The design and implementation of a pilot parallel corpus of Old English

Javier Martín Arista, Universidad de La Rioja

This article presents the pilot corpus on the basis of which the Parallel Corpus of Old English Prose will be compiled. Some conclusions drawn from the pilot corpus may guide the sources, method, and design of the final version. The most important is that the core database has to be organised by textual form so as to enhance the retrievability of information.

1. Introduction

This article discusses the principles that guide the design of a pilot parallel corpus of Old English and presents the preliminary version of the corpus, which is implemented on database software. The relevance of the undertaking lies in the lack of a large collection of texts with parallel translation for the study of Old English. On the theoretical side, the concept of parallel corpus is based on Aijmer and Altenberg (1996, in McEnery and Xiao 2007), while the idea that a pilot corpus should be compiled before the final corpus draws on Biber (1993). These questions are addressed in Section 2, which reviews previous research and sets the standards of the parallel corpus on the basis of the state of the art in parallel corpus design and compilation. On the applied side, the focus of the article is on the selection of the sources that allow for a maximal degree of information retrieval and automatisation. Two types of knowledge bases are distinguished, lexicographical knowledge bases and textual knowledge bases (Section 3), depending on whether they are lemmatised or not. The pilot corpus is shown in Section 4. Then, several aspects of text segmentation, translation, alignment, tagging, and annotation are considered, including the most frequent issues that have been encountered (Section 5). In Section 6, the conclusions insist on the consequences of the implementation of the pilot corpus for the compilation of an aligned parallel corpus of Old English.

2. Previous research and overview of the project

The most authoritative corpora in the field of Old English studies include the Old English segment of the *Helsinki Corpus of English Texts* (hereafter HC), which comprises around 300,000 words; *The York-Helsinki Parsed Corpus of Old English Poetry*, which consists of approximately 70,000 words; *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*, which contains around 1.5 million words (in the remainder of this article, the York corpora together are referred to as YCOE); and the *Dictionary of Old English Corpus* (henceforth DOEC), which was specifically compiled for the *Dictionary of Old English* and comprises about 3 million words. Some relevant aspects of these corpora are reviewed in turn.

The HC, YCOE, and DOEC are segmented by fragment and text, which the DOEC identifies by means of the short title and Cameron number (Mitchell et al. 1975, 1979). As illustration, Figure 1 shows the text name and the first three fragments (tokens) in the *Genesis*.

Short Title: GenA,B
Cameron number: A1.1
[000100 (1)] Us is riht micel ðæt we rodera weard, wereda wuldorcining, wordum herigen, modum lufien.
[000200 (3)] He is mægna sped, heafod ealra heahgesceafta, frea ælmihtig.

[000300 (5)] Næs him fruma æfre, or geworden, ne nu ende cymþ ecean drihtnes, ac he bið a rice ofer heofenstolas.
Figure 1. Text segmentation and identification in the DOEC.

The HC, YCOE, and DOEC provide annotation at text level, which, in the case of the HC, includes, for each fragment file, the abbreviated title, sub-period, manuscript date, dialect, text type, genre, and translation. This can be seen, with respect to *Medicina de quadrupedibus,* in Figure 2.

<B COQUADRU>
<Q O2/3 IS HANDM QUADR>
<N MEDIC QUADRUPEDIBUS>
<A X>
<C O2/3>
<O 850-950>
<M 950-1050>
<K NON-CONTEMP>
<D WS/A>
<V PROSE>
<T HANDB MEDICINE>
<G TRANSL>
<F LATIN>
Figure 2. Extract from the textual information found on *Medicina de quadrupedibus* in the HC.

The HC and the DOEC have been coded with XML, the metalanguage used for markup on the third-generation Internet, as specified for research in the Humanities by the Text Encoding Initiative (TEI). As is shown in Figure 3, the TEI allows corpus compilers to account for the various needs of textual encoding, such as the beginning and the end of a segment written in a foreign language (<foreign> ... .</foreign>), or the graphemes <æ> (&ae), <ð> (&eth), and <þ> (&thorn).

<s id="T02580000100" n="60.1"> <foreign>DOMINICA.I. IN QUADRAGESIMA.</foreign> MEN þa leofostan eow eallum is cuð. þæt ðes gearlica ymryne us gebrincð efne nu þa clænan tid lenctenlices fæstenes. on ðam we sceolon ure gymeleaste and forgægednysse urum gastlicum scrifte geandettan. and us mid fæstene. and wæccum. and gebedum. and ælmesdædum fram synnum aðwean. þæt we bealdlice mid gastlicere blisse ða easterlican mærsunge Cristes æristes wurðian moton. and þæs halgan husles þigene mid geleafan underfon. us to synne forgifennysse. and to gescyldnysse deofellicra costnunga;</s>
Figure 3. The beginning of Ælfric´s Homily for the First Sunday in Lent in XML (DOEC).

The YCOE is the only parsed corpus of Old English. That is to say, it provides morphological tagging and syntactic analysis. Consider, as illustration, the following fragment (in the remainder of this article, Old English textual fragments are identified with the DOEC short title and Cameron number): *Ærest hu Gotan gewunnon Romana rice, & hu Boetius hi wolde eft berædan, & ðeodric þa þæt anfunde, & hine het on carcerne gebringan* [BoHead 000100 (1)]. As can be seen in Figure 4, the morphological tagging displays lexical category and inflectional class, such as the present indicative verbal form *het* (VBDI) and the genitive proper name *Romana* (NR^G).

<T06650000100,1>_CODE +Arest_ADVS^T hu_WADV Gotan_NR^N gewunnon_VBDI
Romana_NR^G rice_N ,_, &_CONJ hu_WADV Boetius_NR^N hi_PRO^A wolde_MDD
eft_ADV^T ber+adan_VB ,_, &_CONJ +Deodric_NR^N +ta_ADV^T +t+at_D^A
anfunde_RP+VBD ,_, &_CONJ hine_PRO^A het_VBDI on_P carcerne_N^D
gebringan_VB ._. coboeth,BoHead:1.2_ID
Figure 4. Morphological annotation in the YCOE.

The YCOE syntactic analysis presented in Figure 5 specifies syntactic hierarchy, dependency and linearisation. For instance, the noun phrase immediately dominated by the node IP and case-marked nominative performs the function of subject, as in *Gotan gewunnon,* while the noun phrase immediately dominated by the node IP and case-marked accusative functions as the direct object, thus *Romana rice*. The noun phrase functioning as subject precedes the direct object and immediately precedes the finite verb.

```
( (CODE <T06650000100,1>)
  (CP-QUE (CP-QUE (ADVP-TMP (ADVS^T +Arest))
                  (WADVP-1 (WADV hu))
                  (C 0)
                  (IP-SUB (ADVP *T*-1)
                          (NP-NOM (NR^N Gotan))
                          (VBDI gewunnon)
                          (NP (NP-GEN (NR^G Romana))
                              (N rice))))
          (, ,)
          (CONJP (CONJ &)
                 (CP-QUE (WADVP-3 (WADV hu))
                         (C 0)
                         (IP-SUB (IP-SUB (ADVP *T*-3)
                                         (NP-NOM (NR^N Boetius))
                                         (NP-ACC (PRO^A hi))
                                         (MDD wolde)
                                         (ADVP-TMP (ADV^T eft))
                                         (VB ber+adan))
                                 (, ,)
                                 (CONJP (CONJ &)
                                        (IP-SUB-CON (ADVP *T*-3)
                                                    (NP-NOM (NR^N +Deodric))
                                                    (ADVP-TMP (ADV^T +ta))
                                                    (NP-ACC (D^A +t+at))
                                                    (RP+VBD anfunde)))
                                 (, ,)
                                 (CONJP (CONJ &)
                                        (IP-SUB-CON (ADVP *T*-3)
                                                    (NP-NOM *con*)
                                                    (NP-ACC-2 (PRO^A hine))
                                                    (VBDI het)
                                                    (IP-INF (NP-ACC *ICH*-2)
                                                            (PP (P on)
                                                                (NP-DAT (N^D carcerne)))
                                                            (VB gebringan)))))))
          (. .)) (ID coboeth,BoHead:1.2))
```

Figure 5. Syntactic parsing in the YCOE.

