstrate language, keeping grammatical features of their substrate so that they become embedded in the superstrate language. Since instances of syntactic change have also been reported in situations without language contact, a complete rejection of any kind of endogenous approach to syntactic change seems to be unfeasible (Willis, 2016:10). This finally brings us to the notion of 'inertia', as formulated by Longobardi (2001) and Keenan (2002):

(6) "Syntactic change should not arise, unless it can be shown to be *caused*" (Longobardi, 2001:278)

(7) "Things stay as they are unless acted upon by an outside force or Decay." (Keenan, 2002:327)

In the context of the acquisition-based model, the Inertial Theory stipulates that a grammar can only change if the conditions in the process of acquisition have changed. According to Willis (2016:11), this then solves the timing part of the actuation problem because reanalysis of a particular structure only occurs at the time something else changes (e.g. phonological erosion or loss of a lexical item). To a certain extent, this 'solution' is no more than a shift of locus of the problem: why would phonology or morphology not be equally inert in this theory? Walkden's (2012) thought experiment about a child failing to acquire V-to-C movement in her grammar (G2), because she never hears wh-questions in the language of her parents (G1) is very insightful in this context. The grammar of her parent(s) (G1) did not change in any way, it just happens to be the case that direct questions were never asked when the child was around, so V-to-C-movement was not part of the PLD. Although this situation might be extremely unlikely, the main argument holds: the 'cause' of change (Longobardi, 2001) consists of the non-occurrence of a particular pattern. The ultimate reasons for this non-occurrence could be a wide variety of extralinguistic events and even chance and human intentionality (i.e. the 'planning' of utterances) needs to be taken into account as well. Walkden (2012:896) thus concludes that the notion of causality in the Inertial Theory is so broad it is rendered entirely vacuous, because it cannot make any useful empirical predictions.

To conclude this section, research on processes of first-language acquisition can help historical linguists characterise the changes more accurately. The acquisition-based approach advocated within the Minimalist Program by the ReCoS project includes typological, cognitive and acquisitional biases (e.g. Input Generalisation and Feature Economy) that not only help predict pathways of changes, but might also shed light on 'what has not happened' and why this is the case. Computational models of acquisition and the competing-grammar approach advocated by C. D. Yang (2002) can give us further insights in predicting changes based on frequencies of patterns containing cues or triggers for a certain innovation. From an empirical point of view, historical linguists should not only describe the syntactic innovation itself, but also the necessary change in conditions (in the acquisition process) that 'triggered' the innovation (how did it happen and why did it happen in this particular way and not vice versa). It is furthermore necessary to try to identify both endogenous and exogenous factors "which might have aided a variant grammar in persisting or becoming more prevalent" (Walkden, 2012:899). This last notion is related to the diffusion of 'reactuation' of syntactic innovations, which is the topic of the next section.

Dynamics of change

Parametric change is traditionally described as having two main characteristics. It is:

1. catastrophic ⇒ when it changes suddenly and irrevocably at a given moment
2. internal to the inquirer ⇒ this means that in principle, it is entirely independent of the child's cultural, social or historical background

Since many syntactic changes observed in diachronic data seem to be gradual, rather than abrupt, the 'catastrophic' nature of parametric change has received much criticism. Language can be transferred in two different ways: transmission in first-language acquisition and diffusion from adult to adult. In this section, I discuss the dynamics of change and possible ways to solve the gradual-abrupt paradox of parametric change.

After the introduction of a novel form (the syntactic innovation or 'actuation' of a change), the form can spread through a speech community. From historical corpora we often observe a period of variation until one system (G2) takes over from the other (G1). This period can be described as individuals having two grammars in competition, formal optionality, diglossia and/or diffusion of the syntactic innovation. This rate of replacement from one grammatical option to another often shows the same 'slow-quick-slow' pattern (as observed by, among others, Osgood and Sebeok (1954:155), Weinreich et al. (1968:113-14)). Kroch (1989) analysed syntactic changes such as the replacement of *have* by *have got* in British English from 1700 to 1935 and the loss of the verb-second constraint in Middle French from 1400 to 1700. He concluded that the 'slow-quick-slow' rate of change can be modelled by a logistic function showing an s-shaped curve when the frequency of new vs. old forms is plotted against time as shown by the equation in Figure 7.2:

$$p = \frac{e^{k+st}}{1 + e^{k+st}}$$

Figure 7.2: S-curve logistic function by Kroch (1989:204) with: p = the frequency of the innovation, t = time, s = the slope of the function, k = the y-intercept (the frequency of the innovation at $t = 0$) and e = Euler's number (approx. 2.71828)

This 'Constant Rate Hypothesis' (CRH) shows the grammars in competition change gradually through a population or within individuals who have access to one of these grammars more readily than others over time. This is in effect a situation of syntactic diglossia (Kroch, 2000:722): speech communities (and individuals) synchronically instantiate several grammatical systems. According to Willis (1998), the same actuation process may be triggered in multiple speakers, in which case apparent diffusion through the speech community may actually be an instance of 'multiple reactuation' (Willis, 1998:47-48). The increase in frequency of the syntactic innovation may furthermore be due to sociolinguistic factors: an abrupt parametric change can therefore appear to be gradual. Lexical diffusion and microparametric changes (Kayne, 2000:3-9) may help keep up the 'mirage of gradualness' as a cushioning effect: "a series of discrete changes to the formal features

of a set of functional categories taking place over a long period and giving the impression of a single, large, gradual change" (Roberts, 2007:300). More studies in first-language acquisition in the context of language contact situations, dialect research, and code-switching will play an important role in refining and explaining the 'S-curved' model. One possible way forward is to include geographical factors into historical syntactic models of change. If geospatial information about the distribution of syntactic innovations is available (e.g. texts from different areas over a certain period of time), this can be integrated into a logistic regression model. Willis (2014) shows how this type of geographically weighted regression model in dialect research can be applied to the innovation and diffusion in the pronominal system of northern varieties of Welsh over the last 150 years. He combines the Constant Rate Hypothesis with geospatial data trying to show the diffusion of syntactic innovations in the speech communities of North Wales. This model can be tested and further refined by studies of 'recent' syntactic innovations in dialect areas for which this type of information is available.

Apart from the speed of change and its geographical diffusion, the direction of syntactic innovations has been the topic of various studies in diachronic linguistics. Especially in studies concerning grammaticalisation, these processes often follow well-defined pathways. However, cases of 'degrammaticalisation' have been reported as well, in which case the directionality of change seems to be reversed (e.g. Willis (2007b), Norde (2009) and Rosenkvist (2010)). The diachronic syntactic 'principles' proposed by Van Gelderen (2009) are similarly laying out certain pathways for change:

(8) **Head Preference Principle** Van Gelderen (2009:136)
Be a head, rather than a phrase.

(9) **Late Merge Principle** Van Gelderen (2009:136)
Merge as late as possible.

These 'principles' are not uncontroversial (cf. Motut (2010)) and, according to Walkden (2014:42) if we adopt the I-language perspective on historical syntax, an independent principle governing the direction of change cannot exist. Willis (2011a:421-424) also notes that if there is any form of universal directionality, it can be reduced to 'local directionality' meaning that the interaction of the acquisition algorithm with the PLD leads to predictable reanalyses. Van Gelderen's Principles, to the extent they are universal, might thus be the result of preferences in the acquisition process. Such acquisitional biases were already discussed in the parametric hierarchy approach above. In this context, a 'pathway of change' is equivalent to the child being pressured to postulate the simplest possible system, for example. Input Generalisation as a principle in acquisition states that a generalisation - if possible - is extended over the widest possible domain (until met with counter-evidence). Functional features may also become less transparent, leading to a complete loss and thus simplification of the system. Changes can occur moving up or down the hierarchy: they might be constrained for cognitive reasons, but change is not unidirectional per se.

To conclude, the exact 'dynamics of change' seem to be specific rather than universal. Since the sociolinguistic context can play an important role in the spread of changes, this should be taken into account (to the extent this is possible in a historic context) in describing the process of transfer in speech communities. Syntactic innovations are not inherently unidirectional, although biases in first-language acquisition in the form of 'local directionality' can lead to predictable reanalyses. The S-curve of the Constant Rate Hypothesis, combined with - where available - geospatial data can further help a well-informed analysis of syntactic changes. However, if we do not have access to ample data (because from the time of 'our British Celt' vernacular texts have not survived or were never written down in the first place), our task of analysing the origin of a particular syntactic pattern is severely complicated. In the next section, I therefore discuss the possibilities and limitations of syntactic reconstruction.

Syntactic reconstruction

In this section I finally turn to the successful 'Comparative Method' in phonological reconstruction mentioned in the introduction. In 1900, Berthold Delbrück, one of the greatest early researchers in the field of historical linguistics expressed his doubts about the possibility of reconstructing syntax in the same way this is done for the lexicon, phonology and morphology (Delbrück, 1900 [1982]:v-vi). Further attempts were nonetheless done by Lehmann (1972), Hopper (1975) and Kiparsky (1995). Various problems arise in the reconstruction of syntax, however, as pointed out by, among others, Lightfoot (1999). I first briefly sketch the fundamentals of the method of comparative reconstruction and then discuss the problems it might cause in the field of historical syntax.

The first step of the comparative method consists of finding a set of corresponding words in (potentially) related languages. In (10) and (11) below, I show a somewhat simplified example from the Indo-European language family for the English adjective 'new'. An important part here is both the formal as well as the semantic similarity to form 'cognates' (form-meaning pairs). In this case, the adjectives in the different Indo-European languages all mean 'new' and can thus be considered proper double (form *and* meaning) cognates. The set below thus qualifies as a proper correspondence set (see Beekes (1995:196) and Schrijver (1995:283ff) for the forms in IE and Celtic respectively):

(10) Sanskrit: *návya-*, *náva-*
Gothic: *niujis*
Hittite: *newa-*
Greek: *néos*
Latin: *novus*
OCS: *novъ*
Tocharian B: *ñuwe*
Old Irish: *núae*
Welsh: *newydd*
Breton: *nevez*
Middle Cornish: *noweth*
Gaulish: *Novio-(magus/dunum)*

The next step involves the proper alignment of the examples, starting from the stem of the adjective, as shown by the sample of languages in (11):

	Sanskrit	n	á	v	-
	Latin	n	o	v	-
(11)	Greek	n	é	-	
	Welsh	n	e	w	-
	PIE	*n	?	?	

From the aligned correspondences, we can then reconstruct the sounds by comparing the forms. For the first letter, this is easy: since initial *n-* appears in all languages, we postulate initial **n-* for the form in the proto-language, in this case Proto-Indo-European (PIE). The vowel and second consonant are less straightforward, because the different languages exhibit different phonemes, or, in the case of the Greek consonant, nothing at all (in that position of the word). To reconstruct these PIE phonemes, we have to find regular sound correspondences in the respective languages, i.e. does Sanskrit short *á* always correspond to Greek *é* and Latin *o*? Does that depend on the phonological context and/or can we find regular sound changes? Sanskrit short *a*, for example, regularly responds to either an *e* or *o* in Greek. However, by regular sound law (Brugmann's Law, cf. Beekes (1995:138)), PIE short *o* in open syllable became a long *ā* in Sanskrit. Since we find a short *a* in an open syllable in the adjective 'new' in Sanskrit *náva-*, it is unlikely this goes back to PIE *o*. The *o* in Latin, however, usually means we have to reconstruct an *o* in PIE as well. However, again by regular sound change PIE **e* became *o* in Latin before *u̯*, ł and *mo* (Beekes, 1995:66). This combined evidence from the regularity and 'exceptionlessness' (*Ausnahmlosigkeit*) of sound changes upon which the Comparative Method heavily relies, forces us to conclude PIE **e* can be the only right vowel to reconstruct. Another final point is the question of orthography and to what extent it is representative of the actual sound. The *v* and *w*, for instance, could both represent the glide or semi-vowel *u̯*. In Greek, furthermore, this **u̯* regularly disappears intervocalically (cf. Beekes (1995:135)). To conclude, the reconstructed form of the stem of the adjective that means 'new' in many different Indo-European languages is PIE **neu̯-*.

The reconstructed 'product' of the Comparative Method by definition does not represent a real language: it is timeless and non-dialectal (cf. Walkden (2014:37)). Successful reconstruction does not need a causal explanation per se: the result is valuable nonetheless, since it shows how the phonological (and morphological) systems of languages and their vocabulary has changed. If we want to apply the same method to syntax, however, we run into problems at the very first step: the 'correspondence problem'. As Calvert Watkins already pointed out in the 1970s, "the first law of comparative grammar is that you've got to know what to compare" (C. Watkins, 1976:312). Walkden (2009) (and subsequent work, Walkden (2014:52), amongst others) conclude that the double cognacy condition (the corresponding form-meaning pair in, for example, the vocabulary item 'new' above) cannot be easily met, because sentences are never the same. Certain idioms or

stock phrases might be compared in several Indo-European languages, but it is impossible to compare whole sentences because sentences are not transmitted as such across generations. Further criticism at attempts to reconstruct syntax were formulated as the 'directionality problem'. As discussed in the previous section, syntactic change is not inherently unidirectional. A change from OV to VO word order could in principle also be reversed. This problem, however, is not necessarily restricted to the syntactic domain. There might be phonetic tendencies, for example, to voice consonants in between vowels, but a change of *o* to *a* could in principle also be reversed. The same could be said about the 'reanalysis problem' stating that grammar must be created again by each new learner: phonological systems need to be learned in the same way (cf. Lightfoot (1979)).

Within the Minimalist Program, a possible solution to the correspondence problem again lies in the Borer-Chomsky Conjecture (BCC). Recall from the beginning of this section that the BCC sees the functional features in the Lexicon as the source of all variation. According to Walkden (2014:55-60), if these functional features take a phonological form as functional items we might reconstruct those in the context of appearance in attested sentences of the daughter languages. An example of this is the reconstruction of the free relative in Brythonic, the predecessor of Welsh, Breton and Cornish in Willis (2011a).

It is difficult, if not impossible to formalise syntactic reconstruction in a framework based on phrase-structure rules (Principles & Parameter theory, Newmeyer's rule-based system or Lexical Functional Grammar) or constraints (e.g. HPSG). An item-based approach like this would in principle work for both derivational as well as representational models (cf. the 'Constructicon' in Construction Grammar and discussion in Walkden (2014)). From the perspective of the Minimalist Program, syntactic primitives are considered to be stored in the lexicon. These functional features form the basis of syntactic variation and can be reconstructed if they take a phonological form in the daughter languages of the proto-form we want to reconstruct.

Interim Summary

In this section I presented several problems in the study of diachronic syntax and how they can be tackled by tools and mechanisms within the framework of Generative Grammar, in particular the most recent version of hierarchical parametric theory in the Minimalist Program. If we assume the existence of an innate capacity for acquiring grammars, we can use insights from synchronic research into formal syntax as well as mechanisms from language acquisitions. This is a considerable advantage in the study of diachronic syntax, because the available data is often limited. Understanding how the grammar of a language is *acquired* helps us understand how grammar can *change*. The question of *why* certain syntactic innovations appeared at a given time and spread through the speech community (problems of actuation and transfer, via transmission and/or diffusion) is more difficult to answer. Evidence for detailed sociolinguistic situations in earlier days is often just as scarce as the extant manuscript sources of the language under investigations.

Adopting a generative acquisition-based approach to diachronic syntax can help us define the exact conditions and/or context in which innovation can and cannot occur and how they can trigger further changes. We could furthermore identify endogenous and exogenous factors playing a role in making variant grammars more prevalent. Finally, the concept of 'multiple reactuation' in speech communities as well as models like the Constant Rate Hypothesis (possibly combined with dialectal (geospatial) data) provide us with a much better understanding not only of specific innovations, but also of the processes involved in syntactic change in general. In the remainder of this chapter, I use the tools and mechanisms of the Minimalist Program discussed in this section to analyse two syntactic innovations in the history of Middle Welsh: the grammaticalisation of the *sef*-construction and the rise of the Abnormal Sentence.

7.3 Diachronic syntax in Middle Welsh

In this section I focus on two syntactic innovations in Middle Welsh: the grammaticalisation process of the identificatory copular clause or '*sef*-construction' and the rise of the Abnormal Sentence. The synchronic syntactic analyses of these constructions were already presented in the previous chapter. Here I present a diachronic analysis in a generative (Minimalist) framework.

7.3.1 Grammaticalisation of the *sef*-construction

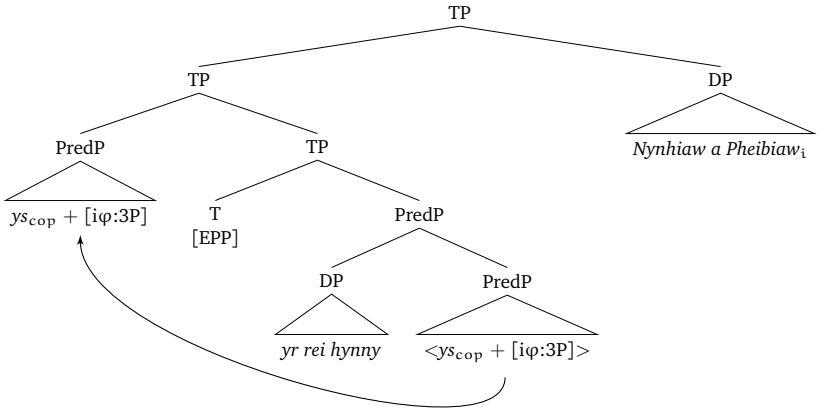
In Chapter 6 I showed various kinds of copular constructions in Middle Welsh. They exhibit different word order patterns and predicate forms (with or without the overt predicate marker *yn*), depending on the information-structural status of the subject or predicate. Predicates that identified the subject could be focussed in Old Welsh by means of a cleft construction, shown in (12a). In Early Middle Welsh, this construction is also attested once with a plural predicate, as in (12b).

(12) a. *issem i anu Genius*
be.PRES.3S.it 3MS name Genius
'that's his name, Genius' (Old Welsh gl. *Genius* in MC - T. A. Watkins (1997:579))

b. *Ys hwy yr rei hynny, Nynhyaw a Pheibyaw*
be.PRES.3S they the ones DEM.P Nynniaw and Peibiaw
'Nynniaw and Peibiaw are those ones' (Lit. 'It's them, those ones, ...)(Middle Welsh CO 598)

I argued that the derivation of (12b) is very similar to the one outlined for Inverted Copular Clauses in Scots Gaelic by Adger and Ramchand (2003). In these sentences, the copula is the head of the Predicate Phrase. It moves to SpecTP to satisfy T's [EPP]-feature pied-piping the complement, in this case the anticipatory predicate third-person plural pronoun *hwy*. The real predicate, co-indexed with the anticipatory predicate, is first-merged adjoined to TP:

(13)



This construction forms the starting point of the reanalyses that occurred in the Middle Welsh period. In texts from this period, we find many variants of this construction. In the following, I argue that these variants show five different stages of the process of grammaticalisation and reanalysis. The examples below represent these five subsequent stages:

Stage 1 - Cleft + focussed predicate (*ys + ef/hwy*)

(14) a. *iss em i anu Genius*
be.PRES.3S it 3MS name Genius
'that's his name, Genius' (Old Welsh gl. *Genius* in MC - T. A. Watkins (1997:579))
b. *Ys hwy yr rei hynny, Nynhyaw a Pheibyaw*
be.PRES.3S they the ones DEM.P Nynniaw and Peibiaw
'Nynniaw and Peibiaw are those ones' (Middle Welsh CO 598)

Stage 2 - Copula + Anticipatory Predicate merge (*ys ef* > *sef*)

(15) a. *Sef gwreic a uynnawd gwreic ieuank*
sef woman PRT want.PAST.3S woman young
'That was the woman he wanted, a young woman.' (YBH 6)
b. *Sef __ a doeth dy nyeint*
sef PRT come.PAST.3S 2S nephews
'That's who came, your nephews.' (WM 89.35)

Stage 3 - Expletive focus marker *sef*

(16) *Sef a wneuthum inheu (...) mynet*
sef PRT do.PAST.1S I (...) go.INF
'This is what *I* did, I went (...)' (WM 492.3 - Watkins 1997:586)

Stage 4 - Loss predicate focus

(17) *Sef a wnaeth pawb yna moli Duw*
sef PRT do.PAST.3S everyone then praise.INF God
'Everyone then did this, they praised God.' (Dewi 4.17)

Stage 5 - Expletive *sef* reanalysed as adverb

(18) *Sef y clywei arueu am ben hwnnw*
sef PRT hear.PAST.3S arms on head that.one
'He could feel armour on that one's head.' (WM 54.28 - Watkins 1997:587)

Schematically, the process with the reanalyses is presented in Figure 7.7:

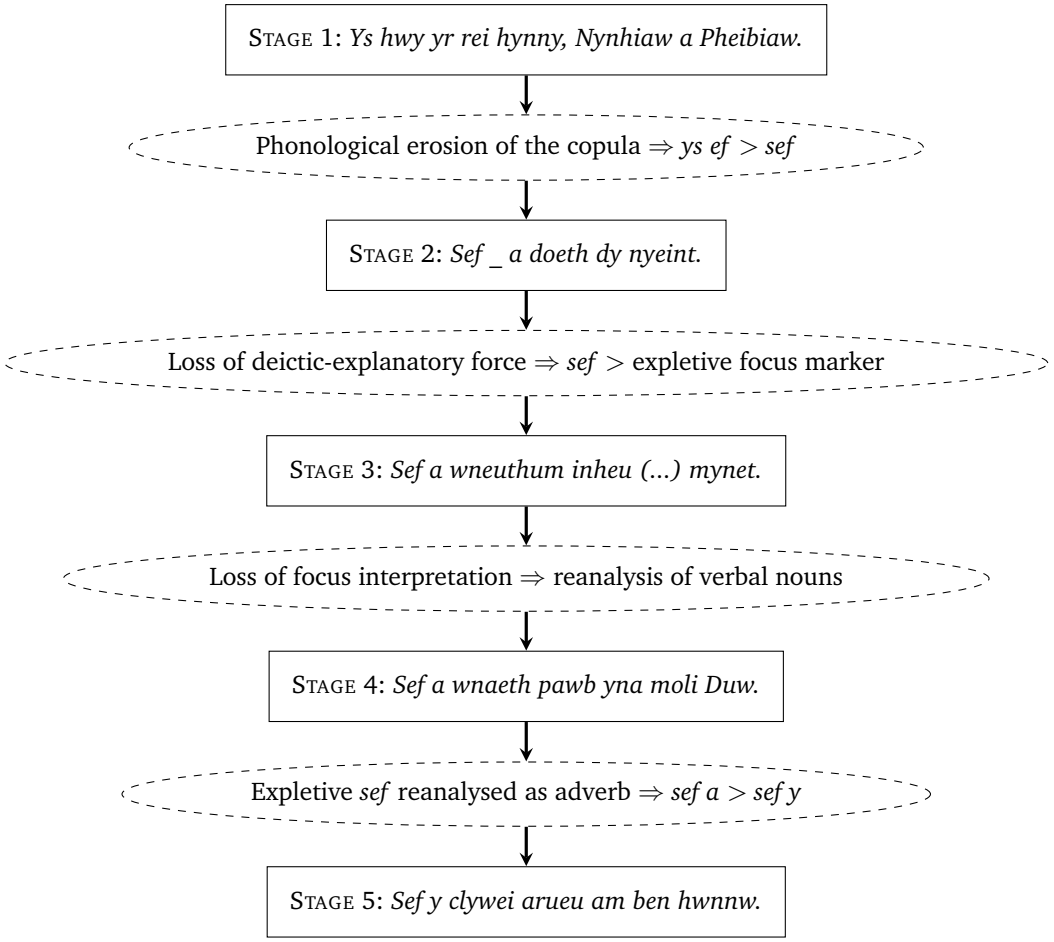


Figure 7.3: Stages of reanalysis of *sef*

From Stage 1 to Stage 2: phonological erosion

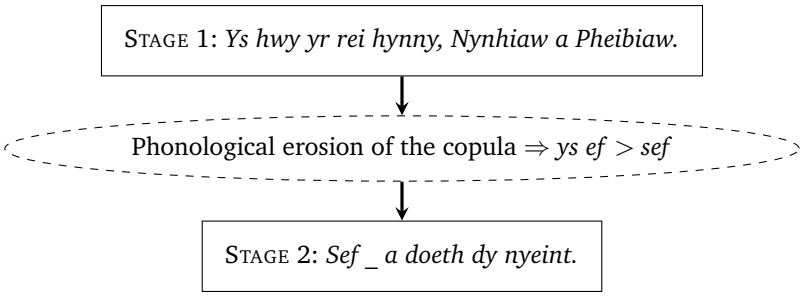


Figure 7.4: Stages 1-2 of reanalysis of *sef*

The derivation of the original cleft sentence with the focussed predicate adjoined to TP was presented in (13) above. T. A. Watkins (1997:579) describes this construction as follows: "In Old Welsh the identificatory copular sentence can be realized as follows: Copula + Anticipatory Predicate + Subject + Postponed Nominal Predicate.". This original *sef*-construction has the following characteristics:
- confined to simple/main clauses of positive declarative sentence types.
- sentences must have nominal (i.e. noun or noun phrase) subject and predicate.
- always identificatory predicates therefore Subject and Predicate must be determinate (definite NPs are inherently so; indefinite NPs may be determinate or indeterminate)
- there is agreement between anticipatory and postponed predicates
- there is agreement between subject and referent
- the only attested tense is present indicative (due to paucity of Old Welsh material, because it is there in Old Irish)
- only attested in 3rd person (since both subject and predicate were obligatorily nominal)
- The subject refers back to a previous (usually immediately preceding) sentence or sentence constituent

Although the full form of the copula is still found in some Early Middle Welsh texts in this construction, there are also signs of phonological reduction. In some cases in Old Welsh already, the copula and anticipatory predicate are written as one word, indicating the start of the merger, as shown in (19a).[11] In Medieval manuscripts, like the *Red Book* of Hergest, the initial vowel of the copula has disappeared, but the double *ss* is still found:

(19) a. *issem* *i* *anu* *Genius*
be.PRES.3S.it 3MS name Genius
'that's his name, Genius' (Old Welsh gl. *Genius* in Martianus Capella - T. A. Watkins (1997:579))

[11] Middle Welsh *ef* 'he, it' was often written as *em* in Old Welsh.

b. *SSef a oruc yr amherawdyr glasowenu.*
sef PRT do.PAST.3S the emperor smile.INF
'The emperor smiled.' (BR 6.25-26)

In most medieval texts, however, the form *sef* is found. This form became structurally ambiguous. It always appeared in the same sequential order *ys + ef* and it was always associated with identificatory predicate focus. The now petrified combination of the copula + anticipatory predicate could thus be reanalysed as one lexical item: the copular focus marker *sef*. This focus marker is then first-merged in the C-domain, satisfying the uninterpretable Focus feature on the C-head.

(20) $[_{\text{PredP}}$ *ys ef*$] > [_{\text{FocP}}$ *sef*$]$

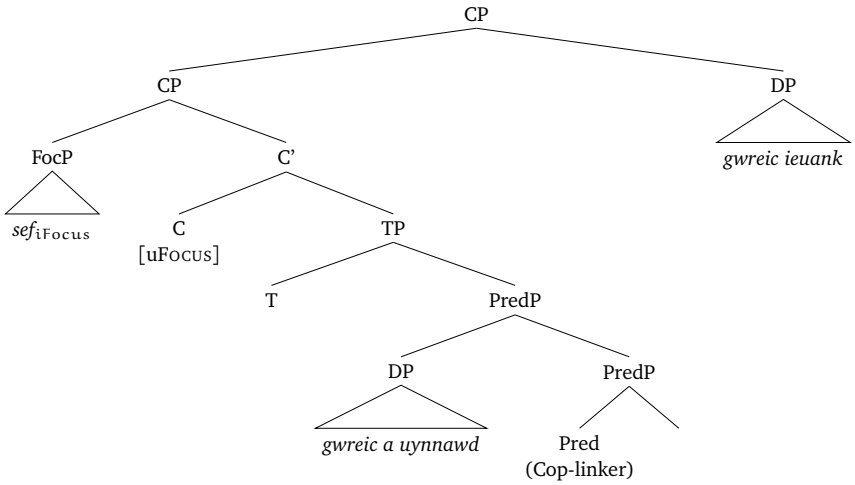
If the subject was not a demonstrative, a relative clause modifying the subject was often used to establish the link with the preceding context. This contextual link was, according to T. A. Watkins (1997) one of the requirements for the *sef*-construction. As shown in example (21), the subject of the clause could be complex, consisting of a DP with a relative clause. The head of the relative could function as the subject, shown in (21a), object (21b) or as an adjunct of the relative verb (21c):

(21) a. *Sef seithwyr a dienghis Pryderi Manawydan (...)*
sef seven.men PRT escape.PAST.3S Pryderi Manawydan (...)
'These were the seven men who escaped, Pryderi, Manawydan (...).' (WM 56.34)
b. *Sef gwreic a uynnawd gwreic ieuank*
sef woman PRT want.PAST.3S woman young
'That was the woman he wanted, a young woman.' (YBH 6)
c. *Sef lle y doethont ygyt y bresseleu*
sef place PRT come.PAST.3P together in Preseleu
'That was the place where they got together, in Preseleu.' (WM 27.28)

In a sentence like (21b), the complex subject DP *gwreic a uynnawd* is in the specifier position of the Predicate Phrase. The head of the PredP is now the phonologically empty copula. This is not a strange stipulation in the context of Middle Welsh, because verbless or 'nominal' copular clauses existed as well (see Chapter 4). The copular focus marker *sef* is then merged in SpecCP and the focussed predicate is adjoined in the same way as before.[12]

[12]Note that adjunction to CP is not necessary to end up with the correct word order *Sef - Subject - Focussed Predicate*. The focussed predicate could also be first-merged (i.e. externally merged) as the complement of the Pred-head and then remain there or be extraposed to end up in the C-domain. I show the derivation with the predicate adjoined to CP here, because adjunction is allowed for the further reanalysis sketched below.

(22)

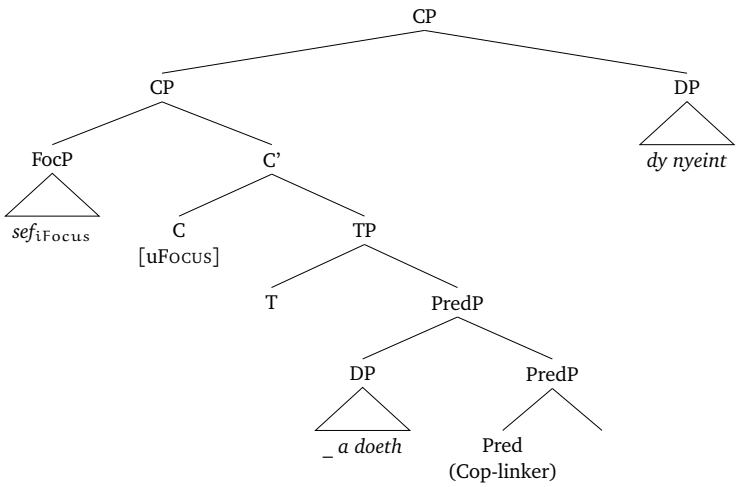


There are furthermore examples of headless relatives.

(23) *Sef __ a doeth dy nyeint*
sef PRT come.PAST.3S 2S nephews
'That's who came, your nephews.' (WM 89.35)

These constructions can be analysed in the exact same way as the above constructions, but they are structurally ambiguous.

