Running head: Word alignment in ParCorOE

Word alignment in a parallel corpus of Old English prose. From asymmetry to inter-syntactic annotation

Javier Martín Arista

Universidad de La Rioja

Abstract: This chapter proposes a model of syntactic annotation for the Parallel Corpus of Old English Prose, an aligned corpus of Old English and Present Day English texts. The research focuses on areas of syntactic divergence between the aligned texts. Syntactic divergence is described in terms of four types of alignment asymmetry (markedness, constituency, order, and configuration) and is represented by means of two components: a structural description and a dependency tree. The main conclusion is that these two components constitute a historical micro-grammar that identifies stability and change with respect to specific categories and constructions.

Key words: Parallel corpus, alignment, syntactic annotation, asymmetry, Old English

1. Introduction

The last decades have witnessed, along with a growing interest in Corpus Linguistics, a thorough investigation into the points of contact between this linguistic discipline and others like Translation and Lexicography. The works by authors such as Hanks (2012), Kübler & Zinsmeister (2014), Schierholz (2015), and Faaß (2017), to cite just a few, explore these regions. Against this background, this chapter intends to be a contribution to corpora and translation research. While its topic, a parallel corpus, does not represent

a completely new advance in Corpus Linguistics, the compilation and alignment of a parallel corpus that involves the modernisation of a previous diachronic stage of the target language stands up as largely virgin territory within the province of Translation.

With these premises, this chapter deals with the syntactic annotation of a parallel corpus with word alignment. The type of correspondence guides the design of syntactic annotation. In a parallel corpus with correspondence based on inter-linguistic translation, the linguistic distance between the source and the target language points to a degree of divergence that advises full syntactic annotation for both the source and the target. On the other hand, a parallel corpus whose correspondence relies on intra-linguistic translation or modernisation (a type of translation that involves the rendering of a text written in a previous diachronic stage of a natural language, as it is the case with Old English-Present Day English, henceforth PDE) necessarily displays a narrower linguistic distance between the source and the target language. This means that the full syntactic annotation of the source and the target language versions of a parallel corpus involving modernisation may present more points of convergence than of divergence, which, in turn, may result in some degree of descriptive inefficiency and redundancy. At the same time, it is predictable that when two texts written in different diachronic stages of the same language are compared, some mismatches arise. This may be of special relevance to English, which has significantly shifted throughout its history from a fully Germanic language to one with an outstanding Romance component, identifiable both in its morpho-syntax and lexicon.

This chapter addresses the research question of how to devise and implement a model of syntactic annotation for a parallel corpus aligned at word level. More specifically, the following sections raise the issues of the identification of the areas of divergence between the syntax of the source and the target language; the definition of

the scope of an inter-syntax is which syntactic divergence is couched in terms of asymmetry; and the development of the components, categories and functions of the inter-syntax.. This said, the chapter is structured as follows. Section 2 reviews previous work in parallel corpora from three perspectives: descriptive, pre-theoretical and theoretical. Section 3 lays the foundations of the Parallel Corpus of Old English Prose (hereafter ParCorOE) and presents the standards that guide its design and compilation. Section 4 proposes an inter-syntax that focuses on the syntactic divergences between the source language and the target language, which are captured in terms of alignment asymmetry. Four types of asymmetry are distinguished: marking, constituency, order and configuration. Section 5 applies this analysis of mismatches, thus identifying the main syntactic phenomena that may resist one-for-one word alignment in ParCorOE. This section also unfolds the inter-syntax of ParCorOE, which is comprised of a structural description and a dependency tree represented with graph theory. To close this work, Section 6 summarises the main conclusions and offers some avenues for future research.

2. Background

A parallel corpus is a type of bilingual or multilingual corpus that contains texts from the source language and their translations (McEnery 2003: 450). In contrast, a comparable corpus *can be defined as a corpus containing components that are collected usidng the same sampling frame and similar balance and representativeness* (McEnery & Xiao 2007a: 3). It is a central requirement of parallel corpora that they align the

source texts and their translations, either at word or sentence level (McEnery & Xiao 2007a: 3).

According to Aijmer and Altenberg (1996, in McEnery and Xiao 2007b: 131), parallel corpora can be used for conducting a wider array of studies than monolingual corpora. Parallel corpora also have various applications to lexicography, language teaching and acquisition, as well as translation. Given that a parallel corpus offers *direct comparability* (Enrique-Arias 2013: 105), diachronic research can benefit from parallel texts because all the target language forms that express a given content from the source language can be analysed.

In spite of the advantages and applications of parallel corpora summarised in this section, a parallel corpus for English Historical Linguistics in general, and for Old English in particular, is not available at the moment. Such an undertaking should consider the state of play regarding the need to automatise corpus annotation. Some authors, such as Lu (2014), underline the importance of Natural Language Processing technology, which allows computers to annotate large corpora at different linguistic levels, so that a minimum of manual revision is required. This aim can be achieved more effectively in corpora of natural languages than in historical corpora because the latter are far smaller, thus resisting statistical processing, and, above all, because historical corpora often raise issues of spelling variation that preclude fully automatic annotation (Johnson 2009). In Historical Linguistics in general and Old English in particular, spelling variation turns the lemmatisation of the corpus -the attribution of the textual forms to the corresponding dictionary forms (Schierholz 2015)- into the central task of corpus tagging and annotation (Martín Arista 2013, 2017a, 2017b): only when a textual form has been assigned to a lemma through a normalisation procedure, is it possible to automatically provide the token in question with the relevant information

from the dictionary database (Tío Sáenz 2015; Metola Rodríguez 2017; Novo Urraca and Ojanguren López 2018; García Fernández fc.).

Therefore, the compilation of an aligned parallel corpus represents a challenging project relevant for corpus and translation research, as well as an investigation with various applications to the linguistic analysis and the lexicography of Old English. From the descriptive point of view, to date there is not a large collection of annotated aligned parallel texts for the study of Old English. In pre-theoretical terms, no parallel corpus has been compiled so far that comprises a text in a historical language and its modernised version. On the theoretical side, the central aspect of a parallel corpus is alignment, in such a way that the more exhaustive tagging and annotation is required the more accurate alignment needs to be. To this effect, this chapter takes the line that the alignment in a corpus that revolves around modernisation has to be guided by the divergences between the old and the modern version of the text, given that diachronic continuity makes allowance for the exclusion of the areas of morphosyntactic convergence. These aspects are discussed in turn in the remainder of this section.

Beginning with the descriptive aspects, the most widely used corpora of Old English include the Old English segment of the *Helsinki Corpus of English Texts* (Rissanen et al. 1991), which contains around 300,000 words; *The York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug 2001), which comprises approximately 70,000 words; *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al. 2003; henceforth YCOE), which files ca. 1.5 million words; and the *Dictionary of Old English Corpus* (Healey et al. 2004), which gathers around three million words and was specifically compiled for the *Dictionary of Old English* (Cameron et al. 2018). These corpora are segmented by fragment and text, with tokens identified by means of a specific number or, in the case of the *Dictionary of Old English Corpus,* by means of

the Cameron number (Mitchell et al. 1975, 1979). The four corpora are marked-up at text level. The *Helsinki Corpus of English Texts*, for instance, provides each fragment file with the abbreviated title, sub-period, manuscript date, dialect, text type, genre, and information on the translation, if relevant. *The York-Helsinki Parsed Corpus of Old English Poetry* and *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (henceforth YCOE) have been tagged morphologically (category and morphological case) and parsed syntactically (hierarchy and linearisation). In spite of the wealth of philological data compiled in these corpora, two major pending tasks remain for Old English linguistics, corpus analysis and lexicography: the lemmatisation of the written records and the compilation of a representative parallel corpus Old English-English.