In spite of the wealth of philological data gathered in these corpora and the depth of the linguistic information available from their annotation and parsing, none of them contains a parallel version of the Old English texts that is aligned with the English. The compilation of such a corpus, therefore, represents a relevant project with various applications to the linguistic analysis and the lexicography of Old English.

According to Aijmer and Altenberg (1996, in McEnery and Xiao 2007: 131), parallel corpora offer a perspective on language comparison that can not be found in monolingual corpora. Indeed, parallel corpora allow researchers to address questions of language universals and typology, to compare source texts and target texts, and to assess the use of language by native and non-native speakers. Parallel corpora, moreover, have

various applications to lexicography, language teaching, and acquisition as well as translation. Several projects have been carried out aiming at the compilation of parallel corpora including English, such as the *English-Norwegian Parallel Corpus*, the *English-Swedish Parallel Corpus,* and the *Oslo Multilingual Corpus: English-Norwegian-German*, an extension of the *Norwegian Parallel Corpus* that also includes French, German, Dutch, Portuguese, Swedish and Finnish. On the other hand, a parallel corpus Old English-Contemporary English is a pending task in the field of English Historical Linguistics.

Such a project may draw on the corpora just mentioned and, above all, on the United Nations Parallel Corpus (UNPC), which constitutes the state of the art in parallel corpus compilation. According to Ziemsky et al. (2016), the UNPC consists of manual translations, from the years 1990-2014, of documents for the six official United Nations languages: Arabic, Chinese, English, French, Russian, and Spanish. The UNPC is based on an index and a concordance and is aimed at translation and language comparison. As the UNPC official web site explains (https://conferences.unite.un.org/uncorpus),

> All documents are organized into folders by language, publication year, and publication symbol. Corresponding documents are placed in parallel folder structures, and a document's translation into any of the official languages (if it exists) can be found by inspecting the same file path in the required language subfolder.

The UNPC has been encoded with XML-TEI markup language, in such a way that every XML file has embedded metadata, including text identifier, translation identifier, publication date, processing place and subject keywords. With a total of 799,276 texts, the UNPC holds 1,727,539 aligned document pairs. A fully aligned sub-corpus has been implemented that contains 86,307 texts, 11,365,709 lines and 334,953,817 English Tokens. Two types of alignment have been carried out. The full corpus compares documents at text level and the sub-corpus aligns documents at sentence level. In other words, the sub-corpus constitutes a fully aligned parallel text in which sentences are aligned across all the languages with the English reference text. As is the case with the HC and the 1981 release of the YCOE, the UNPC is available online in open access.

With these premises, the standards of a parallel corpus of Old English may be set as follows. As a general principle, the various tasks involved in the compilation of the project should be automatised to a extent compatible with accuracy. This said, a parallel aligned corpus Old English-English should comprise a parallel text, that is to say, an Old English text placed along its contemporary English translation, with alignment at word, fragment and text level, so that very source language chunk is paired with a target language chunk. Word, fragment, and text alignment requires segmentation at these three structural levels. The paring should be indicated by means of the highlighting of the source and the target chunk. A concordance and an index by word are needed in order to link the unlemmatised to the lemmatised part of the corpus (Sinclair 1991). The corpus should be provided with morphological tagging (making reference to linguistic information) and lemma annotation (including both linguistic and extra-linguistic information). A parallel aligned corpus should have a search engine that draws on a relational lexical database and offers search options by inflectional form and lemma, as well as by tagging and annotation category. The aim of any given search should be highlighted in the results. Finally, the corpus should be available and searchable online, in open access.

This concept of aligned parallel corpus relies on some basic choices. This project opts for a historical corpus (rather than a translation, comparative linguistics or second language learning corpus); a bilingual corpus Old English-Contemporary English; a unidirectional corpus Old English >>> Contemporary English; a token (textual form) and type (dictionary lemmas) corpus intended for both quantitative and qualitative analysis; and, more importantly, a dictionary and text based corpus that retrieves information from relational lexical databases in order to maximise the automatisation of the tasks of compilation, lemmatisation, annotation, and tagging. The scope, at least at this stage, is restricted to prose texts.

In sum, the ultimate aim of the undertaking is to compile a corpus compatible with theoretical studies as well as applications of Old English lexicography and presentations of Digital Humanities. From the compiler´s point of view, this corpus can combine the philological tradition (text-based) and the new paradigm of Historical Linguistics (corpus-based). The compiler of this kind of corpus, however, is likely to face the problems of data availability and textual transmission characteristic of Historical Linguistics; and to have a research agenda driven by orthographic, dialectal, and diachronic variants. From the user´s point of view, a parallel aligned corpus may provide a useful tool that supplements the information available from dictionaries, glossaries, thesauri and other corpora and can be used at the level of language learner or by the Historical Linguistics researcher.

To meet the standards described above, it may be useful to compile a pilot corpus that guides the design of the final version. In this respect, McEnery (1996: 123) stresses the importance of corpora of historical languages, which, as is the case with the corpora of natural languages, must be quantitatively sufficient and qualitatively representative in order to offer a representation as accurate and faithful as possible of the language of analysis. For Biber (2007), the compilation of a representative corpus must be stepwise. First of all, a pilot corpus must be designed and implemented that gathers as much variation as possible, in such a way that the compilers can identify specific issues and general problems. Then, the pilot corpus must be annotated and an empirical study in the pilot corpus must determine whether the design parameters are adequate. Only then can the compilation of the final corpus begin. Heid (2008: 43) calls the design and implementation of a pilot corpus *preprocessing* and considers it typical of a corpus-based rather than a corpus-driven approach. The following sections deal with these aspects, with a view to drawing conclusions on the adequacy of the corpus architecture as well as the performance of the compilation tasks, which are presented in Figure 6 and discussed in more detain in Section 4.

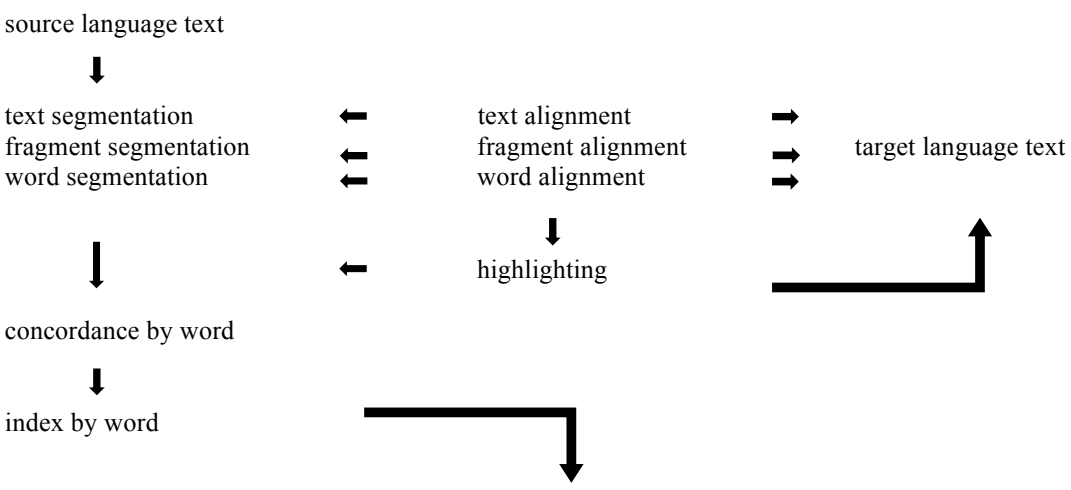


Figure 6. Flow chart of the tasks of corpus compilation.