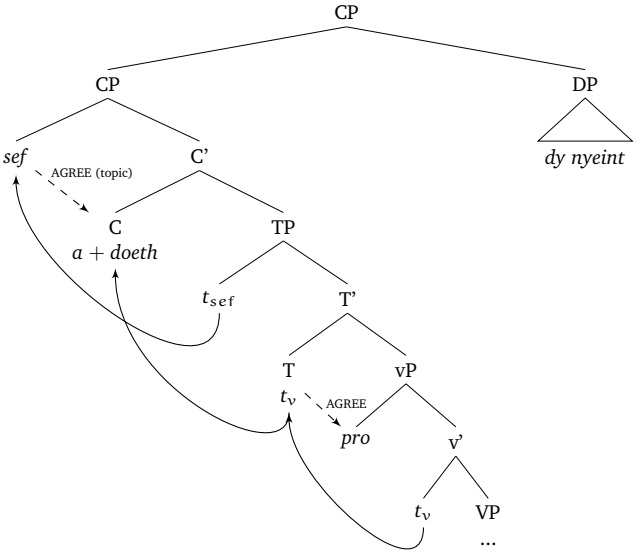
(24)



The ambiguity arises because of the missing head noun in the relative clause that functions as the subject of the copular clause. These subjects were originally in the specifier of the Predicate Phrase. The relative clause *a doeth* 'who came', could

at this stage be reanalysed as the matrix verb. Recall that the most frequently occurring word order pattern in Welsh was the verb-second 'Abnormal Sentence' with the exact same surface structure as relative clauses. The formal focus marker *sef* can now be reanalysed as an expletive merged in SpecTP which subsequently moved up to SpecCP to satisfy C's uninterpretable focus feature. As an expletive, it is considered to be an argument topic and it will thus trigger 'topic agreement', i.e. the complementiser will be realised as *a*, the form it usually takes following core arguments in Abnormal Sentences (instead of *y* following adjuncts). The focussed predicate is then still in the same position adjoining the CP.

(25)



From Stage 2 to Stage 3: loss of deictic-explanatory force

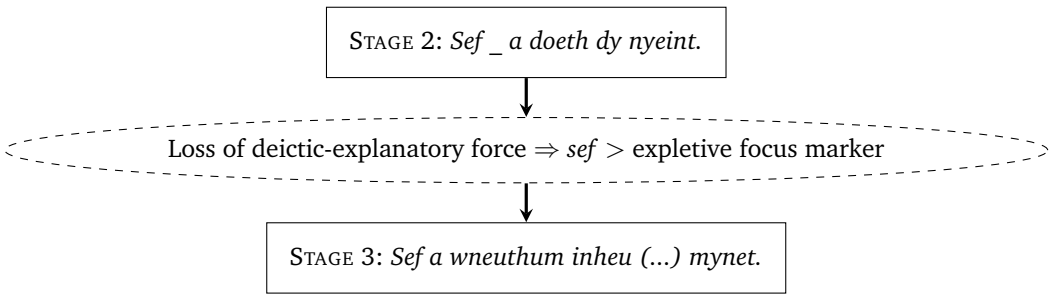


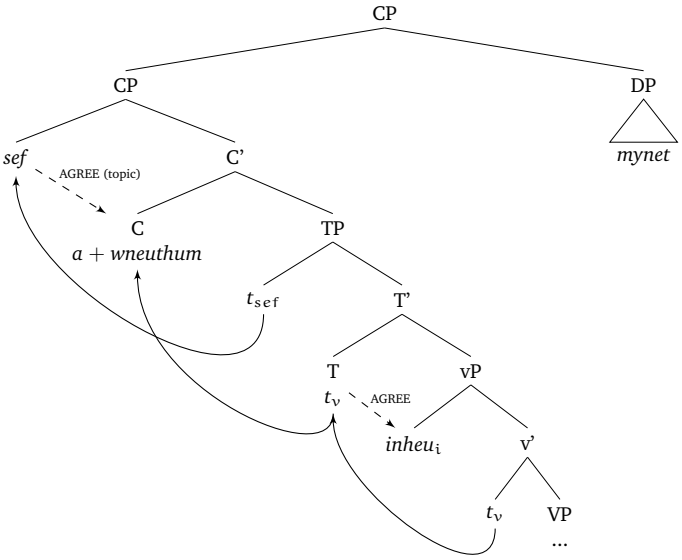
Figure 7.5: Stages 2-3 of reanalysis of *sef*

The next stage of the grammaticalisation process is characterised by the loss of the deictic-explanatory force of *sef*. The new *sef*-construction is no longer necessarily related to the preceding context. The construction could now be used in continuous narrative contexts as well, shown by the example in (26). The construction can in this stage still be parsed in the same way as the examples with headless relative subjects and extraposed predicates above, shown in (27):

Preceding context: "Until it was with difficulty that **I** fled"

(26) *Sef a wneuthum inheu (...) mynet*
sef PRT do.PAST.1S I (...) go.INF
'This is what *I* did, I went (...)' (WM 492.3 - Watkins 1997:586)

(27)



There are two formulaic constructions with unexpressed head-nouns that were very popular and used extensively:

(28) a. *Sef a gausant yn eu kynghor duunaw ar eu llad*
sef PRT get.PAST.3P in 3P council agree.INF on 3P kill.INF
'This is what they decided in their council, they agreed to kill them' (WM 68.8)
b. *Sef a wnaeth y gwraged kyscu*
sef PRT do.PAST.3S the women sleep.INF
'This is what the women did, they slept.' (WM 28.15)

In these sentences, the predicate is a verbal noun: *duunaw* 'agree' or *kyscu* 'sleep'. In non-copular sentences in Middle Welsh, the verbs *cael* 'get' and *gwneuthur* 'do'

could be used as auxiliary verbs. This then paved the way for a further possible ambiguous structure leading to the next stages of the reanalysis.

From Stage 3 to Stage 4: loss of deictic-explanatory force

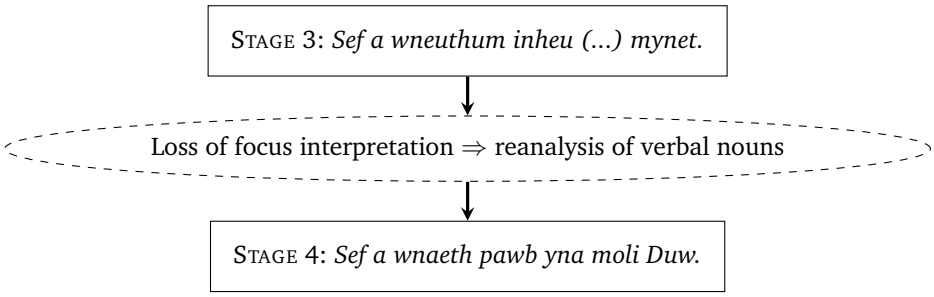
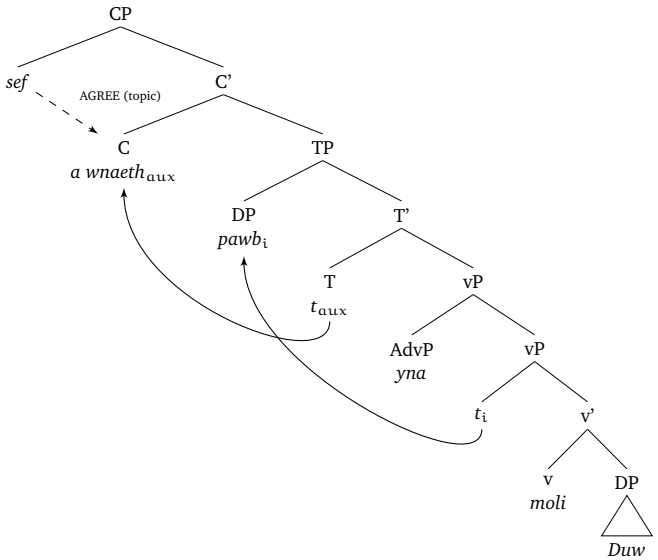


Figure 7.6: Stages 3-4 of reanalysis of *sef*

In the next stage, this structural ambiguity leads to reanalysis of the verbal noun as the matrix verb. The adjoined or extraposed predicate position is lost and along with that the focussed interpretation. The subject moves to SpecTP and agrees with the verb while *sef* is first-merged in SpecCP now. In example (29), the verbal noun *moli* is reinterpreted in this way as the matrix verb and *gwneuthur* 'to do' is the auxiliary (or light verb), resulting in the derivation in (30).

(29) *Sef a wnaeth pawb yna moli Duw*
sef PRT do.PAST.3S everyone then praise.INF God
'Everyone then did this, they praised God.' (Dewi 4.17)

(30)



From Stage 4 to Stage 5: focus marker reanalysed as adverb

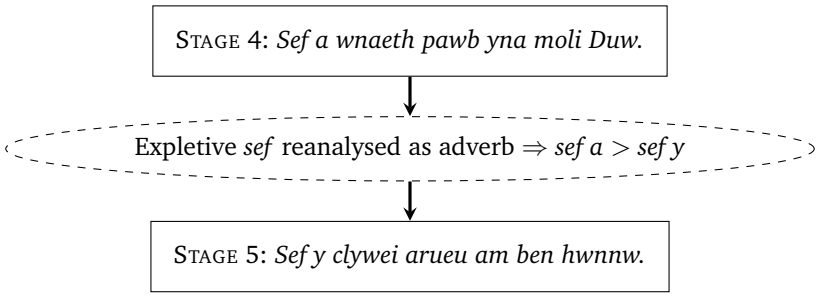


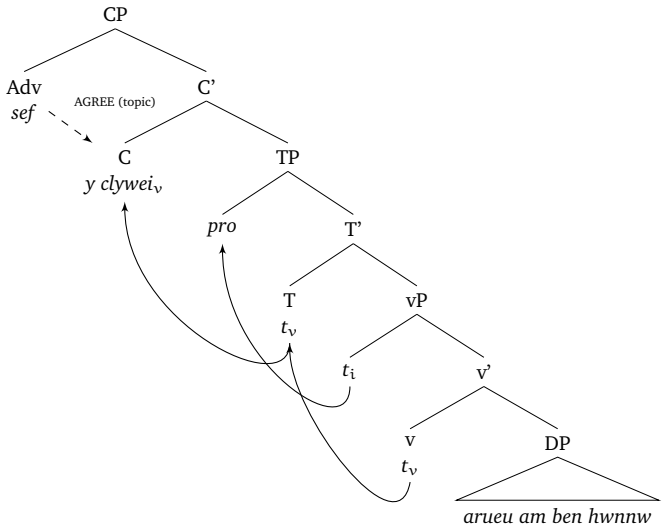
Figure 7.7: Stages 4-5 of reanalysis of *sef*

Eventually the argumental interpretation of expletive *sef* was lost. It was reanalysed as an adverbial element base-generated in SpecCP. Subjects could then move to SpecTP just as they did in any other adjunct-initial Abnormal Sentence (see next section). Adverbs, like all other adjuncts, trigger the pre-verbal particle *y* in the C-head, instead of the particle *a* following argumental DPs as shown in the examples in (31) and the derivation in (32):

(31) a. *Sef y clywei arueu am ben hwnnw*
sef PRT hear.PAST.3S arms on head that.one
'He could feel armour on that one's head.' (WM 54.28)

b. *Sef y kynhelleis inheu y gyuoeth*
sef PRT withhold.PAST.3S I his dominions
'I withheld his dominions.' (WM 394.42 - Watkins 1997:587)

(32)

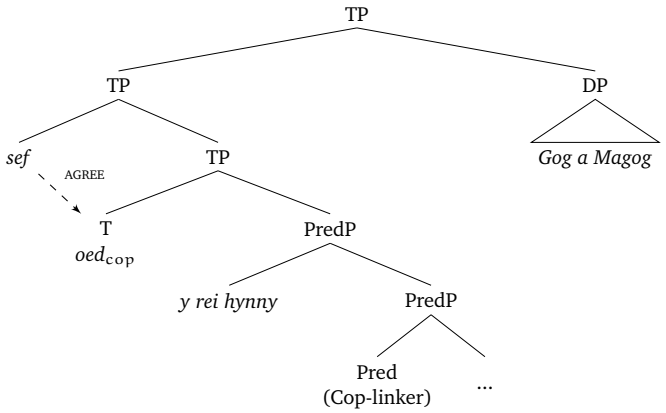


Other *sef*-constructions

The phonological reduction of the copula allowing all subsequent reanalyses described above, also triggered reanalyses of a different kind, creating a further range *sef*-constructions. The cascading pattern of reanalyses described above was specifically possible because of the large number of sentences with subjects consisting of a headless relative (as shown in the section on Stage 2 to Stage 3 above). If the relative clause consisted of a copular clause itself as shown in (33), it could give rise to a further type of reanalysis. Here too the verb from the relative clause could be reinterpreted as the matrix verb as shown by the derivation in (34).

(33) *Sef* ***oed*** *y rei hynny Gog a Magog (...)*
sef be.PAST.3S the ones DEM.P Gog and Magog
'That's what those were, Gog and Magog (...).' (DB 29.11.12)

(34)



Merge of the verb in the T-head could be internal or external, since the verb *bod* 'to be' also functioned as an auxiliary in Middle Welsh. Further movement to the C-head is string-vacuous in this sentence, which is why I only show the TP. The preverbal particle *a* is usually analysed as a complementiser attracting the verb to the C-head, but this *a* could be dropped before *oed*, the imperfect form of the verb *bod* in Middle Welsh. Therefore, with the evidence we have at present we cannot prove it moves up to C or remains in T.

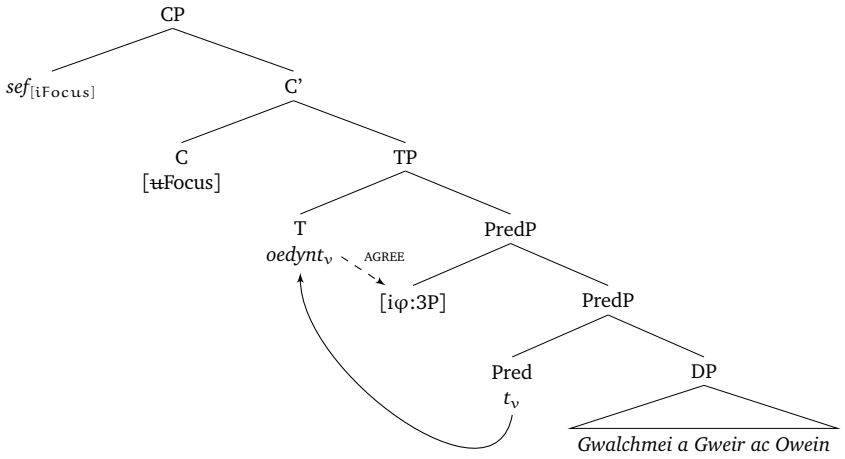
An example with the present-tense verb form *yw* 'is' would now also be a possibility. Note that this could not have been the original form because in the present tense, the verb *bod* 'to be' has a special relative morphology *yssyd* 'that/which is'. Once the verb, either the imperfect form of *bod* or any other verb, was reanalysed as the matrix verb, the medial form of the verb 'to be' could be merged in the T-head as well. This new *sef*-construction with *sef yw/oed...* is called the 'parenthetic-explanatory clause' in traditional Welsh grammars (cf. T. A. Watkins (1997:580-581)):

(35) *Sef **yw*** *honno gwreic doget urenhin*
sef be.PRES.3S DEM.FS wife Doged king
'That's who she is, king Doged's wife.' (WM 453.17 - Watkins 1997:580)

With the advent of medial copular forms like *yw* 'is' above, a further reanalysis could take place: the rise of (dropped) pronominal subjects, as shown in (36). In this construction, *sef* is not interpreted as the expletive. It is externally merged as a focus marker in the specifier of the CP. The verb agrees with the (empty) pronominal subject, as shown in (37). Just as in the above-described stages of reanalysis, here too, the predicate now no longer needs to be in an adjoined position; it can be interpreted in the complement-position of the predicate phrase.

(36) *Sef oed**ynt*** *Gwalchmei (...) a Gweir (...) ac Owein*
sef be.PAST.**3P** Gwalchmei (...) and Gweir (...) and Owein
'That's who they were, Gwalchmei (...) and Gweir (...) and Owein.' (WM 118.19 - Watkins 1997:581)

(37)



Conclusion *sef*-constructions

In this section I presented a detailed analysis of every stage of the process of grammaticalisation of the *sef*-construction in Middle Welsh. For each of the different stages, I presented the characteristics of the ambiguous structures that led to a cascade of new reanalyses. The original trigger was argued to be the phonological erosion of the copula (as already noted by T. A. Watkins (1997)). The predicate + complement *ys ef* first merged into one lexical item that could be externally merged as the expletive in SpecTP or as a focus marker in SpecCP. The relative verb in the complex subject could then be reinterpreted as the matrix verb. From an information-structural point of view, all conditions and characteristics of the original

identificatory focussed predicate were lost. There was no longer a requirement to link the construction to the preceding context and the identificatory interpretation of the predicate as well as its focus marking were lost (semantic bleaching). The *sef*-construction then came to be used in continuous narratives and the variant with the auxiliaries *gwneuthur* 'to do' and *cael* 'to get' became stock phrases. Finally, *sef* lost its argumental status as an expletive and was recategorised as an adverb.

Many of these different forms of the *sef*-construction appear in the same period, sometimes even in the same texts. There has undoubtedly been a period of overlap. We can establish the relative chronology of the different stages in the grammaticalisation process, but since we lack the necessary philological data, it is very difficult to establish a more accurate date for each of the above-sketched stages of reanalysis. There is, however, some supporting evidence for the relative chronology from the *Red Book* of Hergest. The scribe of this manuscript (written around the year 1400) is generally considered to have 'modernised' the text he copied into the *Red Book*. The original was lost, but other older copies of these texts exist, for example, in the *White Book* of Rhydderch, which formed the basis of the present annotated corpus. In comparing certain parallel passages from the *White Book* and the *Red Book*, we see that the 'modernised' *Red Book* more often employs what I described above as the fifth stage of the grammaticalisation process: the adverbial form of *sef* followed by the particle *y* (rather than *a*).

(38) a. *Sef* ***a*** *gausant yn eu kynghor rodi y moch e Wydyon*
sef Csp got in their council give.INF the pig to Gwydyon
'This is what they got in their council: give the pig to Gwydyon' (*White Book*)

b. *Sef* ***y*** *kawssant yn eu kynghor rodi y moch y Wydyon*
sef PRT got in their council give.INF the pig to Gwydyon
'Then giving the pig to Gwydyon was what they got in their council.' (*Red Book*)

To conclude, the identificatory copular clauses with focussed predicates changed dramatically over the Middle Welsh period. Various different forms of this *sef*-construction were found alongside each other, but a careful analysis reveal a clear pattern of a step-by-step reanalysis, with each change triggering the next stage of the process. This relative chronology of the complex grammaticalisation process is to a certain extent confirmed by philological evidence in the form of earlier and later manuscript forms of the same texts.

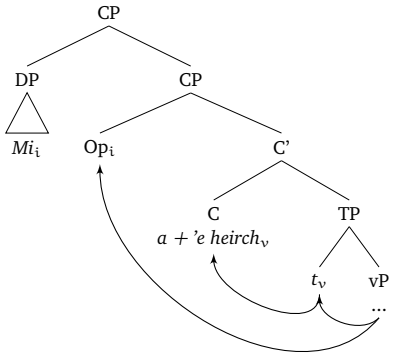
7.3.2 Reanalysis & Extension in the rise and fall of V2

In the previous chapter I discussed the two main types of V2-structures found in Middle Welsh: the so-called Abnormal Sentence and the Mixed Sentence. The traditional distinction between the two is based on Information Structure and agreement patterns: Abnormal Sentences *do* exhibit subject-verb agreement and Mixed Sentences never show subject-verb agreement. Formally, the two can only be

kept apart if the subject preceding the verb is a non-third-person singular pronoun or a plural DP. From an information-structural point of view, the difference is traditionally argued to be Topic (in Abnormal Sentences) vs. Focus (in Mixed Sentences). I have shown in Chapter 6, however, that the IS status of the preverbal constituent cannot be simply divided between these two categories: there are examples of Focus *with* subject-verb agreement and vice versa, examples with preverbal Topics *without* the expected agreement pattern. There are furthermore examples of both agreement patterns in coordinated sentences. A final complication in the data is the 'Complementarity Principle' that holds in all Brythonic languages stating that agreement is only ever found with pronominal elements, never with full DPs. From the point of view of the Complementarity Principle then, agreement with full plural noun phrase subjects in the Abnormal Sentence is unexpected, just as the lack of agreement with pronominal subjects in Mixed Sentences. These Middle Welsh V2-structures are not found in other Celtic languages like Gaulish, Celtiberian or Irish (in any stage of the language). They equally do not occur in Modern Welsh. Modern and Middle Breton as well as Middle Cornish do exhibit the non-agreeing V2 structures equivalent to the Middle Welsh Mixed Sentence. The Abnormal Sentence with subject-verb agreement, however, seems to be a Middle Welsh innovation that was lost again in the Early Modern Welsh period. In the previous chapter, I proposed structures for these 'unexpected' patterns in Abnormal and Mixed Sentences. In the Mixed Sentence, SpecCP is occupied by the relative operator yielding default third-person singular inflection on the verb as shown again in (40):

(39) *Mi a 'e heirch.*
I PRT 3FS seek.3S
'(it is) I who seek her' (Mixed Sentence - WM 479.24)

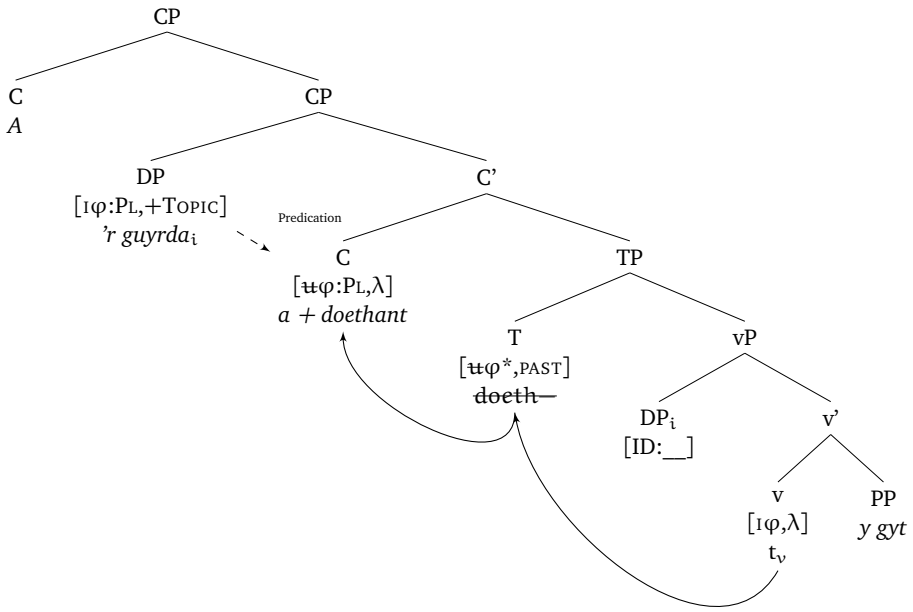
(40)



Plural DPs in agreeing Abnormal Sentences are base-generated in SpecCP. The C-head carries a λ-feature that ensures a predication relation with the DP-topic in its specifier through which agreement can take place. The DP-topic is coindexed with a minimal pronoun subject (a DP without φ-features: [ID:_]).

(41) *A 'r guyrda a doethant y gyt*
and the nobles PRT come.PAST.3P together
'And the nobles came together' (Abnormal Sentence - PKM 90.27)

(42)



The main question from a diachronic syntactic point of view is: where does the Abnormal Sentence with subject-verb agreement come from? Although some Welsh grammarians (e.g. MacCana (1973) and Fife (1991)) have argued that this was merely a literary phenomenon in Middle Welsh, Willis (1998) convincingly argues these V2-structures must have been part of spoken Middle Welsh as well. His arguments are based on language-internal complexity of the V2-rule in various parts of the grammar that would have been hard, if not impossible, to learn as a stylistic feature. Breton and Cornish furthermore also exhibit V2-structures (without subject-verb agreement), so V2 grammar is likely to be inherited from their Common Brythonic ancestor.

A further diachronic question then remains: where do these V2-structures (with and without agreement) come from in general? Richards (1938) and D. S. Evans (1968) already hypothesised that the origin of these Brythonic structure lies in the cleft sentences with contrastive focus. The cleft was followed by a relative clause, introduced by the relative particles *a* or *y*, the exact same particle found in the Mixed and Abnormal V2 orders. Through a process of semantic bleaching, the function of contrastive focus was extended to topics and this then became the basic word order pattern in Middle Welsh (in which the preferred Insular Celtic verb-initial order also found in Irish was lost).

In this section I explore this hypothesis further by examining each of the required syntactic reanalyses and extensions in detail to trace the origin of the Abnormal Sentence. Within the framework of the Minimalist Program, I provide the triggers and linguistic context of every stage in the process that created the right environment for the syntactic reanalyses and extensions we find. In order to describe the first steps that can only be found in reconstructed stages of the languages, it is important to take the sparse Old Welsh data available to us, as well as cross-linguistic evidence from Middle Breton and Middle Cornish into account. Although the focus lies on the rise of V2 structures in Middle Welsh, in the final part of this section I also shed some light on the subsequent loss of V2 with evidence from the 1588 Bible translation.

Overview of syntactic reanalyses & extensions

Figure 7.8 shows an overview of each of the different stages in the process with a description of the possible word order patterns found at that stage specified in the same box. In the following dashed ellipse I describe the trigger(s) that led to a specific change. Any changes in the form of loss/gain of word order patterns in the next stage are presented in the next box. Some of these new patterns may in turn lead to further reanalyses and extensions, until they finally lead to the fifth stage representing Early Modern Welsh when evidence for the acquisition of V2 dropped and the Abnormal Sentence was lost. The Mixed Sentence with contrastive focus on the initial constituent is the only V2-pattern left in positive declaratives in Modern Welsh.

The first stage represents a language that can be reconstructed as the predecessor of Brythonic: 'pre-Common Brythonic'. Following Newton (2006) and Lash (2011), I assume that Insular Celtic had previously lost the articulated CP that was still found in Proto-Indo-European (based on evidence from syntactic reconstruction of Greek, Vedic, Hittite and Latin).

I discuss the labels or languages matching the following stages up to Early Modern Welsh in the context of cross-linguistic evidence from Middle Breton and Cornish. Some word order patterns occur in several stages until they are completely lost or reanalysed. The patterns with optional merger of adjuncts and hanging topics in the C-domain resulting in V2, V3 and V4 orders in Insular Celtic and pre-Common Brythonic, for example, remained until they were replaced by the V2-structures with preverbal particles *a* and *y* in the C-head.

In the same way, patterns with sentence-initial *y(d)* were present from the grammaticalisation of the particle before Stage 2 until Modern Welsh, although during the Middle Welsh period the context in which this sentence-initial *y(d)* was found narrowed down to periphrastic constructions with the auxiliary form of the verb *bod* 'to be'. In the following sections, I discuss each of the stages and the triggers for reanalysis and extension in detail.

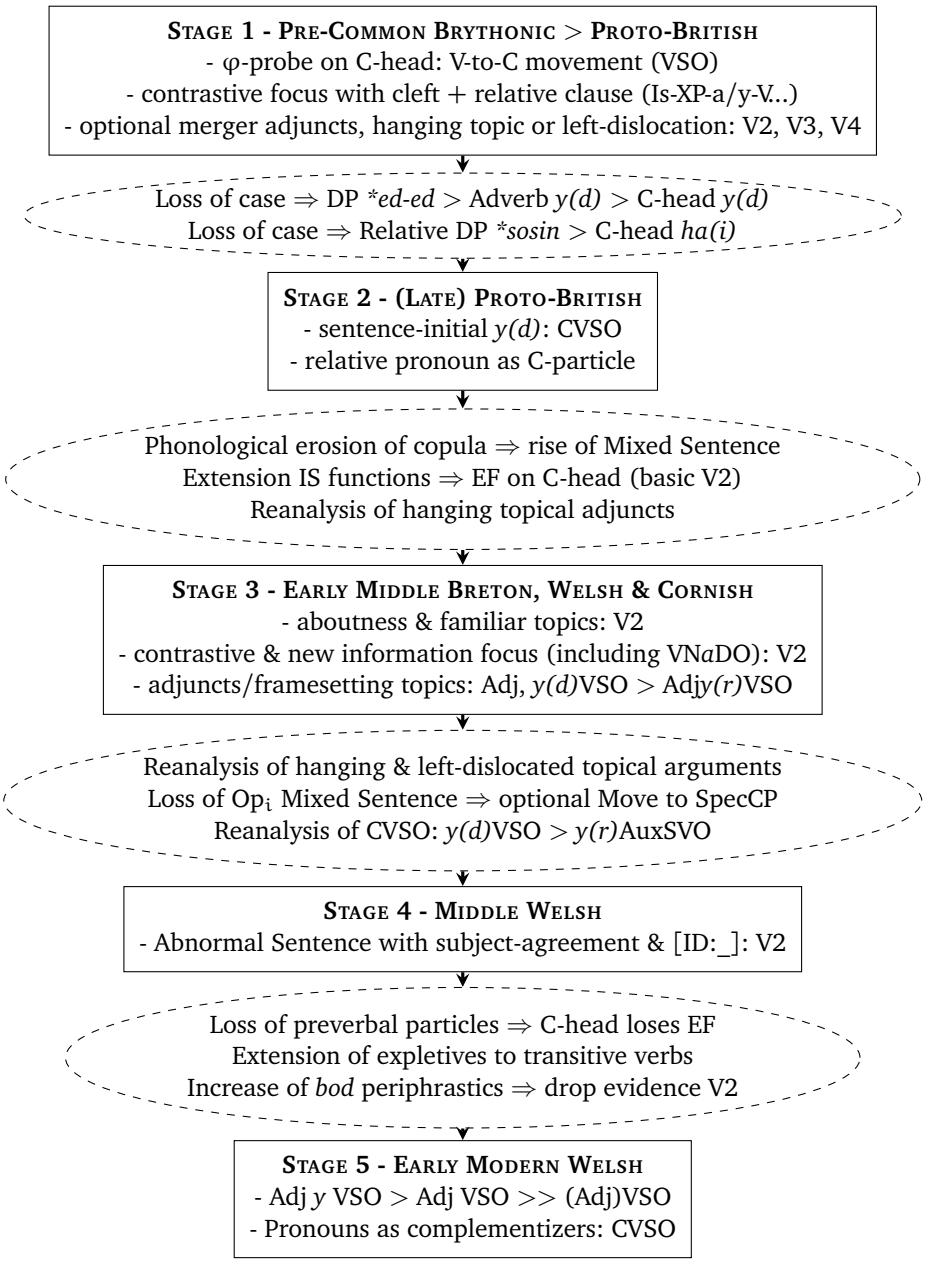


Figure 7.8: Rise & fall V2 from Pre-Common Brythonic to Early Modern Welsh

From Stage 1 to Stage 2: Loss of case

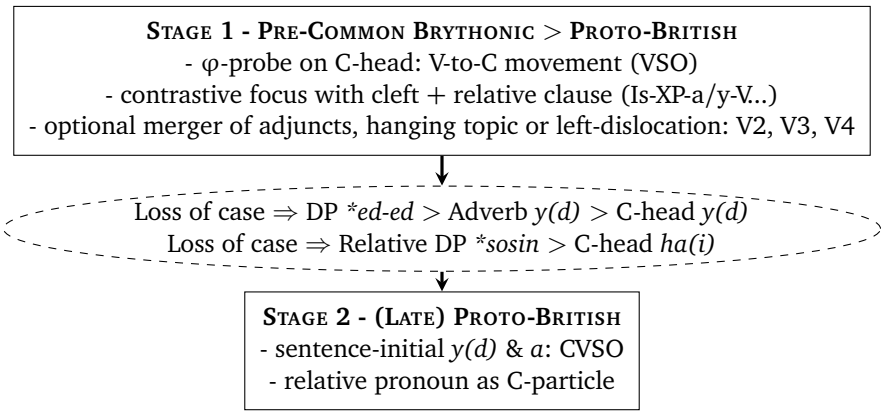


Figure 7.9: Syntactic changes from Stage 1 to Stage 2

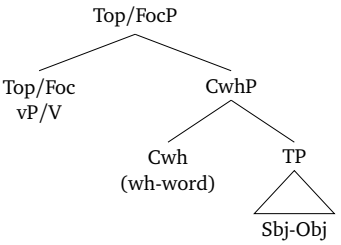
For the first two stages in the process of reanalyses and extensions sketched above we have no written evidence. Stage 1 represents the situation of the language described as Insular Celtic or shortly thereafter, what might be described as 'Pre-Common Brythonic'. This is the form of Celtic spoken in the British Isles before the split of the Irish and the British branches of the Celtic language family. Stage 2 represents the next phase of Common Brythonic, the predecessor of Welsh, Breton and Cornish. Both of these stages can only be described to a certain extent, by means of reconstruction. As discussed in section 7.2.3 above, syntactic reconstruction presents more difficulties than the reconstruction of phonology or morphology. The correspondence problem in particular limits the parts of the grammar that can be reconstructed to those functional items and features that are phonologically overt in the daughter languages. Beyond that, we can still compare syntactic structures and make reasonable assumptions based on plausible patterns of grammaticalisation, reanalysis and local directionality.

