CHAPTER SIX

PILOT STUDY ON OLD ENGLISH SUPERLATIVE ADVERBS LEMMATISATION

YOSRA HAMDOUN BGHIYEL

Abstract

The main of this paper is to describe and present the results of the lemmatisation process of the Old English superlative adverbs through a semi-automatic procedure. The main Old English corpora and dictionaries present certain limitations in this regard as they are not (fully) lemmatised. In addition, the authoritative dictionaries of this language do not list all the attested inflections of headword entries, except for the *Dictionary of Old English* (DOE), which is only available for letters A-I. With this mind, all the forms with the ADVS tag, corresponding to the superlative adverbs, provided by *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* and *The York-Helsinki Parsed Corpus of Old English Poetry* annotated corpora, have been lemmatised. The methodology for the extraction consists of an automatic search of the morphological tag ADVS in the POS (Part-of-Speech) files in the aforementioned corpora. In a second step, the resulting forms have been assigned a lemma through a manual procedure. In order to undertake this task, it has been drawn to the lexical database of Old English *Nerthus*, which has supplied the list of headwords. The standard dictionaries of Old English, including Sweet (1976), Bosworth-Toller (1973) and Clark-Hall (1996), have guided lemmatisation choices, the latter being especially reliable due to its consistent spelling and balance between early and late variants (Ellis 1993). The results obtained after the lemmatisation have been compared with the forms attested by Seelig (1930) and by the DOE. The discussion of the results insist, on the one hand, on the convenience of comparing sources to recognize missing forms and enrich the final inventory, and, on the other, on identifying the difficulties of the process and suggesting solutions.

1. Introduction

This study examines the lemmatisation process of the Old English adverbs in the superlative degree. Old English belongs to the Germanic branch of the Indo-European family. More specifically, it is the language spoken in the British Isles between the 5th and the late 11th centuries. Although the Old English period comprises almost five centuries, most of the surviving texts are copies made in the 9th, 10th and 11th centuries. The study of Old English presents certain limitations, mainly due to the absence of spoken evidences and its inconsistent spelling variation. This remarkable degree of spelling variation has its origins in the lack of a written standard and in the existence of different dialects (Kentish, West Saxon, Mercian and Northumbrian) at that time. All these reasons make it particularly necessary the lemmatisation of the Old English lexicon. Lemmatisation is understood as the process by which a group of words are morphologically related and reduced to a lemma or headword, including both the predictable and the unpredictable forms.

The main Old English lexicographical sources of reference, including Bosworth and Toller's (1973) *An Anglo-Saxon Dictionary,* Clark Hall's (1996) *A Concise Anglo-Saxon Dictionary* and Sweet's (1976) *The Students Dictionary of Anglo-Saxon*, compile neither a full inventory of inflectional forms nor in a systematic way. *The Dictionary of Old English* (DOE) is currently the most complete lexicographical source, however only letters A to I have been published so far. Lemmatisation is also a pending task of Old English historical linguistics, as there is no fully lemmatised corpus of this language. The main Old English Corpora include the *Helsinki Corpus of Old English texts* and the *Dictionary of Old English Corpus* (DOEC) that compile 300,000 words and 3 million words respectively, although they are not lemmatised yet.

The present study is framed within the *Nerthus* Project (Martín Arista et al. 2016), currently concerned with the lemmatisation of the Old English lexicon. Previous works in this area have focused on the lemmatisation of the verbal categories. This is the case of Metola Rodríguez's (2015, 2017) work on strong verbs, Tío Sáenz's (2019) weak verbs and García Fernández's (2018) preterit-present, anomalous and contracted verbs. The three authors employed a semi-automatic methodology that required manual revision. Their methodology guided the first steps of this research on the automatization of the lemmatisation process. This work is concerned with the lemmatisation of the non-verbal categories. To that aim, this pilot study has developed a new lemmatising methodology adapted to the requirements of a lexical class such as the adverbial one. The decision made on choosing inflected adverbs as the first non-verbal class to be lemmatised is conditioned by its relatively low number and the lesser degree of opaqueness if compared with other categories.

Old English adverbs present two inflections, one for the comparative and the other for the superlative degree; this paper will focus on superlative adverbs. The main contribution of this work is thus the identification of a lemma for all the superlative adverbial forms of Old English extracted from the *York-Corpus of Old English* (YCOE) (Taylor *et al*. 2003) and the collation of the resulting forms with a lexicographical and a secondary source in order to verify the results of the process. *The Dictionary of Old English* and *Seelig's* (1930) work on Old English adjectives and adverbs have served this purpose. The DOE is currently the most complete lexicographical source as it is based on the vastest Old English corpus, the *Dictionary of Old English Corpus* (Healey *et al.* 2009), which contains at least one example of every surviving text in this language. On the other hand, Seelig's (1930) work, has grouped together a considerable set of comparative and superlative forms of Old English adjectives and adverbs. Both sources will then be compared with the initial adverb list of inflectional forms and lemmas with the purpose of completing and refining the analysis.

The present papers structured in five sections: The first section introduces the Old English adverbial system. The second section maps this study within the fields of corpus linguistics and electronic lexicography. The third section describes in detail the methodology followed for both the extraction and the lemmatisation processes. The fourth section presents the lemmatised inflectional forms and discusses the contrastive analysis with Seelig and DOE that aims to validate the results. Lastly the fifth section includes the concluding remarks.

2. Old English and its Adverbial System

As remarked in the introduction, the multiple Old English dialects had a significant influence on the spelling variation. Although most of the Old English texts preserved were written in the West-Saxon dialect, texts present a notable orthographical variation. The foremost surviving Old English prose works include *The Anglo-Saxon Chronicle* and Ælfric's and Wulfstan's sermons. Most of the surviving poetry was found in four manuscripts: *The Exeter Book, The Vercelli Book, The Junius Manuscript* and *The Beowulf Manuscript.*

Old English is described as a synthetic language since, as Smith (2009:22) remarks, "there is a close relation between the form and the function of the words that is embodied in its rich use of inflections". Other authors prefer to characterize this language as a "half inflected" (Mitchell and Robinson 1985:62) one because, among other reasons, Old English preserves only four of the eight cases that existed in Indoeuropean and, besides, introduces prepositions in phrases that could stand alone.

