Lemmatising Treebanks. Corpus Annotation with Knowledge Bases

Lematización de *treebanks*. Anotación de un corpus con bases de conocimiento

CARMEN NOVO URRACA
ANA ELVIRA OJANGUREN LÓPEZ
UNIVERSIDAD DE LA RIOJA

Abstract: This article deals with Old English lexicography and corpus analysis. It aims at devising a lemmatisation procedure for a type of annotated and parsed corpus of Old English known as *treebank*. This study addresses two questions, namely where to find the data with which an Old English treebank can be lemmatised; and what procedure should be adopted to link the lemmatisation available from the sources to the treebank. On the grounds of the set of knowledge bases compiled by the *Nerthus* Project, a semi-automatic procedure for annotating *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* with lemma tags is devised, illustrated and assessed.
Keywords: *corpus, lemmatization, treebank, annotation, Old English*

Resumen: Este artículo se centra en la lexicografía del inglés antiguo y el análisis de corpus. El objetivo es definir un procedimiento de lematización para un tipo de corpus del inglés antiguo anotado y parseado conocido como *treebank*. Este estudio se centra en dos cuestiones, concretamente en indicar dónde se encuentran los datos con los que se puede lematizar el *treebank* del inglés antiguo; y qué procedimiento debe adoptarse para enlazar la lematización disponible en las fuentes con el *treebank*. A partir de las bases de conocimiento del Proyecto *Nerthus*, se diseña, pone en práctica y evalúa un procedimiento semiautomático para dotar *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* de etiquetas de lemas.
Palabras clave: *corpus, lematización, treebank/corpus parseado, anotación, Old English*

1. INTRODUCTION

As in other areas of Historical Linguistics (see, for instance, Haug, 2015), corpus compilation and corpus analysis are central tasks in the field of Old English studies.[1] Authoritative corpora like *The Helsinki Corpus of English Texts,* whose Old English segment comprises around 300,000 words, and *The Dictionary of Old English Corpus* (henceforth DOEC)*,* which contains around 3,000,000 words, have undoubtedly accounted for the advances in the study of the Anglo-Saxon language. Other widely used corpora in the field of Old English studies are *The York-Helsinki Parsed Corpus of Old English Poetry* (70,000 words), and *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (hereafter YCOE), which files around 1.5 million words. None of these corpora is lemmatised, though. With the exception of the York corpora, the others are not annotated, either. The poetry and the prose segments of the York Corpus are tagged morphologically and parsed syntactically, although this does not include the assignment of lemma to the inflected attestations that appear in the corpus texts. Put in other words, verbal forms like *blawe, blaweð, blawað, blawen*, etc. are related neither to the class VII strong verb lemma *blāwan* 'to blow' nor to one another.

This said, the other main sources of philological data, along with corpora, are dictionaries. The historical linguist of Old English can resort to corpora and dictionaries, but finds it difficult to use both together. This is so because, on the one hand, Old English dictionaries do not give all the inflections of headword entries and, on the other hand, corpora are not lemmatised, as has been remarked above. This means that, in practice,

[1] This research has been funded through the grant FFI-2017-83360-P, which is gratefully acknowledged.

corpora and dictionaries cannot be exhaustively exploited for researching Old English because the link inflectional form-lemma (or corpus word-dictionary word) is partly missing.

The Dictionary of Old English (DOE) is an exception to what has just been said about the listing of inflectional forms in dictionaries of Old English. It presents its headword entries with all the attested inflections of the headword. For example, in the entry to *blāwan* 'to blow', the DOE also includes canonical forms as well as less predictable forms like *blau, bleowun, blewon, blewan*, etc. This would solve the problem of the link corpora-dictionaries on the side of lexicographical sources if the DOE was complete, but its publication has just reached the letter H. Therefore, when dealing with sets of corpus forms beginning with the letters I-Y, like *oferhogie, oferhogað, oferhogian, oferhogodon, oferhogod, oferhogiað, oferhogienne, oferhogode, oferhogodest, oferhogoden*, etc., the only information available is found in dictionaries (Bosworth-Toller, 1973; Clark-Hall, 1996; Sweet, 1976) that, as a general rule, do not list inflectional forms other than those included in the citations that illustrate the meanings of the word.

