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

Meelen, M.

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.