Most of the everyday vocabulary of Present Day English (PDE) has its Old English correspondence, although there may be considerable differences in the spelling and the pronunciation. New words were coined in the language mainly through affixation, compounding and, to a lesser extent, borrowing. In this regard, Kastovsky (1992: 294) points out the associative character of this language and the semantic and formal transparency that exists in the morphologically related families. Figure one illustrates the distribution of Old English categories according to the *Nerthus* database (Martín Arista et al. 2016).

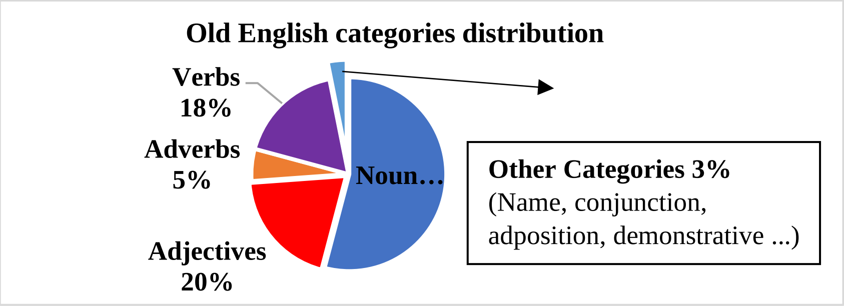


Figure 1: Categories distribution of the Old English Lexicon according to the *Nerthus* database (Martín Arista et al. 2016)

Half of the OE lexicon (54%) are nouns, adjectives and adverbs represent a 20% and a 18% each, whereas adverbs constitute only a five percent of the whole. This fact makes adverbs a suitable category to carry out a pilot study on the lemmatisation of a non-verbal category.

As in PDE, Old English adverbs also behaved as headwords in adverbial phrases and as modifiers of adjectives, adverbs or verbs. Most of these adverbs were created through the addition of the suffix ‘-e’ to an adjectival stem; for example, *dēop* ‘deep’ > *dēope* ‘deeply’; *biter* ‘bitter’ > *bitIre* ‘bitterly’. A substantial number of adverbs were created by adding the suffix *-e* to adjectives ending in *-lic*, hence the ending *-lice* which became widespread as an adverbial suffix. Adjectives as *heard*, for instance, derive adverbs with both endings: *hearde* and *heardlice*.

Adverbs can be classified according to their formation process. Based on this criterion, adverbs are basic, zero derived, affixed and compound

(Maíz Villalta 2010). In brief, basic adverbs are not derived by any productive morphological process; examples of basic adverbs are *hwanon* ‘whence’, *hwarne* ‘(not) at all’ and *nearwe* ‘narrowly, closely, strictly; carefully, exactly; oppressively, forcibly; artfully; anxiously’. Zero derived adverbs undergo category extension and semantic modification without formal change. These adverbs originate, mainly from adjectives, for example *æfter* ‘after’ (<*æfter* adj.), *efen* ‘evenly’ (<*efen* adj.) and *foreweard* ‘continually, always’ (<*foreweard* adj.). Affixation is by far the most productive process of derivation. Affixed adverbs are further divided into prefixed and suffixed. Examples of prefixed adverbs include ***ā***_būfan_ ‘above’, ***and***_ēages_ ‘eye to eye, openly’, ***eall***_rihte_ ‘just, exactly’, ***for***_hwon_ ‘wherefore, why, for what reason’. Examples of suffixed adverbs include _bæc_***ling*** ‘backwards, behind’, _æft_***um*** *‘after’*, _searw_***um*** ‘skilfully’. Finally, compound adverbs are formed by binding two lexeme stems from the same or different categories. The most common patterns identified by Maíz Villalta (2010) are noun+noun and adverb+adverb. Examples of compound adverbs are *ādunweard* ‘downwards’, *bitmǣlum* ‘bit by bit, piecemeal’, *ēastlang* ‘to the east, eastwards, extending east’, *hwīlhwega* ‘for some time’ and *hysewīse* ‘like young men’.

The regular comparative and superlative inflections for adverbs ending in *-e* are *-or* and *-ost* respectively: *gearwe* ‘well, certainly’- *gearor* – *gearost.* Other endings are also possible, for instance *-ur* and *-ar* for the comparative and *-ast, -est* and *-ust* for the superlative. Fulk (2018: 240) remarks that a few Old English adverbs -and also adjectives- form the superlative through double suffixation (*-m-ist-*). This is the case of _inne_***mest*** ‘innermost’ and _yfe_***mest*** ‘uppermost’. In this regard, Campbell (1959: 278) states that ending *-mest* is especially common when a comparative adjective is derived from an adverb, as *inne* ‘inside’ > *innerra* (adv./adj. comp.), *innemest* (adv./adj. superl.).

A group of adverbs (Campbell 1959: 278) undergo mutation in their root vowel when forming the comparative and the superlative. This is the case of *feorr* ‘far’ > _f_***ie***_rr_ – _f_***i***_rrest._ Other adverbs undergo suppletive comparison (Fulk 2018: 240) meaning, they form the comparative and the superlative from a stem that is different from that of the positive adverb. In Old English we find examples such as *yfle* ‘evil’ > *wiers* – *wierst*; *wel* ‘well’ > *bet/sēl* – *bet*I*st/best/sēlest*.

3. The interdisciplinarity of the field

There is no doubt that lexicography and corpus linguistics bear a close relationship. In this regard, any lexicographical work must be necessarily founded on a corpus. The following paragraphs will address the mutual dependence between both disciplines.

The evolution of lexicography has given rise to online historical lexicographical products of different types, including, for example, the third edition of the *Oxford English Dictionary*. As opposed to traditional lexicography, electronic lexicography provides the opportunity to look into "new kinds of evidence, new modes of description, new ways of organizing evidence, new possibilities for exploiting database structure and hypertext links, and the need for new theoretical foundations" (Hanks 2012: 1). The scope of lexicography is not restricted to the production of dictionaries but, as Hied (2008) states, it has numerous applications, including language processing systems, communication and knowledge-oriented purposes, translation, etc.