With this state of play, the field of Old English, and English Historical Linguistics in general, would benefit from advances in the lemmatisation of the existing corpora. This article may be a further step in this direction. Its aim is to devise a lemmatisation procedure for a corpus of Old English. Considering that the YCOE is annotated for morphology and syntax, it represents the best candidate for the undertaking. Therefore, in the rest of this article a semi-automatic procedure for annotating the YCOE with lemma tags is devised, illustrated and assessed. The YCOE is lemmatised with the information available from the knowledge bases of the *Nerthus* Project (www.nerthusproject.com), including a dictionary database, a database of secondary sources and a lemmatiser.

With this aim and method, this article may contribute to the research in the linguistic analysis of Old English with corpus-based lexical databases conducted, among others, by García García (2012, 2013), González Torres (2010a, 2010b, 2011), Martín Arista (2012a, 2012b, 2013a, 2014, 2017a, 2017d.), Martín Arista and Cortés Rodríguez (2014), Martín Arista and Vea Escarza (2016), Mateo Mendaza (2013, 2014, 2015a, 2015b, 2016), Novo Urraca (2015, 2016a, 2016b), Torre Alonso (2011a, 2011b) and Vea Escarza (2012, 2013, 2014, 2016a, 2016b.). The article is also likely to underline points of contact with the treebanks project, as represented, for instance by Taylor, Warner, Pintzuk and Beths (2003) and Taylor, Marcus and Santorini (2003).

2. PREVIOUS RESEARCH

This section reviews the two components of the proposal that is advanced below. In the first place, treebanks, as represented by the YCOE, are considered. Secondly, the sources and steps of annotation are discussed, including lemmatisation with knowledge bases.

A treebank is a corpus annotated with sentence structures (Nivre, 2008: 225), including, among other aspects, the distinction of boundaries between clauses and phrases, constituent structures, and dependency structures (Rosén et al., 2005). Two types of syntactically annotated corpora can be distinguished. The application of advances in syntactic theory to corpus design led to the compilation of parsed corpora, which are explicitly based on a computational model of grammar (Abeillé, 2003). Parsed corpora are often the result of automatic analysis, whether there has been manual post-editing or not. Unlike parsed corpora, tree banks combine *automatic analysis and manual work in order to make the process as efficient as possible while maintaining the highest possible accuracy* (Nivre, 2008: 234). In the compilation of treebanks, in other words, there is agreement on the fact that some degree of manual disambiguation is necessary (Rosén et al., 2005). Apart from the question of automatisation, it is worth pointing out that

treebanks are compatible with grammars and lexicons and admit various layers of annotation (Hajičová et al., 2010).

Marcus et al. (1993) describe the compilation of *The Penn Treebank Corpus*, which files 4.5 million words of American English and is annotated both for part of speech and syntactic structure. Part of speech tagging and syntactic bracketing was automatic, with manual revision. This combined method was preferred for reasons of speed, consistency and accuracy (Marcus et al., 1993: 313). As Taylor et al. (2003) remark, two types of syntactic parsing have been used throughout the project, depending on the degree of complexity: skeletal parsing, which displays standard syntactic labels, and predicate-argument structure, which allows functional labels and null elements.

The aims and method of treebanks have been applied to Old English, so that two much used corpora have been compiled and annotated, the YCOE and its poetry counterpart. As has been said above, these corpora have two levels of annotation, POS (part of speech) annotation and PAS (parsed) annotation. For example, a noun phrase like *lyfiendan gast* 'living spirit' in the context *Hi ealle gelifflæste þurh þone lyfiendan gast* is annotated as presented in Figure 1 (morphological tagging) and Figure 2 (syntactic parsing).

```
&_CONJ hi_PRO^N ealle_Q^N geliff+aste_VBD +turh_P +tone_D^A
lyfiendan_VAG^A Gast_N^A :_. coaelhom,+AHom_1:70.49_ID
```