3. Lexicographical and textual knowledge bases

The final version of the parallel corpus described in the previous section will be implemented on database software, which allows researchers to combine previous findings and new results into an explicit and systematic presentation of a large amount of heterogeneous data. Databases can be adapted to new findings or research aims, while their relational structure maximises the links between related data and enhances the retrievability of significant information. Moreover, databases developed with commercial software like Filemaker have online publication options and guarantee the access to information by means of relations and query functions.

Given these advantages, the *Nerthus* project (www.nerthusproject.com), which engages in the linguistic analysis and lexicography of Old English, has compiled a grid of relational databases by drawing on the available sources as well as on new research conducted with the project databases themselves. They constitute the knowledge bases on which the aligned parallel corpus of Old English prose will rely, both for storing and retrieving information for tagging and annotation, and for the automatisation of tagging, annotation, and lemmatisation.

The most relevant databases for the parallel corpus include *Nerthus* (ca. 30,000 files), which is geared to morphological and lexical analysis; *Freya* (ca. 35,000 files), which is oriented to the indexation of secondary sources; and *Norna* (ca. 190,000 files), a lemmatiser based on the textual attestations of the DOEC.

These databases represent two types of knowledge bases. *Nerthus* and *Freya* draw on the information available from lexicographical sources and, given that they display it as in the reference dictionaries, they are lemmatised. Thus, they can be considered lexicographical knowledge bases. *Norna*, on its part, is a lemmatiser based on a corpus and, therefore, can be described as a textual knowledge base, although it is probably more accurate to say that *Norna* relates, on the one hand, inflectional forms to lemmas; and, on the other hand, it links the types and tokens of such inflectional forms, in such a way that each token is presented in its context through the concordance that generates the index on which the lemmatiser is based. *Nerthus* and *Freya* rely on lexicographical sources (Bosworth-Toller, including the *Addenda* and the *Supplement*; Hall, with the *Supplement* by Merritt; and Sweet) and works in the Philology and Linguistics of Old English and, to a lesser extent, Germanic; whereas *Norna* inputs a concordance and an index to a corpus (the DOEC in this project), so as to assign inflectional forms to lemmas.

As has been remarked above, *Nerthus* is aimed at morphological and lexical analysis. Given a predicate like *linnan*, it is attributed to the lexical class of the verb, and to the morphological class of strong verbs (IIIa); its main forms are listed as *lann* (first preterite), *lunnon* (second preterite), and *lunnen* (past participle). It meaning definition, based on the dictionaries of Old English cited above, is rendered as ‘to cease from, leave off, desist; to yield up; to part from; to lose’. The strong verb *linnan* is described as a lexical primitive that is morphologically related by means of prefixation to the strong verbs *ālynnan* ‘to release’, *oflinnan* ‘to cease’, and *tōlynnan* ‘to take away’; and through prefixation and suffixation to the adverb *unoflinnedlīce* ‘unceasingly’. While basic (unprefixed) strong verbs are the point of departure of lexical derivation (Seebold 1970; Kastovsky 1992), adjectives like *unābindendlic* ‘indissoluble’ often represent the final

stage of word-formation processes. This is stated in *Nerthus* by relating this adjective to the lexical prime *bindan* 'to bind' and, ultimately, to other derivatives of this strong verb such as *bebindan* 'to bind in', *bindere* 'binder', *binding* 'binding', *gebundennes* 'obligation', *ungebunden* 'unbound', etc. The prefix *un-* in *unābindendlic* performs the Oppositive lexical function with respect to the adjective *ābindendlic,* which constitutes a lexical gap, in such a way that morphological relatedness has to be sought through the indirect evidence gathered from alternative spellings and variants such as *unonbindendlic* and *unanbindendlic*. The entry to the database *Nerthus* for *soðfæstness* 'truth' can be seen in Figure 7.

Nerthus. Lexical Database of Old English. Nerthus Project.
www.nerthusproject.com

predicate	sōðfæstnes	status	SUFFIXED
alternative_spellings	sōðfæstness (BT)	lexical_prime	SŌð 2/FÆST 1
category_of_predicate	noun	base	sōðfæst
ge	-	category_of_base	adjective
inflectional_morphology	f.	infl_class_of_base	weak and strong
inflectional_paradigm	e;	status_of_base	SUFFIXED
		derivational_function	PROP('X')
		affix	-NES
		affix_exponent	-nes
predicate_translation	truth, truthfulness, fairness, fidelity; justice	adjunct_of_compounding	
		adjunct_of_compounding_category	
predicate_translation_BT	I. truth, faithfulness, good faith, sincerity; II. truth, righteousness, justice; III. truth of speech or thought	derivational_paradigm	
predicate_translation_Nerthus	truth, truthfulness; faithfulness, sincerity, fidelity; justice, fairness, righteousness		

Figure 7. The database *Nerthus*.

Freya is, above all, a database for the indexing of the secondary sources of Old English. An indicative, non-exhaustive list of topics with just a few references per topic comprises, for example, Germanic etymology and Grammar (Krahe 1967; Orel 2003; Mailhammer 2007), Germanic dialectology (Nielsen 1998), Old English phonology and meter (Fulk 1992; Hogg 2011; Minkova 2014), morphology (Kastovsky 2006; Hogg and Fulk 2011), lexical borrowing (Feulner 2000), syntax (Koopman 1990; Elenbaas 2007; Fischer et al. 2011; Ringe and Taylor 2014), dialectology (Toon 1992); as well as specific topics like complementation (Molencki 1991), reanalysis (Allen 1995), grammaticalisation (Brinton 1996), lexicalisation (Brinton and Traugott 2005), and constructions (Ogura 1986; Möhlig-Falke 2012). Through indexing of the secondary sources, information on individual lexical items is gathered and filed in the database like the following. The class 1 weak verb *āgyltan* (alternative spelling *āgiltan*) 'trespass', is attested, according to the sources, in the inflectional forms *āgyltað, āgyldtæþ, āgyltæð, āgylte, āgulte, agilte, āgulten*, and *āgyltendra*. The secondary sources that deal with *āgyltan* include, among others, Sievers (1903: §405n11b), Hendrickson (1948: 45), Brunner (1965: §405, footnote 11), Horgan (1980: 129), Schwyter (1996: 37, footnote 46; 1996: 107), Dietz (2010: 586), and Hogg and Fulk (2011: §6.93). Figure 8 presents the entry to Freya for *andswarian* 'to answer'.