Newton (2006) compares the C-domain of Old Irish with that of other Indo-European languages like Greek, Vedic, Sanskrit and Hittite. She concludes that Vedic and Hittite only allow two constituents in the left periphery of the clause: in the Topic/Focus head and the C[+wh] head. Greek and Latin on the other hand allowed multiple topics in the C-domain. Proto-Indo-European as well as Proto-Celtic thus seemed to have a C-domain consisting of at least two functional heads: Top/Foc and C (or 'C[+wh]' as Newton calls it). In the stage of the language she calls 'Pre-Old Irish', this (mildly) articulated CP was lost via "clause truncation". This truncation was established by the reanalysis of relative operator XPs in specCwhP as heads of specCwhP and subsequently as affixes on obligatorily fronted verbs, preverbs or negative elements. The triggering diacritic on Top/Foc was reanalysed as an obligatory movement feature resulting in a 'filled-C condition'. The clause-

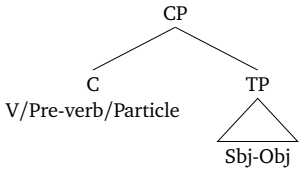
marking suffix **es* linked the verb to the C position. This acquisitional cue then resulted in a reanalysis as V-to-C and pre-verb-to-C movement and conjunct and negative particles occupying the C-head. She then links this new configuration to the development of the Absolute and Conjunct verbal paradigms. Sample derivations for the described sentence structure in PIE and Pre-Old-Irish are given in (43) and (44) below.

Proto-Indo-European ⇒ Pre-Old-Irish

(43)



(44)



The question is: how does Insular Celtic fit in this picture? Does Insular Celtic have an articulated CP like PIE or was this structure already reanalysed in the way Newton has reconstructed for Pre-Old-Irish? The reconstruction of an articulated CP in PIE is based on the possibility of the occurrence of multiple topics or foci alongside other elements in the C-domain (e.g. wh-phrases in CwhP). If we find examples of this in the Brythonic languages, this would be a strong argument to reconstruct an articulate CP in Insular Celtic. The CP truncation could then be postulated as a Pre-Old-Irish innovation only.

As discussed in Chapter 6, in Middle Welsh it was impossible to have both a Topic as well as a Focus constituent preceding the verb. This constraint provides evidence for the strict V2-nature of Middle Welsh word order. The extant data in Old Welsh is extremely limited. Most examples of declarative main clauses exhibit verb-initial order in Old Welsh.

(45) a. *Prinit hinnoid iiii aues*
buy.ABS.3S that four birds
'That buys four birds.' (Old Welsh - Ox 1 B v.234)

b. *Rodesit Elcu guetig equs.*
give.PAST.3S Elcu after horse
'Elcu then gave a horse.' (Old Welsh - Chad2)

There are, however, examples of multiple constituents preceding the verb in Old Welsh yielding V3 or V4 orders, as shown in (46). In these examples, the initial constituents are in fact hanging or left-dislocated topics or adjuncts. As such, they do not provide evidence for an articulated CP. These types of V3 orders are found in Middle Breton and Middle Cornish as well, as shown in (47) and (48) respectively.

(46) a. *Mi telu nit gurmaur*
1S retinue NEG.be.3S very.large
'My retinue, (it) is not very large' (Old Welsh - Juv 3)
b. *Ir pimphet eterin diguormechis lucas hegit hunnoid ...*
the fifth bird add.PAST.3S Lucas go.ABS.3S that.one ...
'the fifth bird that Lucas added, that one goes...'(Old Welsh - Ox 1 B v.234)

(47) a. *breman a crenn me a gouchemen dit*
now PRED express I PRT ask.3S to.2S
'now expressly I ask of you' (Middle Breton - N240)
b. *monet a pret me a preder*
go.INF PRED early I PRT plan.3S
'to go early I plan' (Middle Breton - N64)

(48) a. *Oma ty a ra pedry*
here you PRT do.3S rot
'Here you shall rot.' (Middle Cornish - BMer 3577)
b. *In crist ihesu ny a greys*
in Christ Jesus we PRT believe.3S
'In Christ Jesus we believe.' (Middle Cornish - BMer 1210)
c. *Duk kernov hag oll y dus indan ou threys me as glus.*
Duke Cornwall and all 3MS men under 1S feet I PRT.3P crush
'The Duke of Cornwall and all his men under my feet I shall crush them.'
(Middle Cornish - BMer 2397)

The basic word order in Old Irish was VSO, but similar V2 constructions can be found, as shown in (49):

(49) a. *Cech mab uilc robai ind Éire dochoid chuca.*
every son evil be-rel.PAST.3S in Ireland come.PAST.3S to.3P
'Every son of evil who was in Ireland, he came to them.'(Dindshenchas of Emain Macha - MacCana 1973:96)
b. *Mortlithi márlóchet di doínib dingbatar*
great.plagues great.lightnings from people keep.PRET.PASS.PL.CONJ
'Great plagues and great lightnings are kept from the people.' (AM §12)

It appears then, that in both Brythonic and Irish a specific set of V2, V3 or V4 orders were allowed alongside the basic verb-initial order. There are no overt functional items we can reconstruct for Proto-Insular Celtic, so a perfect correspondence in the form of a double-cognacy condition is impossible to find. We can only compare the extant evidence in the daughter languages and tentatively assume that these V2, V3... orders with adjuncts and hanging and left-dislocated topics were part of the otherwise verb-initial parent language we reconstruct as Proto-Insular Celtic as well. Further comparative evidence could in theory come from Continental Celtic languages like Gaulish in which V2 and V3 orders exist as well.

(50) a. *Ratin briuatiom Frontu Tarbeisonios ieuru*
fort.ACC bridge-dwellers.GEN Fronto Tarbeisu.GEN dedicate.3S
'Frontu, son of Tarbeisu, dedicated the fort of the bridge-dwellers'.(Gaulish OSV - RIG L3)

b. *Buscilla sosio legasit in Alixie Magalu.*
Buscilla this place.3S in Alisia Magalos.DAT
'Buscilla placed this in Alisia to/for Magalos' (Gaulish SOV)

c. *Moni gnatha; gabi budduton imon!*
come.IPV.2S girl take.IPV.2S penis/kiss(?) this
'Come girl; take this penis/kiss(?)!' (Gaulish V1 - St. Révérien[13])

d. *nata vimpi cvrmi da*
girl pretty beer give.IPV.2S
'Pretty girl, bring [me] beer!' (Gaulish OV - spindle-whorl inscriptions)

As the examples in (50) show, however, Gaulish does not only allow hanging and dislocated topics preceding the verb, but also direct objects. These constituents are thus not outside the matrix CP as can be argued for the V2 and V3 structures found in Insular Celtic languages. Instead, these examples show the lack of V-to-C movement (V1 is almost exclusively found in imperatives like (50c)) and cannot tell us much about Insular Celtic. The verb-initial nature seems to be an innovation in the Insular Celtic languages only. For reconstruction of the syntax of Proto-Insular Celtic, we thus have to rely on evidence found in the Irish and Brythonic languages only. In terms of evidence for an articulate CP, we can only reconstruct a phi-probe on C resulting in V-to-C movement and basic verb-initial word order. Since extra-clausal elements such as hanging topics and adverbial phrases can be found in all daughter languages yielding V2, V3 and V4 orders, we can furthermore assume that this was allowed in the Insular Celtic stage of the language as well. It is important to note that allowing these non-verb-initial orders does not exclude the possibility of an articulate CP in Insular Celtic either.

A further reason for Newton (2006) to reconstruct a phi-probe on C yielding verb-initial order in Pre-Old-Irish is the development of the 'double system', i.e. the Absolute-Conjunct paradigms in the verbal system. According to this highly complex system, Old Irish verbs could exhibit different forms according to their position in the sentence. Verbs in absolute sentence-initial position are found with 'absolute' verbal morphology. In Old Welsh, we can still find some examples of absolute verbal endings in the third-person singular. These endings were lost and in Middle Welsh there is no evidence for the Absolute-Conjunct distinction. If we continue to compare Irish and British grammars, we could conclude that this system found in both daughter languages was likely to exist (or to have developed) in their predecessor Insular Celtic as well. However, it is not impossible that the double

[13]There appears to be some discussion on the exact nature and purpose of these sentences with imperative verbs found on spindle-whorls. C. Watkins (1999:542) translates *budduton* as 'penis', but according to Stifter (2011:174n20), the etymology connecting Gaulish *budduton* to Early Irish *bot* 'tail; penis' < *g^{u}*ozdo-* is wrong. Instead, a connection to Middle Irish *bus* 'lip' < **butsu-* "is formally more satisfying". The inscription may thus be of a much more innocent nature, translating 'take that kiss'.

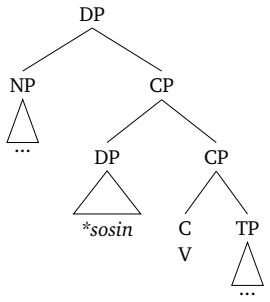
system was a parallel development in both Pre-Old-Irish and Proto-British after the languages split.

To conclude, Old Irish and the earliest attestations in the Brythonic languages show some similar evidence for the development of V-to-C movement in the form of the double inflectional system and sentence-initial particles (declarative and relative). In addition to that, neither Irish nor Brythonic shows explicit evidence for an articulated CP for example in the form of cooccurring topic and focus constituents in preverbal position. However, V2, V3 and V4 orders were possible in both languages with adjuncts and hanging topics. Finally, in Old Irish as well as in Old Welsh, the cleft sentences was the preferred strategy to mark contrastive focus of one single constituent only.

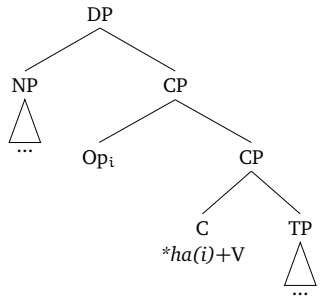
In the course of the Common Brythonic period, some phonological changes triggered further syntactic innovations. Apocope in the 6th century led to a loss of case encoded in morphological endings. As Newton (2006) points out, the loss of morphological endings renders XPs caseless. This in turn can lead to reanalysis of the now indistinct form as a head. This affected the relative pronoun **sosin*, for example. Relative clauses could be formed by the particle **-io* (> Old Welsh absolutive ending *-it*, according to Schrijver (1997)) or with the relative pronoun **sosin* (> Old Welsh *ha(i)*). Since relative pronouns that are inflected (and thus phrases rather than heads) are reconstructed as occupying the specifier of the (lower) CP in PIE (see Newton (2006)), I assume the same position for **sosin* in Brythonic. After case endings were lost the indeclinable relative pronoun was reanalysed as the C-head, as shown in (51) and (52):

Stage 1 relatives ⇒ Stage 2 relatives

(51)



(52)



The loss of case in the neuter pronoun **ed-ed > y(d)* 'it' could have given rise to ambiguity with the adverb *ed* 'thus' (cf. Schrijver (1997:161-176)). Both could be used in relative constructions, for example, in clefts with contrastive focus. Constructions like that are found in Old South-West British (OSWB), shown in (53) and Old Irish as well, as shown in (54).

(53) a. *Is amal it duducer memor.*
COP.3S like thus adduce.REL.3S memory
'It is thus that one adduces memory.' (OSWB - DGVB: Ang477A)

(54) a. *Iss ed dochoid i tir Eogain.*
COP.3S thus go.PAST.REL.3S to land Eogan.GEN
'It is thus that he went into Eogan's land.' (Old Irish - Trip. 150.19)
b. *Is ed fuddera.*
COP.3S this cause.REL.3S
'It is this that causes it.' (Old Irish - Wb 33c12)

The adverbial **ed* > **yd* 'thus' was always found in sentence-initial position. Its reanalysis as a particle in the C-head gave rise to the CVSO orders found in early Breton (*ez*), Cornish (*y(th)* and *as*) and Welsh (*y(d)*) sources.

(55) a. *Yd af i yn agel.*
PRT go.1S I PRED angel
'I shall go as an angel.' (WM 118.27)
b. *Y rodet y march y 'r mab.*
PRT give.IMPERS.PAST the horse to the boy
'The horse was given to the boy.' (PKM 24.4-5)
c. *Y dodym y erchi Olwen.*
PRT come.PAST.1P to seek.INF Olwen
'We have come to ask for Olwen.' (CO 477)

(56) a. *Ez oamp oll, allas, e lastez*
PRT be.1P all alas in suffering
'We are all, alas, in suffering.' (Middle Breton - Nl 328)
b. *Y leferys offeren.*
PRT say.PAST.3S mass
'He said the mass.' (Middle Cornish - BM 4419)
c. *As wrussough cam tremene.*
PRT cause.2P wrong death
'You caused a wrong death.' (Middle Cornish - R40)

There is some further evidence for CVSO orders in this stage in the form of the C-head *a* < *ha(i)* that appears in sentence-initial position in some remnants in Early Welsh poetry, as shown in (57). Schrijver (1997:166) notes, however, that the *a*-particle is merely there to support the cliticised pronoun, which could not occur in sentence-initial position on its own. This could still mean that the particle is the same as the relative marker *a* occupying the C-head in which case we find CVSO order here as well.

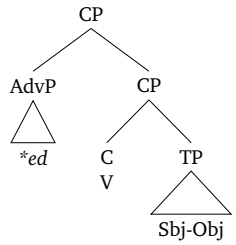
(57) a. *A 's kynnull gwenyn.*
PRT 3MS gather.3S bees
'Bees gather it.' (T 40.8-9)

b. *A 'th uolaf uaur ri*
PRT 2S praise.1S big king
'I praise you, great king.' (HGC II.4)
c. *A 'n gunel iechid*
PRT 1P make.SBJ.3S salvation
'May he make salvation for us.' (HGC III.11)

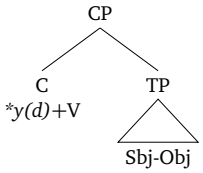
In Early Middle Welsh, the particle *y(d)* could be used in front of any type of verb, although it most commonly appears with *bod* 'to be', *dyuot* 'to come' or *mynet* 'to go'. In Middle Breton, it is only found with those verbs (Hémon, 1975:277). According to Schrijver (1997:171), Old Irish has a very similar adverb *ed* used 'with vb. of motion' (DIL E 58 *ed* VI).

Stage 1 Adv VSO ⇒ Stage 2 CVSO

(58)



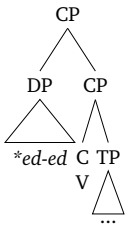
(59)



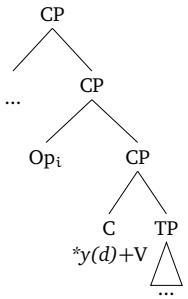
The DP antecedent **ed-ed* also linearly directly preceded the verb in the C-head in all cases. With the loss of the case endings, this phrase could also be reanalysed as a relative marker in the C-head, just like **sosin* in the direct relative clauses. The similarity in form with the adverb-turned-particle *yd* could have given rise to the association of the new relative marker with indirect (rather than direct) relatives, i.e. with non-argumental antecedents. In Middle Breton and Middle Welsh, this relative marker *ez/y(d)* is exclusively found following phrases that are not the subject or object.

Stage 1 (Indirect) relatives ⇒ Stage 2 (Indirect) relatives

(60)



(61)



The main reanalyses between Stage 1 and Stage 2 were triggered by the loss of case morphology due to apocope. Similar to developments in Pre-Old Irish as reconstructed by Newton (2006), phrases occupying the specifier of the CP were reanalysed as particles in the C-head. Relative clauses in the direct predecessors of Welsh and Breton could then be formed in several ways. The relative suffix **-io* became the third-person singular absolute ending (the only absolute ending found in British). Direct relatives were formed by the relative marker **ha(i)* in the C-head. Analogous to this development, the new relative marker **yd* appeared in the C-head following non-argumental antecedents. According to Schrijver (1997), the adverbial phrase **ed* 'thus' was reanalysed as a declarative sentence-initial particle as well. The verb is still moving up to the C-head as well to satisfy the phi-probe. The (relative) particles are like complementisers and have to be merged in the C-head, but they do not carry φ-features and thus cannot satisfy the phi-probe. From a minimalist perspective this means that a similar spec-to-head reanalysis took place here, yielding CVSO orders found in Early Breton and Welsh sources.

From Stage 2 to Stage 3: Loss of copula, rise of V2

A number of changes took place from the first reconstructed stage of the language, (Late) Proto-British, to the earliest attestation in the Brythonic languages. There is evidence for Old Breton and Cornish, but only in the form of lexical glosses (translations) that do not tell us much - if anything at all - about the syntax of these languages. As discussed in Chapter 1, there is more material available in Old Welsh, but even this is very limited. Stage 3 thus also describes the situation as we find in the earliest Medieval stages of the Brythonic languages.

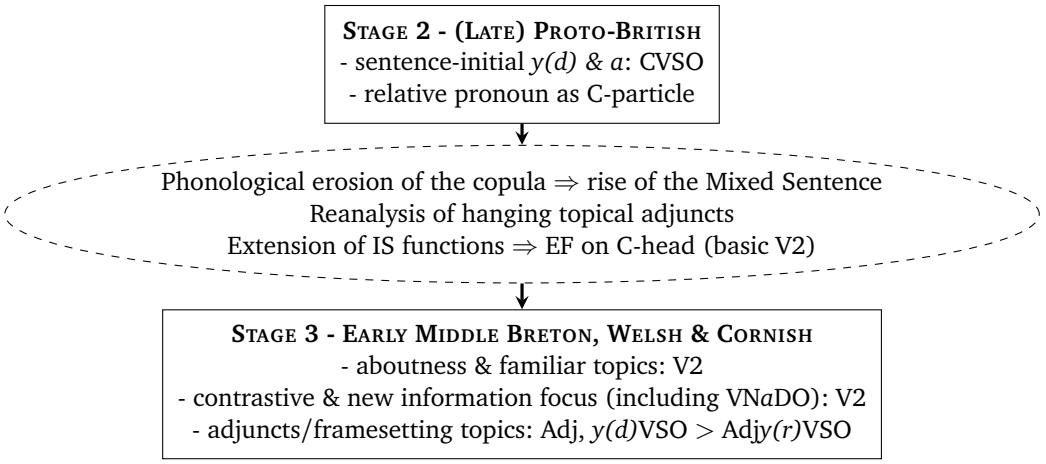


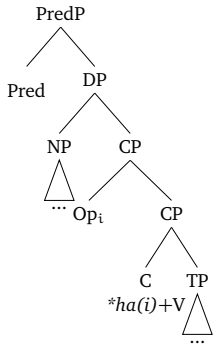
Figure 7.10: Rise & fall V2 from Stage 2 to Stage 3

The phonological erosion of the sentence-initial copula gave rise to the so-called Mixed Sentence, a V2 structure with a relative marker in the C-head carrying an

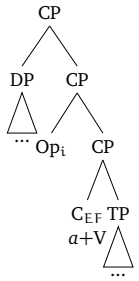
Edge Feature. After the loss of the copula, the sentence was no longer interpreted as a relative clause and the C-head acquired an Edge Feature to ensure its specifier to be occupied at all times.

Cleft Sentence ⇒ Mixed Sentence

(62)



(63)



Initially, the constituent preceding the verb could only be contrastively focussed, but now this information-structural restriction is lost with the emergence of EF on the C-head. Apart from contrastively focussed constituents, contrastive topics can now occupy this first place in the sentence. This was then extended even further to include aboutness and familiar topics until the SpecCP position was a generic position for constituents bearing any kind of IS feature. Non-contrastive focus like new information focus is now also associated with this position. Verbal nouns (with their internal arguments) also belonged to this category now. These sentences with initial verbal nouns (VNs) followed by the inflected form of the verb 'to do' are also frequently found in Middle Breton and Middle Cornish and can were thus likely to exist in Late Proto-British as well, as shown in examples (64), (65) and (66) below.

(64) a. *A dechreu a wnnawn o gyfreith gwlat*
and start.INF PRT do.IPV.1P from law country
'And let us start from the Law of the Country.' (Laws 30)

b. *Agori y drws a oruc ef.*
open.INF the door PRT do.PAST.3S he.
'He opened the door.' (PKM 22.22)

(65) a. *Leuskel a ra hon lestr eun tenn kanol*
fire.INF PRT do.3S 1P boat a gunshot
'Our boat fires a gunshot.' (Middle Breton - MBJJ p.33)

b. *Gervel e zaou vevel a reas ar ronfl.*
call.INF 3MS two servant PRT do.3S the ogre
'The ogre called his two servants.' (Middle Breton - MAV p.34)

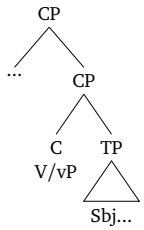
(66) a. *Ty a wra y les.*
you PRT do.3S 3MS width
'You make its width.' (Middle Cornish - O.958)

b. *Oma ty a ra pedry*
here you PRT do.3S rot
'Here you shall rot.' (Middle Cornish - BMer 3577)

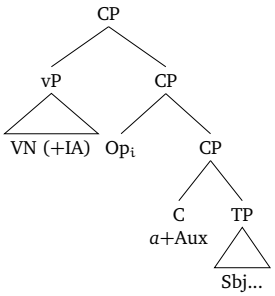
The verb phrases already found in sentence-initial position in earlier stages of the language were now reanalysed as the first constituent in SpecCP with an auxiliary verb in the C-head according to this new V2 requirement.

Celtic verb-initial ⇒ MW VNaDO

(67)



(68)



A further change that took place in Late Proto-British was triggered by the CVSO orders with sentence-initial particle *y(d)* (Breton *ez*, Cornish *y(th)*) in the C-head. Adjuncts in the form of adverbial or prepositional phrases that were originally directly merged outside the matrix CP as hanging topics could now be followed by such a matrix CVSO clause. With the new EF on the C-head, these clause-initial adjuncts could be reanalysed occupying the specifier position of the CP: Adj, *y(d)*VSO > Adj*y(r)*VSO.[14] Schematically, the reanalysis looked like (69) resulting in examples with sentence-initial adjuncts functioning as frame-setting topics followed by the particle *y(r)*, as shown in (70).

(69) $[_{CP}$ PP/AdvP $[_{CP}$ *y(d)* + V $[_{TP}$ ...]]] > $[_{CP}$ PP/AdvP *y(d)* + V $[_{TP}$ ...]]

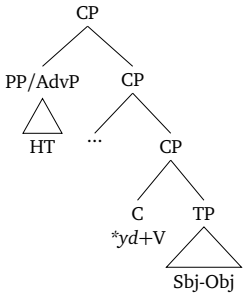
(70) a. *A thrannoeth y talwyt y ueirych idaw.*
and next.day PRT pay.IMPERS.PAST 3MS horses to.3MS
'And on the next day his horses were paid to him.' (PKM 34.23)

b. *Yn Aber Cuawc yt ganant gogeu.*
in Aber Cuawg PRT sing.3P cuckoos
In Aber Cuawg the cuckoos sing.' (CLlH 23.5)

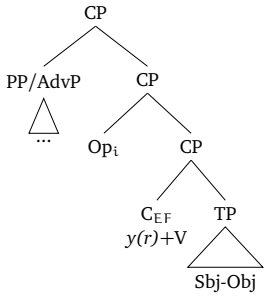
[14]According to Schrijver (1997), *y(d)* changed to *y(r)* in Middle Welsh.

Hanging topic ⇒ Adjunct V2

(71)



(72)



To conclude, the phonological erosion of the copula resulted in a rise of the V2 orders in the so-called Mixed Sentence. The C-head was occupied by the former relative markers *a* or *y(r)* depending on the function of the XP in SpecCP. From an information-structural point of view, there was an extension of the sentence-initial position from contrastive focus to contrastive topic, new information focus and finally also aboutness and familiar topics. The SpecCP position was obligatorily filled by an XP with any of these IS functions because the C-head gained an Edge Feature to attract the verb yielding the preferred verb-second orders in Middle Welsh. When these structures were no longer associated with their relative origin, the Operator that had moved from an adjunct position lower down in the clause to SpecCP was lost and replaced by the PP/AdvP adjuncts.

From Stage 3 to Stage 4: rise of the Abnormal Sentence

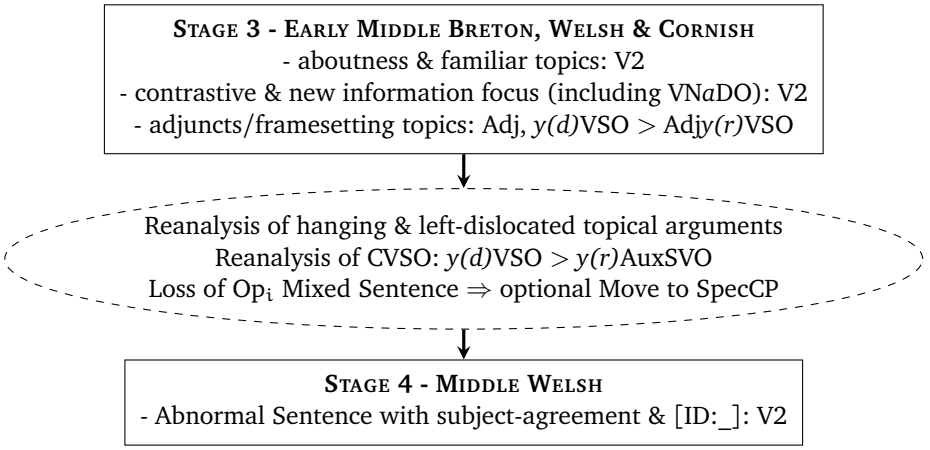


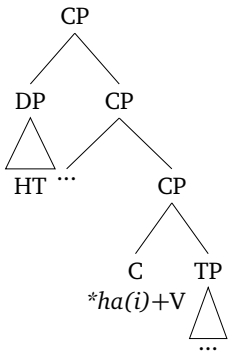
Figure 7.11: Rise & fall V2 from Brythonic to Early Modern Welsh

The changes that took place next resulted in the situation we find in most Middle Welsh literature. The most striking innovation was rise of subject-verb agreement in the so-called Abnormal Sentence. The reanalysis of adjunct phrases formally located outside the matrix CP as hanging topics paved the way for a further reanalysis of arguments as well.

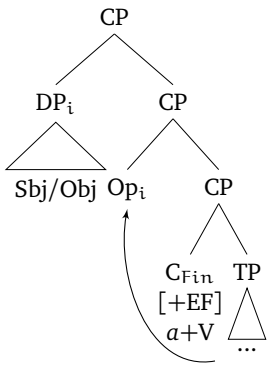
First of all, argumental hanging topics (HTs) that were originally generated outside the matrix CP could be reanalysed as subjects or objects of the matrix. These argumental DPs then occupied the specifier of $C_{Fin}P$, just as their adjunct counterparts.

Hanging topic ⇒ Subject/Object V2

(73)



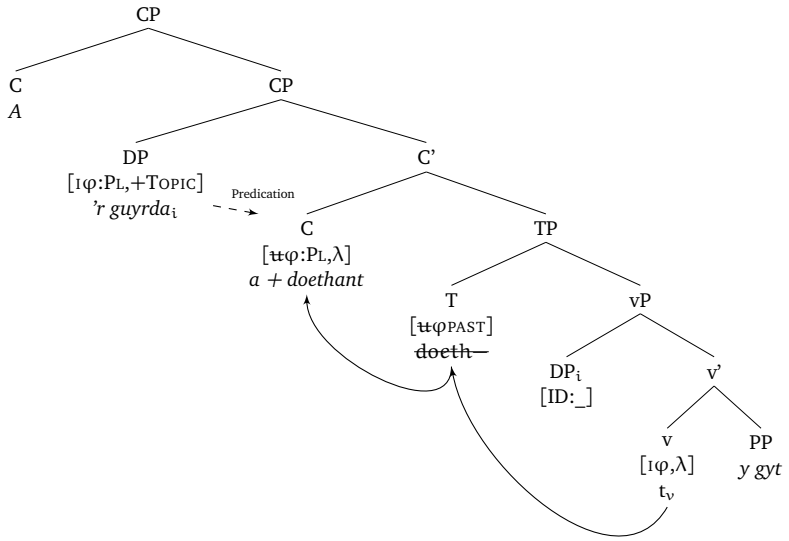
(74)



The original relative markers *a* and *y(r)* that now occupied the C-head had been reinterpreted as positive declarative markers. There was no longer a need to postulate a relative Operator in SpecCP and therefore this was eventually lost as well. Instead, a minimal pronoun [ID:_] entered the derivation as the External or Internal Argument of the verb. With the loss of the relative operator, the base-generated XP in SpecCP could enter a predication relation with the C-head. In addition to the phi-probe and the Edge Feature, the C-head now also bears a λ feature linking the verb in the C-head to the subject DP in its specifier through which the agreement morphology on the verb could be realised. The derivation of these kinds of 'topicalised' Abnormal Sentences is shown again in (76):

(75) *A 'r guyrda a doethant y gyt.*
and the nobles PRT come.PAST.3P together
'And the nobles came together' (Abnormal Sentence - PKM 90.27)

(76)



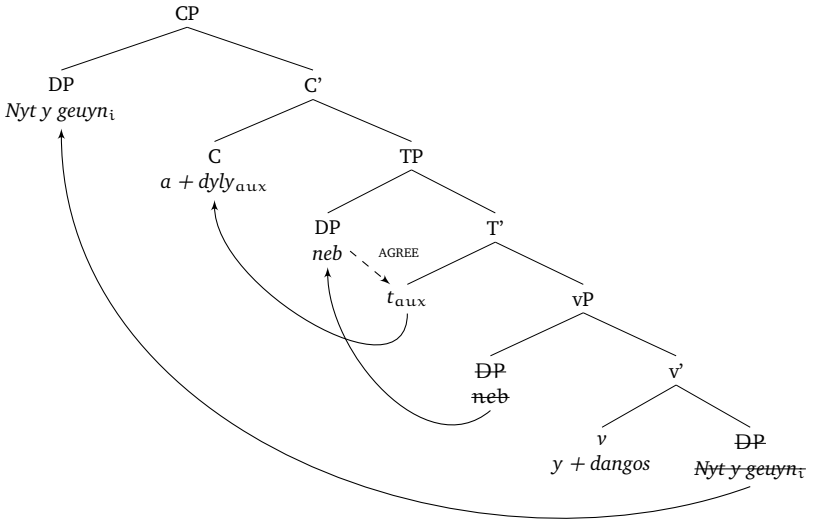
A further development after the loss of the relative operator was the possibility of Moving constituents to SpecCP, rather than externally Merging them in the C-domain with a coindexed minimal pronoun as shown in the Abnormal Sentences above (see Chapter 6 on the minimal pronoun and λ predication in these constructions). Plural DP subjects like *y gwyrda* 'the noblemen' in (75) could not be derived in this way, for the plural agreement goes against the Complementarity Principle that was already well-established in the language by this time. However, pronominal subjects (the most commonly found type of sentence-initial subject) could be analysed either way: both a movement and a base-generated strategy with a minimal pronoun would yield the expected subject-verb agreement as long as no 'trace' of movement is spelled out in the form of an echo pronoun.

From an information-structural point of view, aboutness topics like the full DP subject *y gwyrda* seem to be externally merged at all times, whereas familiar topics like the pronominal subjects could also be internally merged. Constituents representing New Information like verbal nouns or direct objects were gradually lost in the course of the Middle Welsh period. Contrastively focussed constituents are initially externally merged in the typical Mixed Sentence pattern, but in a later stage - after the loss of the relative operator - these could be reanalysed as internally merged constituents as well. This explains the agreement with contrastively focussed pronominal subjects in Late Middle Welsh. Aboutness topics thus seem to be the only constituents towards the end of the Middle Welsh period that were derived via base-generation in SpecCP and coindexed with a minimal pronoun in argument position. These types of topics could remain more associated with their hanging topic origin than familiar topics. Cross-linguistically, there is furthermore

evidence from Italian that indicates a similar base-generation strategy for aboutness topics (cf. Frascarelli (2007)). Constituents with another IS status, like contrastive focus or familiar topics, on the other hand, were fully integrated in the clause and could thus be reanalysed as being derived via a movement strategy instead. An example of such a movement strategy with contrastively focussed constituents from Chapter 6 is given in (78).