Figure 1. ***POS tagging in the YCOE****.*

```
( (IP-MAT (CONJ &)
          (NP-NOM (PRO^N hi) (Q^N ealle))
          (VBD geliff+aste)
          (PP (P +turh)
             (NP-ACC (D^A +tone) (VAG^A lyfiendan) (N^A Gast)))
          (. :))
  (ID coaelhom,+AHom_1:70.49))
```

Figure 2. ***PSD tagging in the YCOE****.*

Clauses in the YCOE are labelled IP, with an additional label that indicates type, like IP-MAT for declarative matrix IPs. The labels in figures 1 and 2 represent the following categories: syntactic categories: NP (noun phrase); lexical categories: N (noun), PRON (pronoun), ADJ (adjective), VB (verb), Q (quantifier), P (preposition), CONJ (conjunction); morphological case at word level: ^N (nominative), ^A (accusative); morphological case at phrase level: -NOM (nominative), -ACC (accusative); tense: D (past); mode: I (indicative); non-finite forms: AG (present participle).

It can also be seen in figures 1 and 2 that neither the morphological tagging nor the syntactic parsing specifies lemma. This differs from the usual practice of treebanks with respect to part of speech annotation, which, according to Hajičová et al. (2010: 168), includes lemma, category, subcategory and inflection.

As regards lemmatisation with knowledge bases, it is necessary, in the first place, to clarify this concept. The term *knowledge base* is used as a further development of a lexical database. A knowledge base is a lexical database that is integrated in a grid of databases, in such a way that certain relations between fields and layouts interpret other data sets (Martín Arista, 2017c). In this line, the *Nerthus* Project has compiled several lexical databases of Old English like *Nerthus* itself (Martín Arista et al., 2016) which, as has just been said, conform a grid of relational databases that can interpret the data from other sources.

Martín Arista (2013b) lays the foundations of a grid of relational databases of Old English comprised of three components: a dictionary database called *Nerthus* (ca. 30,000 files), devised for morphological and lexical analysis; a dictionary database called *Freya* (ca. 35,000 files), aimed to secondary source indexing; and a lemmatiser called *Norna* (ca. 190,000 files), based on the textual attestations of the DOEC. *Nerthus* gathers information on the lemma and its morphology, including inflection and derivation. For example, given a headword entry like *soðfæstness* in Figure 3, it is stated that this is a strong feminine noun whose meaning is defined as 'truth, truthfulness'. It has the spelling variant *soðfæstness* and is morphologically related to the adjectival base of derivation *soðfæst,* so that it is formed by means of the suffixation of *-ness.* This derivation is described from the semantic point of view as an instantiation of the lexical function Property with respect to the adjective *soðfæst.*

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 3. ***The entry to soðfæstness on Nerthus****.*

Figure 4 presents the entry to *andswarian* 'to answer' on the lexical database *Freya*. As is shown in this figure, *andswarian* is a verb from the second weak class with alternative spellings *andswerian, andswerigan, ondswarian, ondswerian, ondsweorian, ondsworian*, and *andwarigan*. Its inflectional forms include *andswarast, andswaraþ, andswarede, andsworede, andswara, andswarigeanne, andswarigende*, etc. This verb is discussed, among other sources, in Sievers (1903), Brunner (1965), Campbell (1987), and Hogg and Fulk (2011).