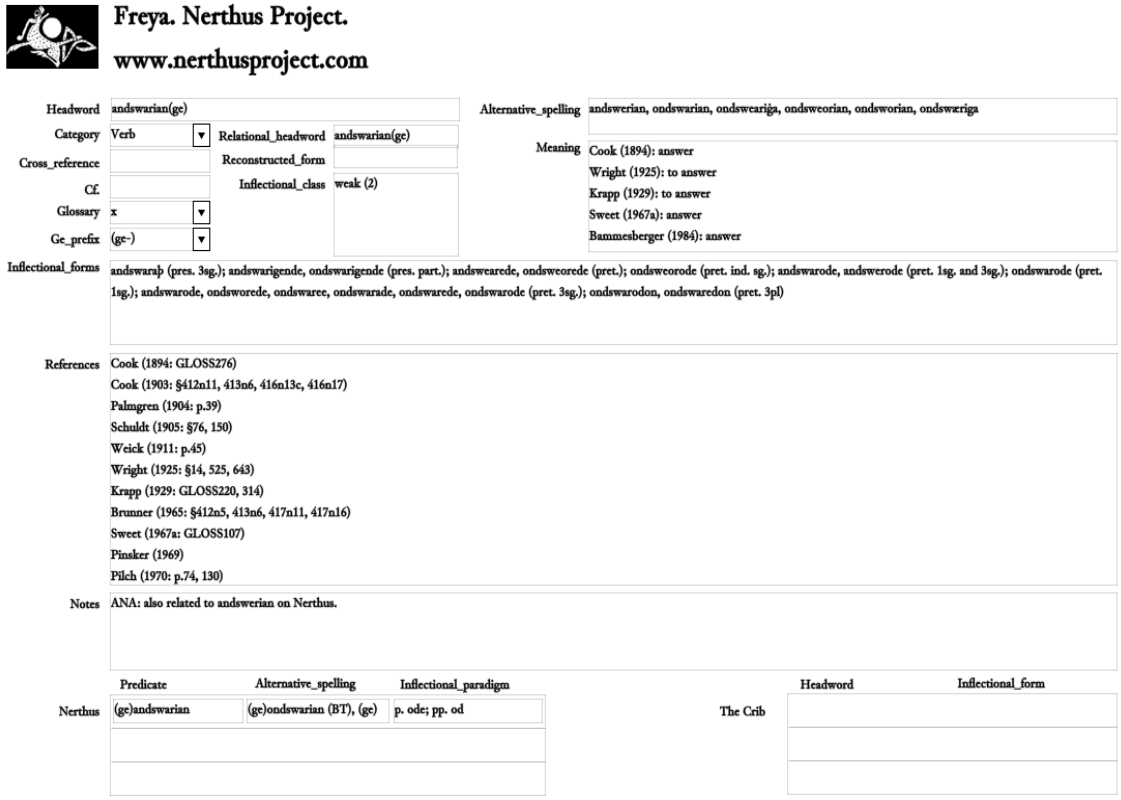
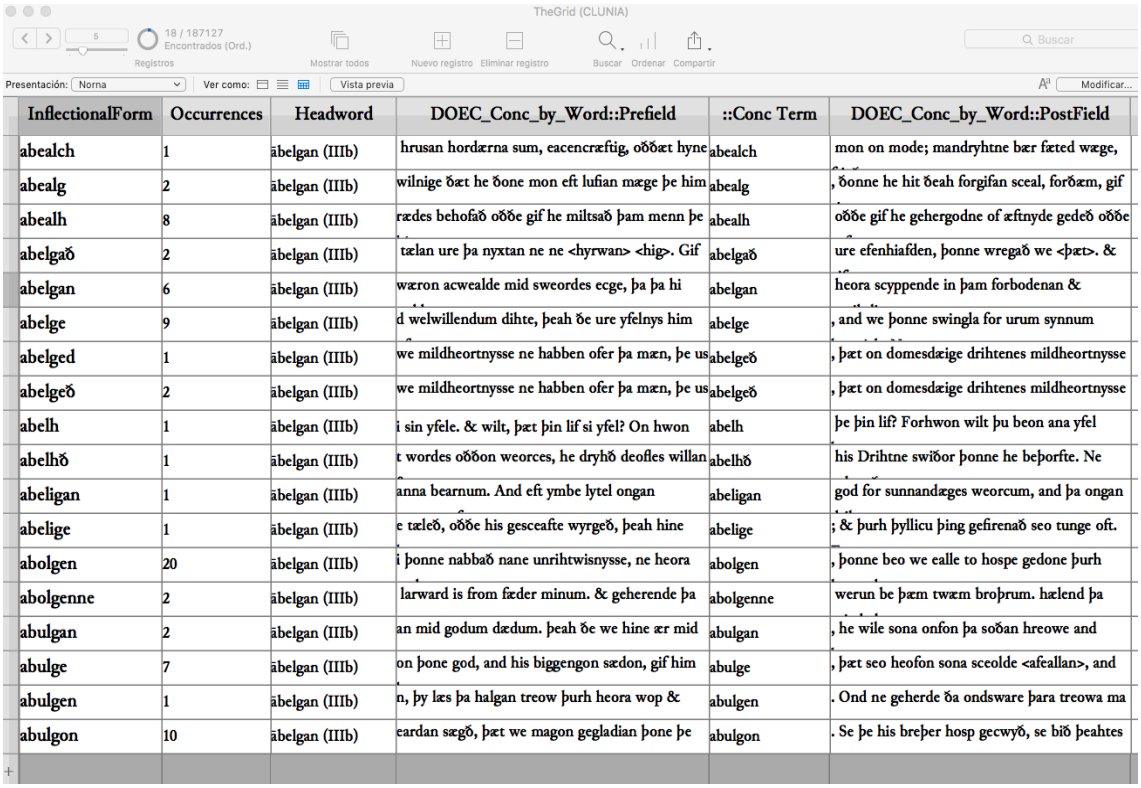


Figure 8. The database *Freya*.

The lemmatiser *Norna* is based on the DOEC. The corpus has been concorded by word and by fragment. The concordance by word consists of three million lines, one per word in the DOEC, while the concordance by fragment contains around two hundred thousand fragments of texts identified with Cameron number and short title. The word concordance has been indexed, in such a way that the resulting index comprises approximately one hundred and ninety thousand inflectional forms, which constitute the input to the lemmatiser. On the grounds of the distinction drawn above between lexicographical and textual knowledge bases, *Norna* belongs to the latter. Indeed, each inflectional form is provided with a textual frequency count, as well as its context, drawn from the concordance by word to the DOEC. This is shown in Figure 9, which displays the inflectional forms corresponding to the strong verb *ābelgan* 'to irritate'. Through a search algorithm targeting prefixes, stems and suffixes, *Norna* can assign lemma to the inflectional forms in the corpus. So far, a maximal accuracy of 80% before manual revision has been reached (Metola Rodríguez 2015). The following inflectional forms have been attributed to the lemma *(ge)tīlian* 'to provide': *getilað, getilaþ, getilian, getilien, getilige, getilod, tila, tilað, tilast, tilaþ, tiliað, tilian, tilianne, tiliaþ, tilie, tilien, tilienne, tiligað, tiligan, tilige, tiligeað, tiligean, tiligen, tiligende, tilode,* and *tilodon*.