(77) *Nyt y geuyn a dyly neb y dangos y*
NEG 3MS.GEN back PRT should.PRES.3S anyone 3SM.GEN show.INF to.3MS.GEN
elynnyon.
enemies
'It is not his back that anyone should show to his enemies' (i.e. 'No one should show his back to his enemies.') (YCM 140.26-7)

(78)



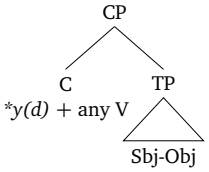
A final syntactic change in this stage was the specification of verbs that were allowed in CVSO contexts. As discussed above, in Breton, CVSO was only possible with verbs of motion and the verb 'to be'. In Middle Welsh there are still some examples with a wider range of verbs like 'to give' or 'to say' etc. In later Middle Welsh, however, the only verb that is allowed to follow the sentence-initial particle *y(d)* is *bod* 'to be'. The particle existed in various forms in front of this verb that was mostly used as an auxiliary, as shown in (79):

(79) a. *Ac y mae matholwch yn rodi brenhinaeth I. y wern*
and PRT be.3S Matholwch PROGR give.INF kingdom I. to Gwern
'And Matholwch is giving the kingdom of I. to Gwern.' (PKM 41.9-10)
b. *Ac y maent yn kyrchu y tir ...*
and PRT be.3P PROGR make.for the land
'And they made for the land...' (PKM 82.16)

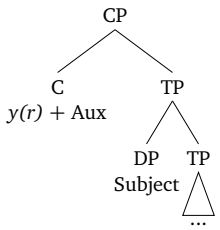
The original CVSO word order pattern thus turned to *y(r)*AuxSVO.

CVSO ⇒ *y(r)*AuxSVO

(80)



(81)



It should be noted that in the above configurations, the C-head does not bear an Edge Feature. If it did, it would trigger the merger of a constituent in SpecCP, which is not what we find in sentences with *y(r)*AuxSVO order. The construction was also found without the particle *y(r)*, but neither of these word order patterns occur frequently in native Middle Welsh tales. The majority of these periphrastic (*y*)AuxSVO orders in the corpus under investigation are found in the 1588 Bible translation in which VSO orders are starting to appear as well. Since there are so few examples in native Middle Welsh literature, it could be argued that those are remnants of the older stage of the language in which C did not yet bear an EF. The increase in frequency in late Middle Welsh could be related to the loss of EF on the C-head again. Alternatively, we have to explain why these Aux-initial orders were possible when C bears a feature that requires its specifier to be filled yielding V2 orders.

One possibility pursuing this argument could be that the particle *y* and the auxiliary are in fact not in the (same) C-head position, but somewhere higher up in a more articulated left periphery. Recall from the previous section that we had no evidence *for* an articulate CP in Common Celtic or Middle Welsh, but - apart from reasons of economy - we also have no conclusive evidence *against* it. If the (former) relative particle *a* is merged in a lower C-head, say C_{Fin} for example, obligatory merger of XP yielding the observed V2 structures would be in SpecC_{Fin}P. The particle *y(d)* found in absolute sentence-initial position could instead be merged in an even higher position in the left periphery, for instance, the head of C_{Force}. A split-CP analysis like this is in fact proposed by, among others, Roberts (2005) (Tallerman (1998) also proposes multiple layers in the CP, but does not label them as 'Fin', 'Force' or 'Topic/Focus' specifically). The verb *bod* 'to be' in particular then also occupies the highest position in the left periphery. Further evidence for this comes from sentences with negation and subordinate clauses in Modern Welsh (cf. Tallerman (1998), Roberts (2004) and Roberts (2005)).

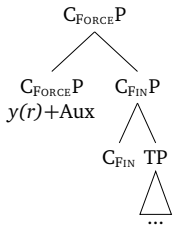
If this is the case, there are two possible scenarios that account for this particle in C_{Force}: the afore-mentioned two forms that both yielded *y(d)* (the neuter pronoun **ed-ed* 'it, this' and the adverb **ed* 'thus') could actually have resulted in two particles each occupying a different C-head. One of those was reanalysed

as the head of C_{Force} bearing a phi-probe to attract the auxiliary. The other one was reanalysed as the head C_{Fin} bearing a phi-probe and an Edge Feature yielding the observed V2 structures (just like the other particle in C_{Fin}, the former relative marker *a*). It is important to note in this context that there was another particle *yd* in Middle Welsh (Middle Cornish *ys-*, Breton *ed-*) that was found before present and imperfect forms of the copula that started with a vowel (MW *ydiw*, MC *vsy*, MB *edy* 'is'). According to Schrijver (1997), this particle must be of yet another Celtic source (that nonetheless had the exact same **VdV* structure yielding *yd*). The origin of this particle (and thus its original syntactic function that is of interest to us here) remains obscure.[15]

The exact etymologies of these particles are important if we want to gain a better understanding of Early Welsh syntax, but a comprehensive investigation goes beyond the scope of the present study. Without further evidence from Old Welsh and OSWB sources, their origin might remain 'obscure'. From a syntactic point of view, however, the following two structures were likely to occur alongside each other in Early Middle Welsh: a periphrastic construction with the auxiliary *bod* 'to be' in C_{Force} and a V2 structure with extended IS functions for the sentence-initial constituent (Subject/Object arguments with *a* or Adjuncts with *y(r)*) in SpecC_{Fin}:

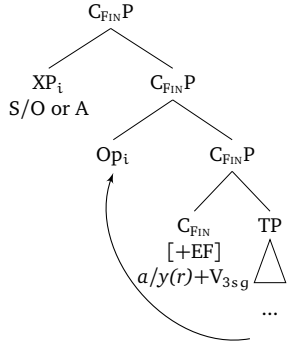
*y(r)*AuxSVO

(82)



V2

(83)



To conclude, Middle Welsh saw the rise of the Abnormal Sentence with subject-verb agreement through the loss of the relative operator and the reanalysis of hanging topics and matrix subjects. The loss of the operator furthermore resulted in a formal split between aboutness topics and constituents with other IS markings. Aboutness topics, for example plural DP subjects, were still base-generated in the C-domain and coindexed with a minimal pronoun in the arguments position of the main clause. Constituents with contrastive focus or familiar topics, on the other hand, were reanalysed as being derived via a movement strategy. Finally, the CVSO order that was possible with all kinds of verbs in early stages of Middle

[15] Schrijver (1997:164) does, however, refer to Pedersen (1913:174, 233) and Morris Jones (1913:288) for what he calls "unconvincing explanations".

Welsh became restricted to constructions with the auxiliary *bod* 'to be'. In these periphrastic sentences, the particle *y(r)* was merged in the head of C_{FORCE} attracting the auxiliary yielding *y(r)*AuxSVO orders as the only remaining alternative to V2 in Middle Welsh positive declarative main clauses. The other declarative particles *a* and *y(r)* occupied the lower C_{Fin}-head bearing an Edge Feature triggering the merger of any XP in its specifier if C_{Force} was not projected (i.e. if the particle *y(r)* associated with C_{Force} was not part of the Numeration.

From Stage 4 to Stage 5: loss of V2

This final stage is without doubt characterised by the loss of V2 word order. The changes involved in this process are described in great detail by Willis (1998) and Willis (2007a). The main triggers for the reanalyses were the loss of the preverbal particles *a* and *y(r)* in the C-head. In combination with the increase in use of periphrastic constructions this led to a significant drop of evidence for the acquisition of V2 word orders.

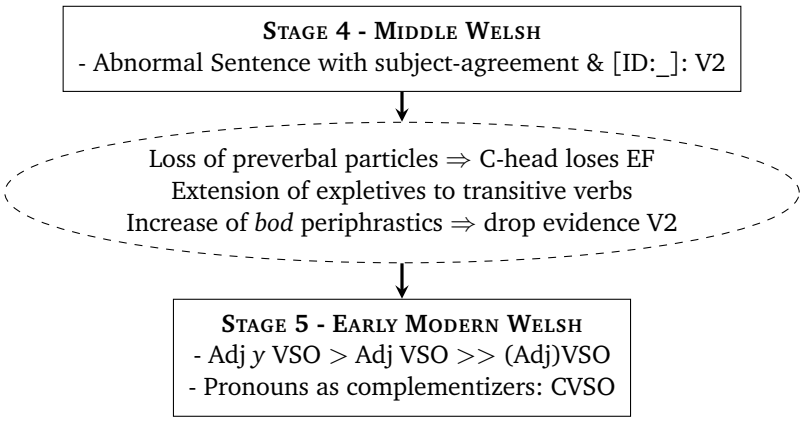


Figure 7.12: Rise & fall V2 from Stage 4 to Stage 5

The loss of the preverbal particles initially resulted in SVO and AdjVSO orders. Recall from Chapter 5 that object-initial sentences were almost completely lost in the late Middle Welsh period, so OVS orders did not arise as the results of the loss of the particles. A further possibility that was more frequently found in the course of the Middle Welsh period put the expletive *ef* in sentence-initial position, even with transitive verbs (see Willis (1998)). Of the SVO sentences, most sentence-initial subjects were pronouns. With the loss of the particle, these pronominal DPs in the specifier of C_{Fin}P were reanalysed as complementisers in the C-head yielding CVSO (again, though now with the former pronouns *mi* and *fe* as C-heads). This type of Spec-to-Head reanalysis was already found in earlier stages of Middle Welsh (the origin of the relative markers *a* and *y(r)*) and is supported by cross-linguistic evidence as well (cf. Willis (2007a)).

Adjunct-initial orders were in turn reanalysed as VSO orders with optional Adverbial or Prepositional Phrases in sentence-initial position. The Edge Feature on the C-head was lost, because children did not receive enough evidence to postulate this feature triggering V2 orders. What constitutes 'enough' in the previous sentence? This brings us back to a point discussed in the introduction: what is the minimum frequency of a cue or trigger needed for a child to postulate a certain grammar? In a language in which the C-head bears a phi-probe, every sentence with a non-subject XP preceding the verb could count as evidence for an EF on the C-head and thus a V2 grammar.

In the Middle Welsh period, there were three different sentence types that could count as this kind of evidence: sentences with initial objects, verbal nouns or adjuncts (adverbial and prepositional phrases). Of those, adjunct-initial orders were most frequently found in almost all Middle Welsh texts in the corpus: as Figure 7.13 shows, these types of non-subject-initial V2 sentences cumulatives make up around 30% of all positive declarative main clauses or even more. In the 1588 Bible translation, however, this is no longer the case: adjunct-initial orders now make up less than 20% and VN-initial and object-initial orders have (virtually) disappeared).

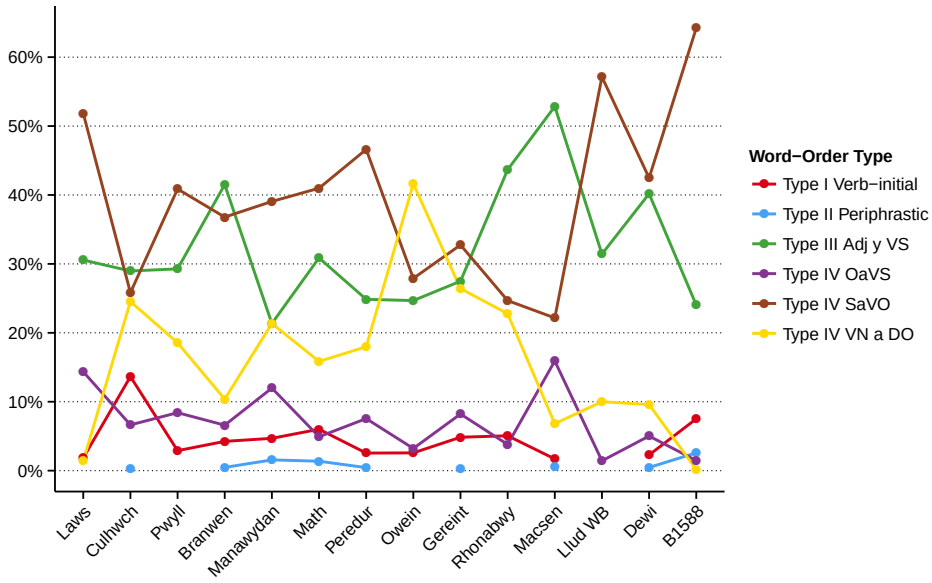


Figure 7.13: Main word order types per text from Early to Late Middle Welsh

Sentences with initial verbal nouns are also frequently found (on average around 20%), but this percentage drops towards the end of the Middle Welsh period. Object-initial orders were never very frequent, remaining around 10%, but again this number drops in the later Middle Welsh texts (BR, LL, Dewi and the Bible translation). If we add up these numbers per text, the 1588 Bible translation already

shows a considerable drop in 'evidence for V2'. Over 60% of the sentences in the Bible are subject-initial, leaving less than 25% adjunct-initial orders. This 25% comes close to the numbers sufficient for V2 acquisition suggested by Lightfoot (1999:154) (30%) and C. D. Yang (2000:114) (23%). With over 60% subject-initial sentences, Late Middle Welsh at first glance looks like it is heading towards SVO like French and English after the loss of V2. Modern Welsh is verb-initial, however, so how did children in the Early Modern Welsh period opt for the loss of the Edge Feature only (keeping the phi-probe and thus verb-initial order)?

Let us first look at the further possible word order types in positive declarative clauses. Object-initial and verb-initial orders in the periphrastic construction with verbal nouns followed by the auxiliary *gwneuthur* 'to do' have almost disappeared completely by the time of the 1588 Bible translation. Absolute verb-initial orders and periphrastic orders with the auxiliary *bod* 'to be' (following the particle *y(r)* in the C_{Force} head) are on the rise, although together they constitute only just over 10% of all positive declarative main clauses. Throughout the Middle Welsh period, however, verb-initial orders were furthermore found in subordinate clauses as well as almost all negative declarative main clauses and yes/no questions. As soon as the pronominal subjects were reanalysed as complementisers, it was no longer necessary to postulate an Edge Feature filling the specifier of C, but the phi-probe on the C-head remained yielding VSO basic word order in Modern Welsh.

Conclusion: the rise & fall of Middle Welsh V2

In this section I have described various processes of reanalysis and extension that led to syntactic innovation from the earliest (reconstructed) stages of the Brythonic languages to Early Modern Welsh. The most striking fact in the history of the Welsh language is that for a period of almost 1000 years (roughly from 600-1600), the grammar seemed to have been defined by a verb-second rule, placing constituents with a specific information-structural status in initial position. Although syntactic change in a generative framework can still be analysed in a parametric context, "V2 grammar" cannot be described as a simple parameter switch. First of all, these verb-second phenomena encompass a wide range of syntactic and information-structural options in the structure of the sentence. This results in varieties within different stages of the language in the case of historical Welsh, but it is also observed in cross-linguistic studies of V2 languages. Not all languages exhibiting a V2 rule have the exact same syntactic structure. This means that a change from 'V2' to 'non-V2' can in fact be the result of a number of smaller reanalyses and extensions in various linguistic domains (see also similar suggestions of changes via small steps by e.g. Haeberli and Ihsane (2015)). In the previous sections, I have given a detailed account of how each of these small syntactic innovations could trigger further extensions and reanalyses, leading to an apparent gradual change in the history of Welsh from verb-initial word order to a preferred V2 order and back again.

I identified possible triggers that led to syntactic innovations, both in the form of reanalyses (e.g. rebracketing or spec-to-head reanalysis) and extensions

(e.g. of the information-structural status of the sentence-initial constituent). I furthermore defined the necessary context in which these changes could take place the way they did, establishing plausible cases for "local directionality" even in reconstructed stages of the language. A prime example of this is the state of the language before the rise of V2. I have argued that a phi-probe on a C-head triggering V-to-C movement, in combination with the existence of clefts indicating contrastive focus as well as optional V2, V3 and V4 orders are a necessary precondition for the development of V2 in the Brythonic languages. In this context, XPs occupying specifier positions in the C-domain (such as the relative pronoun **sosin*) could be reanalysed as functional heads (for instance, triggered by the loss of case morphology due to apocope that turned them into indeclinable relative markers). A change like this is thus wholly in line with Minimalist views on variation stipulated by the Borer-Chomsky Conjecture ("All parameters of variation are attributable to the features of particular items (e.g. the functional heads) in the lexicon." (Baker, 2008:353)). In a hierarchical parametric framework, this would be a 'nanoparametric' change, because it involves the change in the featural make-up of specific lexical items.

The changes that led to generalised V2 in Middle Welsh include further reanalyses in the form of rebracketing of hanging topics to constituents that are part of the matrix occupying the specifier of CP. The extension of the IS function of sentence-initial constituents is a featural change. Alongside an uninterpretable feature probing contrastively focussed constituents, the C-head came to bear a probe for contrastive topics, aboutness topics, familiar topics etc. Along the lines of Minimalist principles of Feature Economy in acquisition, this wide variety of IS-probing features was merged into one generalised Edge Feature (EF) probing any constituent with a specific IS status (i.e. any topicalised or focussed constituent could now be merged in SpecCP).

The combination of a phi-probe and an EF on the C-head yields the so-called 'V2 constraint' that was generalised in this way in Early Middle Welsh (and Breton and Cornish). With an abundance of non-subject-initial V2 orders (in the form of object-initial, verbal-noun-initial and adjunct-initial orders), Middle Welsh children could acquire the V2 rule without any problems. I furthermore argued that the Edge Feature was specifically postulated to be on the lower C-head: $C_{\texttt{Fin}}$, because of the alternative auxiliary-initial periphrastic constructions with the particle *y(r)* in $C_{\texttt{Force}}$. This means that Middle Welsh can be analysed as having reached Stage 3 on the cross-linguistic scale of the Rise of V2 postulated by Wolfe (2015) (see Figure 7.14 below). Modern V2-languages like German or Dutch are characterised by the Edge Feature on the highest C-head: $C_{\texttt{Force}}$. Middle Welsh, however, never reached that stage in the development of V2 in the grammar. The Edge Feature could not be analysed (and thus postulated by children) on the highest C-head, since the periphrastic constructions with the auxiliary *bod* 'to be' were never preceded by other constituents. If Wolfe's 'Stage 3' V2 is a less stable environment for the V2 constraint than his final 'Stage 4' (still existing today in Modern German and Dutch) this could have been a contributing factor in the subsequent loss of V2 in the Early

Middle Welsh period. Much more research in the specific characteristics of each of these stages in a variety of languages is needed, however, before we can draw any such conclusions.

Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
No +EF on C	No +EF on C	C_{Fin} bears an EF	C_{Force} bears an EF
Optional Merger of an XP which is +Top, +Foc, +Neg etc.	Optional Merger of an XP which is +Top, +Foc, +Neg etc.	XP Merger Obligatory	XP Merger Obligatory
CPol, CFoc, CTop probe finite V	C_{Fin} bears an active Phi-Probe	C_{Fin} bears an active Phi-Probe	C_{Force} bears an active Phi-Probe

Figure 7.14: Stages in the Rise of V2 cross-linguistically by Wolfe (2015:44)

Further syntactic innovations in Middle Welsh included the development of the Abnormal Sentence with (unexpected) subject-verb agreement. From the point of view of absolute chronology of the various stages, this is the first structural sign that Middle Welsh is different from its Old South-West British neighbours Breton and Cornish that never developed subject-verb agreement with preposed subjects. A precondition for this further development in Middle Welsh is the existence of sentences with hanging and left-dislocated topics that yielded V2 (and possible V3 and V4 orders in previous stages of the language.[16] These were reanalysed to be in SpecCP position as well satisfying C's Edge Feature, but they were externally merged initially and coindexed with a minimal pronoun lower down in the clause. Apparent subject-verb agreement is the result of the spell-out of the phi-features of the verb in the C-head bearing a λ-feature that allows it to enter into a predication relation with the topical DP in its specifier. It was argued that the situation of Middle Welsh was such that all preconditions were in place allowing this change to happen, including a trigger for the reanalysis of hanging topical arguments analogically to the reanalysis of hanging topical adjuncts. Such an 'analogical trigger' could be viewed as a form of Input Generalisation in which children generalise the structure/interpretation of one construction in all domains or on all levels. This interacted with the extension of IS functions of the sentence-initial constituents at the same time. If we want to answer the question why the situation was such in Middle Welsh and not in Middle Breton or Middle Cornish, a similar thorough investigation of Breton and Cornish word order and information structure is necessary. I leave this - to the extent it is possible with the limited amount of prose data in those languages - for future research at this point.

[16] The Late Latin sources studied by Wolfe (2015) are in Stage 2 of his chronology and these developed into Early Old French, Spanish, Sicilian and Occitan that are argued to be Stage 3 languages with an EF on C_{Fin}. According to this chronology then, Late Latin went through the same process I sketched for Late Proto-British. This might in fact shed some light on the ongoing discussion about language contact after the fall of the Roman empire in Britain (see Schrijver (2002) and Russell (2012) and the discussion on morpho-syntactic similarities in section 7.2.2 above). A detailed investigation of Late British Latin sources is necessary, however, before we can reach any conclusions here.

Since the 'V2 rule' involves at least two different features (both a phi-probe and an EF on the C-head), it is difficult to put this in one single parametric hierarchy. A further complication stems from the fact that the phi-probe is not only associated with attracting the verb, but also with the so-called 'pro-drop' languages. On this highest parametrical level in the hierarchy presented by Biberauer et al. (2014:112), for example, uφ-features are absent from all probes yielding 'radical pro-drop' languages like Chinese or Japanese. Alternatively (and somewhat contrary to their other hierarchies), uφ-features can be present on *all* probes, yielding pronominal arguments, and only then specified to some probes (pro-drop), etc. The phi-probe in the discussion of V2 and verb-initial languages, however, is mainly an indication of verb-movement and can thus be indicative of word order in relation to its subject and direct object. In order to comply with the second condition for V2 (an Edge Feature triggering the merger of an XP to the specifier of its head), we need to complicate the simple hierarchy with further options. A tentative and simplified (i.e. not taking optional/obligatory pied-piping into account, for example) version of this is presented in Figure 7.15:

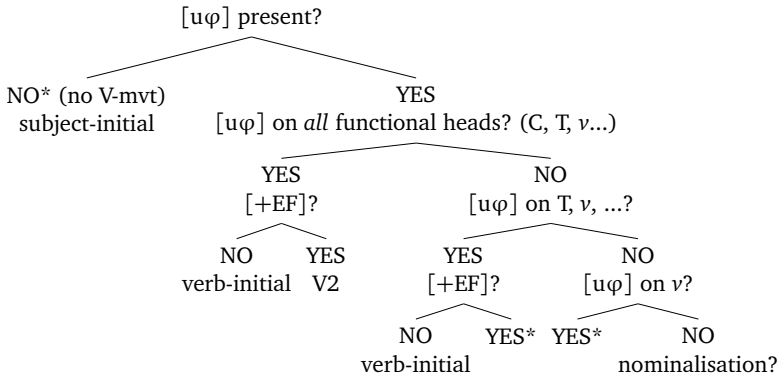


Figure 7.15: Hierarchy for verb-movement via [uφ], including [+EF] yielding V2

This hierarchy shows that a combination of questions need to be answered in order to arrive at V2 word order. In other words, a combination of parametrical settings of the featural hierarchy is necessary to arrive at a grammar with a V2 constraint. Incidentally, if we look at the above tentative figure, we see that the same combination is necessary for verb-initial orders. The subject-initial orders could in theory be further divided into SVO and SOV languages. The asterisk * in the figure here thus actually indicate a link to another parameter hierarchy, namely 'head-finality' (determining, amongst others, OV vs VO orders). Finally, if the verb or V-head is not even probed by little *v*, we could possibly think of languages that involve nominalisation or verb-incorporation. Crucially for our story about the rise and fall of V2 constraints in the grammar, we could arguably insert an extra layer indicating different functional heads in the C-domain. If the Edge Feature is present on the lower C-head (C_{Fin}), for example, the range of languages differs from those

in which EF resides on C_{Force}.

From a diachronic point of view a language can lose or gain two features in this context. The loss of [+EF] on a Macro- or Meso-level is predicted to lead to verb-initial word order (all else being equal), as we see in Welsh. If the phi-probe on the C-head is lost, on the other hand, subject-initial word orders become a possibility as well. The latter arguably happened in the histories of Romance and Present-day English. Much more further research is necessary, however, to test the viability of the above-sketched hierarchy taking more cross-linguistic historical evidence into account. It is furthermore important to investigate the reflexes of possible interaction with other parametrical hierarchies, e.g. the other phi-probe hierarchy for Null Subjects.

For the history of Welsh, the changes in the featural makeup of the C-head facilitated 'the Rise and Fall of V2'. The changes were triggered by the relative markers that came to occupy the C-head and subsequently turned into positive declarative sentence markers. A very similar kind of spec-to-head reanalysis in the CP some 1000 years later then resulted in the loss of the Edge Feature and thus the loss of V2 in Early Modern Welsh.

7.4 Information structure in diachronic syntax

In this chapter I have mainly focussed on structural changes in the history of the Welsh language. One final question that remains concerns the role of information structure in this process of syntactic innovations. In the previous chapter I briefly discussed the 'place' of information structure in the grammar and how it can be encoded in the syntax (rather than other linguistic domains, such as prosody for which we have no historical data). I concluded, following recent Minimalist assumptions that in syntax, information-structural characteristics can be featurally encoded and incorporated as such mainly in the left periphery of the sentence. Do these information-structural features have any influence on changes in the grammar over time? Can they trigger syntactic innovations themselves and/or do they play another role in diachronic syntax?

If we look at the first case study in this chapter on the grammaticalisation of the *sef*-construction, information-structural features definitely played a role in various reanalyses that took place. The original construction only existed in the first place to focus the predicate of identificatory copular clauses. In the course of the process, first the 'identificatory' requirement was lost, leaving a new lexical item *sef* as a specific focus marker. This focussed interpretation was subsequently lost as well and *sef* was eventually reanalysed as the connector of reformulative appositions (like Latin 'id est' still commonly used in abbreviated form in English 'i.e.'). The loss or gain of an information-structural feature like [+FOCUS] could be argued to be the trigger for further syntactic innovations. The question remains, however, what ultimately triggered this loss/gain in the first place.

The same goes for the extension of IS functions of the sentence-initial constituents in verb-second clauses in Middle Welsh. Plausible pathways of extension from, for example, contrastive focus to contrastive topics can easily be identified. The generalisation of probing 'any-IS-marked' constituent rather than probing Focus or Topic specifically is part of a Feature Economy strategy in acquisition. Along such lines, the coexistence of so many different IS features (i.e. Contrastive Topic, Focus, Aboutness Topic, Familiar Topic, etc.) was a necessary prerequisite to postulate a generalised Edge Feature on the C-head. But evidence from languages in which multiple constituents with various IS functions occupy the left periphery shows that this coexistence is not necessarily a trigger for subsequent syntactic innovations. If cross-linguistic diachronic evidence shows that this is a development in more languages, we can get a firmer grip on the role of IS features in this context. For the rise of V2, for example, similar patterns in the history of Romance were discovered by Sam Wolfe. This is a promising start, but much more work is needed before we can reach any final conclusions.

On the basis of much recent literature on diachronic syntax and the case studies presented here in the history of Welsh, we can conclude that information structure definitely plays a role in synchronic variation and thus possible word order patterns. The extent to which it triggers, facilitates or even merely affects changes in the grammar over time is, however, less clear (cf. Taylor and Pintzuk (2015)). As such, this is not a surprising conclusion if we go back to the discussion of possible endogenous and exogenous triggers for language change (see section 7.2.3 and Willis (2016)). In working with historical data (and the extent to which this is available at all) it may not be possible to define the 'ultimate cause' for language change. But in historical syntax, we can describe the exact synchronic state of the grammar in all its detail to explain how and why specific innovations were possible in the first place, how and why they developed in the way they did and why the result is exactly the way we find and not otherwise (cf. Biberauer and Roberts (2015)). Therefore a good understanding of the place of information structure in the syntax of a language as well as a sound methodology to investigate IS functions is important for both synchronic and diachronic research.

7.5 Conclusion

In this chapter I finally turned to diachronic syntax. First of all I discussed various approaches to the study of diachronic syntax, including socio-linguistic variationist, construction grammar and generative approaches. I discussed studies of Welsh historical syntax in these approaches and concluded that they could not give comprehensive accounts or answer all questions in terms of how and why certain changes took place. I argued that adopting a generative acquisitional framework has various benefits in the study of diachronic syntax. First of all it allows us to use insights from various synchronic studies on variation in syntax. The tools and mechanisms tested within the Minimalist Program can furthermore help us define the exact conditions and/or context in which innovation can and cannot occur and

how they can trigger further changes. I have used this to show how innovations were triggered, how children were able to postulate new features or reanalyse the output they are confronted with and why they changed in a certain direction.

I then presented two case studies of syntactic change in the history of Welsh. I described the variation stages and processes of reanalysis and extension and I furthermore examined the role of information-structural features in each of those. The first of these case studies is concerned with a very specific type of focus strategy: identificatory predicate focus. I showed how this construction arose from the cleft construction still found in Old Welsh and how it, after the erosion of the copula, changed in Middle Welsh. First a focus marker *sef* emerged that could be employed in a number of different constructions that were created after reanalysis of the original cleft structure. Then the focussed interpretation was lost and *sef* was reinterpreted as an expletive and, finally, as a linker in reformulative appositional structures ("i.e.").

In the second case study I showed the various stages and innovations involved in the rise of V2 word order in Middle Welsh. A major difficulty in this discussion is the lack of data for the first stages of the language in which the construction originates. This required careful comparison with other Celtic languages such as Gaulish and Irish for the initial stage and other Brythonic languages like Breton and Cornish for the second stage. Since syntactic reconstruction suffers from the correspondence problem of double cognacy (see amongst others, Willis (2007b) and Walkden (2014)), I focussed on the reconstruction of the functional particles in the C-domain that can still be found in the Brythonic languages. I then described the further developments of reanalysis of hanging topics and relative clauses (the 'Mixed Sentence') and the extension of IS functions leading to the postulation of a generalised Edge Feature on the C-head. On the basis of further possible sentence types like the periphrastic construction with the auxiliary *bod* 'to do' in Middle Welsh, I further argued that this Edge Feature must be on a lower C-head, $C_{\texttt{Fin}}$. The phonological erosion of the C-particles in the Early Modern Welsh period finally resulted again in the loss of V2.

I then put these diachronic developments in a wider cross-linguistic context and sketched a tentative feature hierarchy for word order patterns including V2. Finally, I returned to the role of information-structural features in diachronic syntax.

CHAPTER 8

Conclusions

In this thesis I aimed to address the question of the puzzling observations in Middle Welsh word order. First of all, the most-frequently found patterns involve verb-second order. This is 'abnormal' from a Modern Welsh preferred VSO point of view. A further puzzling fact is the large number of possible word orders in Middle Welsh. The verb-second orders alone can take various forms with the sentence-initial constituent and the agreement pattern as the main variables. Finally, it is unclear where these verb-second orders come from, because the limited amount of data available for older stages of the language suggests that sentences with verb-initial orders were more commonly used. In this study, I therefore tried to answer two crucial questions:

1. How can we explain the distribution of the various word order patterns in Middle Welsh?

2. Where do the various verb-second orders (including those with and without subject-verb agreement) come from?

To a certain extent, these questions have been "vexed" and are "by now tormented" by various Welsh scholars in the past decades (see Chapters 1 and 4 in particular). Much progress was made over the years, but there we still find variation in Middle Welsh word order that "frustratingly defies easy explanation" (Poppe, 2014:73). I argue that there are two ways to solve this problem and that we need both if we want to make significant progress in elucidating obscure patterns in word order variation found in any (historical stage of a) language. We first of all need a (large)

digitised corpus that is morpho-syntactically annotated. In second place, we need a consistent methodology to analyse information-structural (or any other) notions that can influence the order of the words in a sentence. Apart from answering the above questions for Middle Welsh, this thesis furthermore presents a sound methodology on how to approach word order phenomena in historical corpora.