Corpus linguistics is defined by Rissanen (2008: 54) as the "linguistic study based on corpora". This author adopts a perspective based on the idea that corpus linguistics is a methodology of research rather than a linguistic discipline. From a historical perspective, a historical corpus aims at intentionally representing and investigating past stages of a language as well as studying language change (Claridge 2008: 242), regardless of whether these corpora are pre-electronic or electronic.

The computerization of electronic corpora has brought many advantages to the study of languages, one is the simultaneous study of both past and present stages of the same language which helps the development of diachronic studies. The use of electronic corpora provides a faster and more accessible path to the collection of evidence, more systematically, allowing for a more accurate analysis, and interpretation of the results. This is highly valuable especially for historical languages like Old English since they rely exclusively on written evidence.

Nevertheless, despite the valuable contributions of electronic lexicography, there are still unsolved problems that affect Old English and that are the consequence of the lack of spoken language evidences and of the enormous spelling variation. Furthermore, the fragmented and inaccurate material that has survived does not represent a full portrait of the society of that time.

It is widely accepted that corpora holding one million words can provide enough evidence of frequent grammatical phenomena, however this may prove insufficient when dealing with infrequent lexical features. There is a need to include more annotated material in Old English corpora, as no software can produce it automatically due to the high variability of the historical languages in contrast with PDE. Two corpora of Old English that have largely contributed to enhancing studies in the area of English historical linguistics and which also constitute an essential part in this research are *The Dictionary of Old English Corpus* and the *York-Toronto-Helsinki Parsed Corpus of Old English*, which comprises the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor *et al*. 2003) and the *York-Toronto-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug 2001).

4. Methodology

This section contextualizes the lemmatisation of the non-verbal categories within the frame of previous lemmatisation studies of verbal categories. Next, the characteristics of the main sources employed in this research are presented. Lastly, the lemmatisation process is described by focusing on both extraction and lemma assignment stages.

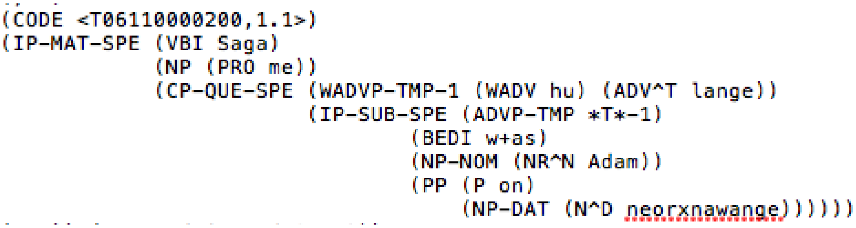
The present work follows the *Nerthus* research line on the lemmatisation of the Old English lexicon. Currently, only the verbal lexicon has been fully lemmatised through a semiautomatic procedure. Metola Rodríguez (2015, 2017), Tío Sáenz (2019) and García Fernández (2019) share a general course of action consisting of launching an automatic search on the database followed by the manual revision of the results obtained and a contrast with the available lexicographical sources. However, the three differ in more specific methodological decisions. Metola Rodríguez (2015, 2017) develops different search algorithms to look for strong verbs by prefix, stem, ablaut or inflectional ending. Tío Sáenz (2019) localised and lemmatised weak verbs based on the inflectional endings associated to their finite and non-finite verbs. Lastly, García Fernández (2018) focuses on the preterite-present verbs, anomalous verbs and contracted verbs, which have been lemmatised through queries that combine underived verbs with the prefixes, both in their canonical and attested variants.

The lemmatised forms have been filed in the lemmatiser *Norna*, a database responsible for the search of inflectional forms, the storage of lemmas and search refinement. *Norna* is part of the relational database *The Grid* (Martín Arista 2010), which also encompasses the lexical database *Nerthus* and the database of secondary sources *Freya*. *Nerthus* contributes with more than 30,000 predicates as well as with information regarding the alternative spellings, category, translation, inflectional morphology and inflectional forms of each citation form. *Nerthus*' list of headwords is based on *An Anglo-Saxon Dictionary* (Bosworth and Toller 1973), *The Student´s Dictionary of Anglo-Saxon* (Sweet 1976) and, above all, on *A Concise Anglo-Saxon Dictionary* (Clark Hall 1996).

4.1. Sources

The main objective of the lemmatising process is the compilation of attested forms under one entry in a corpus or database. Lemmatisation requires a previous collection of a healthy variety of sources that provide all the attested forms to be lemmatised and also a list of headwords. Besides, the eventual use of secondary and tertiary sources both enriches and refines the whole process.

The main sources that nurture this study are *Nerthus* database and the *York-Toronto-Helsinki Parsed Corpus or York Corpus of Old English* (YCOE). *Nerthus* is a lexical database of Old English that compiles headwords, attested spellings, translations, inflectional and morphological information, etc. The list of headwords is supplied by this database. The YCOE, which is divided into *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* and *The York-Helsinki Parsed Corpus of Old English Poetry*, has provided the set of inflectional forms that have been lemmatised. The corpus' prose segment comprises 1.5 million words and the poetry segment contains fifty thousand words approximately; both corpora are morphologically and syntactically annotated. An example of a string of words morphologically annotated is included in Figure 2.



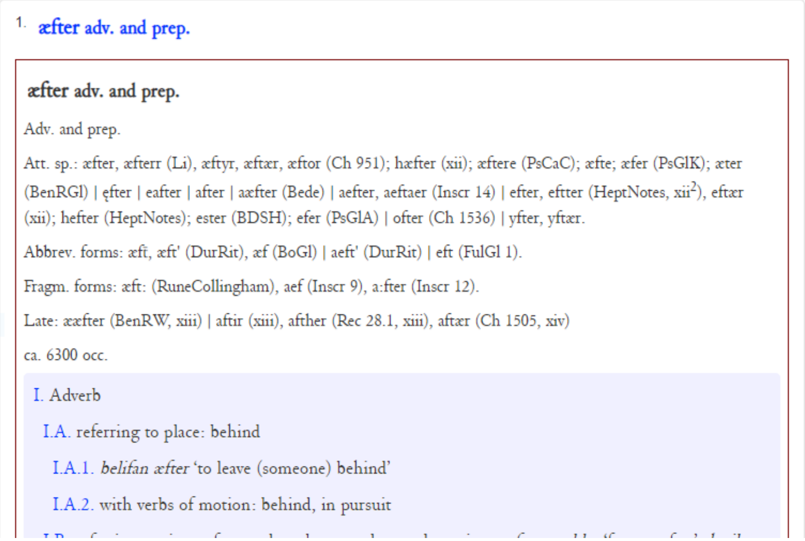
```
(CODE <T06110000200,1.1>)
(IP-MAT-SPE (VBI Saga)
            (NP (PRO me))
            (CP-QUE-SPE (WADVP-TMP-1 (WADV hu) (ADV^T lange))
                        (IP-SUB-SPE (ADVP-TMP *T*-1)
                                   (BEDI w+as)
                                   (NP-NOM (NR^N Adam))
                                   (PP (P on)
                                       (NP-DAT (N^D neorxnawange))))))
```