As for the methodological questions, parallel corpora, as a general rule, compare languages from different linguistic branches, such as Portuguese (Romance) and English (Germanic) with respect to Indo-European; or language belonging to two distinct sub-branches of a linguistic family, as is the case with English (West-Germanic) and Swedish (North-Germanic) within Germanic. Even when it comes to compiling corpora for Historical Linguistics, such corpora tend to be comprised of versions from different languages, rather than presenting two diachronic stages of the same language. For example, the *ENHIGLA* (Old English - Old High German - Latin) parallel corpus contains ca. 21,000 clauses (available at http://pelcra.pl/enhigla/corpus) from the Latin version and the Old English translation of the first twenty-five chapters from the Book of Genesis and the first ten chapters from the Gospel of Luke; the Latin original of and the Old English version of Book I and a fragment of Book II from Bede's *Historia ecclesiastica gentis anglorum*; as well as the Latin version and the Old High German translation of the first seventy-four chapters from Tatian´s *Gospel Harmony*, *De fide catolica contra iudeos* by St. Isidor of Seville, and *Physiologus*. Put differently, parallel

corpora like the ones cited above do not make a claim of continuity on the diachronic axis, which a parallel corpus comparing two diachronic stages of a language certainly does.

With regard to the theoretical aspects mentioned above, it has already been remarked that parallel corpora rely on the correspondence between the source language and the target language texts. This comparison can be established at several levels. Authors such as Kübler and Zinsmeister (2014), as well as Krause and Zeldes (2016), insist on the importance of annotation to achieve the goal of increasing searchability and put forward levels of annotation below the text level that include the sentence and the word level. Sentence level and word level alignment, however, not only beg for additional tokenisation with respect to text alignment but also demand more detailed tagging and annotation. In other words, alignment empirically demonstrates the correspondence between the texts that has been assumed as the point of departure of the research; and, ultimately, relates tokenisation to searchability, which is in need of extensive and accurate tagging and annotation at sentence and word level. Alignment also makes for the adequacy of lemmatisation, which, as pointed out above, constitutes the central task of corpus tagging and annotation. Last but not least, alignment defines the scope of the morphosyntactic comparison between the source text and the target text. Put briefly, alignment can be considered the main characteristic of a parallel corpus.

While alignment determines the scope, asymmetry emphasises certain aspects of the comparison, thus disregarding symmetric parts of the comparison between the source and the target text. Defined in these terms, asymmetry accounts for the divergence between the source and the target under comparison and, conversely, symmetry couches the convergent aspects of the comparison of the source and the target

text, which is basically put aside. As for continuity on the diachronic axis, symmetry corresponds to stability while asymmetry indicates change. An inter-syntax is proposed in sections 4 and 5 that can represent and explain the relevant aspects of morphological case marking, functional relations, argument projection and linearisation.

3. The design of an aligned parallel corpus of Old English prose

Against the background presented in the previous section, ParCorOE, an aligned parallel corpus of Old English prose, is an ongoing project that aims at compiling 300.000 words in the source language, plus the parallel version in the target language. The basic parameters of ParCorOE can be set as follows. As regards general orientation, the corpus will be historical, rather than a corpus devised for translation, comparative linguistics or second language learning. With respect to the number of languages selected, ParCorOE will be bilingual, involving Old English and PDE. As far as directionality is concerned, ParCorOE will be unidirectional: from Old English to PDE. As for the target, ParCorOE will be aimed to textual forms (tokens or inflections), instead of revolving around dictionary words or lemmas. Concerning genre, ParCorOE will select prose texts only.

With these parameters, the following standards guide the design and compilation of ParCorOE. These standards serve the general aim of increasing searchability.

Standard 1: Alignment

An aligned parallel corpus Old English-English consists of a parallel text, that is to say, an Old English text placed along its PDE modernisation, with alignment at text,

sentence and word level, in such a way that very source language segment is paired with a target language segment. Word, sentence, and text alignment is in need of tokenisation at these three structural levels. Alignment parings should be marked by means of the highlighting of the source and the target segment (See Figure 2 below).

Standard 2: Annotation

Three types of annotation must be distinguished: mark up at text level, as well as syntactic annotation and morphological tagging at sentence/word level. Fragments (tokens) are comprised of at least one sentence or one syntactically independent period, identified by means of a text number, such as Mart 55.07.07, corresponding to *Ond monige menn gesegon ðæt ða deadan arison of ðæm byrgennum ond eodon geond ða halgan burh on Hierusalem, oð ðæt Crist eft aras*. (And many people saw the dead arise from their graves and walk through the holy town of Jerusalem until the resurrection of Christ).

Standard 3: Lemmatisation

The corpus must be fully lemmatised, so that all the textual attestations are grouped under the relevant lemma, and each lemma is provided with all its inflections. For example, the following inflections have been attibuted so far to the verba lemma *niman* 'to take': *nam* (ind. pret. 3rd sing.), *naman* (ind. pret. pl.), *name* (subj. pret. sing.), *namon* (ind. pret. pl.), *namon* (subj. pres. pl.), *namon* (subj. pret. pl.), *nim* (imp. sing.), *nimað* (imp. pl.), *nimað* (ind. pres. pl.), *niman* (infinitive), *nimð* (ind. pres. 3rd sg.), *nime* (ind. pres. 1st sg.), *nime* (subj. pres. sing.), *nime* (subj. pret. sing.), *nimeð* (ind. pres. 3rd sg.), *nimen* (infinitive), *nimen* (subj. pres. pl.), *nimenne* (infl. inf.)), *nimest*

(ind. pres. 2nd sg.), *nimine* (infl. inf.), *numen* (pa. part.), *nyme* (subj. pres. sing.). In token analysis, 485 inflections have been lemmatised under the verb *niman*.[1]

Standard 4: Automation

Within the limits imposed by the available written standards and the variation that they present, the annotation of the parallel corpus must be automatic. This includes not only syntactic annotation and morphological tagging, but also the necessary lemmatisation. Lemmas and inflections must be listed dynamically, so that users have access to ablaut patterns, such as *nim-nam-nom-num* in *niman* 'to take'; elision, as in *nymaþ/ nymþ* and other spelling alternatives, like *nimaþ/neomaþ/niomaþ.*

Standard 5: Feeding

The corpus must be fed with the information available from a knowledge base of Old English. The parallel corpus may retrieve information from the relational databases in the knowledge base of Old English in order to maximise the automation of the tasks of tagging, annotation and lemmatisation. For instance, additional spellings and inflections are automatically fed from the knowledge base to the lemmatisation of *niman*, including *neoman* (subj. pres. pl.), *neomendum* (pres. part. dat. pl.), *nimæð, neomaþ, nimaþ, niomað*, *nymþ* (ind. pres. pl.), *nimst, nimest* (pres. 2nd sg.), *niomanne, nimanne, nymenne* (infl. inf.), *nome*, (ind. pret. 2sg.), and *nomon* (ind. pret. pl.).