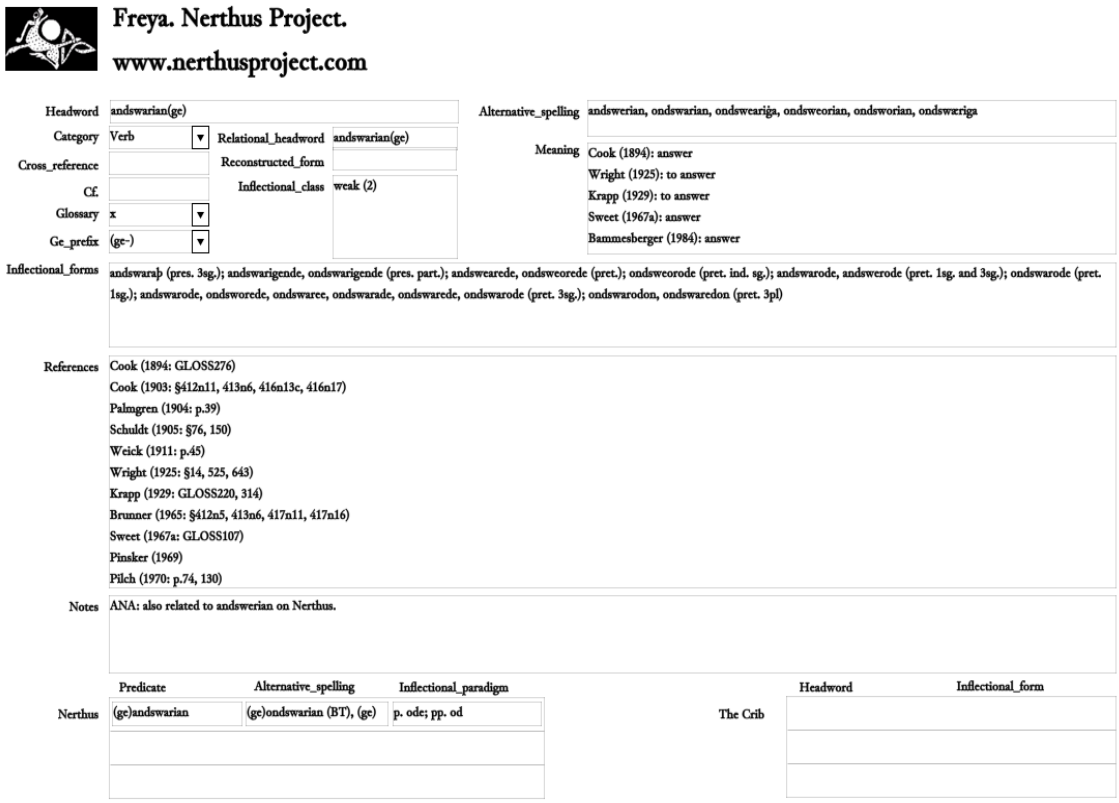
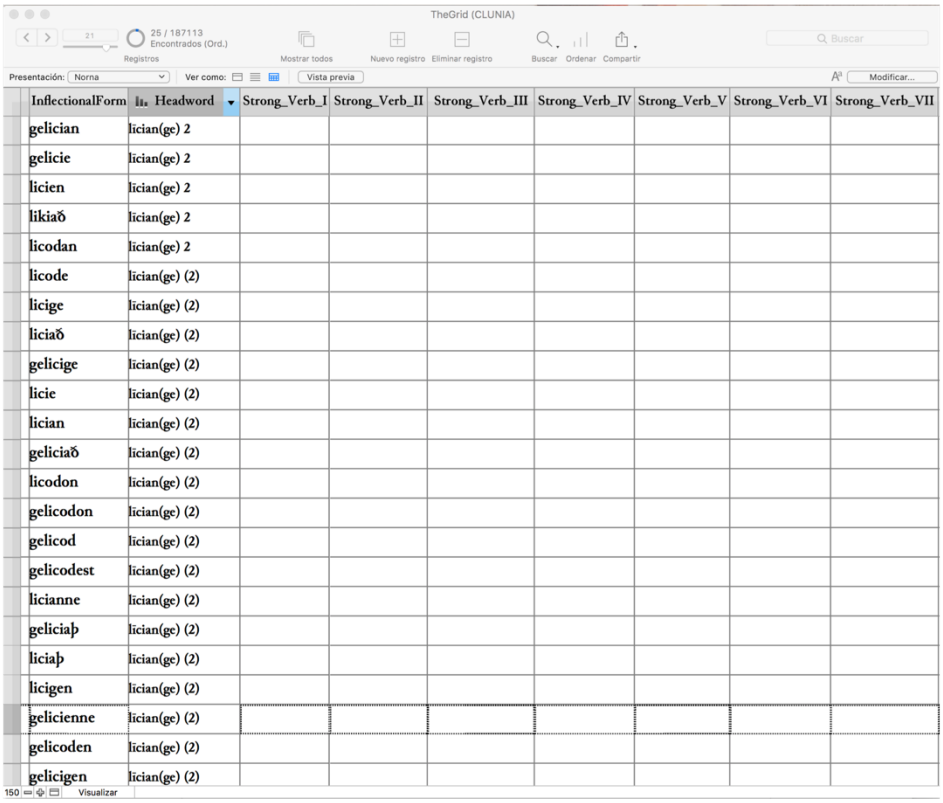