In Chapter 2 I formulated my arguments for the use of annotated corpora in more detail. When conducting historical linguistic research, in particular syntactic research, we can only rely on the distribution of the different forms and constructions that we can find. The extent to which our observations reflect the language at the time is likely to increase when we use larger corpora. If a particular pattern occurs often in one text, we cannot jump to the conclusion that this is the case in all textual evidence. Exactly because the amount of extant data is extremely limited, we must try and retrieve the most information we possibly can. This can be achieved by providing detailed part-of-speech tags. This elaborate morpho-syntactic annotation helps to automatically extract the necessary linguistic information from the corpus. Ideally, we create an annotated corpus containing all extant texts, but building such a corpus is a tremendous task. For the present study, I took the first steps on the way to create a fully annotated Treebank of Middle Welsh by selecting, preprocessing, tagging, correcting and parsing 15 texts from the early to the late Middle Welsh period.

I trained a memory-based part-of-speech (PoS) tagger to automatically assign morpho-syntactic tags to the Middle Welsh texts. The choice of PoS-tagger was mainly based on the good results achieved with minimal preprocessing of the difficult data. The difficulty for any automated task lies mainly in the highly irregular orthography found in the Middle Welsh manuscripts and furthermore, the concept of initial consonant mutation found in all Insular-Celtic languages. I furthermore extended the conventional UPenn tagset tremendously to include highly detailed morpho-syntactic information that can facilitate much more future research. With a Global Accuracy of over 90%, the memory-based tagger performed reasonably well considering the difficult data and large tagset (consisting of $>$200 tags). The amount of time needed for subsequent manual correction was thus fairly limited. I then designed a rule-based grammar for Middle Welsh and used the NLTK regular expression parser to add phrase structure to the corpus based on the corrected PoS-tags. With an extremely detailed grammar and a double loop, the parser assigned hierarchical structures to the corpus. These automatic parses were again manually corrected and subsequently converted to bracketed formats to enable searches via CorpusSearch of XQuery facilitating any queries concerning word order patterns. The main result is a reasonably large corpus (15 texts) from which over 9,000 well-annotated positive declarative main clauses could be extracted. In the future, this corpus can be extended to include more texts from different genres, manuscripts and stages of the Welsh language.

In Chapter 3 I outlined a consistent methodology for the investigation of information structure in historical corpora. I discussed three core information-structural notions in detail: Givenness, Topic (vs. Comment) and Focus (vs. Background). I outlined their main characteristics in a systematic way so that they can be used to annotate a corpus consistently. I annotated the referential status of subjects and objects (i.e. their 'Givenness') in the Middle Welsh corpus according to the Pentaset developed by Komen (2013). In Chapter 5 I showed how this type of annotation can help identify effects in word order distributions in combination with annotated syntactic features. Concerning the second core information-structural notion of Topic, I identified three different kinds of topics in the Middle Welsh corpus: Aboutness, Contrastive and Familiar topics. In the next part of Chapter 3 I presented a detailed overview of different kinds of Focus structures including systematic 'algorithms' to find the right focus articulation (Presentational/Thetic, Predicate or Constituent Focus) and the numerous subtypes of Constituent Focus. I furthermore discussed two further notions that are relevant to information structure: Point of departure and Information Flow. The Principle of Natural information flow stipulates that old information usually precedes new information. In sentences with the reverse order, the 'flow' of information, or in particular the referential status of the core arguments, is 'marked'. This helps to give an accurate description of object-initial word orders in Middle Welsh, as I discussed in Chapter 5. Finally, the 'Points of Departure' of a sentence appear mainly in the form of temporal or circumstantial clauses. In effect, they function as frame setters delimiting the context of the rest of the sentence. The clear definitions and guidelines to find the right labels presented in this chapter facilitate annotation of large corpora. A consistent analysis in turn is indispensable for the type of research historical syntacticians are interested in.

Chapter 4 and 5 presented the data and core observations concerning Middle Welsh word order variation. In the compiled corpus, I found a large number of different word order patterns in positive declarative main clauses. I categorised them based on purely formal reasons into nine different main types:

I Verb-initial (VSO)
  (a) VSO (verb absolute clause-initial)
  (b) particle VSO

II Periphrastic constructions with initial auxiliary (AuxSVO)
  (a) with auxiliary *bod*
  (b) with auxiliary *gwneud*
  (c) with auxiliary *ddaru*

III Verb-second after adjuncts ('Abnormal Sentence')
  (a) AdjP *y* VSO
  (b) PredP *y* VSO
  (c) AspP *y* VSO

(d) AdvP *y* VSO
(e) PP *y* VSO

IV Verb-second after arguments and VNs ('Abnormal Sentence')
(a) S *a* V_{agree} O
(b) O *a* V S
(c) patient *a* $V_{impersonal}$
(d) VN *a* DO_{infl} (*gwneuthur*-periphrasis)

V Verb-second after focussed items ('Mixed Sentence')
(a) *(ys)* focussed noun/pronoun *a* V_{3sg}
(b) *(ys)* focussed adjunct *y* V_{3sg}

VI Bare verbal nouns
(a) VN + agent
(b) VN + *o* + agent
(c) *a(c)* VN (continuing previous finite clause)

VII Copular clauses
(a) SCP
(b) PCS
(c) CPS
(d) C S *yn* P
(e) C S *(ys)sydd* P

VIII Identificational Focus construction
(a) Sef + DP (+ relative)
(b) Sef + *yw/oed*
(c) Sef + *a/y*

IX Non-verbal clauses
(a) *dyma/dyna/llyma/llyna* + S (truncated copular clause)
(b) S *(yn)* P
(c) PS
(d) Absolutive: Ac S P(P)

Sentences with verb-initial word order are rare in Middle Welsh, although variants with sentence-initial conjunctions or declarative particles like *neu(r)* directly followed by the verb are found somewhat more frequently. The second type is a periphrastic construction with the auxiliary form of the verb *bod* 'to be', rendering AuxSVO order. This type is also rarely found. Its frequency increases towards the end of the Middle Welsh period. Word order Types I and II (VSO and AuxSVO) are the predominant patterns found in Modern Welsh. The verb-second pattern (the 'Abnormal Sentence') in one of its various forms (Types III, IV or even the focussed type V, the 'Mixed Sentence') is the most commonly found pattern in

Middle Welsh. The adjunct-initial order can appear in many forms and multiple adjuncts are possible too, as long as the 'topicalised' constituent functions as an adjunct. The other type of 'Abnormal Sentence', Type IV, on the other hand places a core argument (Subject or Direct Object) in sentence-initial position. A variant of this type consists of sentences with verbal nouns in initial position followed by the pre-verbal particle *a* and the auxiliary *gwneuthur* 'to do'. This type most commonly appears in contexts of narrative continuity. In subject-initial sentences, the verb usually agrees with the pre-verbal subject. This is what formally distinguishes the 'Abnormal Sentence' from the 'Mixed Sentence' in which the verb shows default third-person singular inflection (Type V). Sentences with verbal nouns *instead of* finite verbs (Type VI) were mainly possible in (Early) Middle Welsh. In early Middle Welsh texts such as *Culhwch*, the verbal noun could appear in non-finite main clauses on their own followed by the subject. These 'verbal noun + agent' almost disappear in independent main clauses. Only sentence-initial verbal nouns in co-ordinated sentences depending on preceding finite clauses continued to exist much longer. Types VII and VIII only describe sentences with copular verbs. The copula could also be left out in Middle Welsh. These non-verbal sentences were finally labelled as Type IX. If we leave out the copular clauses, we can see a clear trend in the distributional of the various word order patterns presented in rough chronological order in Figure 8.1.

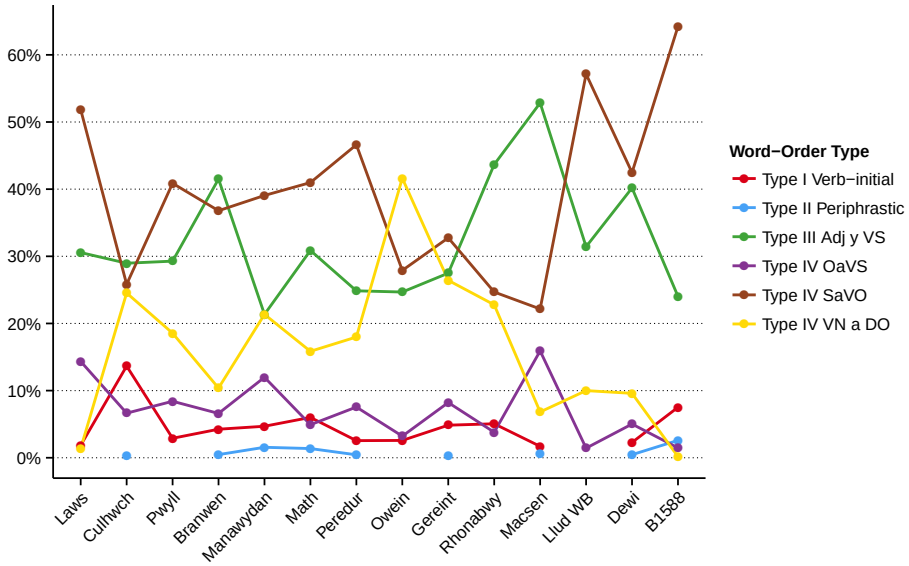


Figure 8.1: Main word order types per text from Early to Late Middle Welsh

It is clear from the above graph that language is already changing at the end of the Middle Welsh period. The preferred word order is still the verb-second 'Ab-

normal' order, but an overwhelming number of sentences are now subject-initial. Verb-initial orders (Type I) and in particularly auxiliary-initial periphrastic orders (Type II) are on the rise. The 1588 Bible translation is particularly interesting, because it was very influential. Most prose texts in Early Modern Welsh are of a religious nature written by people who were very familiar with this translation. As pointed out in the introduction, for the 19th-century Oxford reformers, it was "embarrassing" to hear Jesus and Job speak 'bad Welsh'. The prevalent V2 order in the 1588 translation is indeed different from the Modern Welsh V1-language they spoke. Interesting, however, from this study its becomes clear that the syntax and word order preferences in the 1588 Bible translation also differ from the general patterns in Middle Welsh. The clear preference for subject-initial sentences in the 1588 translation is not found earlier.

In Chapter 5 I systematically presented all possible factors that could influence the word order of the Middle Welsh sentence. Starting with possible grammatical factors, verb-second sentences with verbal nouns in initial position (Type IVc VNaDO) almost exclusively occur with verbs in the preterite tense. The significance of (preterite) tense as a factor is likely to be related to the fact that these verbal-noun patterns are the basic word order in indirect speech passages of narrative tales. In direct speech, on the other hand, subject-initial orders are most frequently attested. The corpus study furthermore shows that impersonal verbs are most frequently found in verb-second sentences with initial adjuncts (Type III). Finally, there seems to be a limited role for Animacy of the core constituents. For subjects, there are no significant results, but inanimate objects tend to appear in object-initial orders more frequently than expected.

Only once all language-internal and language-external factors are systematically tested in this way (to the extent this is possible with the information we have), we can determine whether other factors, such as information-structural notions play a role. The first information-structural notion under investigation was Givenness. Direct objects in initial position almost exclusively convey New information. This indicates that the 'Natural information flow' of the sentence (going from old to new) is reversed and these object-initial sentences are thus marked in this way. The only exceptions to this generalisation are so-called Familiar topics. These are topics that appear in sentence-initial position mainly in the form of demonstrative pronouns. They refer back to the last-mentioned item/person/concept in the immediately preceding context. The corpus study revealed two further observations in terms of textual cohesion. First of all 'points of departure' or frame-setters occur most often in verb-second sentences with adjunct-initial order (Type III) in which they function as the topic. A second observation in this context concerns textual continuity. In order to achieve close cohesion, verbal nouns can be placed in sentence-initial position. They are either relying on an inflected verb in the previous sentence (Type VI) or are continued with an inflected form of the auxiliary 'to do' (Type IVc). Again this is likely to be part of the narrative style in this genre. Finally, focus can first of all be observed in the dedicated (reduced) cleft order called the 'Mixed Sentence'

(Type V). Focus of the identificatory predicate can furthermore be found in the special *sef*-construction (Type VIII), but not all sentences with *sef* are focussed.

Chapter 6 and 7 focussed on the synchronic and diachronic syntactic analysis of the different word order patterns. In Chapter 6 I presented four different case studies related to the most important information-structural features in Middle Welsh. The aim of this chapter was to provide a syntactic analysis for those information-structural phenomena and to see how notions like topic, focus and givenness are implemented in the syntax of the language. Middle Welsh only allowed one topic position, but V3 and even V4 structures are attested. In the discussion I mainly focussed on the puzzling variation and agreement observations in the verb-second 'Abnormal Sentence'. Two different types of analyses were presented and discussed in detail: a movement and a base-generated approach. I argued that agreement with sentence-initial plural DP topics can be explained by adopting a base-generated approach, but not by a movement approach. The topic is base-generated in the left periphery of the clause, but it is co-indexed with a minimal pronoun lower down in the structure. The φ-features on the verb can be checked in the C-domain via λ-predication in the same way this is possible in relative clauses (cf. Kratzer (2009)). The lack of agreement is due to an operator that moves to SpecCP as a remnant of the reduced cleft in an earlier stage of the language. A movement approach (but not a base-generated approach) can account for sentences without subject-verb agreement and I argued that in some particularly difficult coordinate structures exhibiting both plural and singular agreement, a mixed analysis is the best solution. In general, however, movement approaches create problems for sentence-initial subjects, because Middle Welsh seems to adhere to the Complementarity Principle. According to this principle (that is also found in Breton), any form of agreement with plural full DPs is unexpected. The same holds for the lack of agreement with focalised pronouns. Both of these observed structures thus present problems for a movement analysis. Under a base-generated approach, however, these different agreement patterns can be explained. There are, however, also examples that present a greater challenge for a base-generated analysis, such as sentence-initial constituents that must be (locally?) bound by a quantifier and argumental PPs. The Middle Welsh corpus most likely reflects two possible patterns: movement and base-generation.

In the final chapter I discussed various approaches to the study of diachronic syntax, including socio-linguistic variationist, construction grammar and generative approaches. I argued that adopting an generative acquisitional framework has various benefits in the study of Middle Welsh diachronic syntax, because it allows us to use insights from synchronic studies on variation in syntax. The tools and mechanisms tested within the Minimalist Program can furthermore help us define the exact conditions and context in which innovations can and cannot occur and how they can trigger any subsequent changes.

I presented two case studies of syntactic change in the history of Welsh to demonstrate this. The first of these case studies is concerned with a very specific type of focus strategy: identificatory predicate focus. I showed how this construction arose from the cleft construction still found in Old Welsh and led to the emergence of the focus marker *sef*. When the focussed interpretation was lost, *sef* was reinterpreted as an expletive and, finally, as a linker in reformulative appositional structures ("i.e.").

In the second case study I addressed the second research question of the present study. I showed how the the verb-second orders came into existence in Middle Welsh by careful comparison with other Celtic languages. I focussed on the reconstruction of the functional particles in the C-domain that can still be found in the Brythonic languages. I then described the further developments of reanalysis of hanging topics and relative clauses (the 'Mixed Sentence') and the extension of information-structural functions leading to the postulation of a generalised Edge Feature on the C-head. On the basis of further possible sentence types like the periphrastic construction with the auxiliary *bod* 'to do' in Middle Welsh, I further argued that this Edge Feature must be on a lower C-head, $C_{\texttt{Fin}}$. The phonological erosion of the C-particles in the Early Modern Welsh period eventually resulted again in the loss of V2. Finally, I put these diachronic developments in a wider cross-linguistic context and sketched a tentative feature hierarchy for word order patterns including V2:

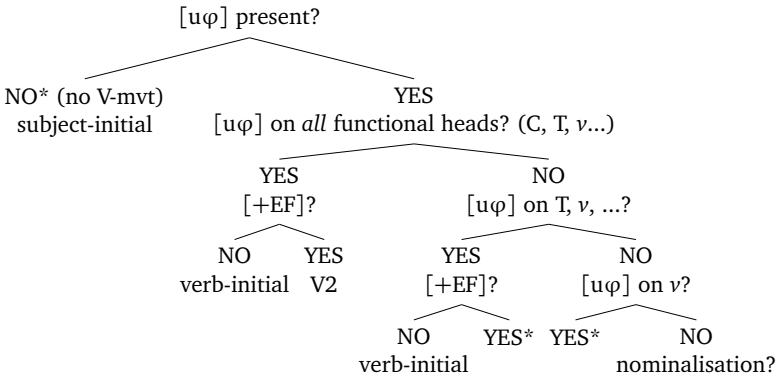


Figure 8.2: Hierarchy for verb-movement via [uφ], including [+EF] yielding V2

This present study finally aimed to investigate the interaction between syntax and information structure and their respective (or combined) effects on word order. From a synchronic point of view, the distribution of word order patterns in Middle Welsh is the result of a combination of both grammatical and information-structural factors. Focus was expressed with a reduced cleft construction, the so-called 'Mixed Order'. In identificatory copular clauses, however, focus was expressed by means of the focus marker *sef* (< *ys* + *ef* 'it is that'). Givenness and textual cohesion

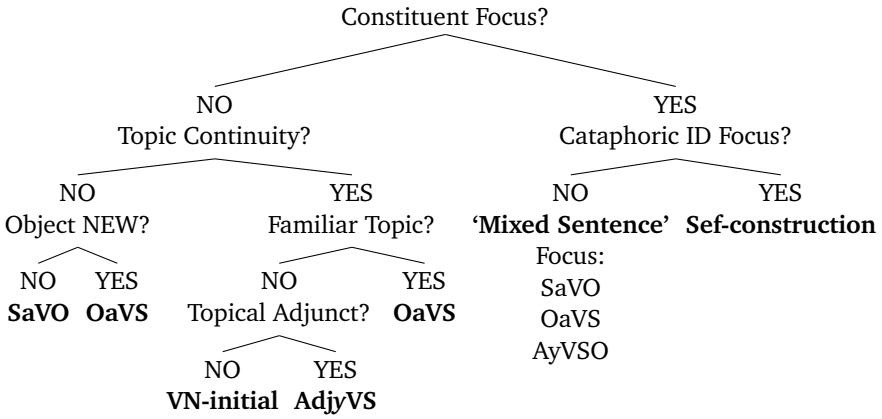


Figure 8.3: Decision algorithm 'predicting' the word order pattern in Middle Welsh

furthermore play a role as described above. Based on the present corpus study,[1] we can establish an algorithm to 'predict' or 'choose' the right option from the wide range of possible word orders; in this way the exact distribution of the various patterns can be explained. With the intended message ready in the Numeration, the syntax can build the sentence that will ultimately yield one of the word order types. In transitive statements in narrative contexts, the basic decision-making algorithm in figure 8.3 can 'predict' the word order (leaving additional adjuncts and extra-clausal constituents like hanging topics aside):

From a diachronic point of view, I showed that information-structural features play a role in syntactic innovations and reanalyses. The extension of information-structural functions of the sentence-initial constituent in verb-second sentences in Middle Welsh (from Contrastive Focus $>$ Contrastive Topics and New Information Focus $>$ Familiar and Aboutness topics) is a good example of this. The ultimate triggers for syntactic changes sometimes remain hard to define, but a detailed and consistent description of the synchronic variation systematically checking different variables as presented in this study is indispensable in diachronic syntactic research.

[1] Needless to say if the corpus is extended with more Middle Welsh texts a similar study needs to be conduct to see if we still arrive at the same result with the extended dataset. I leave this for future research

Appendix - Annotation Manual

1 Introduction

This brief manual describes the guidelines used to add part-of-speech and phrasal annotation to the corpus of Middle Welsh prose. This corpus was initially built for the present investigation in Middle Welsh word order. The focus therefore lies on facilitating queries concerning word order. These query codes are presented at the end of this Appendix.

1.1 Philosophy and goals

The main aim of this project was not to give a correct syntactic analysis or provide a detailed parsed structure. The part-of-speech tags contain highly detailed morphological information, but the phrasal annotation is only a slightly more elaborate shallow parse. In this way, the annotated corpus could remain theory-neutral. At the same time, queries for linear order and hierarchical phrase structure are still possible. And finally, future enrichment of the chunk-parsed corpus is not excluded, because of its flexible XML-format.

Any controversial decisions are avoided as much as possible. The same goes for constructions that are changing over time. A good example is the *sef*-construction in Welsh. The information-structural status of this sentence changes from initial identificatory focus to plain predicate focus in the course of the Middle Welsh period. Since most texts are difficult to date exactly, throughout the corpus I used the specific tag SEF for any occurrence of this type of sentence. In this way, all these sentences can be easily found and investigated by future researchers as well.

1.2 File formats

All mark-up is stripped from the texts, which are then saved as plain text files (.txt). Further preprocessing involved the insertion of sentence-final punctuation (if that was not present in the manuscript already) and the deletion of sentence-internal full

stops (in order to make it readable to the PoS tagger). Finally, utterance boundaries were inserted semi-automatically (automatically after a full stop and manually if the full stop did not exist in the manuscript). The PoS-tagged files created by the Memory-Based Tagger include tags to words in the following fashion: `word/TAG`. These tagged files are saved as text files as well.

The NLTK regular expression chunkparser requires a list of words and tags. Therefore, the PoS-tagged text files were converted to the right format using the script in Figure 4. Chunk-parsed files contain bracketed structures representing phrasal and morphological annotation. The plain text files in this format are thus parsed (.psd). These types of parsed files are searchable with CorpusSearch and other querying tools. The Cesax Software package designed by Erwin Komen (cf. Komen (2013)) converts text and psd files into xml-format (.psdx). In this way, corpus searches are also possible via XQuery.

```
import sys
import os
import re

def make_nltk_readable(file_name):
"""
function takes one argument (file_name), and returns a list
containing (for every sentence) a list of word-pos pairs
"""

all_text = open(file_name)
corpus = []

for line in all_text:
   sentence = []
   pairstrings = re.split("\s", line)
        #split line in word-pos-pair-strings WPPS
        #delete final pairstring
   for p in pairstrings:
            sentence.append(tuple(re.split("\/{1,2}", p)))
        #for each WPPS, split word from PoS and add to sentence
        #print sentence
   sentence = sentence[:-2]

   corpus.append(sentence)
     # add sentence to corpus
   return corpus
```

Figure 4: Script to make output files of the automatic PoS-tagger ready for Chunkparsing

1.3 Text markup

For the markup, I chose the TEI P5 header that is suitable for philological data, translations and annotation in XML format. Any information about the philological background of the text can be stored in this header and easily retrieved for future online usage. In the textual markup, any changes to the annotation, can be indicated as well to trace the history of the annotated text and corpus as a whole. Finally, it would ultimately be possible to combine different versions of the texts (i.e. diplomatic and critical editions) into one xml file to make sure invaluable philological information is not lost.

2 Splitting and joining words

As became clear from the initial pilot, the huge amount of orthographical variation complicates the PoS-tagging task tremendously. The Memory-Based Tagger could filter those out on the basis of the context most of the time. In this way, there was no real need for time-consuming preprocessing of the text in terms of splitting merged tokens. Some tokens, however, were particularly challenging for the automated tagger, since very few generalisations could be made from the small training set (cf. Meelen and Beekhuizen (2013)). Below is a list of items that were split or combined to facilitate automatic tagging.

2.1 Items that are split

- combined words with nasalising prepositions, e.g. *ymwyt* > *y** + *mwyt* 'in food'
- conjunctions combined with definite articles: *ar* > *a** + *r* 'and the'
- particle combined with pronouns, e.g. *ae* > *a** + *e* 'PRT 3MS'

2.2 Combined conjunctions and prepositions

Welsh employs combined prepositions: a combination of a preposition plus a grammaticalised noun. Pronominal objects of these type of prepositions appear in between the two prepositions as a possessive pronoun, e.g. *yn eu herbyn* 'against/towards them' (PKM 65.6-7) from *yn* 'in' + *eu* 'their' + *erbyn* 'opposition'. In this particular case of combined prepositions, a more conservative annotation scheme, acknowledging the nominal origin of the construction yielding the tag sequence 'P 3P N' (preposition third-person plural possessive noun) was preferred to facilitate rule-based chunk-parsing. The most commonly combined prepositions annotated in this way are:

- *ach/ger law* 'beside' (Lit. 'by hand')
- *am law* 'in addition to' (Lit. 'at hand')
- *am/ar/uch ben* 'on top of' (Lit. 'on head')
- *amgylch* 'about' (Lit. 'on circle')

- *ar ffuryf* 'like, as' (Lit. 'in form')
- *ar drws* 'in front of' (Lit. 'at door')
- *ar gefyn* 'on' (Lit. 'on back')
- *ar ol* 'after' (Lit. 'on track')
- *ger/rac bronn* 'by, before' (Lit. 'by breast')
- *heb law* 'past' (Lit. 'without hand')
- *y maes o* 'outside' (Lit. 'in field of')
- *is gil* 'behind' (Lit. 'below back')
- *o achaws* 'because of' (Lit. 'from cause')
- *yn lle* 'instead of' (Lit. 'in place')
- *ym penn* 'after' (Lit. 'in head')

Prepositions in Welsh could also be combined with other prepositions, e.g. *y dan* 'under, below' from *y* 'to' + *tan* 'under'. These complex prepositions were tagged PSUB + PSUB, so they could be recognised as separate, but also as combined prepositions. A further advantage of this is that the automatic tagger looking at the tags preceding and following the focus word, will not encounter the odd sequence of two prepositions. For combined conjunctions, a similar extension was used: *o + herwydd* CONJSUB + CONJSUB meaning 'because'. The most commonly combined prepositions and conjunctions are:

- *hyt ar* 'as far as, up to'
- *hyt at* 'as far as, to'
- *hyt yn* 'until'
- *y am* 'about, towards'
- *y ar* 'on, upon'
- *y gan* 'by, because'
- *y dan* 'under, below'
- *y wrth* 'from
- *y vewn* 'into'
- *o vywn* 'within'
- *y dros* 'for, instead of'
- *y tu ac* 'towards'
- *yr mwyn* 'for the sake of'
- *yn erbyn* 'against'

2.3 Fused forms

Middle Welsh manuscripts exhibit some fused forms as well. The combination found most commonly is the preposition *y* 'to' and the infixed third-person pronoun 'him, her, them' that is often written as *y* as well. These fused forms are annotated with hyphenated tags 'P-PRO'.

3 List of PoS tags

Adjectives and adverbs (ADJ, ADV)

Adjectives appear in various forms:
- positive adjectives, e.g. *coch* 'red' ⇒ ADJ
- comparative adjectives, e.g. *clotuorach* 'more famous' ⇒ ADJR
- superlative adjectives, e.g. *dewraf* 'bravest' ⇒ ADJS
- plural adjectives, e.g. *ieueinc* 'young' ⇒ ADJPL
- ordinal number, e.g. *eil* 'second' ⇒ ADJNUM

Adverbs can appear on their own as true adverbial lexical items, but they can also be adjectives following the predicative particle *yn*, e.g. *yn gyflym* 'quickly'. In these cases, the adjectives are tagged ADJ, but the phrase - a combination of predicative *yn* + ADJ - is labeled as an adverbial phrase ADVP.

Particles (PCL)

There are many different kinds of particles in Middle Welsh:

- preverbal particles, e.g. *a/y* ⇒ PCL
- question particles, e.g. *a* ⇒ PCL-QU
- negative particles, e.g. *ny* ⇒ PCL-NEG
- negative focus particles, e.g. *na* ⇒ PCL-NEG-FOC
- negative question particles, e.g. *oni* ⇒ PCL-QU-NEG
- negative focus question particles, e.g. *onid* ⇒ PCL-FOC-QU-NEG
- focus particles, e.g. *panyw* ⇒ PCL-FOC

Cardinal numbers (NUM)

Cardinal numbers are tagged NUM, regardless of whether they are used as substantives or as adjectives:

- substantives ⇒ *y pedwar hynny* 'those four', *pym mil o wyr* '5,000 men' (lit. '5 thousand of men'), *tri o wyr* 'three men'
- adjective ⇒ *deu wr* 'two men', *teir llong ar dec* 'thirteen ships' (lit. '3 ship on ten'), *pedwar meib ar hugeint* '24 sons' (lit. '4 sons on twenty')

Inflected verbs (VB) and Verbal nouns (VN)

Verbs appear with and without inflection. The uninflected forms can function as nouns or infinitival verbs. To avoid any linguistic interpretation, they are consistently tagged VN. The inflection of the verb is reflected in the tag following VB-. Tense, aspect, mood, person and number are all indicated separately:

- present indicative, e.g. *caraf* 'I love' ⇒ VBPI-1SG
- present subjunctive, e.g. *carhych* 'you would love' ⇒ VBPS-2SG
- preterite verb, e.g. *carawd* 'he loved' ⇒ VBD-3SG
- imperfect indicative, e.g. *carem* 'we loved' ⇒ VBAI-1PL
- imperfect subjunctive, e.g. *carhit* 'was loved' ⇒ VBAS-4
- pluperfect, e.g. *carassewch* 'you (pl) had loved' ⇒ VBG-2PL
- imperative, e.g. *car* 'love!' ⇒ VBI-2SG

Some present and imperfect forms are ambiguous between indicative and subjunctive mood, e.g. *carem* 'we loved'. Whenever they are ambiguous, they are tagged without mood indication: VBA as 'imperfect verb'. Verbs that function as auxiliaries as well have specific tags, e.g. *cael* 'to get' HV-, *bod* 'to be' BE- (unless it is the verbal noun or complementiser, both tagged as BOD), *gwneuthur* 'to do' DO-.

Nominals (N, NPR, PRO)

Singular nouns are N, plural nouns NPL and proper nouns are NPR. There are various types of pronouns in Middle Welsh:

- regular pronouns, e.g. *mi* 'I' ⇒ PRO
- conjunctive pronouns, e.g. *enteu* 'he (too)' ⇒ PROC
- reduplicated pronouns, e.g. *tydi* 'YOU (not him)' ⇒ PROR
- accusative pronouns (infixed clitics), e.g. *e* 'her' ⇒ PRO-A
- genitive pronouns (infixed clitics), e.g. *fy* 'my' ⇒ PRO-G
- indefinite pronoun, *un* 'one' ⇒ ONE

Prepositions (P)

Some prepositions can be inflected in Welsh. The inflection is tagged like verbal endings, e.g. *iddo* 'to him' P-3SGM, *amdanaf* 'about me' P-1SG.

Wh-words

There are various wh-words in Middle Welsh:

- wh-adverbs, e.g. *pryd* 'when?', *sut* 'how?' ⇒ WADV
- wh-determiners, e.g. *pa* 'which, what X' ⇒ WD
- wh-pronouns, e.g. *pwy* 'who?' ⇒ WPRO
- wh-quantifiers, e.g. *sawl* 'how many?' ⇒ WQ
- unidentified wh-item, e.g. *beth* 'what?' ⇒ W

Other tags

Finally, there are some remaining tags:

- Demonstratives, e.g. *hwnnw* 'that' ⇒ DEM
- Determiners, e.g. *yr* 'the' ⇒ D
- Conjunctions, e.g. *a* 'and', *pan* 'when' ⇒ CONJ
- Complementisers, e.g. *y* 'that' ⇒ C
- Quantifiers, e.g. *rai* 'some' ⇒ Q
- Foreign words, e.g. *lama* 'why?' (Aramaic) ⇒ FW
- Predicative markers, e.g. *yn* ⇒ PRED
- Progressive markers, e.g. *yn* ⇒ PROGR
- Reflexives, e.g. *hun* '-self' ⇒ REFL
- Interjections, e.g. *o* 'oh' ⇒ INTJ
- Punctuation ⇒ PUNC

Generating a Middle Welsh PoS-tagger

The tagger is first of all created with the standard parameter settings. Each of these settings can be adjusted, according to what works best for the corpus used. The optimal settings for a certain corpus could be retrieved automatically by running

a script trying all possible options and evaluating the results with a 10-fold cross-validation (see results below).