Standard 6: Searchability

The corpus must be searchable by text, fragment and word, as well as by morphological tag and syntactic annotation. Combined searches by inflectional form and lemma are also required. The corpus must be based on a concordance and an index, so that the main layouts are interconnected (see figures 1 and 2).

[1] The following abbreviations are used in this section: ind. (indicative); sub. (subjunctive); imp. (imperative); infl. inf. (inflected infinitive); pres. part. (present participle); pa. part. (past participle); pres. (present); pret. (preterite); sg. (singular); pl. (plural); dat. (dative).

Standard 7: Dissemination

The corpus must be available online in open access (see Figure 2).

To recapitulate, the background, parameters and standards presented so far point to a corpus compatible with theoretical studies as well as applications of Old English lexicography and presentations of Digital Humanities. Turning to the question of representativeness, McEnery (1996: 123) stresses the importance of the corpora of historical languages and remarks that, exactly like the corpora of natural languages, historical corpora must be quantitatively sufficient and qualitatively representative so as to offer an accurate representation of the language of analysis. Biber (2007) suggests that a corpus that has been compiled in various stages is more likely to be representative. This author recommends to design and implement a pilot corpus that gathers as much variation as possible, so that the compilers can identify specific issues and general problems. Heid (2008: 43) calls the design and implementation of a pilot corpus *preprocessing* and holds that for an approach to be corpus-based rather than corpus-driven, preprocessing is necessary.

In this line, a ten-thousand-word pilot corpus was compiled and annotated (Martín Arista 2017a, 2017b, 2018). The texts, as well as their modernisations, were extracted from Fernández Cuesta et al. (1997). The aim of the pilot corpus was to find design inadequacies and compilation shortcomings. From the quantitative point of view, ten thousand fully tokenised and annotated words suffice to raise issues in the corpus architecture as well as inconsistencies to the tokenisation and the annotation. From the qualitative point of view, a variety of prose texts were chosen. The selection of texts comprised fragments from the *Anglo-Saxon Chronicle*, *Orosius*, Ælfric´s *Lives of Saints*, *Cura Pastoralis,* and Bede´s *Ecclesiastical History*, thus including the major

genres of historical prose, religious prose and translations from Latin. This set of texts is representative of the dialect of the vast majority of the records of Old English, which are written in the West Saxon variety. As to datation, Bede's *Ecclesiastical History,* and *Cura Pastoralis* can be dated to the 9th. century (early Old English); *Orosius* and the fragments from the *Anglo-Saxon Chronicle* can be dated to the 10th. century (classical Old English); while the *Lives of Saints* corresponds to the 11th. century (late Old English).

The pilot corpus has two main components: the concordance (including a word index) to the texts and the parallel corpus layouts. Two layouts have been distinguished: the static presentation and the dynamic presentation. The static presentation offers the running texts Old English-PDE, aligns them by fragment and word and provides word-for-word gloss as well as fragment modernisation. This is presented in Figure 1.

Figure 1: The static presentation of the pilot corpus.

The dynamic presentation of the parallel corpus is aligned at word level. Each word is highlighted in the source and in the target text. Full tagging and annotation are fed from the Knowledge-Base of Old English (Martín Arista and Ojanguren López 2018). The information that has been imported from the databases includes lemma, alternative spellings, lexical category, morphological class, inflectional paradigm, derivational paradigm, meaning definition, and the references of secondary sources that deal with the lemma or the inflectional form in question. As shown in Figure 2, there are two basic query options, by inflectional form and by lemma. Full inventories of inflectional forms and lemmas are available. The database software that files the corpus guarantees information retrieval through simple, combined and stepwise searches. It also facilitates open access because it makes allowance for an online publication that can be accessed and searched with an Internet browser.

Figure 2. The dynamic presentation of the pilot corpus.

In its present state, ParCorOE consists of ca. 160,000 word files, with the new layout presented in Figure 3.

Figure 3: Tokenisation, tagging and lemmatisation of ParCorOE.

Quantitatively speaking the final corpus will comprise 300,000 words, the first half being due by March 2021. This amount represents about one tenth of all the written records of Old English. From the qualitative point of view, all the major prose genres of Old English have already been included. The words processed so far by category and text are tabulated in Table 1.

Category	Texts	Word count
Historical prose	*The Anglo-Saxon Chronicle*	25,000
Religious prose	*Ælfric´s Homilies*	25,000
The Bible	*St. Mark*	25,000
Translations from Latin	*Benedictine Rule, Martyrology*	25,000
Legal prose	*Laws*	12,500
History	*Orosius*	12,500
Philosophy	*Boethius*	12,500
Medicine and herbaries	*Leechbook*	25,000

Table 1. Word count by category and text.

The source language texts and the target language translations draw on the editions cited below. Put in other words, the texts are not modernised *ad hoc*, but rather follow available PDE translations. The choice of the edition and translation at the present state of the research has been guided by copyright status. All texts are free of copyright.

The tokenisation, tagging or annotation of the texts that have been processed so far have pointed out some instances and areas of mismatch between the source language

and the target language that hamper the word-for-word correspondence required for syntactic annotation. The solution proposed in the following sections is the implementation of an inter-syntax that has two functions: (i) to focus on the areas of syntactic divergence between the Old English text and its translation; and (ii) to map the source language tokens onto the target language ones by means of a set of tags that represent hierarchy and dependency. All the local instances of mismatch have been filed in an asymmetry bank.

4. Inter-syntax and asymmetry

The inter-syntax of ParCorOE can be described as an intermediate step between tokenisation, on the one hand, and glossing and annotation, on the other. The inter-syntax has two components, namely a structural description and a dependency tree. The structural description, in turn, comprises two labeled bracketing representations, one for the source language (extracted from the YCOE) and another one for the target language (based on the same categories as the YCOE). The dependency tree displays functional tags that relate dependent elements to their heads.

The definition of the scope of the inter-syntax requires the previous identification of the areas of divergence between the syntax of the source and the target language. This is tantamount to saying that the task is defined gradually and unidirectionally: in the search for mismatches, symmetrical and asymmetrical parings are considered, although the inter-syntax focuses on asymmetry. Old English is always the source language.

The mismatches that cause asymmetry between the source language text and its translation are described on structural grounds, but explained on a functional basis. For this reason, two sets of syntactic tags are required, categorial tags and functional tags, so that categorial tags account for hierarchy and functional tags for dependency. As has just been said, the structural description relies on the YCOE and the functional tags have been adapted from https://universaldependencies.org/u/dep/. While the structural description of the source and target language segment accounts for hierarchy and linearisation, the dependency tree aims to the relations that hold both in the source and the target language segment, as well as those that, applying in the source or the target only, constitute a description of variation on the synchronic axis or an explanation for change on the diachronic axis. The annotation in this framework, therefore, is couched in terms of an inter-syntax that maps the source language segment onto the corresponding target language segment. It must be stressed from this point that the terms *syntax* and *syntactic* are used comprehensively, so that morphological phenomena with impact on syntax, such as the assignment of morphological case, are considered.