Figure 2: YCOE morphological annotation

The YCOE associates a morphological tag to each inflectional form. A simple search by tag in the corpus offers all the attested forms available. For the purpose of this study, only those items containing the tag ADVS, which stands for superlative adverb, have been extracted. It should be noted that although the extraction process was carried out in both the prose and poetry segments of the YCOE, only the prose segment attests comparative or superlative adverbs. No superlative adverbs were found in the twenty poetry texts of the YCOE.

The Dictionary of Old English and Seelig's work (1930) *Die Komparation der Adjektiva und Adverbien im Altenglischen* have served as contrastive secondary sources to validate the results of the lemmatisation.

The DOE has published A-I entries. Each of these entries offers detailed information related to part of speech, gender, grammatical class, attested spellings, dialectal variations, number of occurrences in the DOEC and citations from the corpus. Figure 3 below contains a screenshot of the DOE's entry for the word *æfter* as an adverb.



1. **æfter adv. and prep.**

æfter adv. and prep.

Adv. and prep.

Att. sp.: æfter, æfterr (Li), æftyr, æftær, æftor (Ch 951); hæfter (xii); æftere (PsCaC); æfte; æfer (PsGlK); æter (BenRGl) | ęfter | eafter | after | aæfter (Bede) | aefter, aeftaer (Inscr 14) | efter, eftter (HeptNotes, xii^2), eftær (xii); hefter (HeptNotes); ester (BDSH); efer (PsGlA) | ofter (Ch 1536) | yfter, yftær.

Abbrev. forms: æft̄, æft' (DurRit), æf (BoGl) | aeft' (DurRit) | eft (FulGl 1).

Fragm. forms: æft: (RuneCollingham), aef (Inscr 9), a:fter (Inscr 12).

Late: ææfter (BenRW, xiii) | aftir (xiii), afther (Rec 28.1, xiii), aftær (Ch 1505, xiv)

ca. 6300 occ.

I. Adverb

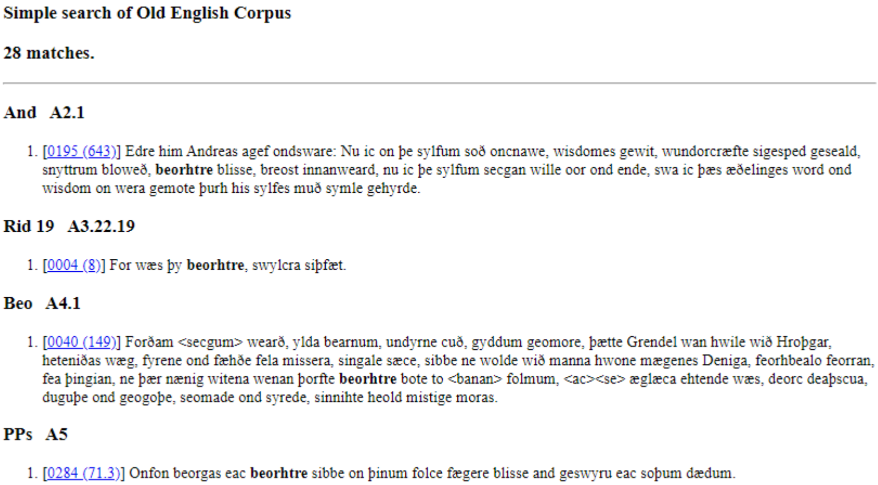
I.A. referring to place: behind

I.A.1. *belifan æfter* 'to leave (someone) behind'

I.A.2. with verbs of motion: behind, in pursuit

Figure 3: The entry for the adverb *æfter*.

The DOE draws its information from *The Dictionary of Old English Corpus*, a textual source that can also be accessed online and compiles at least one copy of every Old English surviving text. The DOEC registers around three million words in the language and almost a third, of this amount are Latin words. As displayed in Figure 4, a simple query for adverb *beorhte* retrieves all the occurrences of the adverb in the corpus.



Simple search of Old English Corpus

28 matches.

And A2.1

1. [0195 (643)] Edre him Andreas agef ondsware: Nu ic on þe sylfum soð oncnawe, wisdomes gewit, wundorcræfte sigesped geseald, snyttrum bloweð, **beorhtre** blisse, breost innanweard, nu ic þe sylfum secgan wille oor ond ende, swa ic þæs æðelinges word ond wisdom on wera gemote þurh his sylfes muð symle gehyrde.

Rid 19 A3.22.19

1. [0004 (8)] For wæs þy **beorhtre**, swylcra siþfæt.

Beo A4.1

1. [0040 (149)] Forðam <secgum> wearð, ylda bearnum, undyrne cuð, gyddum geomore, þætte Grendel wan hwile wið Hroþgar, heteniðas wæg, fyrene ond fæhðe fela missera, singale sæce, sibbe ne wolde wið manna hwone mægenes Deniga, feorhbealo feorran, fea þingian, ne þær nænig witena wenan þorfte **beorhtre** bote to <banan> folmum, <ac><se> æglæca ehtende wæs, deorc deaþscua, duguþe ond geogoþe, seomade ond syrede, sinnihte heold mistige moras.

PPs A5

1. [0284 (71.3)] Onfon beorgas eac **beorhtre** sibbe on þinum folce fægere blisse and geswyru eac soþum dædum.