There are many possible parameter settings (see the MBT reference manual Daelemans et al. (2010)). You can first of all choose which features you would like to take into account when assigning tags to known or unknown words. The letter sequences following -p (known words) and -P (unknown words) indicate the specific context and characters at the beginning and or the end of the word that the tagger should take into account. For the Middle and Modern Welsh taggers, the following features gave the best results:

```
-p dfa -P sssdFawchn
```

The letter 'F' is the focus word that can be examined with the following features. The letter 's', for instance, indicates that the final character should be taken into account. The triple repetition of the letter 's' means that it will take the last three characters into account. Not surprisingly for a language that relies on inflectional suffixes, the last three final characters were important to guess the correct tag for unknown words. 'd' and 'a' refer to the tag of the left and the right context words respectively; 'w' is used for the left or right context words themselves. The letters 'c', 'h' and 'n' stand for capital letters, hyphens or numbers. Features like these help the tagger assign the correct tag for a word it has not 'seen' in the training set and is thus labeled as 'unknown'.

On the basis of this MBTg (the tagger generator) first creates an ambitag lexicon. This is a list of words associated with the different tags it can have according to the training corpus. When a word-tag combination occurs less than 5% (by default, this too is an adjustable setting), it is not included.

Then it creates a frequency list of the 100 (by default, but 200 gave better results for Welsh) most frequent words in the corpus. All words not in the most-frequent-words list are transformed into special symbols: HAPAX-<code> (<code> is either 0, or a combination of H (hyphen), C (capital letter), and N (number)). Instances are created using the specified information sources for known words (as indicated with -p in the parameter settings), then the case base is generated from that (see Daelemans et al. (2010) for further technical details on this process).

On the basis of this, the case base for known words is generated by TiMBL. By default, a lazy-learning algorithm like IGTREE is used, but for this particular corpus, I got better results with the alternative IB1 algorithm for both known and unknown words.

For unknown words, the tagger uses a k-nearest-neighbour algorithm (based on Aha, Kibler, and Albert (1991) but with added *Information Gain* weighting). In addition to that, the selected feature metric is set to `-mM` 'MVDM' (Modified Value difference metric), which allows for partial feature matches (cf. Stanfill and Waltz (1986), Cost and Salzberg (1993) and Daelemans and Van den Bosch (2005)). Finally, weighting of features can be done in an inverse linear fashion with the parameter setting `-dIL`. This means that the neighbour with the smaller distance

is weighted more heavily than the one with a greater distance. From all this, a settings file is created that can be used to annotate unseen texts in the rest of the corpus.

Since there is no need to understand or adjust any of the above-mentioned algorithms or parameter settings to generate a tagger, the MBTg offers a simple and quick way to generate a tagger for any new language or corpus. Thousands of words can be tagged per second and there is no need for any additional smoothing for sparse data since this is already part of the similarity-based model (Zavrel and Daelemans (1997)). Spelling, morphology, context and the words themselves are all sources of information integrated in the weighted similarity metric.

4 List of phrasal tags

The following phrasal tags were used for chunkparsing the corpus:

- verb phrase, combining the preverbal particle and the verb (including direct object) ⇒ VP
- noun phrase, projection of any noun ⇒ NP
- determiner phrase, any determiner/adjective/demonstrative + noun (no internal hierarchy) ⇒ DP
- prepositional phrase, any preposition with a following NP or DP ⇒ PP
- inflected prepositional phrase, projection of inflected prepositions ⇒ PPROP
- adjectival phrase, projection of any adjective ⇒ ADJP
- adverbial phrase, projection of any adverb or adjective + predicative marker ⇒ ADVP
- aspectual phrase, combination of aspectual marker + verbal noun ⇒ ASPP
- numeral phrase, projection of any numeral ⇒ NUMP
- numeral determiner phrase, NUMP + determiners/demonstratives ⇒ NUMDP
- complementiser phrase, main or subordinate clause ⇒ CP or CP-SUB
- quantifier phrase, projection of any quantifier ⇒ QP

Chunking Middle Welsh

The NLTK modules are based on Python; their rule-based regular expression parser works best under version 2.7. In order to chunkparse the PoS-tagged texts, the (manually corrected) output of the Memory-Based Tagger needs to be converted to a format that is readable to the parser. The text files were automatically converted with a Python-based text-preparation script ('chunkprep.py')[2]:

[2]Many thanks to Barend Beekhuizen for helping me develop the Python scripts presented here.

```
import sys
import os
import re

def make_nltk_readable(file_name):
"""
function takes one argument (file_name), and returns a list
containing (for every sentence) a list of word-pos pairs
"""

all_text = open(file_name)
corpus = []

for line in all_text:
   sentence = []
   pairstrings = re.split("\s", line)
        #split line in word-pos-pair-strings WPPS
        #delete final pairstring
   for p in pairstrings:
            sentence.append(tuple(re.split("\/{1,2}", p)))
        #for each WPPS, split word from PoS and add to sentence
        #print sentence
   sentence = sentence[:-2]

   corpus.append(sentence)
     # add sentence to corpus
   return corpus
```

Figure 5: Script to make output files of the automatic PoS-tagger ready for Chunkparsing

The chunkparser was originally not meant to perform parses with such extensive hierarchical structures as required for the present study, but by adjusting the option to loop through the grammar multiple times, these structures can be created nonetheless.

The python module 'pprint' can finally be used to ensure the newly parsed text is printed in the right .psd format to enable search queries via, for example, CorpusSearch. Figure 6 shows the step-by-step commands in Python to chunkparse text X. 'Xgold' refers to the gold standard, the version of the PoS-tagged text that has been manually corrected.

```
>>>import nltk, re, pprint, chunkprep
>>>grammar = r"""
...
VP:{<PCL-PRO-G|PCL-PRO-A|PCL|PCL-NEG|PCL-NEG-PRO-A>?|VBPS-2PL|...>}
    PROP: {<PRO|PROC|PROR|PROX>}
    VNP: {<PRO-G>?<VN|DON|HVN><PROP>?}
    ASPP: {<PROGR|PERF><PRO-G>?<VNP|HVN|DON|BOD>}
    DEMP: {<D><DEM>}
    NUMP: {<NUM>?<NUM|ONE><P><NUM>}
    WDP: {<WD><N|NPL|ONE|QP>}
    NP: {<N|NPL|NPR>}
    NUMDP: {<NUM>?<NUM|ONE><NP><P><D>?<NUM|NP>}
    NUMP: {<NUM|ONE><NP><ADJP>?}
    DP: {<D><NUMP>}
    DP: {<PRO-G><ADJP>?<NP><PROP>}
    DP: {<D><NUM><NUM>?<DEM>?}
    DP: {<NP|D><ADJP>}
    REFLP: {<PRO-G><REFL>}
    P: {<PSUB><PSUB>}
    P: {<P><P>}
    PP: {<P><PP><VNP>}
    PWP: {<P><WPRO|WDP>}
    ONEP: {<ONE><PP|OTHER>}
    PP: {<P><ONEP>}
    DEMP: {<DEM>}
    ADJQP: {<ADJQ><ADJQ>}
    PREDP: {<PRED><ADV|PRO-G>?<NP|ADJP|ADJQP|QP|DP>}
    ADJP: {<ADJP><PP>}
..."""
>>>cp = nltk.RegexpParser(grammar, loop=2)
>>>text = chunkprep.make_nltk_readable('Xgold.txt')
>>>results = []
>>>for t in text:
... result = cp.parse(t)
... results.append(result)
>>>f = open('Xchunked.psd', 'w')
>>>for r in results:
... f.write(r.pprint())
... f.write('\n')
>>>f.close()
```

Figure 6: Adopting & implementing the Python NLTK Chunkparser

5 Known annotation issues

In the current stage, the annotation of the corpus was done in such a way to optimise the search queries specific to the present thesis. The focus lies on the part-of-speech annotation. The highly detailed tag set facilitates future research in any linguistic framework. The same goes for the relatively flat structure of the chunk-parsed files. This can be extended to a full parse on the basis of the manually corrected .psd(x) files or on the basis of the PoS-tagged .txt files.

From a syntactic point of view, the difference between subject and object constituents was initially not indicated. Since Middle Welsh allows subject- and object-initial orders as well as pro-drop it was impossible to do this automatically. The DP-initial orders that could be ambiguous were manually disambiguated at a later stage, dividing them into SVO and OVS orders.

The most important elements that are not included in the current annotated corpus are empty categories. The main reason for not including these at this stage was because they were not necessary for the present investigation in word-order. Furthermore, the aim was to keep the annotated corpus as theory-neutral as possible and empty categories are very theory-specific. The flexible xml-based nature (compatible with the psd file structure) means that those can be added at a later stage as well. This can be done by developing a context-free grammar and/or manual insertion.

Finally, at various stages in the process of creating the corpus, manual correction was necessary. Since there was only one annotator available to build the present corpus, checking cross-annotator agreement was no option. Although an effort has been made to double-check all the corrected versions, some errors no doubt remain. In future, when making the annotated files accessible for everyone online, a final check will be done to filter out any possible mistakes and/or inconsistencies.

6 Coding queries

Figure 7 shows a sample of algorithms in Xquery code used to retrieve values for features like Negation, Focus or Tense, Aspect and Mood for different kinds of verbs (DO- 'to do', BE- 'to be', HV- 'to get', VB any other verb) from the PoS-tagged and Chunkparsed database (converted to XML format). The queries employ standard XQuery code plus additional functions built into the software package CorpusStudio (cf. Komen (2009b)), like `ru:matches` to match labels of PoS-tags indicated in the query with those in the database.[3]

[3]This is just a sample excerpt of the entire query that works with an accompanying definition file in which specific variables like `$vp` and `$sentence` are defined.

```
(: Find Focus particles :)
  let $Foc := $sentence/descendant::eTree[ru:matches
   (@Label, '*FOC*')][1]
  let $strFoc := ru:NodeText($Foc)
  let $feat_Foc := if ($strFoc = '') then '-' else $strFoc

(: Find Negation :)
  let $Neg := $sentence/descendant::eTree[ru:matches
   (@Label, '*NEG*')][1]
  let $strNeg := ru:NodeText($Neg)
  let $feat_Neg := if ($strNeg = '') then '-' else $strNeg

(: Find Mood :)
  let $feat_Mood :=
   if (exists($vp/descendant::eTree[ru:matches
(@Label, 'VBI*|DOI*|BEI*|HVI*')]))
then 'Imperative'
   else if(exists($vp/descendant::eTree[ru:matches
(@Label, 'VBPS*|VBAS*|BEPS*|BEAS*|DOPS*|DOAS*|HVPS*|HVAS*')]))
then 'Subjunctive' else 'Indicative'

(: Find Tense and Aspect :)
  let $feat_TenseAspect :=
   if (exists($vp/descendant::eTree[ru:matches
(@Label, 'VBP-*')]))
then 'Perfect'
else if(exists($vp/descendant::eTree[ru:matches
(@Label, 'VBAI*|VBAS*|DOAI*|DOAS*|BEAI*|BEAS*|HVAI*|HVAS*')]))
then 'Imperfect'
else if(exists($vp/descendant::eTree[ru:matches
(@Label, VBG*|DOG*|BEG*|HVG*')]))
then 'Pluperfect'
   else if(exists($vp/descendant::eTree[ru:matches
(@Label, 'VBD*|DOD*|BED*|HVD*')]))
then 'Preterite' else 'Present'
```

Figure 7: XQuery code to retrieve Focus, Negation and Tense/Aspect feature values

Figure 8 shows some excerpts of the complex query to find the various word order types.[4]

```
(: Look through each text for [‘S’] :)
 for $search in //eTree[ru:matches(@Label, ‘S’)]

(: Determine what the first constituent is, excl. CONJ and C :)
  let $initialCns :=
  $search/child::eTree[not(ru:matches
  (@Label, ‘CONJ*|INTJ|C|PCL-QU’))][1]

(: Determine the main type of this sentence :)
  let $mainType :=
  tb:WelshMainCat($initialCns, $search)

 return ru:back($search, ‘’, $cat)

(...)

(: Get the VP :)
   let $vp := tb:WelshVP($sentence)

(: Determine the main category :)
   let $mainCat :=  if ($initialCns/@Label = ‘SEF’)
then ‘Type VI Sef’
else if (ru:matches($initialCns/@Label,‘W*|PCL-QU’)
and not(exists($sentence/child::eTree[ru:matches
(@Label, ‘QP’)])) )
then ‘Type X Question’
else if (ru:matches($initialCns/@Label, ‘*FOC*’))
then ‘Type XI Focus’
(...)
else if ($initialCns/@Label = ‘VNP’ and
(some $ch in $initialCns/following-sibling::eTree satisfies
        ($ch is $vp and exists($vp/child::eTree[ru:matches
        (@Label, ‘DO*’)])) ) )
then ‘Type IIIc VNaDO’
```

Figure 8: Sample XQuery definition & query algorithm to find the main word order type

[4]Many thanks to Erwin Komen for teaching me how to use XQuery.

References

Aarts, J. (1991). Intuition-based and Observation-based Grammar. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics.* London: Longman.

Aboh, E. O. (2010). Information structuring begins with the numeration. *Iberia: An International Journal of Theoretical Linguistics*, *2*(1), 12-42.

Abraham, W., & de Meij, S. (Eds.). (1986). *Topic, Focus and Configurationality: Papers from the 6th Groningen Grammar Talks, Groningen, 1984*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Adams, J. N. (2007). *The regional diversification of Latin 200 BC-AD 600*. Cambridge: Cambridge University Press.

Adger, D. (2011). Clefted situations: A note on expletives in Scottish Gaelic clefts. In A. Carnie (Ed.), *Formal Approaches to Celtic Linguistics* (p. 3-15). Cambridge Scholars Publishing.

Adger, D. (2013a). Constructions and grammatical explanation: comments on Goldberg. *Mind & Language*, *28*(4), 466-478.

Adger, D. (2013b). *Constructions are not explanations.* Available online on `http://ling.auf.net/lingbuzz/001675`.

Adger, D., & Ramchand, G. (2003). Predication and equation. *Linguistic inquiry*, *34*(3), 325-359.

Adger, D., & Ramchand, G. (2005). Merge and move: *Wh*-dependencies revisited. *Linguistic Inquiry*, *36*(2), 161–193.

Aelbrecht, L., Haegeman, L., & Nye, R. (2012). *Main clause phenomena: New horizons* (Vol. 190). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, *6*(1), 37-66.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383-409.

Allen, C. (1995). *Case marking and reanalysis: grammatical relations from Old to Early Modern English*. Oxford: Oxford University Press.

Alter, S. G. (1999). *Darwinism and the linguistic image: language, race, and natural*

theology in the nineteenth century. Johns Hopkins University Press.
Ambridge, B., Pine, J. M., & Lieven, E. V. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, *90*(3), e53-e90.
Andersen, H. (1973). Abductive and deductive change. *Language*, *49*, 765-793.
Anderson, S. (1982). Where's morphology? *Linguistic Inquiry*, *13*(4), 571–612.
Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics*, *1*(1), 93-111.
Anwyl, E. (1899). *A Welsh Grammar for Schools, part II - Syntax* (Vol. 2). London: Swan Sonnenschein & Co. Ltd.
Ariel, M. (1999). *Accessing Noun-Phrase Antecedents*. London/New York: Routledge.
Badan, L., & Del Gobbo, F. (2011). On the syntax of topic and focus in Chinese. In P. Beninca & N. Munaro (Eds.), *Mapping the left periphery* (p. 63-91). New York & Oxford: Oxford University Press.
Bailey, N. A. (2009). *Thetic constructions in koine greek*. Unpublished doctoral dissertation, Amsterdam: Vrije Universiteit.
Baker, M. (2008). The macroparameter in a microparametric world. In T. Biberauer (Ed.), *The limits of syntactic variation* (Vol. 132, p. 351-373). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Baker, M. C. (2009). Is head movement still needed for noun incorporation? *Lingua*, *119*(2), 148-165.
Baker, P. (2006). *Using corpora in discourse analysis*. Bloomsbury Publishing.
Barðdal, J. (2011). The rise of dative substitution in the history of Icelandic: A diachronic construction grammar account. *Lingua*, *121*(1), 60-79.
Barðdal, J., & Eythórsson, T. (2012). Reconstructing syntax: Construction grammar and the comparative method. In H. C. Boas & I. Sag (Eds.), *Sign-based construction grammar* (pp. 257–308). Stanford, California: CSLI Publications.
Barðdal, J., Smirnova, E., Sommerer, L., & Gildea, S. (2015). *Diachronic Construction Grammar* (Vol. 18). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Baxendale, T. (2009). *Y rhyfel oeraf* (T. D. Jones, Trans.). London: Rily Publications.
Beaver, D. I. (2004). The optimization of discourse anaphora. *Linguistics and Philosophy*, *27*(1), 3–56.
Beekes, R. S. (1995). *Comparative Indo-European linguistics: an introduction*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
Beekhuizen, B. (2015). *Constructions Emerging*. Utrecht: LOT Dissertation Series.
Beekhuizen, B., Bod, R., & Verhagen, A. (2014). The linking problem is a special case of a general problem none of us has solved: Commentary on Ambridge, Pine, and Lieven. *Language*, *90*(3), e91-e96.
Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405.
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biberauer, T. (2015). *The limits of syntactic variation: en emergentist comparative perspective. (Invited talk given at the Workshop on Language Variation and Change and Cultural Evolution)*. Centre for Linguistics History and Diversity, York University, 14 February 2015.

Biberauer, T., Holmberg, A., & Roberts, I. (2014). A syntactic universal and its consequences. *Linguistic Inquiry*, *45*(2), 169-225.

Biberauer, T., Holmberg, A., Roberts, I., & Sheehan, M. (2014). Complexity in comparative syntax: the view from modern parametric theory. In F. J. Newmeyer & L. B. Preston (Eds.), *Measuring grammatical complexity* (p. 103-127). Oxford: Oxford University Press.

Biberauer, T., & Richards, M. (2006). True optionality: When the grammar doesn't mind. *Minimalist essays*, *91*, 35-67.

Biberauer, T., & Roberts, I. (2008). Cascading parameter changes: internally driven change in Middle and Early Modern English. In T. Eythórsson (Ed.), *Grammatical change and linguistic theory: the Rosendal papers* (p. 79-113).

Biberauer, T., & Roberts, I. (2015). *The significance of what hasn't happened (Invited talk given at the Workshop on Language Variation and Change and Cultural Evolution)*. Centre for Linguistics History and Diversity, York University, 14 February 2015.

Biberauer, T., Sheehan, M., & Newton, G. (2010). Impossible changes and impossible borrowings. In A. Breitbarth, C. Lucas, S. Watts, & D. Willis (Eds.), *Continuity and change in grammar* (p. 35-60). Amsterdam: John Benjamins.

Biberauer, T., & Walkden, G. (2015). *Syntax Over Time: Lexical, Morphological, and Information-structural Interactions* (Vol. 15). Oxford: Oxford University Press.

Birch, S., & Clifton, C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and speech*, *38*(4), 365-391.

Birner, B. (2006). Inferential relations and noncanonical word order. In *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn.* (p. 31-51). Amsterdam: John Benjamins.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, *21*(4), 491–504.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 14-28). London: Routledge.

Borer, H. (1984). *Parametric syntax*. Dordrecht, Holland: Foris Publications.

Bornkessel, I., Schlesewsky, M., & Friederici, A. (2003). Contextual information modulates initial processes of syntactic integration: The role of inter-versus intrasentential predictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 871-882.

Borsley, R. (1989). A note on ellipsis and case. *Linguistic Inquiry*, *20*(1), 125–130.

Borsley, R., & Stephens, J. (1989a). Agreement and the position of subjects in

Breton. *Natural Language and Linguistic Theory*, *7*(3), 407–427.
Borsley, R., & Stephens, J. (1989b). Agreement and the position of subjects in Breton. *Natural Language & Linguistic Theory*, *7*(3), 407-427.
Borsley, R., Tallerman, M., & Willis, D. (2007). *The syntax of welsh*. Cambridge University Press.
Bresnan, J. (2001). *Lexical-functional syntax*. Oxford: Blackwell.
Bromiley, G. (1997). *The International Standard Bible Encyclopedia: Vol I: A-D*. Grand Rapids, MI: William B. Eerdmans.
Brown, J. S. (1988). Patch use as an indicator of habitat preference, predation risk, and competition. *Behavioral Ecology and Sociobiology*, *22*(1), 37–47.
Brugmann, K. (1876). Zur Geschichte der stammabstufenden Deklinationen, Erste Abhandlung: Die Nomina auf *-ar-* und *-tar-*. *Curtius' Studien*, *9*, 361-406.
Bucholtz, M. (2008). Theories of discourse as theories of gender: Discourse analysis in language and gender studies. In J. Holmes & M. Meyerhoff (Eds.), *The handbook of language and gender* (p. 43-68). Oxford: Blackwell.
Büring, D. (1997). *The meaning of topic and focus: the 59th Street Bridge accent*. London: Routledge.
Büring, D. (2003). On D-trees, beans, and B-accents. *Linguistics and philosophy*, *26*(5), 511-545.
Bury, D. (2002). A reinterpretation of the loss of verb-second in Welsh. In D. Lightfoot (Ed.), *Syntactic effects of morphological change* (p. 215-231). Oxford: Oxford University Press.
Busa, R. (1992). Half a Century of Literary Computing: Towards a 'New' Philology. *Literary and Linguistic Computing*, *7*, 69-73.
Campbell, L. (2000). What's wrong with grammaticalization? *Language sciences*, *23*(2), 113-161.
Campbell, L., & Janda, R. (2000). Introduction: conceptions of grammaticalization and their problems. *Language sciences*, *23*(2), 93-112.
Cappelle, B. (2009). Can we factor out free choice? *Describing and modeling variation in grammar*, *204*, 183.
Cardinaletti, A., Cinque, G., & Giusti, G. (1988). *Constituent structure*. Dordrecht: Foris Publications.
Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In L. C. N. (Ed.), *Subject and topic* (p. 27-55). New York: Academic Press.
Chambers, E. (1728). *Cyclopaedia; or, an universal dictionary of arts and sciences*. London.
Charles-Edwards, T. (2001). The Textual Tradition of Medieval Welsh Prose Tales and the Problem of Dating. In B. Maier, S. Zimmer, & C. Bakte (Eds.), *150 Jahre "Mabinogion"-Deutsch-Walisische Kulturbeziehungen* (p. 23-39). Tübingen: De Gruyter.
Charles-Edwards, T. M. (2013). *Wales and the Britons, 350-1064*. Oxford: Oxford University Press.
Cheng, L. L.-S., & Downing, L. J. (2012). Against FocusP: arguments from Zulu.

In *Information structure. contrasts and positions* (p. 247-267). Cambridge: Cambridge University Press.

Choi, H.-W. (1999). *Optimizing structure in context: Scrambling and information structure*. Stanford: CSLI Publications.

Chomsky, N. (1964). *Current issues in linguistic theory*. The Hague: Mouton, Den Haag.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: M.I.T. Press.

Chomsky, N. (1969). Quine's Empirical Assumptions. In D. Davidson & J. Hintikka (Eds.), *Words and Objections* (Vol. 21, p. 53-68). Springer Netherlands.

Chomsky, N. (1986). *Barriers*. Cambridge, Massachusetts: MIT Press.

Chomsky, N. (1995). *The minimalist program*. Cambridge, Massachusetts: MIT Press.

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik* (pp. 89–156). MIT Press.

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (Ed.), *Ken hale: A life in language* (p. 1-52). Cambridge, MA: MIT Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, *36*(1), 1-22.

Chomsky, N. (2013). Problems of projection. *Lingua*, *130*, 33-49.

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, *19*(1), 1-24.

Cinque, G. (1977). The movement nature of left dislocation. *Linguistic inquiry*, 397-412.

Cinque, G. (1999). *Adverbs and functional heads: a cross-linguistic perspective*. New York: Oxford University Press.

Clahsen, H. (1991). *Child language and developmental dysphasia: Linguistic studies of the acquisition of German* (Vol. 2). Amsterdam, Philadelphia: John Benjamins Publishing.

Clark, R., & Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry*, *24*, 299-345.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Conrad, S. (2010). What can a corpus tell us about grammar. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 227-240). London: Routledge.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, *10*(1), 57-78.

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge: Cambridge University Press.

Cowles, H. W. (2012). The psychology of information structure. In M. Krifka & R. Musan (Eds.), *The expression of information structure* (Vol. 5, p. 287-317).

Berlin: Walter de Gruyter.

Cowles, H. W., Walenski, M., & Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. *Topoi*, *26*(1), 3-18.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Crisma, P., & Longobardi, G. (2009). Introduction: change, relatedness and inertia in historical syntax. In P. Crisma & G. Longobardi (Eds.), *Historical syntax and linguistic theory* (p. 1-13). Oxford: Oxford University Press.

Culy, C. (1996). Formal properties of natural language and linguistic theories. *Linguistics and Philosophy*, *19*(6), 599–617.

Curme, G. (1978). *A Grammar of the English Language*. Verbatim Books.

Currie, O. (2000). Word order stability and change from a sociolinguistic perspective. *Stability, Variation, and Change of Word-order Patterns Over Time*, *213*, 203-230.

Currie, O. (2013). The history of gradual change and continual variation. In A. G. Ramat, C. Mauri, & P. Molinelli (Eds.), *Synchrony and diachrony: A dynamic interface* (Vol. 133, p. 43-78). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Curzan, S. (2008). Historical corpus linguistics and evidence of language change. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook* (Vol. 2, p. 1097-1109). Berlin: Walter de Gruyter.

Cutler, A., & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition*, *7*(1), 49-59.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, W., Zavrel, J., Van den Bosch, A., & Van der Sloot, K. (2010). Mbt: memory-based tagger - version 3.2 Reference Guide. *ILK Technical Report - ILK 10-04*.

D'Alessandro, R., & Van Oostendorp, M. (2016). Prosody, phi-features and deixis in Southern Italian: what vocatives can tell us on the architecture of language. *Available from* `http://ling.auf.net/lingbuzz/002721`.

Daneš, F. (1970). One instance of Prague school methodology: functional analysis of utterance and text. *Method and theory in linguistics*, 132-146.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In F. Daneš (Ed.), *Papers on functional sentence perspective* (p. 106-128).

Davies, J. (1621[1809]). *Antiquae linguae britannicae, nunc communiter dictae cambro-britannicae, a suis cymraecae vel cambricae, ab aliis wallicae rudimenta: Juxta genuinam naturalemque ipsius linguae proprietatem, qua fieri potuit accurata methodo et brevitate conscripta*. London: Oxonii.

Davies, P., & Deuchar, M. (2010). Using the Matrix Language Frame model to measure the extent of word-order convergence in Welsh-English bilingual speech. In A. Breitbarth, C. Lucas, S. Watts, & D. Willis (Eds.), *Continuity and change in grammar* (p. 77-96).

Davies, S. (1995). *Crefft y Cyfarwydd: Astudiaeth o dechnegau naratif yn Y Mabinogion*. Caerdydd: Gwasg Prifysgol Cymru.

Davies, S. (1998). Written text as performance: the implications for Middle Welsh prose narratives. In *Literacy in medieval celtic societies* (p. 133-48). Cambridge.

Davies, W. D., & Dubinsky, S. (2004). *The grammar of raising and control: A course in syntactic argumentation*. John Wiley & Sons.

Davis, H., Gillon, C., & Matthewson, L. (2014). How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language*, *90*(4), e180-e226.

Delbrück, B. (1900 [1982]). *Einleitung in das Sprachstudium: ein Beitrag zur Geschichte und Methodik der vergleichenden Sprachforschung*. Cambridge: Cambridge University Press.

Déprez, V., & Pierce, A. (1993). Negation and functional projections in early grammar. *Linguistic Inquiry*, *24*(1), 25–68.

De Saussure, F., Bailly, C., & Séchehaye, A. (1916). *Cours de linguistique générale*. Paris: Grande Bibliothèque Payot.

Destruel, E., & Velleman, L. (2014). Refining Contrast: Empirical Evidence from the English it-Cleft. In *Empirical issues in syntax and semantics* (pp. 197–214). Colloque de syntaxe et sémantique à Paris.

De Swart, H., & De Hoop, H. (1995). Topic and focus. *Glot international*, *1*(7), 3-7.

Deutscher, G. (2002). On the misuse of the notion of 'abduction' in linguistics. *Journal of Linguistics*, *38*(03), 469-485.

Di Eugenio, B. (2003). Discourse processing. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (Vol. 1, p. 976-983). London: Nature Publishing Group.

Dik, S. C. (1997). *The theory of functional grammar: the structure of the clause*. Berlin/New York: Mouton de Gruyter.

Dreschler, G. (2015). *Passives and the loss of verb second: A study of syntactic and information-structural factors*. Utrecht: LOT dissertation series.

Dresher, B. E. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, *30*(1), 27–67.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61-74.

Eckhoff, H., & Haug, D. (2011). Personal pronouns with articles: a quantitative approach. In *Information structure and corpus annotation: theoretical and practical perspectives.* Oslo, Lysebu: University of Oslo.

É.Kiss, K. (1995). *Discourse configurational languages*. Oxford: Oxford University Press.

É.Kiss, K. (1998). On generic and existential bare plurals and the classification of predicates. In S. Rothstein (Ed.), *Events and grammar* (pp. 145–162). Dordrecht: Kluwer Academic Publishers.

É.Kiss, K. (2001). Parasitic chains revisited. In P. Culicover & P. Postal (Eds.), *Parasitic gaps* (pp. 99–124). Cambridge, Massachusetts: MIT Press.

Ellis, N., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N.

(2001). *Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh.* Available online from `http://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en`. Last accessed d.d. 20 September 2013.

Engdahl, E., & Vallduví, E. (1996). Information packaging in HPSG. *Edinburgh working papers in cognitive science*, *12*, 1-32.

Erteschik-Shir, N. (2007). *Information structure: The syntax-discourse interface* (Vol. 3). Oxford: Oxford University Press.

Evans, D. S. (1968). The sentence in early Modern Welsh. *Bulletin of the Board of Celtic Studies*, *22*, 311-337.

Evans, D. S. (1971). Concord in Middle Welsh. *Studia Celtica*, *6*, 42-56.

Evans, D. S. (1990). Insular Celtic and the emergence of the Welsh language. In *Britain 400-600: Language and history.* Heidelberg: Carl Winter.

Evans, D. S. (2003 [1964]). *A grammar of Middle Welsh*. DIAS, Dublin.

Evans, E. (1958). Cystrawennau 'sef' mewn Cymraeg Canol. *Bulletin of the Board of Celtic Studies*, *18*, 38-54.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(05), 429-448.

Evers, A., & Van Kampen, J. (2008). Parameter setting and input reduction. *The Limits of Syntactic Variation*, *132*, 483-515.

Faarlund, J. T. (1990). *Syntactic change: toward a theory of historical syntax* (Vol. 50). Berlin & New York: Mouton de Gruyter.

Fairclough, N. (1992). Discourse and text: Linguistic and intertextual analysis within discourse analysis. *Discourse & Society*, *3*(2), 193-217.

Falileyev, A. (2003). Languages of Old Wales: A Case for Co-existence. *Dialectologia et Geolinguistica*, *2003*(11), 18-38.

Fanselow, G., & Lenertová, D. (2011). Left peripheral focus: mismatches between syntax and information structure. *Natural Language & Linguistic Theory*, *29*(1), 169-209.

Ferstl, E. C., & von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: an event-related fMRI study. *Cognitive Brain Research*, *11*(3), 325-340.

Féry, C., & Ishihara, S. (2016). *The Oxford Handbook of Information Structure*. Oxford: Oxford University Press.

Féry, C., & Krifka, M. (2008). Information structure. Notional distinctions, ways of expression. In *Unity and diversity of languages* (p. 123-136). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Fife, J. (1986). The semantics of gwneud inversions. *BBCS*, *33*, 133-144.

Fife, J. (1988). *Functional syntax: a case study in Middle Welsh*. Lublin: Redakcja Wydawnictw Katolickiego Uniwersytetu Lubelskiego.

Fife, J. (1991). Some constituent-order frequencies in Classical Welsh Prose. In J. Fife & E. Poppe (Eds.), *Studies in Brythonic word order* (Vol. 83, p. 251-274). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Fife, J. (2010). Typological aspects of the Celtic languages. In *The Celtic Languages* (p. 1-21). London: Routledge.

Fife, J., & King, G. (1991). Focus and the Welsh 'Abnormal Sentence': A cross-linguistic perspective. In J. Fife & E. Poppe (Eds.), *Studies in Brythonic word order* (Vol. 83, p. 81-153). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (Vol. 82, p. 35-60).

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 501-538.