Figure 4. ***The entry to (ge)andswarian in Freya.***

With respect to *Norna*, this lemmatiser assigns lemma on a semi-automatic basis by means of searches for the prefix, stem or ending of words in the DOEC. A concordance and an index have been made to this corpus, in such a way that the index consists of a list of types with the number of occurrences of each type (or number of tokens). This is illustrated in Figure 5, which presents part of the inflectional forms lemmatised under *(ge)līcian* 'to like' in the lemmatiser including *geliciað, gelician, geliciaþ, gelicie, gelicienne, gelicige, gelicigen, gelicod, gelicoden, gelicodest, gelicodon, licað, liciað, lician, licianne, liciaþ, licie, licien, liciende, licige, licigen, licodan, licode, licodon,* and *likiað*.



InflectionalForm	Headword	Strong_Verb_I	Strong_Verb_II	Strong_Verb_III	Strong_Verb_IV	Strong_Verb_V	Strong_Verb_VI	Strong_Verb_VII
gelician	lician(ge) 2							
gelicie	lician(ge) 2							
licien	lician(ge) 2							
likiað	lician(ge) 2							
licodan	lician(ge) 2							
licode	lician(ge) (2)							
licige	lician(ge) (2)							
liciað	lician(ge) (2)							
gelicige	lician(ge) (2)							
licie	lician(ge) (2)							
lician	lician(ge) (2)							
geliciað	lician(ge) (2)							
licodon	lician(ge) (2)							
gelicodon	lician(ge) (2)							
gelicod	lician(ge) (2)							
gelicodest	lician(ge) (2)							
licianne	lician(ge) (2)							
geliciaþ	lician(ge) (2)							
liciaþ	lician(ge) (2)							
licigen	lician(ge) (2)							
gelicienne	lician(ge) (2)							
gelicoden	lician(ge) (2)							
gelicigen	lician(ge) (2)							

Figure 5. ***The inflectional forms of the lemma (ge)līcian in Norna.***

To recapitulate, the lexical databases *Nerthus* and *Freya* as well as the lemmatiser *Norna* are configured as a grid of interconnected knowledge bases that, as such, can be used for interpreting other data sets. For example, these knowledge bases are being used for the annotation of a parallel corpus of Old English (Martín Arista, in preparation), including lemmatisation.

Martín Arista (2017b, 2017c) presents the tasks and components required for annotating a corpus with information from two knowledge bases. The steps of this process can be described as follows. In the first place, the input corpus is concorded by word and by fragment. Then, an index is built on the resulting concordance. The inflected forms that belong in the index need lemmatisation, or assignment of lemma (dictionary word). Two lists have been obtained so far, the inflectional form list and the lemma list. To mark up words in these lists, a basic distinction has to be borne in mind between contextual information and context-free information. Inflectional forms need to be marked up with respect to their context, whereas lemmas can be marked up without making reference to specific contexts. For this reason, two types of markup are distinguished: one that makes reference to the context, and another which is relatively independent from context. The mark up of inflectional forms is called *tagging* and the one of lemmas is dubbed *annotation*, although both terms ultimately refer to the process of enriching a corpus with information on the words that it contains. A further distinction is drawn between linguistic and extra-linguistic information. The descriptions based on linguistic levels as well as linguistic categories and functions are linguistic, as opposed to the information related to secondary sources of the language of analysis, which can be considered metalinguistic. Whereas tagging and annotation convey linguistic information, metadata provide metalinguistic information on the words in question. Such information can be retrieved from knowledge bases, which in this model include two types: dictionary knowledge bases and secondary source knowledge bases. The information from dictionaries and knowledge bases has to be extracted and interpreted but, once it has been gathered,

classified and stored in a database, it is ready for the automatisation of the markup of the corpus. The various tasks and components just described are presented in Figure 6.