InflectionalForm	Occurrences	Headword	DOEC_Conc_by_Word::Prefield	::Conc Term	DOEC_Conc_by_Word::PostField
abealch	1	ābelgan (IIIb)	hrusan hordærna sum, eacencræftig, oððæt hyne	abealch	mon on mode; mandryhtne bær fæted wæge,
abealg	2	ābelgan (IIIb)	wilnige ðæt he ðone mon eft lufian mæge þe him	abealg	, ðonne he hit ðeah forgifan sceal, forðæm, gif
abealh	8	ābelgan (IIIb)	rædes behofað oððe gif he miltsað þam menn þe	abealh	oððe gif he gehergodne of æftnyde gedeð oððe
abelgað	2	ābelgan (IIIb)	tælan ure þa nyxtan ne ne <hyrwan> <hig>. Gif	abelgað	ure efenhiafden, þonne wregað we <þæt>. &
abelgan	6	ābelgan (IIIb)	wæron acwealde mid sweordes ecge, þa þa hi	abelgan	heora scyppende in þam forbodenan &
abelge	9	ābelgan (IIIb)	d welwillendum dihte, þeah ðe ure yfelnys him	abelge	, and we þonne swingla for urum synnum
abelged	1	ābelgan (IIIb)	we mildheortnysse ne habben ofer þa mæn, þe us	abelgeð	, þæt on domesdæige drihtenes mildheortnysse
abelgeð	2	ābelgan (IIIb)	we mildheortnysse ne habben ofer þa mæn, þe us	abelgeð	, þæt on domesdæige drihtenes mildheortnysse
abelh	1	ābelgan (IIIb)	i sin yfele. & wilt, þæt þin lif si yfel? On hwon	abelh	þe þin lif? Forhwon wilt þu beon ana yfel
abelhð	1	ābelgan (IIIb)	t wordes oððon weorces, he dryhð deofles willan	abelhð	his Drihtne swiðor þonne he beþorfte. Ne
abeligan	1	ābelgan (IIIb)	anna bearnum. And eft ymbe lytel ongan	abeligan	god for sunnandæges weorcum, and þa ongan
abelige	1	ābelgan (IIIb)	e tæleð, oððe his gesceafte wyrgeð, þeah hine	abelige	; & þurh þyllicu þing gefirenað seo tunge oft.
abolgen	20	ābelgan (IIIb)	i þonne nabbað nane unrihtwisnysse, ne heora	abolgen	, þonne beo we ealle to hospe gedone þurh
abolgenne	2	ābelgan (IIIb)	larward is from fæder minum. & geherende þa	abolgenne	werun be þæm twæm broþrum. hælend þa
abulgan	2	ābelgan (IIIb)	an mid godum dædum. þeah ðe we hine ær mid	abulgan	, he wile sona onfon þa soðan hreowe and
abulge	7	ābelgan (IIIb)	on þone god, and his biggengon sædon, gif him	abulge	, þæt seo heofon sona sceolde <afeallan>, and
abulgen	1	ābelgan (IIIb)	n, þy læs þa halgan treow þurh heora wop &	abulgen	. Ond ne geherde ða ondsware þara treowa ma
abulgon	10	ābelgan (IIIb)	eardan sægð, þæt we magon gegladian þone þe	abulgon	. Se þe his breþer hosp gecwyð, se bið þeahtes

Figure 9. The lemmatiser *Norna*.

To summarise, lexicographical and textual databases contribute to the compilation of the parallel corpus in the following way. First of all, the lemmatiser *Norna* assigns lemma to the attested forms in the selection of texts. Then, the database *Freya* provides most of the information for the tagging (at inflectional form level) and annotation (at lemma level). The information on the meaning definition and the derivational morphology of the lemma is available from the database *Nerthus*. The contents of the annotation and tagging are discussed in more detail in the next section.

4. The pilot corpus

A pilot corpus has been compiled that contains around 10,000 words in the Old English part, as well as a gloss (word for word translation), and a translation into Contemporary English. The selection of texts includes fragments from the *Anglo-Saxon Chronicle*, *Orosius*, Ælfric´s *Lives of Saints*, *Cura Pastoralis,* and Bede´s *Ecclesiastical History*. The texts as well as the translation have been extracted from Fernández Cuesta et al. (1997). The choice of texts includes historical prose, religious prose and translations from Latin. From the dialectal point of view, all the texts in the selection belong to the West Saxon variety of Old English. Chronologically, Bede´s *Ecclesiastical History,* and *Cura Pastoralis* are early texts (9th. century); *Orosius* and this segment of the *Anglo-Saxon Chronicle* can be dated to the 10th. century; while the *Lives of Saints* is considered a late text, dated to the 11th. century. Quantitatively, 10,000 words may suffice to identify design shortcomings and, at least, the main issues of compilation. From the qualitative point of view, most written records of Old English correspond to the West Saxon variety, which is consistent with the choice made for the parallel corpus; moreover, the selection includes texts from three different centuries as well as various text genres.

Given the standards and the choices presented above that should govern the compilation of a parallel corpus, a crucial decision has to be made as to the structural level at which the corpus is organised. This is not only a theoretical question regarding

linguistic analysis, but also an applied aspect that determines the implementation of the database, whose architecture may revolve around the text, the fragment, or the word. Concerning the pilot corpus, the target of the description is the word, this structural unit representing the target of segmentation, alignment and highlighting. Nevertheless, in order to limit the number of files inputted to the database, the architecture of the pilot corpus is based on the fragment. In other words, the number of files in the database is equal to the number of fragments into which the texts have been segmented. This should be compatible with segmentation, alignment and highlighting at word level.

This said, the compilation of the pilot corpus has entailed the tasks described below.

Segmentation, in the first place, is the division of the source text into the units that are mapped to the target text. In this project, segmentation is required at two levels. Once texts have been divided into fragments, fragments are further segmented into orthographical words, with the exception of sets of two words from the same lexical category and adjacent to each other. The segmentation of the texts has been based on the DOEC so as to guarantee correspondence with the grid of databases presented above, whose core is a word and fragment concordance to the DOEC. With this segmentation, the database architecture consists of one file per (sub)fragment and several fields for the tagging and annotation of each word in the fragment, as is presented in Figure 10.

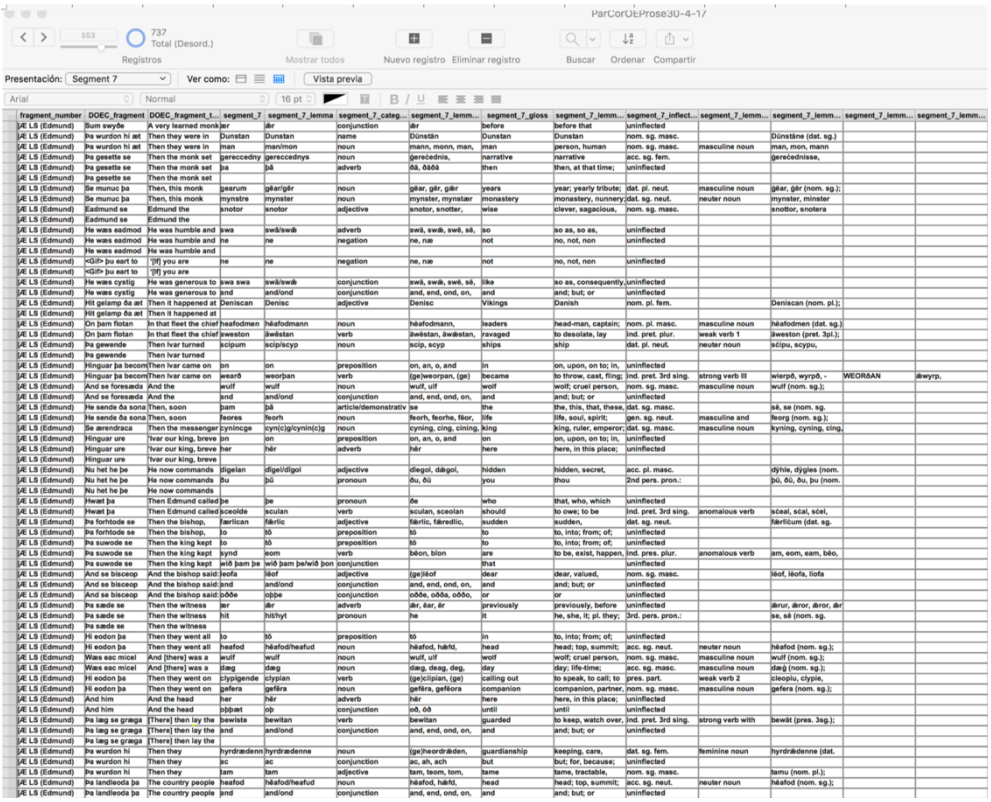


Figure 10. Database architecture, files and fields.