To summarise, the structural change of the target language segment with respect to the source language segment can be derived from the tree diagrams or the labeled bracketing, but explanations based on phrasal and clausal functions require dependency relations. The areas of divergence between the source and the target language reflect a variety of syntactic phenomena that have been discussed from different angles in works like Visser (1963-73), Mitchell (1985), Denison (1993), Martín Arista (2000a, 2000b), Hogg and Fulk (2011), and Ringe and Taylor (2014).

The position held in this respect is that syntactic divergences can be captured in terms of alignment asymmetry. In this line, Scrivner (2015: 2) distinguishes the following schemas of alignment at word level: between two single words (one-to-one),

between a single word and a multi-word unit (one-to-many), between a multi-word unit and a single word (many-to-one) and zero alignment. Alignment schemas are thus coached in terms of (a)symmetry: the number of slots in the source language is equal to or different from the number of slots in the target language. However, the asymmetry between the source and the target language cannot be restricted to quantity. Rather, it may be the result of marking, constituency, order or configuration. Markedness asymmetry involves more or less marked slots in the source language. Constituency asymmetry is the result of the presence of fewer slots in the source language. Order asymmetry is a consequence of a relative order in the target language different from the source language. Configuration asymmetry conveys a syntactic configuration substantially different from the source language. These four types of asymmetry may involve categories and relations and can be found at phrasal or sentential (inflectional phrase) level, although substantial changes to morphosyntactic configuration may often affect the inflectional phrase, whilst the type of asymmetry involving markedness is more likely to arise within units of the phrasal level. Consider the following example (quoted with the DOEC number).

(1) [LawWi 000500 (5)]

Gif ðæs geweorþe gesiþcundne mannan ofer þis gemot,

þæt he unriht hæmed genime ofer cyngæs bebod & biscopes & boca dom,

se þæt gebete his dryhtne C scillinga an ald reht;

If after this meeting, a nobleman presumes to enter into an illicit union, despite the command of the king and the bishop, and the written law, he shall pay 100 shillings compensation to his lord, in accordance with established custom. (Attenborough 1922: 25)

In (1), several instances arise of the four types of asymmetry distinguished in this work. Beginning with markedness asymmetry, the verbal forms *geweorþe* 'please' and *gebete* 'pay' are inflected for the subjunctive, which is no longer possible in PDE by morphological means. In the noun phrase, the Old English dative *dryhtne* 'lord' requires a morphologically unmarked noun governed by a preposition, which is compatible with the definition of constituency asymmetry given above. The same can be said of the case-marked genitive noun *boca* in the noun phrase *boca dom* 'judgement of books'. It is also of relevance for constituency asymmetry that the noun phrases *cyngæs bebod & biscopes & boca dom* 'the command of the king and the bishop, and the judgement of books' require a definite article functioning as determiner in PDE. Focusing on order asymmetry, the coordinate modifier in the genitive case in *cyngæs bebod & biscopes* cannot be extraposed in the PDE counterpart, thus 'the king's and the bishop's command'). As regards configuration asymmetry, the verb *geweorþan* 'to please' selects the thematic roles Theme *ðæs* 'of that' (case-marked genitive) and Experiencer *gesiþcundne mannan* 'noble man' (inflected for the dative). Moreover, no introductory *hit* is found in the Old English fragment in sentence-initial position and the verb is complemented by a *þæt*-clause with the dependent verb in the subjunctive (*þæt he unriht hæmed genime* 'to enter into an illicit union'), rather than by a *to*-infinitive clause realising a linked predication that shares the first argument with the matrix predication, as in 'if it pleases a noble man to enter into an illicit union'.

Given this kind of evidence, the discussion that follows in the next section puts aside markedness asymmetry (which is rather predictable when it comes to comparing a more inflective and a less inflective language) to concentrate on constituency, order and configuration asymmetry.

5. The scope and components of the inter-syntax

This section identifies the areas that present alignment asymmetry and proposes an inter-syntax that incorporates the labeled bracketing of the YCOE but that crucially hinges around a dependency tree. Asymmetry phenomena are described with respect to the structural levels of the noun phrase and the inflectional phrase. While all the fragments have been extracted from the segment of PacCorOE that has been processed so far, they are headed by the DOEC text name and number.

In the noun phrase, a frequent asymmetry type is constituency asymmetry. It is often the case that the noun phrase is morphologically marked by means of case in the source language and by means of prepositional government in the target language. This can involve the accusative, the genitive, the dative and the instrumental, as in *lytle werede* 'with a small force' in (2).

(2) [ChronA (Bately) 036100 (871.30)]

Þa feng Ęlfred Ęþelwulfing his broþur to Wesseaxna rice, & þæs ymb anne monaþ gefeaht Ęlfred cyning wiþ alne þone here lytle werede æt Wiltune & hine longe on dæg gefliemde, & þa Deniscan ahton wælstowe gewald.

Then his brother Alfred, son of Æthelwulf, succeeded to the kingdom of Wessex. And one month later king Alfred fought with a small force against the entire host at Wilton, and for a long time during the day drove

them off, and the Danes had possession of the place of slaughter. (Garmonsway 1972: 72).

Order asymmetry can also arise in the noun phrase. For instance, the genitive *ðæs cyninges* 'of the king' follows the nominal head *þegn* 'thane' in (3), in contradistinction to the proper name genitive *Ecgferðes* 'Ecgfrith's', which precedes the head.

(3) [Mart 5 (Kotzor) 018700 (Ma 7, B.5)]

He wæs Ecgferðes þegn ðæs cyninges, ac he forlet þa wæpna ond ða woruldlican wisan ond eode on þæt mynster ond wæs þær mæssepreost ond abbod.

He was a thane of King Ecgfrith, but he gave up his weapons and his secular life and joined the monastery and was a priest there and an abbot. (Rauer 2013: 62)

In the inflectional phrase, the lack of do-support in the source language causes constituency asymmetry, as can be seen in (4), where *lifde* 'lived' is negated with the negative word *ne* 'not' only.

(4) [Or 3 036900 (11.82.18)]

Þagiet ne mehte se nið betux him twæm gelicgean, þeh heora na ma ne lifde þara þe Alexandres folgeras wæron, ac swa ealde swa hie þa wæron hie gefuhton: Seleucus hæfde seofon & seofontig wintra, & Lisimachus hæfde þreo & seofontig wintra.

Still the hostility between those two could not end, even though they were the only ones of Alexander´s followers left, but, as old as they were, they went on fighting: Seleucus was seventy-seven years old and Lysimachus was seventy-three. (Godden 2016: 218)

The contraction of negative *bēon* 'to be', *habban* 'to have', *willan* 'will', *wītan* 'to know' and *āgan* 'ought to' also causes constituency asymmetry, given that the written representation takes one more slot in the target language than in the source language. The question is illustrated with respect to *witan* 'to know' in (5).

(5) [LawICn 003100 (6.2)]

Full georne hig witan, þæt hig nagon mid rihte þurh hæmedþingc wifes gemanan.