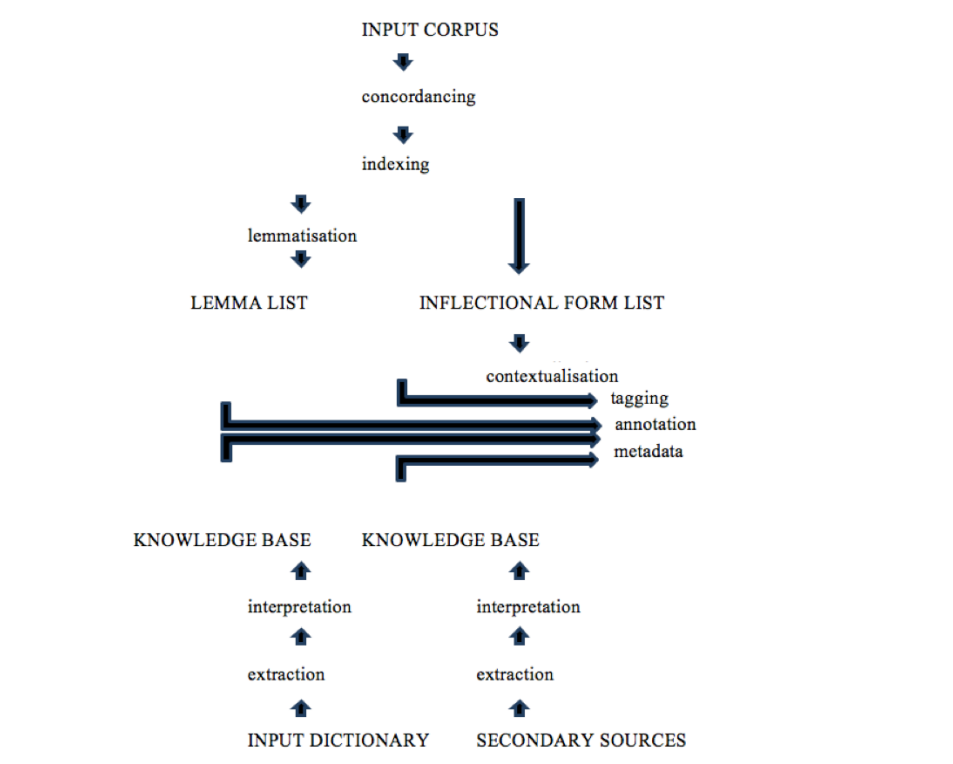


Figure 6. ***Data flow in the annotation of a corpus with knowledge bases****.*

Consider, as illustration of the process depicted in Figure 6, the following fragment (text file and number as in the DOEC).

(1) [Bo 097700 (33.79.26)]
Ne bisnode þe nan man, forþamþe nan ær þe næs þara þe auht oððe nauht worhte.
No man set you an example, because no one was before you, who anything or nothing may make.