Figure 4: Simple search of the adverb *beorhte* in the DOEC

The secondary source employed in this study to verify the results of the lemmatisation is Seelig's (1930) work *Die Komparation der Adjektiva und Adverbien im Altenglischen*. The author compiles a list of lemmatised comparative and superlative adjectives and adverbs from a variety of sources. In the second chapter of this work, Seelig addresses the comparative and superlative adverbs and divides them according to three categories, namely adverbs undergoing regular comparison, adverbs with vowel stem change and adverbs subjected to irregular comparison. According to Seelig (1930: 57-70), adverbs undergoing regular comparison form the comparative and superlative through the addition of the suffixes -*or* and -*ost*. He attests a total of 170 regular adverbs, among which we find *faestlice* 'fast' (*fæstlicor, fæsðlicor, fæstlicost, fæstlicast*), *gelómlíce* 'often' (*gelómlícor, gelómlícost*), *smale, smæle* 'small' (*smælor, smalost*). The second group consists of those adverbs whose comparative or superlative forms experiment vowel change. Only 9 paradigms belong to this group; examples of this type are *heah, hea* 'high' (*hearor, hyhst*) and *softe* 'soft' (*seft, softor, softost*), among others. Finally, the third group addresses irregular comparison, according to which the comparative and the superlative are formed from a different stem[1]; only six adverbs have been

[1] Seelig's (1930) irregular comparison coincides with the suppletive formation of the comparative and the superlative degree that Fulk draws on Campbell (Fulk 2018: 240)

identified by the author as undergoing irregular comparison, including *wyrs* ‘worse’ (*wærse, wiers, wirs, wyrs, wierst, wyrrest, wyrst*) or *sēl* ‘best’ (*sǣl, sēlast, sēlest, sēlost*).

4.2. Lemmatisation Procedure

The lemmatisation process requires, in the first place, a list of headwords or lemmas that can be assigned to the inflectional forms in question. The lexical database *Nerthus* has provided the list of adverbial headwords including two additional fields, alternative spellings and headword translation, which facilitate the task of lemma assignment. As Martín Arista (2011: 10) points out, spelling variants are neither independent predicates nor morphologically contrastive forms but, variants of the predicate they appear with. The translation field, in turn, provides an equivalent of the Old English word in Present Day English. The full list of adverbial headwords amounts to 1,755.

Secondly, *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* and *The York-Helsinki Parsed Corpus of Old English Poetry* constitute the primary source for the lemmatisation of these inflectional forms. This corpus is annotated both morphologically and syntactically. Figure 5 exemplifies morphological annotation through POS (Part-of-Speech) labels, while Figure 6 represents syntactic annotation through PAS (parsed) labels:

```
<T06950000200,3>_CODE He_PRO^N s+ade_VBD +I_ADV ,_, +t+at_C
+t+are_ADV^L w+aren_BEDS swy+de_ADV feawe_Q^N o+d+de_CONJ
nan_NEG+Q^N ,_, +te_C swa_ADV frig_ADJ^N w+are_BEDS ._.
Coaugust,Aug:3.3_ID
```

Figure 5: Part-of-speech (POS) annotation in the YCOE

```
        ((CODE <T06950000200,3>)
        (IP-MAT (NP-NOM (PRO^N He))
              (VBD s+ade)
              (ADVP (ADV +I))
              (, ,)
              (CP-THT (C +t+at)
                          (IP-SUB (ADVP-LOC (ADV^L +t+are))
                                      (BEDS w+aren)
                                      (NP-NOM (QP-NOM (ADV swy+de) (Q^N
feawe))
                                                  (CONJP (CONJ o+d+de)
                                                              (QP-NOM (NEG+Q^N nan)))
      (, ,)
                                                  (CP-REL (WNP-NOM-1 0)
                                                              (C +te)
                                                              (IP-SUB (NP-NOM *T*-1)
      (ADJP-NOM-PRD (ADV swa) (ADJ^Nfrig)) (BEDS w+are))))))
                                                                                    (. .)) (ID
Coaugust,Aug:3.3))
```

Figure 6: Parsed (PAS) annotation in the YCOE

The process of lemmatisation begins with the extraction of the material that will be eventually lemmatised. In order to extract the desired forms, all the words from the YCOE that contain the ADVS label have been exported with their corresponding tag and contextual information, i.e. the text name, code and genre, into an Excel file (see Table 1 below).

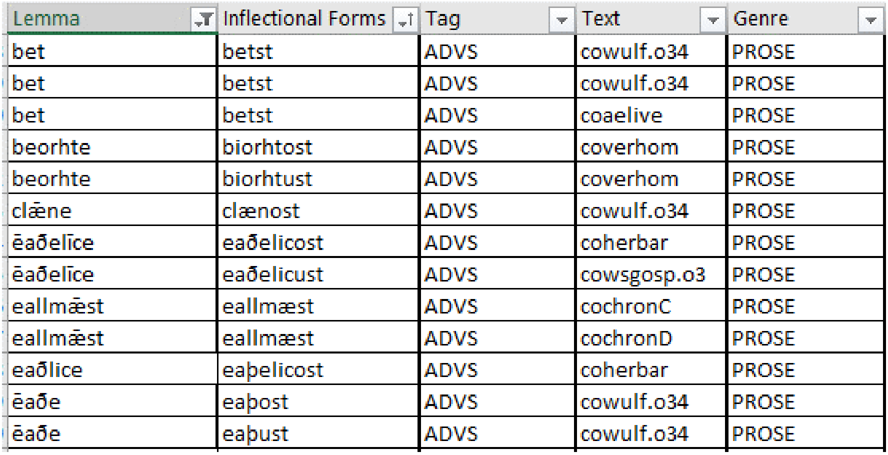
Form	Tag	Text	Genre
Ærest	ADVS^T	Covinsal	PROSE
Ærost	ADVS^T	Covinsal	PROSE
Oftost	ADVS^T	Covinsal	PROSE
Seldost	ADVS^T	coboeth.o.02	PROSE
Selest	ADVS	COBENRUL	PROSE
Teonlycost	ADVS	conicodA	PROSE
Ytemest	ADVS^T	cogregdH.o23	PROSE
Ytemest	ADVS^L	COBENRUL	PROSE

Table 1: Sample of extracted superlative adverbial forms

The first column in Table 1 lists all the extracted inflectional forms. The second column gives the morphological tag as presented in the YCOE. This tag may further specify the type of adverbial by adding ^L if the meaning is locative and ^T if it is temporal. This information proves particularly useful when translating adverbs as *ytemest*, which has a locative value, meaning covering a specific distance, and a temporal one, referring to lasting or taking a great amount of time. To recapitulate, only the words with the ADVS, ADVS^L and ADVS^T tags have been extracted.