Firbas, J. (1964). On defining the theme in functional sentence analysis. *Travaux linguistiques de Prague*, *1*, 267-280.

Fischer, O. (1992). Syntax. In N. F. Blake, R. Lass, & S. Romaine (Eds.), *The Cambridge history of the English language: Volume II, 1066-1476* (p. 207-408). Cambridge: Cambridge University Press.

Fischer, O., Van Kemenade, A., Koopman, W., & Van der Wurff, W. (2000). *The syntax of early English*. Cambridge: Cambridge University Press.

Fodor, J. A. (1966). How to learn to talk: Some simple ways. In *The genesis of language* (p. 105-122). Cambridge Massachusetts: MIT Press.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, *29*(1), 1–36.

Foley, W. A. (1994). Information Structure. In R. Asher (Ed.), *The Encyclopedia of Language and Linguistics* (p. 1678-1685). New York: Pergamon Press.

Fortson, B. (2010). *Indo-European language and culture: an introduction (Second Edition)*. Cambridge: Cambridge University Press.

Fox, D. (2002). Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry*, *33*(1), 63–96.

Francis, W., & Kučera, H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island, USA.

Frascarelli, M. (2000). *The syntax-phonology interface in focus and topic constructions in italian*. Dordrecht: Kluwer.

Frascarelli, M. (2007). Subjects, topics and the interpretation of referential pro. *Natural Language & Linguistic Theory*, *25*(4), 691-734.

Frascarelli, M., & Hinterhölzl, R. (2007). Types of topics in German and Italian. In *On information structure, meaning, and form* (p. 87-116). Amsterdam: John Benjamins.

Fried, M. (2009). Construction grammar as a tool for diachronic analysis. *Constructions and Frames*, *1*(2), 262-291.

Galves, C., Cyrino, S., Lopes, R., Sandalo, F., & Avelar, J. (2012). *Parameter Theory and Linguistic Change* (Vol. 2). Oxford: Oxford University Press.

Garrett, A. (2012). The historical syntax problem: Reanalysis and directionality. In D. Jonas, J. Whitman, & A. Garrett (Eds.), *Grammatical change: Origins, nature, outcomes* (p. 52-72). Oxford: Oxford University Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gianollo, C., Guardiano, C., & Longobardi, G. (2008). Three fundamental issues in parametric linguistics. In T. Biberauer (Ed.), *The limits of syntactic variation* (pp. 109–142). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*, 407-454.

Givón, T. (1984). *Syntax: A functional-typological Introduction* (Vol. 1). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and language*, *26*(1), 69-83.

Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago, IL: Chicago University Press.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., ... Stoel, R. (2007). Information structure. In S. Dipper, M. Götze, & S. Skopeteas (Eds.), *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS)* (p. 147-187). Potsdam: Universitätsverlag Potsdam.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371-395.

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, *37*(3), 538-546.

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 212-226). London: Routledge.

Green, J. (2005). Reading poetic texts in isaiah. *Leaven*, *13*(2), 60-62.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language.* Cambridge, Massachusetts: MIT Press.

Gregory, M. L., & Michaelis, L. A. (2001). Topicalization and left-dislocation: A functional opposition revisited. *Journal of pragmatics*, *33*(11), 1665-1706.

Grice, H. P. (1989). *Study in the way of words*. : Cambridge, MA: Harvard University Press.

Gries, S. T., & Stefanowitsch, A. (2007). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. Berlin: Walter de Gruyter.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203-225.

Guest, L. C. (1849). *The Mabinogion. From the Llyfr Coch o Hergest, and other ancient Welsh manuscripts, with an English translation and notes*. London: Longmans.

Gundel, J. K. (1974). *The role of topic and comment in linguistic theory*. Unpublished doctoral dissertation, Austin: University of Texas.

Gundel, J. K. (1988). Universals of topic-comment structure. *Studies in syntactic typology*, *17*, 209-239.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274-307.

Gwenogvryn Evans, J. (1898-1910). *Report on Manuscripts in the Welsh Language*. London: Royal Commision on Historical Manuscripts.

Haeberli, E., & Ihsane, T. (2015). Revisiting the loss of verb movement in the history of English. *Natural Language and Linguistic Theory*, 1-46.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441.

Hale, K. (1983). Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, *1*(1), 5–48.

Hale, M. (1998). Diachronic syntax. *Syntax*, *1*(1), 1–18.

Halliday, M. A. (1967). Notes on transitivity and theme in english: Part 2. *Journal of linguistics*, *3*(02), 199-244.

Handford, M. (2010). What can a corpus tell us about specialist genres? In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 255-269). London: Routledge.

Hannay, M., & Mackenzie, J. L. (2002). *Effective writing in English: a sourcebook*. Bussum: Coutinho.

Harbour, D. (2011). Mythomania? methods and morals from 'the myth of language universals'. *Lingua*, *121*(12), 1820-1830.

Harlos, A., Poppe, E., & Widmer, P. (2014). Decoding Middle Welsh clauses or "Avoid Ambiguity". *Indogermanische Forschungen*, *119*(1), 125-148.

Harris, A. C., & Campbell, L. (1995). *Historical syntax in cross-linguistic perspective* (Vol. 74). Cambridge: Cambridge University Press.

Hartmann, K., & Zimmermann, M. (2007). Focus strategies in Chadic–the case of Tangale revisited. *Studia Linguistica*, *61*(2), 95-129.

Haug, D. (2009). Info-structural annotation in the PROIEL corpus. In *Annotating and analysing information structure in historical corpus texts.*

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry*, *21*(2), 223–262.

Hedberg, N., & Sosa, J. M. (2007). The prosody of topic and focus in spontaneous english dialogue. In C. Lee, M. Gordon, & D. Bŭuing (Eds.), *Topic and focus* (p. 101-120). Springer.

Heine, B., & Kuteva, T. (2002). *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

Heller, D. (1999). *The syntax and semantics of specificational pseudoclefts in hebrew*. Unpublished doctoral dissertation, Tel-Aviv University.

Hémon, R. (1975). *A historical morphology and syntax of Breton* (Vol. 3). Dublin institute for advanced studies.

Hinterhölzl, R. (2009). The role of information structure in word order variation and word order change. In R. Hinterhölzl & S. Petrova (Eds.), *Information structure and language change: New approaches to word order variation in germanic* (p. 45-66).

Hirt, B. (2012). *Measuring the Award's impact*. London: The Duke of Edinburgh's International Award Foundation.

Hirt, H. (1913). Fragen des Vokalismus und der Stammbildung im Indogermanischen. *Indogermanische Forschungen*, *32*, 236-247.

Hirt, H. (1921). *Indogermanische Grammatik II: Der indogermanische Vocalismus*. Heidelberg.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Psychology Press.

Holmberg, A. (2005). Is there a litte pro? evidence from Finnish. *Linguistic Inquiry*, *36*(4), 533–564.

Holmberg, A. (2013). Verb second. In T. Kiss & A. Alexiadou (Eds.), *Syntax: an International Handbook of Contemporary Syntactic Research*. Berlin: Walter de Gruyter Verlag.

Holmberg, A., & Roberts, I. (2005). On the role of parameters in Universal Grammar: A reply to Newmeyer. In H. Broekhuis, N. Corver, R. Huybregts, U. Kleinhenz, & J. Koster (Eds.), *Organizing grammar: Linguistic studies in honor of henk van riemsdijk* (p. 538-553). Berlin: Mouton de Gruyter.

Hopper, P. J. (1975). *The syntax of the simple sentence in Proto-Germanic* (Vol. 143). Den Haag: De Gruyter Mouton.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.

Horvath, J. (1981). *Aspects of Hungarian syntax and the theory of grammar*. Unpublished doctoral dissertation, University of California, Los Angeles.

Horvath, J. (2010). "discourse features", syntactic displacement and the status of contrast. *Lingua*, *120*(6), 1346-1369.

Hruska, C., Alter, K., Steinhauer, K., & Steube, A. (2000). Can wrong prosodic information be mistaken by the brain. *Journal of Cognitive Neuroscience*, Supplement 122: E82.

Hunston, S. (2010). How can a corpus be used to explore patterns? In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 152-166). London: Routledge.

Huws, D. (1991). Llyfr Gwyn Rhydderch. *CMCS*, *21*, 1-37.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, Massachusetts: MIT Press.

Jamison, S. W. (1983). *Function and Form in the -áys-formations of the Rig Veda and Atharva Veda*. Göttingen: Vandenhoeck & Ruprecht.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Olso/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, Oslo University.

Johnson, K. (2003). In search of the English middle field.

Johnson, S. (1755). *A dictionary of the English Language*. London, Consortium.
Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N. (1989). Mental models. In *Foundations of cognitive science* (p. 469-499). Cambridge, MA: MIT Press.
Johnson-Laird, P. N. (2013). Mental models and cognitive change. *Journal of Cognitive Psychology*, *25*(2), 131-138.
Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological review*, *99*(3), 418-439.
Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International journal of corpus linguistics*, *15*(1), 106-131.
Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*(2), 113-147.
Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical linguistics*, *1*(1-3), 181-194.
Kay, P., & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. *Language*, *75*, 1-33.
Kayne, R. S. (2000). *Parameters and universal grammar*. Oxford: Oxford University Press.
Keenan, E. L. (2002). Explaining the creation of reflexive pronouns in English. In D. Minkova & R. Stockwell (Eds.), *Studies in the history of the english language: a millennial perspective.* (p. 325-355). Berlin: Mouton De Gruyter.
Kelly, E. P., & Sikora, M. (2011). *Reading the Fadden More Psalter: an introduction.* Dublin: National Museum of Ireland.
Kennedy, G. (1998). *An introduction to corpus linguistics*. London, New York: Longman.
King, G. (1993). *A comprehensive grammar of Modern Welsh*. Oxford: Oxford University Press.
Kintsch, W. (1989). The representation of knowledge and the use of knowledge in discourse comprehension. *Language processing in social context*, 185-209.
Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (p. 209-226). Oxford: Blackwell.
Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363-394.
Kiparsky, P. (1995). Indo-european origins of Germanic syntax. In A. Battye & I. Roberts (Eds.), *Clause structure and language change* (pp. 140–172). Oxford: Oxford University Press.
Kirk, A. (2012). *Word order and information structure in New Testament Greek*. Utrecht: LOT dissertation series.
Koch, J. (1991). On the prehistory of Brittonic Syntax. In J. Fife & E. Poppe (Eds.), *Studies in Brythonic word order* (Vol. 83, p. 1-43). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Koch, J. (1992). "Gallo-brittonic" Vs. "insular Celtic": The Inter-relationships of the Celtic Languages Reconsidered. In *Bretagne et Pays Celtiques: Langues, Histoire, Civilisation.* Rennes: Presses Universitaires Rennes, Saint-Brieuc:Skol.

Komen, E. (2009a). CESAC: Coreference Editor for Syntactically Annotated Corpora. In *7th York-Newcastle-Holland Symposium on the History of English Syntax (SHES7)* (Vol. 8). Nijmegen: CLS Department English Language and Culture.

Komen, E. (2009b). *Corpus Studio manual*. Nijmegen: Radboud University Nijmegen.

Komen, E. (2013). *Finding Focus*. Utrecht: LOT Dissertation Series.

Komen, E., & Los, B. (2012). *The pentaset: Annotating information state primitives.*

Koopman, H. (1984). *The syntax of verbs*. Dordrecht, The Netherlands: Foris Publications.

Koster, J. (2000). Extraposition as parallel construal. *Ms. University of Groningen*.

Krámsky, J. (1972). A Contribution to the Investigation of the Frequency of Occurrence of Nominal and Verbal Elements in English. *Prague Studies in Mathematical Linguistics*, *4*, 35-45.

Kratzer, A. (2009). Making a Pronoun: Fake Indexicals as Windows into the Properties of Pronouns. *Linguistic Inquiry*, *40*(2), 187–237.

Krifka, M. (1999). Additive particles under stress. *Semantics and Linguistic Theory (SALT)*, *IX*, 111-128.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3), 243-276.

Krifka, M., & Musan, R. (2012). Information structure: Overview and linguistic issues. In M. Krifka & R. Musan (Eds.), *The expression of information structure.* Berlin: De Gruyter Mouton.

Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language variation and change*, *1*(03), 199-244.

Kroch, A. (2000). Verb-object order in early Middle English. In S. Pintzuk, S. Tsoulas, & A. Warner (Eds.), *Diachronic syntax: Models and mechanisms* (p. 132-163). Oxford University Press.

Kroch, A., Santorini, B., & Delfs, L. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, `http://www.ling.upenn.edu/hist-corpora`.

Kroch, A., Santorini, B., & Diertani, A. (2010). *The Penn-Helsinki parsed corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, `http://www.ling.upenn.edu/hist-corpora`.

Kroch, A., & Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, `http://www.ling.upenn.edu/hist-corpora`.

Kruijff, G.-J., & Duchier, D. (2003). Information structure in topological dependency grammar. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (p. 219-226).

Kučerová, I., & Neeleman, A. (2012). *Contrasts and positions in information structure*. Cambridge: Cambridge University Press.

Kuno, S. (1987). *Functional Syntax: Anaphora, Discourse, and Empathy*. Chicago: University of Chicago Press.

Kuryłowicz, J. (1927). Les effets du ə en indoiranien. *Prace Filologiczne, 11*, 201-243.

Kuryłowicz, J. (1949). La nature des procès dits «analogiques». *Acta linguistica*, *5*(1), 15-37.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified ii (1994-2005). In M. Gernsbacher & M. Traxler (Eds.), *Handbook of psycholinguistics* (p. 659-724). New York: Elsevier.

Kytö, M. (1991). *Manual to the Diachronic Part of "The Helsinki Corpus of English Texts" Coding Conventions and Lists of Source Texts.* University of Helsinki, Department of English.

Kytö, M., & Rissanen, M. (1992). A language in transition: The Helsinki Corpus of English Texts. *ICAME Journal*, *16*, 7-27.

Labov, W. (2001). *Principles of linguistic change, ii: social factors*. Oxford: Blackwell.

Lambrecht, K. (1994). *Information structure and sentence form: A theory of topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Langacker, R. W. (1988). A usage-based model. *Topics in cognitive linguistics*, *50*, 127-163.

Lash, E. J. F. (2011). *A Synchronic and Diachronic Analysis of Old Irish Copular Clauses*. Unpublished doctoral dissertation, Cambridge University.

Lass, R. (1980). *On explaining language change*. Cambridge: Cambridge University Press.

Lass, R. (1997). *Historical linguistics and language change* (Vol. 81). Cambridge:Cambridge University Press.

Ledgeway, A. (2016). Introduction. In A. Ledgeway & I. Roberts (Eds.), *Cambridge Handbook of Historical Syntax.* Cambridge: Cambridge University Press.

Lee, D. Y. (2010). What corpora are available? In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 107-121). London: Routledge.

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991*, 105-122.

Leech, G. (2004a). *Meaning and the english verb*. London: Pearson Education.

Legate, J. A. (2002). *Warlpiri: theoretical implications*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Lehmann, W. P. (1972). Proto-Germanic syntax. *Toward a grammar of Proto-Germanic*, 239-268.

Lehmann, W. P. (1973). A structural principle of language and its implications. *Language*, *49*, 42–66.

Levinsohn, S. (2009). *Self-instruction materials on narrative discourse analysis.*

SIL-International, available at `www.sil.org/ levinsohns/narr.pdf`.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S. C., & Evans, N. (2010). Time for a sea-change in linguistics: Response to comments on the myth of language universals. *Lingua*, *120*(12), 2733-2758.

Li, C. N., & Thompson, S. A. (1976). Subject and Topic: A New Typology of Language in Subject and Topic. In C. N. Li & S. A. Thompson (Eds.), *Subject and topic* (p. 458-489). New York: Academic Press.

Lidz, J., & Williams, A. (2009). Constructions on holiday. *Cognitive linguistics*, *20*(1), 177-189.

Lightfoot, D. (1979). *Principles of diachronic syntax*. Cambridge: Cambridge University Press.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. Cambridge: Cambridge University Press.

Lightfoot, D. (1999). *The development of language: Acquisition, change, and evolution*. Oxford: Blackwell.

Lightfoot, D. (2002). Myths and the prehistory of grammars. *Journal of Linguistics*, *38*(01), 113-136.

Lightfoot, D. (2006). *How new languages emerge*. Cambridge: Cambridge University Press.

Lightfoot, D., & Westergaard, M. (2007). Language Acquisition and Language Change: Inter-relationships. *Language and Linguistics Compass*, *1*(5), 396-416.

Lions, J. (1996). *Ariane 5: Flight 501 Failure.* Available from `http://sunnyday.mit.edu/accidents/Ariane5accidentreport.html`. Last accessed d.d. 3 September 2014.

Lipták, A. (2011). The structure of the topic field in Hungarian. In P. Beninca & N. Munaro (Eds.), *Mapping the left periphery* (p. 163-200).

Longobardi, G. (2001). How comparative is semantics? a unified parametric theory of bare nouns and proper names. *Natural Language Semantics*, *9*(4), 335–369.

Longobardi, G. (2003). Methods in parametric linguistics and cognitive history. In P. Pica (Ed.), *Linguistic variation yearbook* (Vol. 3, p. 101-138). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Longobardi, G., & Roberts, I. (2010). Universals, diversity and change in the science of language: Reaction to "The Myth of Language Universals and Cognitive Science". *Lingua*, *120*(12), 2699-2703.

López, L. (2009). *A derivational syntax for information structure* (Vol. 23). Oxford: Oxford University Press.

Lozano, C. (2006). Focus and split-intransitivity: the acquisition of word order alternations in non-native Spanish. *Second Language Research*, *22*(2), 145–187.

Lubotsky, A. (1990). La loi de Brugmann et $*H_3e$. *La reconstruction des laryngales*, *CCLIII*, 129–136.

Lubotsky, A. (1997). Review of: Marianne Volkart, Zu Brugmanns Gesetz im Altindischen. *Kratylos*, *42*, 55-59.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Lüdeling, A., & Kytö, M. (2008). Introduction. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook. Volume 1* (p. i-xii). Berlin: Walter de Gruyter.

MacCana, P. (1973). On Celtic Word and the Welsh 'Abnormal' Sentence. *Ériu*, 90-120.

MacCana, P. (1976). Latin influence on British: the pluperfect. In J. J. O'Meara & B. Naumann (Eds.), *Latin Script and Letters, AD 400-900: Festschrift presented to Ludwig Bieler on the Occasion of his 70th Birthday* (p. 194-203). Leiden: Brill.

MacCana, P. (1991). Further notes on constituent order in Welsh. In J. Fife & E. Poppe (Eds.), *Studies in Brythonic word order* (Vol. 83, p. 45-80). Amsterdam, Philadelphia: John Benjamins Publishing Company.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, *26*(4), 657-657.

Manning, H. P. (1995). Fluid intransitivity in Middle Welsh: Gradience, typology and 'unaccusativity'. *Lingua*, *97*(2), 171-194.

Manning, H. P. (1997). The geology of railway embankments: Oxford Welsh and the 'abnormal sentence'. In *Papers from the Panels on Linguistic Ideologies in Contact* (Vol. 33, p. 59-74). Chicago Linguistic Society.

Manning, H. P. (2004). The geology of railway embankments: Celticity, Liberalism, the Oxford Welsh reforms, and the word order (s) of Welsh. *Language & Communication*, *24*(2), 135-163.

Marriott, K., Meyer, B., & Wittenburg, K. B. (1998). A survey of visual language specification and recognition. In *Visual language theory* (p. 5-85). Springer.

Marty, A. (1884). Über subjektlose sätze und das verhältnis der grammatik zur logik und philosophie, iii: Von gewissen unterschieden der sprachlichen ausdrücke und speziell der aussagen, die nicht den durch sie bezeichneten gedanken betreffen ('innere sprachform'und deren wirkungen).". *Vierteljahrsschrift für wissenschaftliche Philosophie*, *8*, 292-340.

Mathesius, V. (1929 [1983]). Functional linguistics. In J. Vachek (Ed.), *Praguiana* (p. 121-42). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Maynard, C., & Leicher, S. (2007). Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In *Corpus Linguistics beyond the Word. Corpus Research from Phrase to Discourse.* (p. 107-115). Amsterdam: Rodopi.

McCloskey, J. (1992). Adjunction, selection and embedded verb second. *Ms University of California at Santa Cruz*.

McEnery, T., & Hardie, A. (2012a). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2012b). *Corpus linguistics: Method, theory and practice*.

Cambridge: Cambridge University Press.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
McEnery, T., & Xiao, Z. (2004a). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In M. Lino, M. Xavier, F. Ferreire, R. Costa, & R. Silva (Eds.), (p. 1175-8).
McFadden, T. (2014). *Corpus research methodology.* Handout from seminar on Corpus research for historical syntax, Utrecht, April 2014.
Meelen, M., & Beekhuizen, B. (2013). PoS-tagging and chunking historical Welsh. In *Proceedings of the scottish celtic colloquium 2012.*
Meillet, A. (1958 [1912]). L'évolution des formes grammaticales. In *Linguistique historique et linguistique générale* (p. 130-58). Paris: Champion.
Meisel, J., Elsig, M., & Rinke, E. (2013). *Language Acquisition and Change: A Morphosyntactic Perspective*. Oxford: Oxford University Press.
Mereu, L. (2009). Introduction. In L. Mereu (Ed.), *Information Structure and its Interfaces* (Vol. 19, p. 1-11). Berlin: Walter de Gruyter.
Meurman-Solin, A., López-Couso, M. J., & Los, B. (2012). On the interplay of syntax and information structure: synchronic and diachronic considerations. In A. Meurman-Solin, M. J. López-Couso, & B. Los (Eds.), *Information structure and syntactic change in the history of English* (p. 3-18). Oxford: Oxford University Press.
Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.
Meyer, C. F. (2008). Pre-electronic corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook. Volume 1* (p. 1-13). Berlin: Walter de Gruyter.
Meyer, C. F., & Tao, H. (2005). Response to Newmeyer's 'grammar is grammar and usage is usage'. *Language*, *81*(1), 226–228.
Milroy, J. (1992). *Linguistic variation and change: on the historical sociolinguistics of English*. Oxford: Blackwell.
Mithun, M. (1987). Is basic word order universal? In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (p. 281-328). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Morris Jones, J. (1913). *A Welsh grammar: historical and comparative*. Oxford: Clarendon Press.
Morris-Jones, J. (1931). *Welsh syntax: an unfinished draft*. Cardiff: the University of Wales Press.
Motut, A. (2010). Merge over Move and the empirical force of economy in Minimalism. *Toronto Working Papers in Linguistics*, *33*, 1-54.
Neeleman, A., & Van de Koot, H. (2008). Dutch scrambling and the nature of discourse templates. *Journal of Comparative Germanic Linguistics*, *11*(2), 137-189.
Nekula, M. (1999). Vilém mathesius. In J.-O. Verschueren, J. Östman, J. Blommaert, & C. Bulcaen (Eds.), *Handbook of Pragmatics* (p. 1-14). Amster-

dam/Philadelphia: John Benjamins Publishing Company.
Nelson, M. (2010). Building a written corpus: what are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 53-65). London: Routledge.
Nevalainen, T., & Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.
Newmeyer, F. J. (2004). Against a parameter-setting approach to typological variation. In P. Pica, J. Rooryck, & J. Van Craenenbroek (Eds.), *Linguistic variation yearbook* (Vol. 4, p. 181-234). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Newmeyer, F. J. (2005). A reply to the critiques of 'grammar is grammar and usage is usage'. *Language*, *81*(1), 229–236.
Newton, G. (2006). *The development and loss of the Irish double system of inflection*. Unpublished doctoral dissertation, Cambridge University.
Nivre, J. (2008). Treebanks. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook. Volume 1* (Vol. 1, p. 225-241). Berlin: Walter de Gruyter.
Niyogi, P., & Berwick, R. C. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, *106*(25), 10124-10129.
Nolda, A. (2004). Topics detached to the left. On 'left dislocation', 'hanging topic' and related constructions in German. *ZAS Papers in Linguistics*, *35*, 423-448.
Norde, M. (2009). *Degrammaticalization*. Oxford: Oxford University Press.
Nurmio, S. (2015). *Studies in grammatical number in Old and Middle Welsh*. Unpublished doctoral dissertation, St John's College, University of Cambridge.
Nurmio, S., & Willis, D. (2016). The rise and fall of a minor number: The case of the Welsh numerative. *Unpublished Ms.*.
O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: a memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1200-1210.
O'Keeffe, A., & McCarthy, M. (2010). Historical perspective: what are corpora an how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 3-13). London: Routledge.
O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Onea, E., & Beaver, D. (2011). Hungarian focus is not exhausted. *Semantics and Linguistic Theory*, *19*, 342-359.
Osgood, C. E., & Sebeok, T. A. (1954). *Psycholinguistics: a survey of theory and research problems* (Vol. 49). American Psychological Association.
Osthoff, H., & Brugmann, K. (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen* (Vol. 1). Leipzig: S. Hirzel.
Östman, J.-O., & Virtanen, T. (1999). Theme, comment, and newness as figures in information structuring. *Amsterdam studies in the theory and history of linguistic science series*, *4*, 91-110.

Ouhalla, J. (1994). Verb movement and word order in Arabic. In D. Lightfoot & N. Hornstein (Eds.), *Verb movement* (pp. 41–72). Cambridge, Great Britain: Cambridge University Press.

OUP. (2014). *History of the OED*. Available online from `http://public.oed.com/history-of-the-oed` last accessed d.d. 3 September 2014.

Parker, W. (2007). *The Four Branches of the Mabinogi*. Dublin: Bardic Press.

Paul, H. (1920 [1880]). *Prinzipien der Sprachgeschichte*. Tübingen: Niemeyer.

Payne, D. (1987). Information structuring in Papago narrative discourse. *Language*, *63*, 783-804.

Pearl, L. (2014). Evaluating learning-strategy components: Being fair (Commentary on Ambridge, Pine, and Lieven). *Language*, *90*(3), e107-e114.

Pedersen, H. (1913). *Vergleichende Grammatik der keltischen Sprachen* (Vol. 2). Göttingen: Vandenhoeck und Ruprecht.

Pinker, S. (1984). Visual cognition: An introduction. *Cognition*, *18*(1), 1-63.

Pintzuk, S. (1991). *Phrase structures in competition: variation and change in Old English word order*. Unpublished doctoral dissertation, University of Pennsylvania.

Pintzuk, S. (2002). Verb-object order in Old English: variation as grammatical competition. In D. Lightfoot (Ed.), *Syntactic effects of morphological change* (p. 276-299). Oxford: Oxford University Press.

Pintzuk, S., & Plug, L. (2002). *The York-Helsinki parsed corpus of Old English poetry*. Department of Linguistics, University of York. Oxford Text Archive, first edition, `http://www-users.york.ac.uk`.

Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, 59-72.

Plein, K., & Poppe, E. (2014). Patterns of verbal agreement in "Historia Gruffud vab Kenan": norm and variation. *Études celtiques*(40), 145-164.

Pollard, C., & Sag, I. (1987). *Information-based Syntax and Semantics* (Vol. 1). Stanford: CSLI publications.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Poppe, E. (1989). Constituent ordering in 'Breudwyt Maxen Wledic'. *BBCS*, *36*, 43-63.

Poppe, E. (1990). Word-order patterns in Breudwyt Ronabwy. In M. Ball, J. Fife, E. Poppe, & J. Rowland (Eds.), *Celtic Linguistics. Ieithyddiaeth Geltaidd. Readings in the Brythonic Languages. Festschrift for TA Watkins* (p. 445-460). Amsterdam: John Benjamins.

Poppe, E. (1991). Word order in Cyfranc Lludd a Llefelys: note on the pragmatics of constituent-ordering in MW narrative prose. In J. Fife & E. Poppe (Eds.), *Studies in Brythonic word order* (Vol. 83, p. 155-205). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Poppe, E. (1993). Word order in Middle Welsh: the case of Kedymdeithyas Amlyn ac Amic. *Bwletin y Bwrdd Gwybodau Celtaidd*, *40*, 95-117.

Poppe, E. (2009). The pragmatics of Middle Welsh word order: Some conceptual and descriptive problems. In *Pragmatische Kategorien. Form, Funktion und Diachronie.* (Vol. 24, p. 247-264).

Poppe, E. (2014). How to Achieve an Optimal Textual Fit in Middle Welsh Clauses. *Cambrian Medieval Celtic Studies*, *68*, 69-100.

Popper, K. (1935). *Logik der Forschung: zur Erkenntnistheorie der modernen Naturwissenschaft*. Vienna: Springer.

Popper, K. (1968). *The Logic of Scientific Discovery*. New York: Harper Torch Books.

Preminger, O. (2011). *Agreement as a fallible operation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (p. 223-255). Academic Press, New York.

Ramble, C. (2013). Both Fish and Fowl? Preliminary Reflections on Some Representations of a Tibetan Mirror-World. In F.-K. Ehrhard & P. Maurer (Eds.), *Nepalica-Tibetica: Festgabe for Christoph Cüppers* (Vol. 2, p. 75-89). International Institute for Tibetan and Buddhist Studies GmbH.

Randall, B., Taylor, A., & Kroch, A. (2005). *Corpussearch 2.* Available at: `http://corpussearch.sourceforge.net/credits.html`.

Reinhart, T. (1981). Definite NP anaphora and C-command domains. *Linguistic Inquiry*, *12*(4), 605–635.

Reinhart, T. (1982). Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica*, *27*(1), 53-94.

Repp, S. (2010). Defining 'contrast' as an information-structural notion in grammar. *Lingua*, *120*(6), 1333-1345.

Rhys, J., & Jones, B. (1902). *The Welsh People: Chapters on Their Origin, History, Laws, Language, Literature, and Characteristics*. T. Fisher Unwin.

Richards, M. (1938). *Cystrawen y frawddeg Gymraeg*. Caerdydd: Gwasg Prifysgol Cymru.

Riester, A., Lorenz, D., & Seemann, N. (2010). A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the seventh international conference of language resources and evaluation (LREC), Valletta, Malta* (p. 717-722).

Rissanen, M. (1989). Three problems connected with the use of diachronic corpora. *ICAME journal*, *13*, 16-19.

Rissanen, M. (1998). Towards an integrated view of the development of English: Notes on causal linking. *Trends in linguistics studies and monographs*, *112*, 389-406.

Rissanen, M. (2008). Corpus linguistics and historical linguistics. In *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter.

Ritz, J., Dipper, S., & Götze, M. (2008). Annotation of Information Structure: an Evaluation across different Types of Texts. In *Proceedings of the 6th LREC-2008 Conference.* Marrakech, Morocco.

Rizzi, L. (1997). The fine structure of the left periphery. In L. Haegeman (Ed.), *Elements of grammar* (pp. 281–337). Dordrecht: Kluwer Academic Publishers.

Rizzi, L. (2004). On the cartography of Syntactic Structures. In L. Rizzi (Ed.), *The structure of CP and IP* (pp. 3–15). Oxford: Oxford University Press.

Roberts, I. (1997). Restructuring, head movement, and locality. *Linguistic Inquiry*, *28*(1), 423–460.

Roberts, I. (2004). The C-system in Brythonic Celtic languages, V2, and the EPP. In L. Rizzi (Ed.), *The structure of CP and IP: the Cartography of syntactic Structures* (Vol. 2, pp. 297–328).

Roberts, I. (2005). *Principles and parameters in a VSO language: A case study in Welsh*. Oxford: Oxford University Press.

Roberts, I. (2007). *Diachronic syntax*. Oxford: Oxford University Press.

Roberts, I. (2009). A deletion analysis of null subjects. In T. Biberauer, A. Holmberg, I. Roberts, & M. Sheehan (Eds.), *Parametric variation: Null subjects in minimalist theory* (p. 58-87). Cambridge: Cambridge University Press.

Roberts, I. (2010). *Agreement and head movement: Clitics, incorporation, and defective goals*. MIT Press.

Roberts, I. (2012). Macroparameters and minimalism. In C. Galves, S. Cyrino, R. Lopes, F. Sandalo, & J. Avelar (Eds.), *Parameter Theory and Linguistic Change* (Vol. 2, p. 320-335). Oxford: Oxford University Press.

Roberts, I., & Holmberg, A. (2005). On the role of parameters in Universal Grammar: A reply to Newmeyer. In H. Broekhuis, N. Corver, M. Everaert, & J. Koster (Eds.), *Organizing grammar: Linguistic studies in honor of Henk van Riemsdijk* (p. 538-553). Berlin: Mouton de Gruyter.