For instance, *bisnode* is the third person of the singular number of the preterite indicative of the weak 2 class verb *bīsnian* 'to set an example'. It has the alternative spellings *bisenian, bisnigan, bȳsnian*. Its inflectional paradigm includes *bȳsniað* (present indicative plural); *bisnige* (present subjunctive singular); *bisnian* (infinitive; present subjunctive plural); *bisnode* (preterite indicative third person singular); *bisnodon* (preterite indicative plural); *bysna* (imperative singular). This verb is morphologically related to the lexical prime *bisen* 'example', as well as to the compounds and derivatives *bisenung* 'example', *forebisen* 'example', *lārbysn* 'proof', *misbȳsnian* 'to set a bad example'. So far, the information has been obtained from the dictionary knowledge bases (*Nerthus* and *Freya*), whereas the corpus knowledge base (*Norna*) relates to this lemma

the following inflectional forms (without morphological tagging, unlike the ones given above): *bisne, bisna, bisnian, bisniað, bisnode, bisnodon, bisnade, bisneden, bisnien, bisnan, bisnige, bysnian, bysnað, bysnode, bysnon, bysnan, bysnigende*. The metadata of the verb include the references to the following secondary sources: Sedgefield (1899: 220), Hargrove (1902: 80), Fowler (1972: 48), Wilcox (1994: 168), Mitchell and Robinson (1995: 309), O'Neill (2001: 284), and Marsden (2004: 412). With this tagging, annotation and metadata, each corpus token displays context-dependent and context-free information that permits several types of generalisations, like the textual frequency of each spelling variant of the lemma, the predictable and unpredictable forms in inflectional paradigms, the inventory of lexical items by morphological class, etc.

3. THE LEMMATISER *NORNA*

As Martín Arista and Metola Rodríguez (forthcoming) explain, the lemmatiser *Norna*, has three functions: searching for inflectional forms, storing the assignment of lemmas and refining subsequent searches through comparison with lexicographical and textual sources.

As regards the search for inflectional forms, *Norna* uses query strings aimed to the prefixes, stems and suffixes of corpus words, so that search hits are potential inflectional forms of the lemma under analysis. In the lemmatiser *Norna*, inflectional forms are assigned a lemma on the basis of reference lists of headwords retrieved from the lexical database *Nerthus*. The inflectional forms of verbs are being lemmatised at the moment, including weak verbs (Tío Sáenz, 2015), strong verbs (Metola Rodríguez, 2015, 2017), and preterite-present, anomalous and contracted verbs (García Fernández, fc.). Three different procedures have been used for the definition of query strings.

Firstly, strong verbs have been searched for prefix, stem, ablaut and inflectional ending. For instance, the query required for finding the canonical inflections of the class II strong verb *bēodan* comprises (notice that the wildcard * stands for any segment in preverbal or postverbal position; and that the interchangeable letters ð and þ have to be duplicated): **bead*, *beod*, *beodað*, *beodan*, *beodaþ*, *beode*, *beodeð*, *beoden*, *beodeþ*, *biedest*, *biedst*, *biest*, *bietð*, *bietst*, *biett*, *bietþ*, *bude*, *buden**, **budon.* The results are the following: *bead, beada, beadas, beod, beodað, beodan, beodanne, beodaþ, beode, beodeð, beoden, beodendan, beodende, beodenne, beodest, beodeþ, boden, bude, budon, gebead, gebeodan, gebeode, gebeodenne, geboden, gebodene, gebodenes, gebodenne, gebodenum, gebude, gebuden*.

In the second place, weak verbs have been searched for inflectional ending. A search based on the canonical endings of the first weak class for the stem *(ge)bǣd-* produces the following results: *bædað, bædde, bæddon, bæde, bædeð, bæden, bædendum*. For its part, a search for the second class verb stem *(ge)wiln-* turns out the following inflectional forms: *wilnast, wilniað, wilnian, wilnianne, wilniaþ, wilnie, wilniende, wilnode, wilnoden, wilnodest, wilnodon*, *gewilnast, gewilnian, gewilniaþ, gewilnie, gewilniende, gewilnige, gewilnod, gewilnode, gewilnodest, gewilnodon.*

