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

Meelen, M.

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.

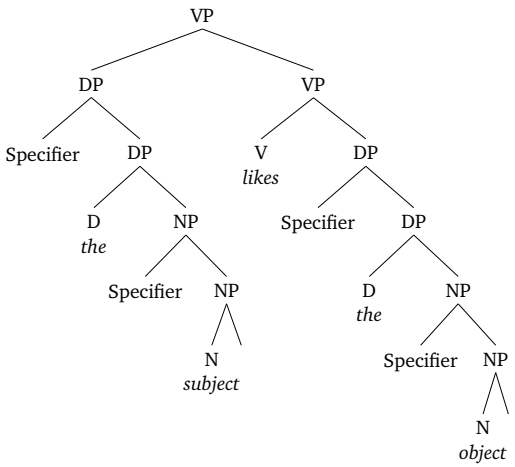
The two core operations of the Minimalist Program are Merge and Agree. Merge is the main structure-building operation that simply takes two syntactic objects α and β and forms a new object $\gamma = \{\alpha, \beta\}$ (Chomsky, 2001:3). The syntactic items can be drawn from the set of items in the Numeration (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.