They know full well that they have no right to marry. (Attenborough 1922: 22)

Another source of asymmetry in the inflectional phrase is the verbal conjugation of the source language, which does not have continuous, periphrastic or compound tenses, but presents a morphologically distinct subjunctive. This subjunctive translates as a modal periphrasis or as an indicative, as is the case with *læge* 'lay' and *bude* 'lived' in example (6). This usually causes constituency asymmetry, but may also produce configuration asymmetry.

(6) [Or 1 008000 (1.14.5)]

He sæde þæt he æt sumum cirre wolde fandian hu longe þæt land norþryhte læge, oþþe hwæðer ænig mon be norðan þæm westenne bude.
He said that on one occasion he decided to find out how far the country extended northward, or whether anyone lived to the north of that uninhabited region. (Godden 2016: 36)

In the inflectional phrase, various phenomena cause order asymmetry, beginning with extraposition, which may involve a multiple subject or a relative clause, such as *þe ær wæs forslagen* 'which had been cut through before' in example (7).

(7) [Æ LS (Edmund) 004700 (176)]
And his swura wæs gehalod þe ær wæs forslagen, and wæs swylce an seolcen þræd embe his swuran ræd, mannum to sweotelunge hu he ofslagen wæs.
And his neck, which had been cut through before, was healed and there was something like a red silken thread around his neck for men to remember how he had been killed. (Skeat 1881: 326)

Stranded prepositions also convey order asymmetry, as is the case with *Him com þa gangende to Godes engel* 'God's angel came to him walking' in (8).

(8) [Judg 006600 (13.3)]
Him com þa gangende to Godes engel, & cwæð ðæt hi sceoldon habban sunu him gemæne;

‘And an angel of the Lord appeared to her, and said: Thou art barren and without children: but thou shalt conceive and bear a son. (*Douay Rheims Bible*: 475).

Order asymmetry also results from various fronting phenomena, including the fronting of the auxiliary verbs *bēon* and *habban*, illustrated in (9.a) and (9.b) respectively, as well as the fronting of nominative complements, such as *Themestocles* ‘Themistocles’ in (9.c), accusative and dative objects.

(9)

a. [MkGl (Li) 000500 (1.4)]
Wæs iohannes in woestern gefulwade & bodade fulwiht hreownisses on forgefnisse synna.
John was in the desert baptizing, and preaching the baptism of penance, unto remission of sins. (Leonard 1881: 17)

b. [ChronE (Irvine) 026610 (658.3)]
Hæfde hine Penda adrefedne & rices benumene forþan þet he his swustor forlet.
Penda had expelled him and deprived him of his kingdom because he had repudiated his sister. (Garmonsway 1972: 32)

c. [Or 2 012700 (5.47.18)]
Themestocles hatte Atheniensa ladteow.
The leader of the Athenians was Themistocles. (Godden 2016: 127)

Two further characteristics of the source language bring about order asymmetry with respect to the target language, namely the V2 Rule, which places the subject after the verb in the context of an initial adverbial, as happens in *Þa comon þa menn* 'Then these men came' in (10.a); and the relatively generalised final verb in dependent clauses, which is illustrated by means of *gefultumade* 'may help' and *gehiersumade* 'may subject' in (10.b).

(10)

a. [ChronA (Bately) 007400 (449.9)]
Þa comon þa menn of þrim mægþum Germanie, of Ealdseaxum, of Anglum, of Iotum.
These men came from three nations of Germany: from the Old Saxons, from the Angles, from the Jutes. (Garmonsway 1972: 12)

b. [ChronA (Bately) 032600 (853.1)]
Her będ Burgred Miercna cyning & his wiotan Ęþelwulf cyning þæt he him gefultumade þæt him Norþwalas gehiersumade.
In this year Burgred, king of Mercia, and his councilors besought king Æthelwulf that he would help them to subject the Welsh. (Garmonsway 1972: 66)

The existence of double negation in Old English, comprising both the phrasal and the sentential levels, causes configuration asymmetry with respect to the target language. This may be due not only to the negative words themselves, but also to the lack of *do*-support in Old English. For instance, *ne* negates at sentential level and *naht* at phrasal level in (11).

(11) [Mart 5 (Kotzor) 023900 (Ma 21, B.4)]

Ond on sumum þara mynstra þe he ofergeseted wæs þa broðor him woldon sellan attor drincan forðon þe hi ne mostan for him naht unalyfedlices begangan.

And in one of the monasteries over which he presided, the brothers tried to give him poison to drink, because with him they were not allowed to do anything illicit. (Rauer 2013: 71)

Various phenomena that may be grouped under the heading of omission result in constituency asymmetry with respect to the target language. The status of the elements which are required in the target language considerably varies. In Old English, the formal subjects *there* and *it* are not compulsory, as is shown in (12.a) and (12.b), respectively.

(12)

a. [Mart 2.1 (Herzfeld-Kotzor) 016700 (De 0, A.1)]

On þam twelftan monðe on geare byð an ond XXX daga.

'There are thirty-one days in the twelfth month of the year.'

(Rauer 2013: 223)

b. [Mart 2.1 (Herzfeld-Kotzor) 000800 (Ju 24, B.3)]

Þonne gelympeð þæt wundorlice on þæs sumeres sungihte on mydne dæg þonne seo sunne byð on þæs heofones mydle, þonne nafað seo syl nænige sceade.

Then amazingly, it happens during the summer solstice at midday, that when the sun is in the middle of the sky, the column does not have any shadow. (Rauer 2013: 125)

Fully lexical subjects can also be omitted, as is illustrated in (13.a), but the omission of lexical verbs is restricted, as a general rule, to *bēon*, as is presented in (13.b).

(13)

a. [Mart 5 (Kotzor) 094900 (Au 30, A.2)]
Wæs in ðære ceastre þe is nemned Tubsocensi.
He lived in the city which is called Thibiuca. (Rauer 2013: 171)

b. [ÆCHom I, 7 003800 (234.79)]
Swutel is þæt ða tungelwitegan tocneowon crist. soðne man: þa ða hi befrunon. hwær is se ðe acenned is.
It is manifest that the astrologers knew Christ to be a true man, when they inquired, 'Where is he who is born?' (Thorpe 1844: 107).

As in PDE, the subject of a coordinate construction is, as a general rule, omitted in Old English. However, the object of a construction of coordination is left unexpressed far more often in the source language than in the target language of the corpus. An instance of the omission of the object of a coordinate construction is given in (14), in which the object *hine* 'him' is shared by *gebringan* 'bring' and *belucan* 'lock up'

(14) [Bo 001200 (1.7.23)]

þa þæt ongeat se wælhreowa cyning ðeodric, þa het he hine gebringan on carcerne & þærinne belucan.

When that cruel king Theoderic discovered this, he ordered him to be put into a prison and locked up there. (Godden et al. 2009: 5)

The omission of a complementiser, such as the one depending on *secge* 'say' in (15), involves constituency asymmetry too.

(15) [Mk (WSCp) 009900 (3.29)]

Soþlice ic eow secge, se þe ðone halgan gast bysmerað, se næfð on ecnysse forgyfenesse, ac bið eces gyltes scyldig.