To achieve a systematic and efficient extraction of the desired forms, I first opened the POS files with the text editor *Notepad++* because it allows to manage heavier files than other text editors. The first task is to carry out a preliminary search in each text in order to quantify the number of adverbs to be extracted per text. In this step, the twenty poetry texts were discarded as they did not register any hit containing the ADVS tag. In the next step, the files comprising more than ten or fifteen words underwent the following adjustments: by using the search and replace engine, the sequences +a, +d, +t were respectively replaced with æ, ð, þ; next, both small and capital RP+ and $ sequences were replaced by nothing; then, single spaces were replaced with a paragraph mark, giving rise to a column; next, double paragraph marks were replaced by a single paragraph mark. Furthermore, letter ‘þ’ was normalised to ‘ð’. Once all these procedures were carried out, the resulting list was sorted alphabetically. All the undesired results were manually eliminated, including the text code, stops, semicolons, commas, etc. The resulting list was then copied and pasted into the first column of the Excel file, each Excel page corresponding to a different text. Finally, the pasted column was divided into two columns, one containing the inflectional forms and the other the tag. Then all the data were selected and sorted by the morphological tag column. Two additional columns were added that include the name of the text and its genre so that the inflectional form can be easily tracked in case further clarification is required.

A total of 1,267 superlative adverbs were extracted from the YCOE. Lemmatisation is a lexicographical task that, as aforesaid, is far from being fully automatic. In order to tackle this task, each of the extracted inflectional forms have been manually assigned a lemma from the *Nerthus* headword list. In a first lemmatisation round, almost 80% of the inflectional forms were assigned a lemma, whereas 20% of these required deeper examination in order to find the appropriate lemma. In these cases, the DOE (for letters A to I) and the rest of Old English dictionaries of reference, above all Bosworth and Toller, were consulted to disambiguate doubtful cases. Table 2 exemplifies this stage of the process. The left-most column lists the lemmas assigned to the extracted forms, the second column comprises the extracted forms as provided by the YCOE, the third column corresponds to the morphological tag of each form and the right-most columns present the text code where the form is attested and its genre.



Lemma	Inflectional Forms	Tag	Text	Genre
bet	betst	ADVS	cowulf.o34	PROSE
bet	betst	ADVS	cowulf.o34	PROSE
bet	betst	ADVS	coaelive	PROSE
beorhte	biorhtost	ADVS	coverhom	PROSE
beorhte	biorhtust	ADVS	coverhom	PROSE
clǣne	clænost	ADVS	cowulf.o34	PROSE
ēaðelīce	eaðelicost	ADVS	coherbar	PROSE
ēaðelīce	eaðelicust	ADVS	cowsgosp.o3	PROSE
eallmǣst	eallmæst	ADVS	cochronC	PROSE
eallmǣst	eallmæst	ADVS	cochronD	PROSE
eaðlice	eaþelicost	ADVS	coherbar	PROSE
ēaðe	eaþost	ADVS	cowulf.o34	PROSE
ēaðe	eaþust	ADVS	cowulf.o34	PROSE

Table 2: Superlative adverbs assigned a lemma

The last stage of the process consists of analysing and validating the lemmatisation results. The DOE and Seelig have contributed to this task.

5. Results and Discussion

In this section, I will provide a brief quantitative summary of the results of the lemmatisation process. Next, I will discuss the contrastive analysis of the results with the DOE and Seelig (1930) for letters starting A to I forms in comparison with the ones collected by the two lexicographical sources. Finally, I will also discuss some of the difficulties and inconsistencies that arose during the lemmatisation process and its validation.

As a result, a total of 80 lemmas, out of the 1,755 headwords provided by *Nerthus*, were assigned to the 1,267 superlative adverbs attested in *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. The distribution based on the tags reveals that 477 forms are assigned the tag ADVS, while 762 superlatives have a temporal meaning (ADVS^T) and twenty-five a locative one (ADVS^L). Superlative adverbs with locative meaning include the forms *feorst, firrest, nehste, next* and *ytemest*. Superlatives with temporal meaning include the superlative forms of the adverbs *ǣr*, *fyrmest*, *fyrst*, *lange*, *late*, *leng*, *nēah, oft*, *seldor*, *sīð* and *ūt.*

Despite the fact that *Nerthus'* lemma list has proved to be suitable overall, it was necessary to resort to other sources when it was not possible to find an appropriate lemma or when more than one lemma could be assigned to a form. In these cases, the contrastive analysis with the DOE and Seelig (1930) helped to shed light on these doubtful cases. Table 3 illustrates how this contrastive was carried out. Notice that the DOE is the only source that maintains the distinction between spellings ð and þ.

Lemma: *hraðe*	**YCOE**	**DOE**	**Seelig (1930)**
Superlative	*hraðost, raðost, raðust, raðosð*	*hraþost, hraðost, raþost, raðost, raðosð; hradost, radost; hraþust, raþust, raðust, raðes*	*hraðost*

Table 3: Different inflectional forms attested by the sources

In Table 3 it can be perceived that the forms extracted from the YCOE are fully attested by DOE and not by Seelig (1930). In addition, DOE shows that there are forms that are not attested by the YCOE either and thus it explains how the contrastive analysis offers as well new forms to be considered in the inventory of the Old English Adverbial forms.

After having completed the lemmatisation of the superlative adverbs, the following forms remained unlemmatised: *eðost, leofost, liffest, liofast, suiðusð, suiðust,* and *ytemest*. To contrast all the superlative adverbs systematically, two additional columns were added in the Excel file comprising the superlative adverbs. Both columns served to indicate whether a lemma and an inflectional form were attested by the corresponding sources.