Secondly, aligning is mapping the units in the source text to the corresponding units in the target text. In general, parallel corpora opt for paragraph alignment, sentence alignment, expression alignment or word alignment, or a combination. Fragment alignment (one or more sentences) and word alignment have been chosen for this project. A file has been created for each fragment, whereas a different field has been defined for each word, in such a way that the correspondence between Old English and Contemporary English is established between the Anglo-Saxon word and its gloss. All the words in a

fragment provide the original text and the translation for the whole fragment. The source word and its gloss are highlighted. This can be seen in Figure 12.

In the third place, lemmatisation has been carried out to relate textual forms to dictionary forms. Once the textual form and the lemma have been identified, the former has been tagged and the latter has been annotated. The difference, therefore, between tagging and annotation in this project lies in context. Tagging is contextual, whereas annotation is non-contextual. Annotation also includes non-linguistic information, as is the case with the references of the secondary sources on the lemma in question. Tagging includes labels for morphological analysis, lexical class and gloss, while annotation offers information on the lemma as to alternative spellings, inflectional class, inflectional paradigm, derivational paradigm and secondary sources. For example, *swungon* is tagged for lemma (*swingan*), category (verb), inflectional class (strong III), inflectional form (2nd. person plural, preterite indicative), and gloss ('to beat'); and annotated for alternative spellings (*swingean, swyngean, swyngan*), meaning ('to swinge, beat, strike, smack, whip, scourge, flog, give a blow with the hand; to chastise, afflict; to swing oneself; to strike, dash; to beat the wings'), inflectional paradigm (first preterite *swang, swong*; second preterite *swungon*; past participle *swungen*; weak forms *geswinged, gesuincged, geswungdon, gesuuingde*), lexical prime (SWINGAN), and derivational paradigm (*āswengan, āswingan, beswingan, feorhsweng, framswengan, geswing, handgeswing, heaðusweng, heorusweng, hetesweng, oferswingan, ofswingan, swangettung, sweng, swengan, swenge, sweordgeswing, swingan, swinge, swingell, swingere, swinglung, tōswengan, tōswung, ðrēaswinge, wælsweng, windswingel,* and *wīteswinge*).

Fourthly, task automatisation has been maximised by retrieving the information required for the tagging of the inflectional form and the annotation of the lemma from the knowledge bases *Nerthus* and *Freya*. In practise, all the fields in the annotation part have been filled automatically, once the lemma has been assigned, by means of importation from the the knowledge bases. Automatisation has also been enhanced by assigning the lemmas available from the lemmatiser *Norna*. The highlighting of the Old English word and its Contemporary English counterpart has been fully automatic.

One of the main principles governing this project is the searchability of the corpus, which should be available, along with the corpus itself, in open access. A distinction has been made in this respect between the static and the dynamic presentations of the corpus. The static presentation coincides with the running texts Old English-English, including fragment, word-by-word gloss and translation. This can be seen in Figure 11.

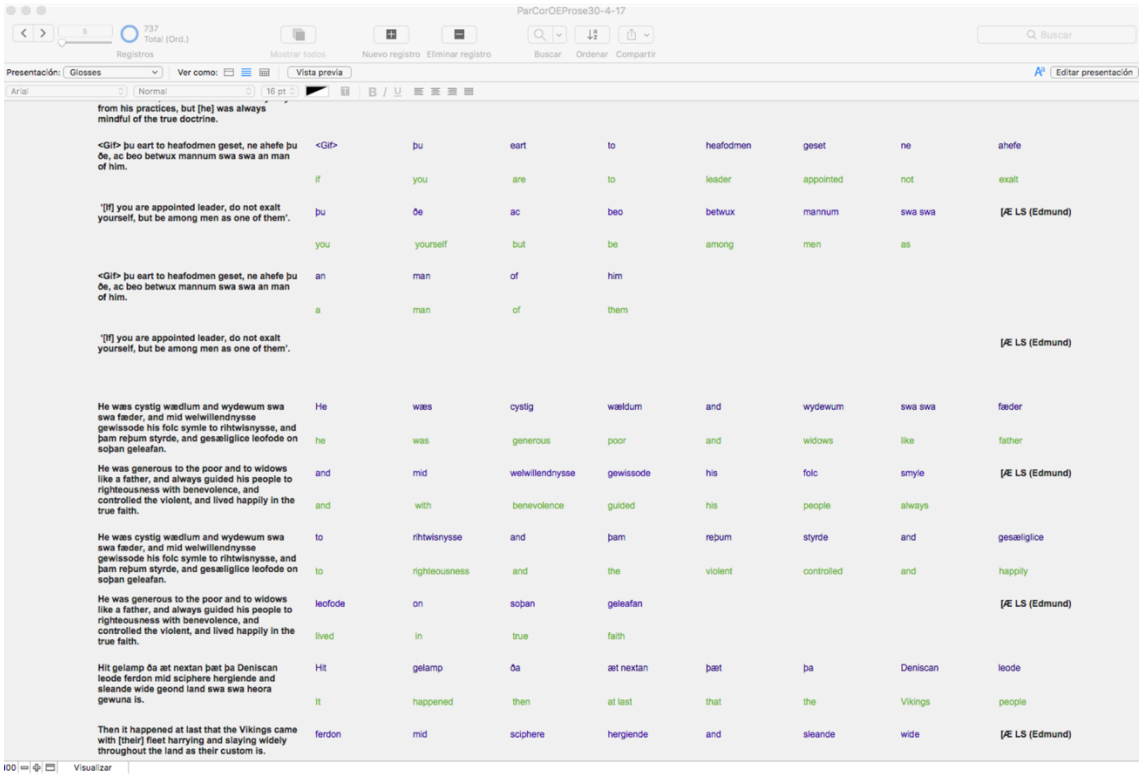


Figure 11. The static presentation.

The dynamic presentation of the corpus is geared to searches but also provides alignment and word highlighting. As can be seen in Figure 12, there are two basic query options, by inflectional form and by lemma. In order to facilitate user searches, a list of inflectional forms and lemmas in the corpus has been provided. The database software that files the corpus guarantees information retrieval because it permits simple searches (one criterion), combined searches (two or more criteria), and stepwise searches (the search of previous results). Not least, the query language includes wildcards.

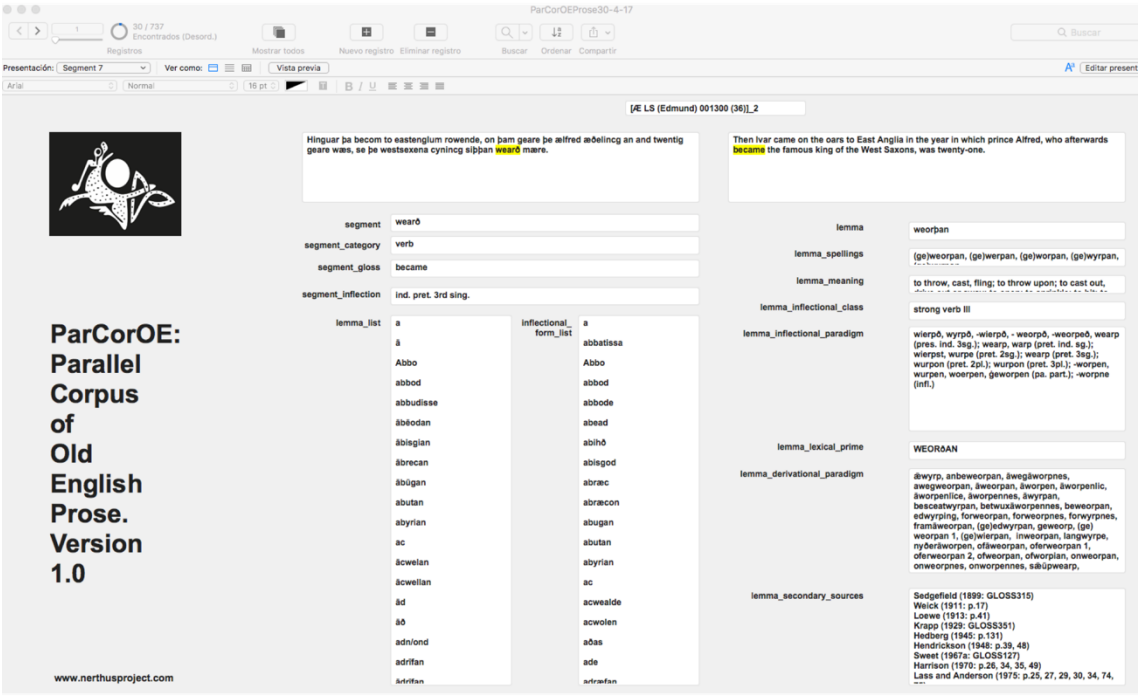


Figure 12. The dynamic presentation.