Roberts, I., & Roussou, A. (2003). *Syntactic change: A minimalist approach to grammaticalization* (Vol. 100). Cambridge: Cambridge University Press.

Rodway, S. (2004). The Red Book Text of "Culhwch ac Olwen": A Modernising Scribe at Work. *Studi Celtici*, 93-161.

Rodway, S. (2013). *Dating Medieval Welsh Literature: Evidence from the verbal system*. CMCS.

Rooth, M. E. (1985). *Association with focus*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Rosenkvist, H. (2010). A case of degrammaticalization in northern Swedish. In A. Breitbarth, C. Lucas, S. Watts, & D. Willis (Eds.), *Continuity and change in grammar* (Vol. 159, p. 303-320). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Ross, J. R. (1986). *Infinite syntax*. New Jersey: Ablex Norwood.

Rouveret, A. (1994). *Syntaxe du gallois: principes généraux et typologie*. CNRS.

Rowland, T. (1876). *A grammar of the Welsh language*. Wrexham: Hughes & son.

Rühlemann, C. (2010). What can a corpus tell us about pragmatics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 288-301). London: Routledge.

Russell, P. (1999). What did medieval Welsh scribes do? The scribe of the 'Dingestow Court Manuscript'. *Cambrian Medieval Celtic Studies*(37), 79-96.

Russell, P. (2003). Texts in Contexts: Recent Work on the Medieval Welsh Prose Tales. *CMCS*, 59-72.

Russell, P. (2012). An habes linguam Latinam? Non tam bene sapio: Views of Multilingualism from the Early Medieval West. In A. Mullen & P. James (Eds.), *Multilingualism in the graeco-roman worlds* (p. 193-224). Cambridge: Cambridge University Press.

Sampson, G. R. (2007). Grammar without grammaticality. *Corpus linguistics and linguistic theory*, *3*(1), 1-32.

Sasse, H.-J. (1987). The thetic/categorical distinction revisited. *Linguistics*, *25*(3), 511–580.

Schmalhofer, F., Friese, U., Pietruska, K., Raabe, M., & Rutschmann, R. (2005). Brain processes of relating a statement to a previously read text: Memory resonance and situational constructions. In *Proceedings of the XVII Conference of The Cognitive Science Society* (p. 1949-1954).

Schmidt, K. H. (1990). Gallo-Brittonic or Insular Celtic. In *Studia indogermanica et palaeohispanica in honorem A. Tovar et L. Michelena* (p. 255-267). Universidad del País Vasco/Universidad de Salamanca.

Schrijver, P. (1995). *Studies in British Celtic historical phonology* (Vol. 5). Amsterdam: Rodopi.

Schrijver, P. (1997). *Studies in the History of Celtic Pronouns and Particles* (Vol. 2). Department of Old Irish, National University of Ireland.

Schrijver, P. (2002). The rise and fall of British Latin: Evidence from English and Brittonic. In M. Filppula, J. Klemola, & H. Pitkänen (Eds.), *The Celtic Roots of English* (p. 87-110). University of Joensuu.

Schrijver, P. (2007). What Britons spoke around 400 AD. In N. Higham (Ed.), *Britons in Anglo-Saxon England* (p. 165-71). Woodbridge.

Schrijver, P. (2014). *Language contact and the origins of the Germanic languages* (Vol. 13). London: Routledge.

Schumacher, S. (2011). Mittel- und Frühneukymrisch. In E. Ternes (Ed.), *Brythonic Celtic - Britannisches Keltisch: from Medieval British to Modern Breton* (Vol. 11, p. 85-236). Munich Studies in Historical Linguistics.

Schütze, C. T., Sprouse, J., & Caponigro, I. (2015). Challenges for a theory of islands: A broader perspective on Ambridge, Pine, and Lieven. *Language*, *91*(2), e31-e39.

Schwarzschild, R. (1999). Givenness, avoid F and other constraints on the placement of accent. *Natural Language Semantics*, *7*(2), 141–177.

Scott, M. (2010). What can corpus software do? In A. O'ÂĂÂŹKeeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (p. 136-151). London: Routledge.

Sheehan, M. L. (2007). *The EPP and null subjects in Romance*. Unpublished doctoral dissertation, University of Newcastle upon Tyne.

Sheehan, M. L. (2015). On the lack of consistency in (Romance) consistent null subject languages. *Paper presented at Cambridge Comparative Syntax (CamCos) 4, May 2015*.

Silk, J. (2014). Keeping Up With the Joneses: From William Jones to John James Jones. *Annual Report of The International Research Institute for Advanced*

Buddhology at Soka University(17), 427-441.
Sims-Williams, P. (2003). *The Celtic Inscriptions of Britain: Phonology and Chronology, c. 400-1200*. Publications of the Philological Society, Blackwell Publishing.
Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
Speyer, A. (2008). Doppelte Vorfeldbesetzung im heutigen Deutsch und im Frühneuhochdeutschen. *Linguistische Berichte*, *2008*(216), 455-485.
Sproat, R. (1983). Vso languages and welsh configurationality. In *Proceedings of the harvard celtic colloquium* (Vol. 3, pp. 39–68).
Sproat, R. (1985). Welsh syntax and VSO structure. *Natural Language and Linguistic Theory*, *3*(2), 173–216.
Sprouse, J., & Lau, E. F. (2013). Syntax and the brain. In *The Cambridge Handbook of Generative Syntax* (p. 971-1005). Cambridge: Cambridge University Press.
Stalnaker, R. (1974). Pragmatic presuppositions. In P. K. Unger & M. K. Munitz (Eds.), *Semantics and philosophy* (p. 197-214). New York: New York University Press.
Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5), 701-721.
Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, *29*(12), 1213-1228.
Stifter, D. (2011). The textual arrangement of Alise-Sainte-Reine [L-13]. *Zeitschrift fur Celtische Philologie*, *58*, 165-181.
Strachan, J. (1909). *An introduction to early Welsh*. Manchester: Manchester University Press.
Stump, G. T. (1989). Further remarks on Breton agreement. *Natural Language and Linguistic Theory*, *7*(3), 429–471.
Sturzer, N. (2001). How Middle Welsh expresses the unexpected. *CMCS*(41), 37-53.
Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (p. 150-164). Longman Harlow, Essex.
Szendrői, K. (2001). *Focus and the syntax-phonology interface*. Unpublished doctoral dissertation, University College London.
Takahashi, N., Miner, L. L., Sora, I., Ujike, H., Revay, R. S., Kostic, V., ... Uhl, G. R. (1997). VMAT2 knockout mice: heterozygotes display reduced amphetamine-conditioned reward, enhanced amphetamine locomotion, and enhanced MPTP toxicity. *Proceedings of the National Academy of Sciences*, *94*(18), 9938–9943.
Tallerman, M. (1996). Fronting constructions in Welsh. In R. Borsley (Ed.), *The syntax of the Celtic languages: a comparative perspective* (p. 97). Cambridge: Cambridge University Press.
Tallerman, M. (1998). Celtic word order: Some theoretical issues. In *Constituent order in the languages of Europe* (Vol. 20, p. 599-648). Berlin, New York: Mouton de Gruyter.
Tallerman, M. (2011). *Understanding syntax: third edition*. London: Hodder

Education.

Tallerman, M., & Wallenberg, J. (2012). *The Middle Welsh historic infinitive.* University of Wales, July 2012.

Taylor, A., & Pintzuk, S. (2014). Testing the theory. In *Information Structure and Syntactic Change in Germanic and Romance Languages* (Vol. 213, p. 53-77). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Taylor, A., & Pintzuk, S. (2015). Verb order, object position, and information status in Old English. In (Vol. 15, p. 318-335). Oxford: Oxford University Press.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). *The York-Toronto-Helsinki parsed corpus of Old English prose (YCOE)*. Department of Linguistics, University of York. Oxford Text Archive, first edition, `http://www-users.york.ac.uk`.

Teleman, U. (1974). *Manual för grammatisk beskrivning av skriven och talad svenska.* Lund: Studentlitteratur.

Thomason, S. G., & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. University of California Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words*. New York: Columbia University Press.

Thorne, D. (1993). *A comprehensive Welsh grammar*. Oxford: Blackwell.

Thurneysen, R. (2003 [1946]). *A grammar of Old Irish*. Dublin Institute for Advanced Studies.

Traugott, E. C., & König, E. (1991). The semantics-pragmatics of grammaticalization revisited. *Approaches to grammaticalization*, *1*, 189-218.

Traugott, E. C., & Pintzuk, S. (2008). Coding the York-Toronto-Helsinki Parsed Corpus of Old English Prose to investigate the syntax-pragmatics interface. In S. M. Fitzmaurice & D. Minkova (Eds.), *Studies in the History of the English Language IV: Empirical and Analytical Advances in the study of English language change* (p. 61-80).

Travis, L. (1984). *Parameters and effects of word order variation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Vallduví, E. (1992). *The information component*. New York: Garland.

Vallduví, E., & Vilkuna, M. (1998). On rheme and kontrast. *Syntax and semantics*, 79-108.

Van Berkum, J. J., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158-171.

Van der Wal, J. (2009). *Word order and information structure in Makhuwa-Enahara*. Utrecht: LOT dissertation series.

Van der Wal, J. (2015). What you see is (not) what you get: information structure on the interface between syntax and discourse. *Lecture at MFil, 6-7 November 2015, University of Manchester*.

Van der Wurff, W., & Foster, T. (1997). Object-verb word order in 16th century English: A study of its frequency and status. In *Language history and language modelling: a festschrift for Jasek Fisiak on his 60th birthday* (Vol. 101, p. 439-

453). Berlin: Mouton De Gruyter.
Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
Van Gelderen, E. (2004). *Grammaticalization as economy* (Vol. 71). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Van Gelderen, E. (2009). Feature economy in the linguistic cycle. In P. Crisma & P. Longobardi (Eds.), *Historical syntax and linguistic theory* (p. 93-109).
Van Gelderen, E. (2011). *The linguistic cycle: Language change and the language faculty*. Oxford: Oxford University Press.
Van Kemenade, A. (2007). Formal syntax and language change Developments and outlook. *Diachronica*, *24*(1), 155-169.
van Koppen, M. (2007). Agreement with (the internal structure of) copies of movement. *The copy theory of movement*, *107*, 327.
Van Valin, R. (1993b). A synopsis of Role and Reference Grammar. In R. D. Van Valin (Ed.), *Advances in Role and Reference Grammar* (p. 1-164). Amsterdam, Philadelphia: John Benjamins Publishing Company.
Vendryes, J. (1912). La place du verbe en celtique. *Mémoires de la Société de Linguistique de Paris*, *17*, 337-51.
Vennemann, T. (1974). *Topics, Subjects, and Word Order: from SXV to SVX via TVX* (Vol. 2). Amsterdam: North-Holland Publishers.
Vincent, N. (1988). Latin. In M. Harris & N. Vincent (Eds.), *The romance languages* (p. 26-78). Londen: Routledge.
Vitt, A. (2011). *Peredur vab Efrawc: Edited Texts and Translations of the MSS Peniarth 7 and 14 Versions*. Unpublished doctoral dissertation, Aberystwyth University.
Volkart, M. (1994). *Zu Brugmanns Gesetz im Altindischen*. Universität Bern. Institut für Sprachwissenschaft. Arbeitspapier 33.
Von Fintel, K. (2000). What Is Presupposition Accommodation? *Unpublishe Manuscript, MIT*.
Vriezen, T. C., & Van der Woude, A. (2000). *Oudisraëlitische en vroegjoodse literatuur*. Kampen: Uitgeverij Kok.
Wade-Evans, A. (1909). *Welsh Medieval Law: Being a Text of the Laws of Howel the Good, Namely the British Museum Harleian Ms. 4353 of the 13th Century, with Translation, Introduction, Appendix, Glossary, Index, and a Map*. Oxford: Clarendon Press.
Wagner, H. (1959). *Das Verbum in den Sprachen der Britischen Inseln*. Tübingen: Niemeyer.
Walkden, G. (2009). *The comparative method in syntactic reconstruction.* MPhil dissertation: University of Cambridge.
Walkden, G. (2011). Abduction or Inertia? The logic of syntactic change. In C. Cummins, C.-H. Elder, T. Godard, M. Macleod, E. Schmidt, & G. Walkden (Eds.), *Proceedings of the Sixth Cambridge Postgraduate Conference in Language Research* (p. 230-239). Cambridge Institute of Language Research.
Walkden, G. (2012). Against inertia. *Lingua*, *122*(8), 891-901.
Walkden, G. (2014). *Syntactic Reconstruction and Proto-Germanic*. Oxford: Oxford

University Press.

Wallenberg, J., Ingason, A. K., Sigurdsson, E. F., & Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)*. `www.linguist.is/icelandic_treebank`.

Wallis, S. (2008). Searching treebanks and other structured corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook* (Vol. 1, p. 738-758). Berlin: Walter de Gruyter.

Wasow, T. (2002). *Postverbal behaviour*. Stanford, California: CSLI Publications.

Watkins, C. (1976). Towards Proto-Indo-European syntax: problems and pseudo-problems. In S. Steever, C. Walker, & S. Mufwene (Eds.), *Papers from the parasession on diachronic syntax* (p. 306-326). Chicago Linguistic Society.

Watkins, C. (1999). Two Celtic notes. In P. Anreiter & E. Jerem (Eds.), *Studia celtica et indogermanica: Festschrift für wolfgang meid zum 70. geburtstag* (p. 539-543). Archaeolingua Alapítvány.

Watkins, T. A. (1977). Trefn yn y frawddeg Gymraeg. *Studia Celtica*, *12/13*, 367-395.

Watkins, T. A. (1987). Constituent order in the Old Welsh verbal sentence. *Bulletin of the board of Celtic Studies*, *34*, 51-60.

Watkins, T. A. (1988). *Constituent order in the positive declarative sentence in the medieval Welsh tale 'Kulhwch ac Olwen'* (Vol. 41). Institut für Sprachwissenschaft der Universität Innsbruck.

Watkins, T. A. (1993). Constituent order in main/simple verb clauses of Pwyll Pendeuic Dyuet. *Language Sciences*, *15*(2), 115-139.

Watkins, T. A. (1997). The sef[...] Realization of the Welsh Identificatory Copular Sentence. In A. Ahlqvist, C. R. Ó. Cléirigh, & V. Čapková (Eds.), *Dán do oide: essays in memory of Conn R. Ó Cléirigh*. Linguistics Institute of Ireland.

Watson, W. G. (1973). *The language and poetry of the book of Isaiah in the light of recent research in Northwest Semitic*. Unpublished doctoral dissertation, University of Aberdeen.

Weber-Wulff, D. (1992). *Rounding error changes Parliament makeup* (Vol. 13). Available from `http://catless.ncl.ac.uk/Risks`. Last accessed d.d. 3 September 2014.

Weinreich, U., Labov, W., & Herzog, M. (1968). *Empirical foundations for a theory of language change*. Austin: University of Texas Press.

West, M. (1953). *A general service list of english words*. London: Longman.

Westergaard, M. (2009). *The acquisition of word order: micro-cues, information structure, and economy* (Vol. 145). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied linguistics*, *21*(1), 3-25.

Williams, S. (1980). *A Welsh grammar*. Cardiff: the University of Wales Press.

Williams ab Ithel, J. (1856). *Dosparth edeyrn davod aur, or The ancient Welsh grammar*. London.

Willis, D. (1998). *Syntactic Change in Welsh: A Study of the Loss of the Verb-second*.

Oxford: Oxford University Press.
Willis, D. (2006). Negation in Middle Welsh. *Studia Celtica*, *40*(1), 63-88.
Willis, D. (2007a). Specifier-to-head reanalyses in the complementizer domain: evidence from Welsh. *Transactions of the Philological Society*, *105*(3), 432-480.
Willis, D. (2007b). Syntactic lexicalization as a new type of degrammaticalization. *Linguistics*, *45*(2), 271-310.
Willis, D. (2011a). Reconstructing last week's weather: Syntactic reconstruction and Brythonic free relatives. *Journal of Linguistics*, *47*(02), 407-446.
Willis, D. (2011b). The limits of resumption in Welsh wh-dependencies. In A. Rouveret (Ed.), *Resumptive pronouns at the interfaces* (pp. 189–222). Amsterdam: John Benjamins.
Willis, D. (2014). Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun *chdi*. *Unpublished ms. University of Cambridge*.
Willis, D. (2015). Two predicate-phrase heads in Welsh copular clauses. *Paper presented at the Workshop on Copulas across Languages, Greenwich*.
Willis, D. (2016). Endogenous and exogenous theories of syntactic change. In A. Ledgeway & I. Roberts (Eds.), *Cambridge Handbook of Historical Syntax*. Cambridge: Cambridge University Press.
Wiltschko, M. (2014). *The universal structure of categories: Towards a formal typology* (Vol. 142). Cambridge University Press.
Winford, D. (2005). Contact-induced changes: Classification and processes. *Diachronica*, *22*(2), 373–427.
Wolfe, S. (2015). *Syntactic Microvariation in the Medieval Romance Languages and the Nature of V2 Reconsidered.* SyntaxLab Talk, Tuesday 9th June 2015, University of Cambridge.
Wright, S. (1993). In Search of History: English Language In the Eighteenth Century. In *English Language Corpora: Design, analysis and exploitation* (p. 25-39). Amsterdam: Rodopi.
Yang, C. D. (2000). Internal and external forces in language change. *Language variation and change*, *12*(03), 231-250.
Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.
Yang, C. L., Perfetti, C. A., & Schmalhofer, F. (2007). Event-related potential indicators of text integration across sentence boundaries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 55-89.
Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 217-235.
Young, R. W., & Morgan, W. (1980). *The Navajo Language: A Grammar and Colloquial Dictionary*. Albuquerque: University of New Mexico Press.
Yule, G. (1981). New, current and displaced entity reference. *Lingua*, *55*(1), 41-52.
Zaring, L. (1996). "Two BE or not two BE": Identity, Predication and the Welsh Copula. *Linguistics and Philosophy*, *19*(2), 103-142.

Zavrel, J., & Daelemans, W. (1997). Memory-based learning: Using similarity for smoothing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 436-443.

Zimmermann, M., & Féry, C. (2010). *Information structure: theoretical, typological, and experimental perspectives*. Oxford: Oxford University Press.

Zinsmeister, H., Hinrichs, E., Kübler, S., & Witt, A. (2008). Linguistically annotated corpora: Quality assurance, reusability and sustainability. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International handbook. Volume 1* (p. 759-776). Berlin: Walter de Gruyter.

Zipf, G. K. (1935). *The psycho-biology of language*. Boston, Houghton-Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Oxford: Addison-Wesley Press.

Zubizarreta, M. L. (1998). *Prosody, focus, and word order*. Cambridge, Massachusetts: MIT Press.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162-185.

Samenvatting in het Nederlands

Dit proefschrift gaat over de woordvolgorde in de Middelwelshe zin. De meest voorkomende woordvolgorde is de zogenaamde 'V2' volgorde, waarbij het werkwoord net als in het Nederlands op de tweede plaats in de zin staat. Deze volgorde wijkt af van de normale woordvolgorde van de zin in de moderne Welshe taal waarin het werkwoord voorop staat. Een tweede opvallende observatie in de Middelwelshe syntaxis betreft de grote variatie in woordvolgorde. Het werkwoord op de tweede plek kan diverse constituenten volgen en de congruentie tussen onderwerp en persoonsvorm neemt vaak ongebruikelijke vormen aan. Tot slot is het onduidelijk waar deze zinnen met V2 volgorde vandaan komen. De beperkte data uit eerdere stadia van de Welshe taal lijkt te wijzen op een voorkeur voor zinnen met het werkwoord voorop (VSO). Ik probeer in deze studie daarom de volgende vragen te beantwoorden:

1. Hoe kan de distributie van de verschillende patronen in woordvolgorde in Middelwelsh verklaard worden?

2. Waar komen de verscheidene V2 volgordes vandaan (inclusief de patronen met en zonder congruentie)?

Om deze vragen te kunnen beantwoorden is het om te beginnen noodzakelijk om een digitale database te creëren zodat alle teksten doorzoekbaar zijn en daarnaast voorzien van morfo-syntactische annotatie. Daarnaast is het belangrijk om een consistente methodologie te hanteren voor de analyse van de informatiestructuur en andere factoren die de volgorde van de constituenten in de zin kunnen bepalen.

In hoofdstuk 2 presenteer ik de argumenten voor een geannoteerd corpus. In historisch taalkundig onderzoek, met name onderzoek naar syntaxis, kunnen we enkel kijken naar de distributie van verschillende constructies die we vinden in manuscripten. Hoe meer manuscripten en teksten we kunnen digitaliseren en systematisch onderzoeken, hoe meer informatie we krijgen over de taalsituatie in die tijd. Als een bepaalde constructie in slechts een enkele tekst voorkomt, kun-

nen we niet direct concluderen dat deze constructie ook daadwerkelijk onderdeel was van de gangbare gesproken of geschreven taal. Omdat de hoeveelheid data die is overgeleverd vaak erg beperkt is, is het belangrijk om optimaal gebruik te maken van de data die ter beschikking is. Dit kan onder andere door de toevoeging van grammaticale informatie over de woordsoort en de daarbijbehorende vervoeging en verbuiging in zogenaamde 'Part-of-Speech' (PoS) tags. Deze gedetailleerde morfosyntactische informatie faciliteert de automatische extractie van de nodige taalkundige informatie. Idealiter voegen we alle beschikbare teksten en manuscripten toe aan een dergelijk geannoteerd corpus, maar dit kost heel veel tijd en moeite. Voor het onderzoek in dit proefschrift heb ik daarom de eerste stappen gezet om een dergelijke geannoteerde database te creëren voor het Middelwelsh: 15 teksten werden hiervoor zorgvuldig geselecteerd, klaargemaakt voor automatische annotatie, getagd en gecorrigeerd en uiteindelijk ook voorzien van de meest basale syntactische informatie.

Ik heb een PoS-tagger getraind die automatisch morfosyntactische labels toekent aan woorden in Middelwelshe teksten. Met een Global Accuracy van meer dan 90% presteerde de tagger redelijk goed gezien de complexe data en de zeer gedetailleerde tagset (bestaande uit meer dan 200 tags). De tijd die nodig was voor de handmatige correctie na de automatische toekenning van de tags werd hierdoor aanzienlijk beperkt. Naast een PoS-tagger heb ik ook een basale syntactische grammatica ontworpen voor het Middle Welsh met behulp van de NLTK parser. Na de handmatige correctie werden de bestanden geconverteerd naar diverse vormen die systematisch doorzocht kunnen worden met CorpusSearch of XQuery. Het belangrijkste resultaat in hoofdstuk 2 is een geannoteerde database van 15 teksten waarmee diverse taalkundige bijzonderheden systematisch onderzocht kunnen worden.

In hoofdstuk 3 werk ik een methodologie uit voor onderzoek naar informatiestructuur in historische corpora. Hierbij stonden drie onderwerpen centraal: Referentialiteit (Oude of Nieuwe Informatie), Topic (vs. Comment) en Focus (vs. Achtergrondinformatie). Ik heb de kenmerken van elk van deze onderwerpen systematisch omschreven, zodat ze gebruikt kunnen worden om historische corpora te annoteren. Er zijn daarnaast nog twee andere factoren die een rol spelen in de informatiestructuur van de zin: zogenaamde 'Points of Departure', het uitgangspunt van de zin, en de Information Flow, de manier waarop oude en nieuwe informatie elkaar volgen. De duidelijk uitgewerkte en omschreven definities en algoritmes faciliteren de annotatie van grote corpora. Een dergelijke consequente analyse is onmisbaar in elk historische syntactisch onderzoek.

In hoofdstuk 4 presenteer ik de data en de belangrijkste observaties betreffende de variatie in woordvolgorde in Middelwelsh. In de genannoteerde database heb ik diverse patronen gevonden in positieve hoofdzinnen. Ik heb deze onderverdeeld in negen hoofdtypen op basis van hun formele structuur. De meest voorkomende woordvolgorde aan het begin van de Vroegmiddelwelshe periode is nog steeds de

'Abnormal order' met het werkwoord op de tweede plaats van de zin, maar het overgrote deel van de zinnen in de bijbelvertaling van 1588 begint nu met het onderwerp. Zinnen met het werkwoord op de eerste plaats en met name zinnen met een hulpwerkwoord nemen in frequentie toe. De syntactische structuur van de bijbelvertaling wijkt dus af van zowel Middel- als Modernwelsh.

In Hoofdstuk 5 onderzoek ik systematisch all factoren die mogelijk invloed kunnen hebben op de woordvolgorde van de Middelwelshe zin. Beginnende met de grammaticale factoren zien we dat de 'Abnormal order' met een perifrastische constructie en een verbaalnomen vrijwel alleen maar voorkomt in de verleden tijd. Dit is waarschijnlijk gerelateerd aan het feit dit type woordvolgorde met name voorkomt in narratieve context. In direct taalgebruik, zoals in dialogen, staat meestal het onderwerp vooraan. Een grondige studie van het corpus laat verder zien dat werkwoorden met onpersoonlijke/passieve vervoeging vooral voorkomen in V2-zinnen met initiële adjuncten. Tot slot is er een beperkte rol voor 'Animacy' van het onderwerp en lijdend voorwerp. Als het lijdend voorwerp 'inanimate' is, staat het vaker in het begin van de zin dan verwacht.

Alleen als alle interne en externe factoren systematisch zijn onderzocht, kunnen we bepalen of andere factoren, zoals informatiestructuur werkelijk een rol spelen. De eerste factor betreffende de informatiestructuur die ik heb bekeken is 'referentialiteit' - de informatiestatus van het onderwerp en het lijdend voorwerp. Uit mijn onderzoek blijkt dat het lijdend voorwerp vrijwel alleen vooraan kan staan (OVS volgorde) als het Nieuwe Informatie bevat. Dit betekent dat de natuurlijke informatieflow van de zin (normaal van Oude naar Nieuwe informatie) omgedraaid is. Deze lijdende voorwerpen worden dus gemarkeerd door de natuurlijke informatieflow om te draaien. De enige uitzonderingen op deze generalisatie zijn de zogenaamde Familiar Topics. Dit zijn met name aanwijzende voornaamwoorden die de eerste plaats van de zin innemen. Ze verwijzen naar het laatstgenoemde concept of persoon in de voorafgaande context.

De corpusstudie leverde nog twee andere resultaten op die verband houden met de informatiestructuur, met name wat betreft de samenhang van de tekst ('tekstcohesie'). Framesetters of 'points of departure' komen om te beginnen meestal voor in V2-zinnen met adjuncten op de eerste plaats waar ze als topic fungeren. Een tweede observatie in deze categorie betreft de continuïteit. Om een sterke link tussen twee zinnen te bewerkstelligen, konden verbale nomina op de eerste plek van de zin geplaatst worden. Deze kunnen ofwel afhankelijk zijn van een vervoegd hulpwerkwoord in de voorafgaande zin of ze worden ondersteund door een vervoegde vorm van het hulpwerkwoord *gwneuthur* 'doen'. Dit maakt opnieuw deel uit van het narratieve karakter van de teksten in dit genre. Focus kan ten slotte gevonden worden in de speciale focusconstructie: de (gereduceerde) cleftzinnen ('Mixed Sentence'). Focus van een 'identity' predicaat komt meestal tot uiting in de specifieke *sef*-constructie, maar niet alle zinnen met *sef* bevatten focus, zeker niet aan het eind van de Middelwelshe periode.

In hoofdstuk 6 en 7 ligt de nadruk op de synchrone en diachrone syntactische analyse van de verschillende types woordvolgorde. Hoofdstuk 6 presenteert vier casussen die allemaal met informatiestructuur te maken hebben. Ik heb onderzocht hoe topics, focus en de referentiële status van constituenten samenhangen met de syntactische structuur. Het Middelwelsh kende maar een topicpositie, maar zinnen met V3- en V4-structuren komen ook voor. Ik heb twee verschillende analyses gepresenteerd voor V2-zinnen: een gebaseerd op verplaatsing van een constituent en een waar de constituent pas later aan de zin wordt toegevoegd. De constructie waarbij de constituent zelf verplaatst naar de eerste plek van de zin ontstond pas later in het Middelwelsh.

In het laatste hoofdstuk beschrijf ik verschillende manieren om historische syntaxis te analyseren: socio-linguistiek, constructiegrammatica en generatief syntactische benaderingen. Ik laat zien dat een generatief framework diverse voordelen heeft in de studie van Middelwelshe historische syntaxis, omdat het gebruik kan maken van inzichten van studies naar synchrone variatie in andere talen. Deze technieken die grondig getest zijn in het Minimalisme (Chomsky's 'Minimalist Program') helpen ons de exacte condities en context te definiëren waarin syntactische vernieuwingen wel en niet kunnen plaatsvinden en hoe ze naar bepaalde innovaties kunnen leiden.

Ik presenteer opnieuw verschillende casussen, dit keer op het gebied van syntactische innovaties in de geschiedenis van het Welsh. De eerste gaat over een speciale focusconstructie: identificatiefocus van het predikaat. Ik laat zien hoe deze constructie is voortgekomen uit de cleftconstructie in het Oudwelsh en hoe de focusmarker *sef* is ontstaan. Toen de focus verdween werd *sef* opnieuwe geïnterpreteerd als een expletief element en, uiteindelijk als een linker in reformulatieve appositieconstructies ("i.e."). In de tweede casus ga ik in op een van de belangrijkste onderzoeksvragen van dit proefschrift: het ontstaan van de V2-constructies. Door zorgvuldige vergelijking met andere Keltische talen en de reconstructie van de functionele partikels in het C-domein laat ik zien hoe de V2-zinnen in het Welsh zijn ontstaan. Na deze reconstructie volgde de herinterpretatie van zogenaamde 'hanging topics' en de uitbreiding van de functionele informatiestructuur van de constituenten op de eerste plaats van de zin. De fonologische erosie van de partikels in Vroegmodernwelsh leidde uiteindelijk tot het verdwijnen van de V2-constructie. Ten slotte vergelijk ik de resultaten van deze syntactische studie in het Middelwelsh met andere Middeleeuwse taalgroepen met het V2-fenomeen zoals het Romaans en Germaans.

In dit proefschrift probeer ik in het algemeen de interactie tussen syntaxis en informatiestructuur te analyseren en de invloed die beide uitoefenen op de woordvolgorde. De distributie van verschillende types woordvolgorde in het Middelwelsh het resultaat van zowel grammaticale factoren als informatiestructuur. Focus werd uitgedrukt door een gereduceerde cleftconstructie, de zogenaamde 'Mixed Order'. In identificerende copulazinnen daarentegen werd focus uitgedrukt door de focusmarker *sef* (< *ys* + *ef* 'dit is het'). Referentiële status en textcohesie speelden ook

een rol. Op basis van de huidige corpusstudie kunnen we een basaal algoritme ontwerpen om de juiste woordvolgorde te bepalen in transitieve zinnen in het Middelwelsh.

Vanuit historisch oogpunt heb ik laten zien dat informatiestructuur een rol kan spelen in syntactische innovaties, maar niet noodzakelijk de directe aanleiding vormt. De uitbreiding van informatiestructurele functies van constituenten op de eerste plaats van de zin is een goed voorbeeld hiervan. De uiteindelijke trigger voor syntactische veranderingen blijven soms moeilijk te achterhalen, maar een gedetailleerde en consequente beschrijving van de synchrone variatie waarbij alle variabelen systematische worden gecontroleerd is onmisbaar in historisch syntactisch onderzoek.

Curriculum Vitae

Marieke Meelen was born on 10 September 1986 in Maastricht, The Netherlands. She started studying Comparative Indo-European Linguistics and Hebrew and Aramaic studies at Leiden University in 2004, minoring in Comparative Religious Studies. She obtained her Bachelors degree in linguistics in 2008 and was admitted to the two-year Research Master programme in linguistics in Leiden that same year. During her masters, she completed various post-graduate modules in Celtic Studies at Utrecht University in The Netherlands and she spent a year as an Erasmus student at Aberystwyth University in Wales to learn Welsh. In 2010, she completed her MPhil (with distinction) and received a PhD fellowship from Leiden University for her project on syntactic change in Welsh. This thesis is the result of this research.