Lemma	Inflectional Forms	Tag	Text	Genre	Seelig	DOE
eallmǣst	eallmæst	ADVS	cochronC	PROSE	0	1
eallmǣst	eallmæst	ADVS	cochronD	PROSE	0	1
endemest	ændemest	ADVS	coboeth.o.02	PROSE	0	1
endemest	endemest	ADVS	coboeth.o.02	PROSE	0	1
ende-nēxt	endenexð	ADVS	coaelholm	PROSE	0	1
fægre	fægerost	ADVS	coverhom	PROSE	0	1
fæste	fæstost	ADVS	coaelive	PROSE	1	2
fæstlice	fæsðlicost	ADVS	cocuraC	PROSE	2	2
fullīce	fullicost	ADVS	cocuraC	PROSE	2	2
fyrmest	fyrmest	ADVS	coboeth.o.02	PROSE	1	2
fyrmest	fyrmest	ADVS	cocathom1	PROSE	2	2

Table 4: Contrastive analysis with sources

Three different colours were used to indicate similarities and differences in the attestation of forms by the sources. Red colour expresses that neither the headword nor the inflectional form have been attested by the source, as occurs with lemma *endemest* and the inflectional forms *endemest* and *ændemest*, none of which has been attested by Seelig. Green indicates that the lemma or the inflectional form have been attested by the source. For example, *eallmæst* has been attested by the DOE but not by Seelig. Finally, yellow reveals that a form is attested by a source but has not been associated with the same lemma. For instance, Seelig attests the inflectional form *fyrmest* but it is assigned the lemma *fyrm* instead. In those cases in which the DOE or Seelig suggest a different lemma for a form, the one suggested by *Nerthus* prevails over the rest.

In order to validate the results, the superlative forms starting with letters A-I have been contrasted with both the DOE and Seelig's work, while for those forms beginning with letters L-W only Seelig's work is available.

Lemmas starting with letters A to I amount to 886 tokens that have been mapped into 80 lemmas. 856 have been attested by DOE and only 36 by Seelig. The forms attested by both sources include the following inflectional forms of lemma *ær* 'before that, soon, formerly, before-hand, previously, already': *ærœs, ærast, æresð, ærest, ærst, ærust, æryst, ærost, æst, arest, erest, erost*; *deopost*, which is included under lemma *deope* 'deeply, thoroughly, entirely' ; and *geornost*, included under lemma *georne* 'eagerly, zealously, earnestly, gladly'.

Some the inconsistencies found during the lemmatisation of superlatives include the assignment of different lemmas if compared with the sources consulted. An example is the form *fægerost*, which is assigned the *Nerthus* lemma *fægre* 'fairly, elegantly', although DOE opts for *fægere.* Irregular or suppletive comparison posed a real problem in a few forms. In these cases, the help of the DOE and of Old English grammars such as Campbell (1959) made it easier to establish associations between *fyrrest* and lemma *feor* 'far, far away' or between *betest* and *betesð* and the lemma *wel* 'well'.

Another case that deserves attention is the form *inmest. Nerthus'* list suggests two possible lemmas, namely *in* and *inne*, meaning 'in, inwards, into inside'. In order to resolve this ambiguity, the form was searched in the DOE to verify the lemma that had been assigned to *inmest*, however the DOE presents it as an inflectional form in both entries *in, inn* and *inne*. The inflectional form was searched in the DOEC and the only occurrence belongs to a text from the Cura Pastoralis: *tihð his fet sua he* ***inmest*** *mæg* (CP B9.1.3 [1149 (35.241.7)]), which is one of the citations that appear under headword *in, inn* in the DOE.

This study contributes with three lemmatised inflectional forms that are attested by neither the DOE nor Seelig; these are: *eðost* (*ēað*e ‘easily, lightly’), *gewissost* (*wise* ‘wisely’) and *ðwyrlicost* (*ðwēorlīce* ‘insolently’). The DOE, in turn, has identified a total of six lemmas that are unattested by *Nerthus*; these are *andgietfullīce, bet, eallmǣst, endemest, endenēxt* and *fyrst*.

As for adverbs starting with letters L to W, they make up a total of 82 tokens that have been mapped into 40 lemmas. Most of these forms have been attested by Seelig. The contrastive analysis reveals that, on the one hand, Seelig (1930) contributes with the lemma, *wyrs* ‘worse’, which was not attested in *Nerthus* and which has been associated with the superlative forms *wierst, wyrest.* On the other hand, this research has identified new superlative forms unattested by Seelig, although their lemma has been collected by the author; these forms are: *geornlicast, geornlicest* (*geornlīce* ‘openly, manifestly’), *hatust* (*hate* ‘hotly’), *healicast* (*hēalīce* ‘highly, aloft’), *hluddost* (*hlūde* ‘loudly, aloud’), *længast, længest* (*lange* ‘long’), *nearwlicast* (*nearolīce* ‘narrowly, closely’), *raðust* (*hraðe* ‘hastily, quickly’), *rihtlicost* (*rihtlice* ‘justly, uprightly’), *swiðest, swiðosð* (*swiðe* ‘very much’), *teartlicost* (*teartlīce* ‘sharply, severly’), *ungeredelicost* (*ungerǣdlīce* ‘sharply, roughly’) and *widdast* (*wide* ‘widely’).

The comparison of the results obtained with Seelig’s work evinces that 78% of the inflectional forms are attested by the author. A total of fifteen forms have not been identified by Seelig, although their lemma is available in his work. Finally, 31 types are not compiled by the author.

The analysis and the sources available found it difficult to assign a lemma to the following forms: *eðost, suiðusð, suiðust* and *ytemest.* Their formal opacity and the lack of evidences in the sources consulted required deeper investigation. The solution adopted was to search the forms in *Freya*, a database of secondary sources that is part of the *Nerthus*, which revealed that *eðost* is a superlative form derived from *ēð*, which is an alternative spelling of *ēaðe*. In like manner, the form *utemest* derives from lemma *ut*, through double suffixation.