5. Some compilation issues

Some issues have arisen throughout the compilation of the pilot corpus, including aspects of alignment and translation. Beginning with alignment, certain linguistic aspects of Old English hinder alignment at word level. In some cases, the order of the words in Old English and English do not coincide, as is the case with verbs that take separable preverbal particles like *ūt* 'out' in example (1a) and compound tenses, as in example (1b). In other cases, the Old English sentence is in the active voice but a passive is required in the English version, as in example (1c), which stages the indefinite pronoun *mon* 'someone'. It is also sometimes the case that one more word is required in the English version than in the Old English one. This happens, for instance, in the absence of the formal subject *hit* 'it'.

(1)

a. [Or 1 014500 (1.17.27)]: *& þonne hys gestreon beoð þus eall aspended, þonne byrð man hine ut, & forbærneð mid his wæpnum & hrægle.*
'And when his goods are all spent in this way, then he (the dead man) is carried out and cremated with his weapons and his clothing.'

b. [ChronA (Bately) 042900 (893.11)]: *Hæfde se cyning his fierd on tu tonumen, swa þæt hie wæron simle healfe æt ham, healfe ute, butan þæm monnum þe þa burga healdan scolden.*
'The king had divided his army into two [sections], so that half of them were always at home and half out, besides those men who had to man the fortresses.'

c. [CPLetWærf 001800 (33)]: *Her mon mæg giet gesion hiora swæð, ac we him ne cunnon æfterspyrigean.*
'Their tracks can still be seen here (lit. here someone may yet see their tracks), but we are not able to follow them.'

d. [Æ LS (Edmund) 003900 (145)]: *Wæs eac micel wundor þæt an wulf wearð asend, þurh Godes wissunge to bewerigenne.*
'And [there] was a great miracle, that a wolf was sent, through God's guidance, to protect the head against the other wild animals by day and night.'

Turning to issues of translation, it must be remarked that translation-driven analysis is avoided, so that a segment such as *æt nextan* is glossed as 'at next' and translated as 'afterwards' but analysed as a preposition and an adverb. Apart from the relationship with analysis, issues of translations have to do with the lack of a direct correspondence between the source language and the target language. As shown in (2a), the Old English preterite with the adverb *ær* 'before' translates as the pluperfect. Similarly, in the absence of continuous tenses in Old English there is no word for word correspondence with English when a continuous tense is required by the translation. This is illustrated in (2b). Also related to verbal morphosyntax is the question of the subjunctive. Old English has a morphologically distinct subjunctive, which is lacking in English, so that the translation has to rely on modal auxiliaries, as *wære* 'would be' in (2c). Non-nominative subjects, such as the dative *me* 'to me' in (2d) also pose some problems for translation into English, because a formal subject is often needed.

(2)

a. [Æ LS (Edmund) 002700 (94)]: *Þa gewende se ærendraca ardlice aweg, and gemette be wæge þone wælhreowan Hingwar mid eallre his fyrde fuse to Eadmunde, and sæde þam arleasan hu him geandwyrd wæs.*
'Then the messenger went away quickly, and on the way he met the bloodthirsty Ivar with all his army hastening to Edmund, and told the wicked [man] how he had been answered.'

b. [Æ LS (Edmund) 003300 (123)]: *Betwux þam þe he clypode to Criste þagit, þa tugon þa hæþenan þone halgan to slæge, and mid anum swencge slogon him of þæt heafod, and his sawl siþode gesælig to Criste.*
'While he was still calling to Christ, the heathens drew away the saint to kill him, and struck off his head with one stroke; and his soul went blessed to Christ.'

c. [Æ LS (Edmund) 002500 (83)]: *Witodlice þu wære wyrðe sleges nu, ac ic nelle afylan on þinum fulum blode mine clænan handa, forðanþe ic Criste folgie, þe us swa gebysnode, and ic bliðelice wille beon ofslagen þurh eow gif hit swa god foresceawað.*
'Truly, you would be now worthy of death, but I will not defile my clean hands with your foul blood, because I follow Christ who thus set an example for us. And I will happily be slain by you if God so ordains.'

d. [CPLetWærf 002500 (49)]: *Forðy me ðyncð betre, gif iow swæ ðyncð, ðæt we eac sumæ bec, ða ðe niedbeðearfosta…*
'Therefore, it seems better to me, if it seems so to you as well, that we should also translate some books, those which are most needful for all men to know.'

6. Summary and conclusion

This article has presented the pilot corpus on the basis of which the Parallel Corpus of Old English Prose will be compiled. As was expected, the implementation of the pilot corpus has contributed in a significant way to some facets of the final sources, method and design. Given that searchability and automatisation have been identified throughout the article as principles guiding the corpus, the conclusions of the article necessarily make reference to these aspects.

With a view to limiting the number of files inputted to the database, the architecture of the pilot corpus is based on the fragment. However, for technical reasons related to the maximal number of fields that can be created for a given file, fragments have been further divided into sub-fragments containing a maximum of fifteen words. As it has turned out, this architecture does not allow for a direct link with the concordance by word, which undermines the searchability of the corpus. Therefore, the main conclusion that can be drawn from the implementation of the pilot corpus is that the core database has to be organised by textual form (word) rather than by fragment of the DOEC.

Regarding automatisation, this article has identified the sources of linguistic and metalinguistic information that will be used for the annotation of the lemmas of the corpus, including morphological and semantic aspects as well as the references to the secondary sources that deal with the lemmas in question. The information from these sources can be fed automatically into the database by means of relations between layouts and fields. Annotation is largely automatic even at the present stage but the automatisation of alignment is pending for future research. Lemmatisation and morphological tagging may be fully automatic in the near future, once the database of secondary sources *Freya* and the lemmatiser *Norna* have been completed.

Once compiled, the Parallel Corpus of Old English Prose may have several advantages over the corpora that have been reviewed in Section 2. Apart from the parallel text, the lemmatisation as well as the richness of tagging and annotation constitute assets of the project. On technical side, database software guarantees not only the retrievability of information, but also open access, user-friendly format and, above all, the possibility of revising and updating the corpus. On the other hand, some of the corpora reviewed above have strong points that are at the present stage out of the scope of this project, notably the syntactic parsing of the YCOE. Other questions, like the creation of a downloadable version of the corpus in XML markup language, remain a question for future research.

Acknowledgement
This research has been funded through the grants Q2618002F and FFI2017-83360-P (MINECO), which are gratefully acknowledged.

References

Corpora

Aijmer, K., B. Altenberg, M. Johansson, and M. Svensson (comp.)
1993-2001 *The English-Swedish Parallel Corpus*. Department of English, University of Lund and University of Göteborg.