But he that shall blaspheme against the Holy Ghost, shall never have forgiveness, but shall be guilty of an everlasting sin. (Leonard 1881: 28)

Complex conjunctions, such as *mid þæm þe* 'when' in (16.a), and complex relatives, like *þær ðær* 'where' in (16.b), take fewer slots in the target than in the source language. This is also a matter of constituency asymmetry.

(16)

a. [Or 2 003400 (2.39.6)]

Hi swaþeah heora unðances mid swicdome hie begeaton, mid þæm þe hie bædon þæt hie him fylstan mosten ðæt hie hiera

godum þe ieð blotan mehten: þa hie him þæs getygðedon, þa hæfdon hi him to wifum, & heora fæderum eft agiefan noldon.

The Romans got them anyway by trickery, despite the opposition of the fathers, when they asked the Sabines to help them sacrifice, to their gods. When the Sabines agreed to this, the Romans seized the daughters as their wives and would not return them to their fathers. (Godden 2016: 107)

b. [ChronE (Irvine) 031600 (679.1)]

Her man ofsloh Ælfwine be Trentan þær ðær Egferð & Æðelred gefuhton.

In this year Ælfwine was slain beside the Trent, at the place where Ecgfrith and Æthelred fought. (Garmonsway 1972: 38)

Impersonal verbs require one more argument (a formal subject) in the target language in order to realise the Patient, Recipient or Beneficiary, which, in the source language is case marked accusative, as *hine* 'him' in (17.a) or dative, like *him* 'them' in (17.b). These are instances, therefore, of constituency asymmetry.

(17)

a. [Bo 045300 (16.39.20)]

Hine lyste eac geseon hu seo burne, hu lange, & hu leohte be þære oðerre.

He wanted also to see how it burnt, how long and how brightly in comparison with the other city. (Godden et al. 2009: 26)

b. [CP 123900 (36.261.3)]

Him is to secgeanne ðæt hie unablinnendlice geðencen hu monig yfel ure Dryhten & ure Alisend geðolode mid ðam ilcan mannum ðe he self gesceop, & hu fela edwites & unnyttra worda he forbær, & hu manige hleorslægeas he underfeng æt ðæm ðe hine bismredon.

They are to be told to consider incessantly how many evils our Lord and Redeemer suffered among the same men whom he himself had created, and how much reproach and how many vain words he endured, and how many blows he received from his revilers. (Sweet 1881: 260)

Reflexives with intransitive verbs also cause constituency asymmetry. They take one more argument in the source than in the target language, either case-marked accusative, such as *hine* 'himself' in (18.a), or dative, like *him* 'themselves' in (18.b)

(19)

a. [Mk (WSCp) 016800 (5.22)]

& ða com sum of heahgesamnungum Iairus hatte, & þa he hine geseah he astrehte hine to his fotum.

And there cometh one of the rulers of the synagogue named Jairus: and seeing him, falleth down at his feet. (Leonard 1881: 34)

b. [Or 1 029500 (10.29.12)]

Hi þa þæt lond forleton, & him hamweard ferdon.

Then they left that land and went home. (Godden 2016: 79)

As can be seen in (19), there is configuration asymmetry between an instance of the verb *hātan* ‘to be called’, which occurs in active sentences in the source language, and the corresponding passive in the target language.

(19) [Or 1 003600 (1.11.1)]

Seo Ægyptus þe us near is, be norþan hire is þæt land Palastine, & be eastan hiere Sarracene þæt land & be westan hire Libia þæt land, & be suþan hire se beorg þe mon hæt Climax.

The part of Egypt that is nearer to us has Palestine to the north, and to the east is the Saracen land, and to the west is Libya, and to the south is a mountain called Climax. (Godden 2016: 30)

Example (19) also illustrates the configuration asymmetry holding with respect to the indefinite pronoun *mon* ‘someone’ in the source language and in the target language, which frequently calls for a passive.

Considering the areas of asymmetry presented in this section, the inter-syntax of ParCorOE comprises a set of dependency relations that links the hierarchy and linearisation of the source language representation to the target language. Hierarchy and linearisation are displayed by labeled bracketing representations of the type adopted by the YCOE, which is illustrated in Figure 4, representing *Aristoteles hit gerehte on þære bec þe Fisica hatte* ‘Aristotle explained it in the book which is entitled Physics’ (Godden et al. 2009). Figure 5 shows the structural description of the target language segment.

```
(IP-MAT-SPE (NP-NOM (NR^N Aristoteles)
                    (NP-ACC (PRO^A hit))
                    (VBPS gerehte)
                    (PP (P on)
                        (NP-DAT (D^D +t+are) (N^D bec)
                                (CP-REL-SPE (WNP-NOM-1 0)
                                            (C +te)
                                            (IP-SUB-SPE (NP-NOM *T*-1)
                                                        (NP-PRD (NR Fisica))
                                                        (VBD hatte))))))
            (. .)) (ID coboeth,Bo:40.140.8.2794))
```

Figure 4: Source language labeled bracketing from the YCOE.

```
(IP-MAT-SPE (NP (NR Aristotle)
                (VBPS explained)
                (NP (PRO it))
                (PP (P in)
                    (NP (D the) (N book)
                            (CP-REL-SPE (WNP-1 0)
                                        (C that)
                                        (IP-SUB-SPE (NP *T*-1)
                                                    (VBD is entitled))))))
                                                    (NP-PRD (NR Physics))
```

Figure 5: Target language bracketing based on the YCOE.

With the relations of hierarchy and linearisation that arise in figures 4-5, the representation of dependency put forward in Figure 8 has two main properties: explicitness and compatibility with the labeled bracketing provided by the YCOE. At

the present stage, the tags and relations of dependency include the ones listed in Figure 6.

DET	Determiner of
MOD	Modifier of
QUANT	Quantifier of
SUB	Subject of
OBJ	Object of
EXPLSUBJ	Expletive Subject of
EXPLOBJ	Expletive Object of
COMP	Complement of
ADV	Adverbial of
ADP	Adposition to
GVN	Governed by

Figure 6: Dependency tags and relations.

These tags and relations rely on a concept of dependency that involves argumenthood (Subject of, Object of, Expletive Subject of, Expletive Object of), complementation (Complement of), government (Governed by) and obligatoriness (Determiner of, Modifier of, Adposition to).

The annotation procedure calls for the manual selection of the relevant tag and relation from a scroll-down menu on a database implemented in Filemaker. This procedure, which is fully manual at the moment, will be partly automatised once a larger segment of the corpus has been processed, so that certain associations between lexical items and dependency tags and relations can be predicted on a statistical basis.

The inter-syntactic representation shown in Figure 7 resorts to graph theory in order to increase searchability and favour visualisation. In graph theory, a graph consists of vertices and tokens. Binary graphs relate two tokens to each other. In directed graphs the relationship holds in one direction only. In the inter-syntactic representation presented in Figure 8, each of the two constituents between which a relation of dependency holds is a node. The arc represents the dependency type. It is directed, which means that it points from the dependent to the head of the dependency relation. In Figure 7, higher arcs represent main sentence relations, while lower arcs are used for dependent clausal relations. The structural level of the clause is displayed over the linguistic segment and the level of the phrase is represented under the linguistic segment.