The other forms, *suiðusð* and *suiðust,* were examined in context so as to discover the appropriate lemma. From a formal perspective, two lemmas are possible: *sið* ‘late, afterwards’ or *suð* ‘southwards, south’. The sentences in which these forms confirm that these forms are associated with lemma *sīð*: *ðonne ðonne hie hie selfe suiðusð eaðmedað* (CP B9.1.3 [1457 (41.301.14)]) ‘Then they humbled themselves the latest’ and *ðeah ða tunga suiðust mænde* (CP B9.1.3 [1517 (43.309.8)]) ‘Still the tongues declare the latest’.

Finally, some searches in the online version of Bosworth and Toller's dictionary reveal that the forms *leofost, liffest* and *liofast*, which have been tagged as superlative adverbs by the YCOE, were not correctly analysed. These examples were examined in their context to check their lexical class: *þonne hit wære* ***leofost*** *gehealden* (Whom 13 B2.3.1 [0004 (12)]) 'and often it is more quickly lost when it is held dearest'; *min bearn* ***liffest*** *gedoan* (Ch 1510 (Rob 6) B15.6.27 [0002 (4)]) 'my child has done the quickest'; *swæ him* ***liofast*** *sie* (Ch 1510 (Rob 6) B15.6.27 [0004 (11)]) 'as it may best please them'. In all three cases, each form is the only adverbial attestation that appears in the corpus, whereas the rest of occurrences are adjectives. This may lead to think that these examples are also adjectives, though in these sentences they perform an adverbial function. In fact, a search in Bosworth and Toller's dictionary confirms this hypothesis.

6. Conclusions

The work presented here has aimed at defining and implementing a methodology for the lemmatisation of Old English adverbs in the superlative. A total of 1,267 superlative adverbs have been lemmatised into 80 lemmas provided by the *Nerthus* database through an automatic extraction and a manual lemma assignment procedure. A relevant contribution of this study is the identification of three inflectional forms that were attested by neither DOE nor Seelig; these are *eðost, gewissost* and *ðwyrlicost*. The analysis has also demonstrated that superlative adverbs are restricted to prose texts, as no evidences have been obtained from poetry. Once the methodology for superlative adverbs has proved viable, it remains for further research to apply this methodology to the rest of non-verbal categories that are unlemmatized. Ultimately, the more lemmatised forms a corpus has the more valuable it will be for a variety of studies, including textual frequency, spelling variation, syntactic complementation, etc.

References

Bosworth, J. and T. N. Toller. 1973 (1898). *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.

Campbell, A. 1959. *Old English Grammar*. Oxford: Oxford University Press.

Claridge, C. 2008. Historical corpora. In A. Lüdeling, M. Kytö and T. McEnery (eds.), *Corpus Linguistics. An international Handbook*. Berlin: Mouton de Gruyter. 242-254.

Dictionary of Old English: A to I online, ed. Angus Cameron, Ashley

Crandell Amos, Antonette diPaolo Healey et al. (Toronto: Dictionary of Old English Project, 2018

Fulk, R. D. 2018. *A Comparative Grammar of the Early Germanic Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 237-240

García Fenández, L. (2018). *The lemmatisation of the verbal lexicon of Old English on a relational database. Preterite-present, contracted, anomalous and strong VII verbs*. PhD Dissertation, Department of Modern Languages, University of La Rioja.

Hall, J. R. C. 1984 (1960). *A Concise Anglo-Saxon Dictionary.* With a supplement by H. D. Meritt. Toronto: University of Toronto Press.

Hanks, P. 2012. Corpus Evidence and Electronic Lexicography. In S. Granger and M. Paquot (eds.), *Electronic Lexicography*. Oxford University Press.1- 4.

Healey, A. diPaolo *et al.* (ed.) 2018. *Dictionary of Old English: A to I online.* Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Healey, A. diPaolo with John Price Wilkin and Xin Xiang. 2009. *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 A. Lüdeling, M. Kytö and T. McEnery (eds.), Corpus Linguistics. An international Handbook. Berlin: Mouton de Gruyter. 131-153.

Kastovsky, D. 1992. Semantics and Vocabulary. In R. M. Hogg (ed.), *Cambridge History of English Language. Volume I: The Beginnings to 1066*. Cambridge: Cambridge University Press. 290-408.

Maíz Villalta, G. 2012. *The Formation of Adverbs in Old English: A Functional View*. Master's Dissertation, Department of Modern Languages, University of La Rioja.

Mitchell, B. and F. Robinson. 1985 (1964). A Guide to Old English. Oxford: Blackwell.

Martín Arista, J. 2010. Building a Lexical Database of Old English: issues and landmarks. In Current Projects in Historical Lexicography. Newcastle: Cambridge Scholars Publishing. 1-33.

Martín Arista, J. 2011 Morphological relatedness and zero derivation in Old English. In C. Butler and P. Guerrero (eds.), Morphosyntactic Alternations in English. London: Equinox.

Martín Arista, J. (ed.), L. García Fernández, M. Lacalle Palacios, A. E. Ojanguren López and E. Ruiz Narbona. 2016. *Nerthus V3. Online Lexical Database of Old English. Nerthus* Project. Universidad de La Rioja [www.*Nerthus*project.com]´

Metola Rodríguez, D. 2015. *Lemmatisation of Old English Strong Verbs on a Lexical Database.* Ph Dissertation, Department of Modern Languages, University of La Rioja.

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

Rissanen, M. 2008. Corpus Linguistics and historical linguistics. In A. Lüdeling, M. Kytö and T. McEnery (eds.), *Corpus Linguistics. An international Handbook*. Berlin: Mouton de Gruyter. 53-68.

Seelig, F. 1930. *Die Komparation der Adjektiva und Adverbien im Altenglischen*. Heidelberg: Winter.

Smith, J. J. 2009. Old English. A Linguistic Introduction. Cambridge: Cambridge University Press.

Sweet, H. 1976 (1896). *The Student's Dictionary of Anglo-Saxon.* Cambridge: Cambridge University Press.

Taylor, A., A. Warner, S. Pintzuk and F. Beths (eds.) 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose.* Department of Language and Linguistic Science, University of York.

Tío Sáenz, M. 2019. A semi-automatic lemmatisation procedure: Old English class 1 and class 2 weak verbs. *Electronic Lexicography of Old English* 1: 67-76.