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

Johansson S., K. Hofland, J. Ebeling, and S. Oksefjell (comp.)
1994-1997 *The English-Norwegian Parallel Corpus*. Department of British and American Studies, University of Oslo.

Johansson, S., and C. Fabricius-Hansen (comp.)
1999-2008 *The Oslo Multilingual Corpus*. The Faculty of Humanities, University of Oslo.

Pintzuk, S. and L. Plug (comp.)

2001 *The York-Helsinki Parsed Corpus of Old English Poetry*. Department of Language and Linguistic Science, University of York.

Rissanen M., M. Kytö, L. Kahlas-Tarkka, M. Kilpiö, S. Nevanlinna, I. Taavitsainen, T. Nevalainen and H. Raumolin-Brunberg (comp.)

1991 *The Helsinki Corpus of English Texts*. Department of Modern Languages, University of Helsinki.

Taylor, A., A. Warner, S. Pintzuk and F. Beths (comp.)

2003 *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Department of Language and Linguistic Science, University of York.

Studies

Allen, C.

1995 *Case Marking and Reanalysis. Grammatical Relations from Old to Early Modern English*. Oxford: Clarendon Press.

Altenberg, B. and K. Aijmer

1999 "The English-Swedish parallel corpus: A resource for contrastive research and translation studies." Paper presented at the ICAME XX Conference, held in Freiburg, Germany in May 1999.

Biber, D.

2007 "Representativeness in corpus design", in: W. Teubert and R. Krishnamurthy (eds.), *Corpus linguistics: Critical concepts in linguistics* (Vol. II). London: Routledge. 134-165.

Bosworth, J. and T. N. Toller

1973 (1898) *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.

Brinton, L.

1996 *Pragmatic Markers in English. Grammaticalization and Discourse Functions*. Berlin: Mouton de Gruyter.

Brinton, L. and E. C. Traugott

2005 *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Brunner, K.

1965 (1942) *Altenglische Grammatik*. Tübingen: Niemeyer.

Campbell, A.

1972 *An Anglo-Saxon Dictionary: Enlarged addenda and corrigenda*. Oxford: Clarendon Press.

Dietz, K.

2010 "Die Altenglischen Präfixbildungen un ihre Charakteristik", *Anglia* 128.3: 561-613.

Elenbaas, M.

2007 *The Synchronic and Diachronic Syntax of the English Verb-Particle Combination*. Utrecht: LOT.

Fernández Cuesta, J., N. Rodríguez Ledesma, and G. Alvárez Benito

1997 *Prosa anglosajona*. Sevilla: Universidad de Sevilla.

Feulner, A. H.

2000 *Die Griechischen Lehnwörter im Altenglischen*. Frankfurt am Main: Peter Lang.

Fischer, O., A. van Kemenade, and W. Koopman

2011 *The Syntax of Early English*. Cambridge: Cambridge University Press.

Fulk, R. D.
1992 *A History of Old English Meter*. Philadelphia: University of Pennsylvania Press.

Hall, J. R. C.
1996 *A Concise Anglo-Saxon Dictionary*. Supplement by H. D. Merritt. Toronto: University of Toronto Press.

Healey, A. diPaolo (ed.)
2016 *The Dictionary of Old English in Electronic Form A-H*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Heid, U.
2008 "Corpus linguistics and lexicography", in: A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook* (Volume 1). Berlin: Mouton de Gruyter. 132-153.

Hendrickson, J. R.
1948 *Old English Prepositional Compounds in Relationship to their Latin Originals*. Philadelphia: Linguistic Society of America.

Hogg, R. M.
2011 *A Grammar of Old English. Volume 1: Phonology*. Oxford: Wiley-Blackwell.

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

Horgan, D. M.
1980 "Patterns of Variation and Interchangeability in some Old English Prefixes", *Neuphilologische Mitteilungen* 81.2: 27-130.

Kastovsky, D.
1992 "Semantics and vocabulary", in: R. Hogg (ed.), *The Cambridge history of the English language I: The beginnings to 1066*. Cambridge: Cambridge University Press. 290-408.
2006 "Typological Changes in Derivational Morphology", in: A. van Kemenade and B. Bettelou Los (eds.) *The Handbook of the History of English.* Oxford: Blackwell publishing. 151-177.

Koopman, W. F.
1990 *Word Order in Old English: With Special Reference to the Verb Phrase*. Amsterdam: University of Amsterdam.

Krahe, H.
1967 *Germanische Sprachwissenschaft*. Berlin: Walter de Gruyter.

Mailhammer, R.
2007 *The Germanic Strong Verbs. Foundations and Development of a New System*. Berlin: Mouton de Gruyter.

Martín Arista, J. (ed.), L. García Fernández, M. Lacalle Palacios, A. E. Ojanguren López and E. Ruiz Narbona
2016 *NerthusV3. Online Lexical Database of Old English*. Nerthus Project. Universidad de La Rioja. [www.nerthusproject.com]

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

McEnery, T. and Z. Xiao
2007 "Parallel and Comparable Corpora-The State of Play", in: Y. Kawaguchi, T. Takagaki, N. Tomimori, and Y. Tsuruga (eds.), *Corpus-Based Perspectives in Linguistics*. Amsterdam: John Benjamins. 131-146.
Metola Rodríguez, D.
2015 *Lemmatisation of Old English Strong Verbs on a Lexical Database*. PhD Dissertation, Department of Modern Languages, University of La Rioja.
Minkova, D.
2014 *Historical Phonology of English*. Edinburgh: Edinburgh University Press.
Mitchell, B., C. Ball, and A. Cameron
1975 "Short titles of Old English texts", *Anglo-Saxon England* 4: 207-221.
1979 "Short titles of Old English texts: addenda and corrigenda", *Anglo-Saxon England* 8: 331-333.
Möhlig-Falke, R.
2012 *The Early English Impersonal Construction*. Oxford: Oxford University Press.
Molencki, R.
1991 *Complementation in Old English*. Katowice: Uniwersytet Śląski.
Nielsen, H. F.
1998 *The Continental Background of English and its Insular Development until 1154*. Odense: Odense University Press.
Ogura, M.
1986 *Old English 'Impersonal' Verbs and Expressions*. Copenhagen: Rosenkilde and Bagger.
Oksefjell, S.
1999 A Description of the English-Norwegian Parallel Corpus. Compilation and Further Developments. *International Journal of Corpus Linguistics* 4(2): 197–219.
Orel, V.
2003 *A Handbook of Germanic Etymology*. Leiden: Brill.
Ringe, D. and A. Taylor
2014 *A Linguistic History of English Volume II: The Development of Old English*. Oxford: Oxford University Press.
Schwyter, J. R.
1996 *Old English Legal Language - The Lexical Field of Theft*. Gylling: Odense University Press.
Seebold, E.
1970 *Vergleichendes und etymologisches Wörterbuch der germanischen starken Verben*. The Hague: Mouton.
Sievers, E.
1903 (1885) *Old English Grammar*. Boston: The Athenaeum Press. Translated by A. S. Cook.
Sinclair, J.
1991 *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sweet, H.

1976 *The student's Dictionary of Anglo-Saxon*. Cambridge: Cambridge University Press.

Toller, T. N.

1921 (1898) *An Anglo-Saxon Dictionary: Supplement*. Oxford: Clarendon Press.

Toon, T. E.

1992 "Old English Dialects", in R. Hogg (ed.) *The Cambridge History of the English Language I: The Beginnings to 1066*. Cambridge: Cambridge University Press. 409-451.

Ziemski, M., M. Junczys-Dowmunt, and B. Pouliquen

2016 The United Nations Parallel Corpus v1.0. Paper delivered at the *Language Resources and Evaluation Conference* (LREC'16), held in Portorož, Slovenia in May 2016.