Graphs are generated with RAWGraphs from an Excel spreadsheet displaying the following columns: dependent, dependent token number, head, head token number, dependency tag and structural level. The data filed in the Excel spreadsheet is then imported to Filemaker, which allows for searches by lexical item (e.g. *hit*) and by dependency relation. Both types of searches can be simple (e.g. GVN) or complex (e.g. GVN and phrase level)

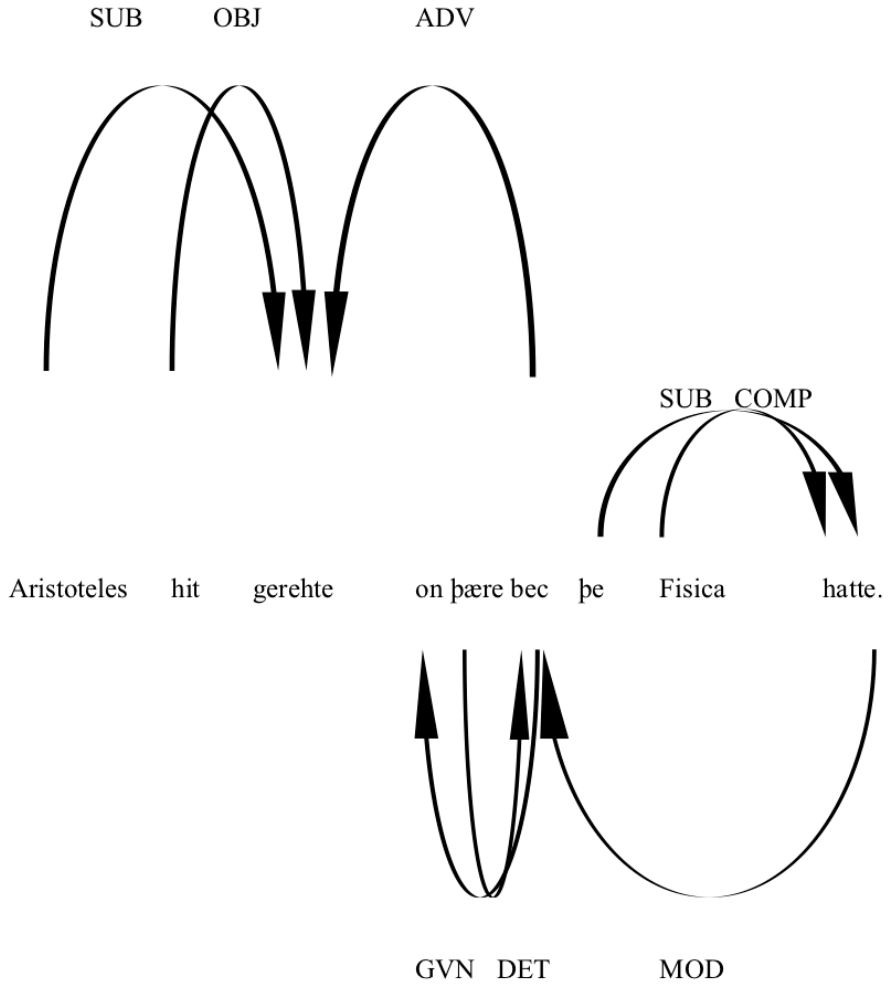

Figure 7: Inter-syntactic representation by means of a dependency tree.

The comparison of the structural description in figures 4, 5 and the dependency tree in Figure 7 indicates areas of stability as well as areas of change: whereas the phrasal and clausal relations of dependency remain, thus SUB, OBJ, ADV, COMP, GVN, DET, and MOD; change concentrates on the areas of morphological case (the accusative *hit*, the dative *bec* and the nominative *Fisica* are marked in the source language version) and linearisation (the Object *hit* and the Complement *Fisica* precede their respective verbs in the Old English text). The examples discussed above also display changes to the relations of dependency presented in Figure 7, which stresses the

need for an inter-syntactic model that specifies both clausal and phrasal dependency relations.

6. Conclusion and further research

This chapter has addressed the question of how to devise and implement an inter-syntactic model for ParCorOE that equips the corpus with syntactic annotation compatible with word alignment. The fact that ParCorOE is a corpus of intra-linguistic translation has guided this solution. On the one hand, two diachronic stages of English are compared, which predicts a considerable amount of convergence between the source language and the target language. On the other, alignment at word level calls for a level of correspondence that excludes local mismatches. The balanced solution described in this chapter restricts the syntactic annotation to the areas of divergence between the source and the target language.

Syntactic divergences have been explained on the basis of asymmetry and with respect to all structural levels: markedness asymmetry (generalised); constituency asymmetry (noun phrase, reflexive pronominal phrase, inflectional phrase, complementiser, conjunction); order asymmetry (noun phrase, prepositional phrase, inflectional phrase, adverbial phrase); and configuration asymmetry (noun phrase, inflectional phrase both active and passive, complementiser). The inter-syntax comprises the structural description of the source and the target language segments (with YCOE labels, in order to guarantee compatibility) as well as a dependency tree. The comparison of the structural description the dependency tree constitutes a historical micro-grammar, in the sense that it distinguishes syntactic stability from change,

including the change of dependency relations. The dependency tree is represented by means of graph theory so as to increase explicitness and to facilitate searchability. Overall, ParCorOE can be searched for text, fragment and token and, above all, for lexical items, morphological categories and dependency relations.

This model of inter-syntax has been found adequate to represent all the local mismatches that have arisen so far, but more research will be necessary as the corpus processing advances. In this line, the identification of more areas of asymmetry is pending. It also remains for future research to determine whether alignment at word level may increase the exhaustivity of annotation and boost automation: although the syntactic annotation procedure is manual at the moment, it is expected that it will be partially automatised in the near future.

Acknowledgements

This research has been funded through the grant FFI2017-83360P, which is gratefully acknowledged.

References

Aijmer, Karin, Bengt Altenberg, Mats Johansson, and Mikael Svensson (comp.) 1993-2001. *The English-Swedish Parallel Corpus*. Department of English, University of Lund and University of Göteborg.

Attenborough, Frederick L. (ed. and trans.). 1922. *The Laws of the Earliest English Kings*. Cambridge: Cambridge University Press.

Cameron, Angus, Ashley C. Amos, and Antonette diPaolo Healey (eds.). 2018. *The Dictionary of Old English in Electronic Form A-I*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Denison, David. 1993. *English Historical Syntax: Verbal Constructions*. London: Longman.

Enrique-Arias, Andrés. 2013. "On the usefulness of using parallel texts in diachronic investigations: insights from a parallel corpus of Spanish medieval Bible translations." In *New Methods in Historical Corpora,* ed. by Paul Durrell, Martin Scheible, Silke Whitt, and Richard J. Bennett, 105-116. Tübinguen: Gunter Narr.

Faaß, G. 2017. "Lexicography and corpus linguistics." In *The Routledge Handbook of Lexicography*, ed. by Pedro A. Fuertes-Olivera, 123-137. Abingdon: Routledge.

Fernández Cuesta, Julia, Nieves Rodríguez Ledesma, and Gloria Alvárez Benito (eds. and trans.). 1997. *Prosa anglosajona*. Sevilla: Universidad de Sevilla.

García Fernández, Laura. 2018. "Preterite-present verb lemmas from a corpus of Old English." In *Verbs, Clauses and Constructions: Functional and Typological Approaches*, ed. by Pilar Guerrero Medina, Roberto Torre Alonso, and Raquel Vea Escarza, 59-76. Newcastle: Cambridge Scholars Publishing.

Garmonsway, George N. (ed. and trans.). 1972. *The Anglo-Saxon Chronicle*. London: Dent & Sons LTD.

Godden, Malcom. 2016. *The Old English History of the World. An Anglo-Saxon Rewriting of Orosius*. Cambridge, Massachusetts: Dumbarton Oaks.

Godden, Malcom, Susan Irvine (eds.), with Mark Griffith, and Rohini Jayatilaka. 2009. *The Old English* Boethius*. Volume II*. Oxford: Oxford University Press.

Hanks, Patrick. 2012. “Corpus Evidence and Electronic Lexicography.” In *Electronic Lexicography,* ed. by Sylviane Granger, and Magali Paquot, 57-82. Oxford University Press.

Healey, A. diPaolo (ed.) with John P. Wilkin, and Xin Xiang. 2004. *The Dictionary of Old English Web Corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Heid, Ulrich. 2008. “Corpus linguistics and lexicography.” In *Corpus Linguistics. An International Handbook* (Volume 1), ed. by Anke Lüdeling and Merja Kytö, 132-153. Berlin: Mouton de Gruyter.

Hogg, Richard M. and Robert D. Fulk. 2011. *A Grammar of Old English. Volume 2: Morphology.* Oxford: Blackwell.

Johnson, B. 2009. *Using the Levenshtein algorithm for automatic lemmatization in Old English*. MA Thesis, The University of Georgia.

Krause, Thomas, and Amir Zeldes. 2016. ANNIS3: “A new architecture for generic corpus query and visualization.” *Literary and Linguistic Computing* 31(1): 118–139.

Kübler, Sandra, and Heike Zinsmeister. 2014. *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.

Leonard, Henry C. (ed. and trans.). 1881. *A Translation of the Anglo-Saxon Version of St. Mark’s Gospel*. London: James Clarke & Co.

Lu, Xiaofei. 2014. *Computational Methods for Corpus Annotation and Analysis*. Dordrecht: Springer.

Martín Arista, Javier. 2000a. “Sintaxis medieval inglesa I: complementación, caso y sintaxis verbal.” In *Lingüística histórica inglesa*, ed. by Isabel de la Cruz Cabanillas and Javier Martín Arista, 224-312. Barcelona: Ariel.

Martín Arista, Javier. 2000b. “Sintaxis medieval inglesa II: funciones, construcciones y orden de constituyentes.” In *Lingüística histórica inglesa*, ed. by Isabel de la Cruz Cabanillas and Javier Martín Arista, 313-377. Barcelona: Ariel.

Martín Arista, Javier. 2013. Nerthus. “Lexical Database of Old English: From word-formation to meaning construction.” Lecture delivered at the Research Seminar, School of English, University of Sheffield.

Martín Arista, Javier. 2017a. “Toward a parallel corpus of Old English prose. Preliminary questions and initial design.” Lecture delivered at the Departmental Colloquium Series at the Department of Language and Linguistic Science, University of York.

Martín Arista, Javier. 2017b. “The Nerthus Project at the crossroads. From lexical database to parallel corpus of Old English.” Lecture delivered at the 2017 International Conference of SELIM, held at the University of Málaga.

Martín Arista, Javier. 2018. “The design and implementation of a pilot parallel corpus of Old English.” In *Aspects of Medieval English Language and Literature*, ed. by Michiko Ogura and Hans Sauer, 111-134. Berlin: Peter Lang.

Martín Arista, Javier, and Ana E. Ojanguren López. 2018. “Doing Electronic Lexicography of Old English with a Knowledge-Base.” Workshop delivered at the CLASP Project (University of Oxford).

McEnery, Tony. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony 2003. “Corpus linguistics.” In *Oxford handbook of computational linguistics*, ed. by Ruslan Mitkov, 448-463. Oxford: Oxford University Press.

McEnery, Tony & Richard Xiao. 2007a. Parallel and comparable corpora: What are they up to? *Incorporating Corpora: Translation and the Linguist. Translating Europe*. Clevedon: Multilingual Matters.

McEnery, Tony, and Richard Xiao. 2007b. "Parallel and Comparable Corpora-The State of Play." In *Corpus-Based Perspectives in Linguistics*, ed. by Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori, and Yoichiro Tsuruga, 131-146. Amsterdam: John Benjamins.

Metola Rodríguez, Darío. 2017. "Strong Verb Lemmas from a Corpus of Old English. Advances and issues." *Revista de Lingüística y Lenguas Aplicadas* 12: 65-76.

Mitchell, Bruce. 1985. *Old English Syntax* (2 vols.). Oxford: Oxford University Press.

Novo Urraca, Carmen, and Ana E. Ojanguren López. 2018. "Lemmatising Treebanks. Corpus Annotation with Knowledge Bases." *RAEL* 17: 99-120.

Oksefjell, Signe. 1999. "A Description of the English-Norwegian Parallel Corpus." Compilation and Further Developments. *International Journal of Corpus Linguistics* 4(2): 197–219.

Pintzuk, Susan, and Leendert Plug (comp.). 2001. *The York-Helsinki Parsed Corpus of Old English Poetry*. Department of Language and Linguistic Science, University of York.

Rauer, Christiane. (ed. and trans.). 2013. *The Old English Martyrology*. Cambridge: D. S. Brewer.

Ringe, Don, and Ann Taylor. 2014. *A Linguistic History of English Volume II: The Development of Old English*. Oxford: Oxford University Press.

Rissanen Matti, Merja Kytö, L. Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Tertu Nevalainen and Helena Raumolin-Brunberg (comp.). 1991.

The Helsinki Corpus of English Texts. Department of Modern Languages, University of Helsinki.

Schierholz, Stefan J. 2015. "Methods in Lexicography and Dictionary Research." *Lexikos*: 25(1): 323-352.

Scrivner, Olga. 2015. "Tools for Digital Humanities: Parallel Corpus and Visualization." Paper presented at the Conference Corpora 2015, held at Saint-Petersburg, Russia.

Skeat, Walter W. (ed.) 1881. *Ælfric's Lives of Saints*. *Volume I*. Oxford: Oxford University Press.

Sweet, Henry (ed.). 1881. *King Alfred's West-Saxon Version of Gregory's* Pastoral Care. London: Trübner & Co.

Taylor, Ann, Anthony Warner, Susan Pintzuk and Frank Beths (comp.) 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Department of Language and Linguistic Science, University of York.

The Holy Bible Translated from the Latin Vulgate (Douay Rheims Version) 1971 (1899). Rpt. Rockford, Illinois: Tan books.

Thorpe, Benjamin. (ed. and trans). 1844. *The Homilies of the Anglo-Saxon Church. Volume I*. London: Red Lion Court.

Tío Sáenz, Marta. 2015. "The Regularization of Old English Weak Verbs". *Revista de lingüística y lenguas aplicadas* 10: 78-89.

Visser, Ferdinand. 1963-1973. *An Historical Syntax of the English Language* (4 vols.). Leiden: Brill.