In the third place, preterite-present verbs, anomalous verbs and contracted verbs have been lemmatised by means of searches aimed to morphological relations, especially the relation between simplex and prefixed verbs. That is to say, the inflectional forms of the underived verbs are combined with the prefixes to define the queries. For example, the list of attestations of *willan* 'to want' given by the grammars of Old English (Sievers, 1903; Wright and Wright, 1925; Brunner, 1965; Campbell, 1987; Hogg and Fulk, 2011) consists of these forms: *nællað*, *nællas*, *nælle*, *nælleð*, *nælles*, *nalde*, *naldon*, *naldun*, *nallað*, *nallan*, *nallas*, *nalles*, *nallo*, *nallon*, *nellað*, *nellan*, *nellaþ*, *nelle*, *nille*, *noldan*, *nolde*, *nuillic*, *nyl*, *nyle*, *nylt*, *nyllað*, *nyllan*, *nyllaþ*, *nylle*, *nyllic*, *uaelle*, *ualde*, *uil*, *wælde*,

wælle, *wællo*, *walde*, *wallað*, *wallas*, *wallon*, *wellaþ*, *welle*, *wellende*, *wil*, *will*, *wile*, *wilein*, *wileina*, *wileis*, *wili*, *wilt*, *wille*, *willa*, *willað*, *willan*, *willaþ*, *wille*, *willen*, *willende*, *willio*, *willo*, *wolde*, *wolden*, *woldest*, *woldon*. For its part, the inventory of prefixes and preverbs comprises (Kastovsky, 1992) *ā-*, *āgēn-*, *āweg-*, *adūn-*, *æfter-*, *æt-*, *and-*, *be-*, *beforan-*, *betwux-*, *dyrn-*, *ed-*, *efen-*, *eft-*, *for-*, *fore-*, *forð-*, *fram-*, *ful-*, *ge-*, *geond-*, *hearm-*, *in-*, *mān-*, *mǣg-*, *mis-*, *niðer-*, *nyd-*, *of-*, *ofer-*, *oft-*, *on-*, *onweg-*, *oð-*, *riht-*, *tō-*, *twi-*, *ðri-*, *ðurh-*, *ūp-*, *ūt-*, *un-*, *under-*, *wið-*, *wiðer-*, *wyrg-*, *ymb-*. These preverbal forms are searched in their canonical form as well as in their attested variants, thus *æfter-*, *æft-*, *æftyr-*, *efter-*, *eftyr-*, *after-*. The results obtained for the derivatives of *willan* with this method include the lemmas that follow, with the inflectional forms given between brackets: *andwillan (andwalde), anwillan (annwille, anwælde, anwalde, anwilla, anwillan, anwille), bewillan (bewillan), edwillan (adwellaþ, eadwolde, eduaelle, edwelle), gewillan (gewælde, gewalde, gewil, gewile, gewill, gewillað, gewille, gewilt, iwill), onwillan (onwælde, onwalde, onwillan), ungewillan (ungewill, ungewille), unrihtwillan (unrihtwillan), unwillan (unwilla, unwillan, unwillende), ymbwillan (ymbwælde)*.

After these results have been filed in *Norna*, the lemmatiser is used for comparison with the available lexicographical sources. Such a comparison is ultimately intended to provide feedback that allows the researcher to refine the queries gradually. For example, the automatic lemmatisation of the class II strong verb *bēodan* misses, according to the entry to *bēodan* in the DOE, some relatively predictable variations of stem, such as the i-mutated forms, for instance, *byt, bytt*; relatively predictable ending assimilations like *beot*; unpredictable stem spellings such as *bed, bedon, beadande, biodan, bud*; and unpredictable ending variation or weakening like *beoda, budan, budun, beodum, beodonne*. As regards the class 1 weak verb *bǣdan*, the DOE also lists forms with relatively predictable assimilations of endings like *bædt*; forms with unpredictable consonant gemination to the stem, such as *bæddan*; and forms with unpredictable stem vowels like *baedde, baedendrae, baedendre, bedændræ, beadætþ.* With respect to the remaining verbal classes, the anomalous verb *āgān* may be representative. The comparison of *āgān* with the DOE confirms the assignment of the inflectional forms *aeode*, *aga*, *agað*, *agæð*, *agæn*, *agæþ*, *agan*, *agane*, *aganne*, *aganre*, *ageð*, *agen*, *agena*, *ageodest*, *ageþ* and *agon* to the lemma. By contrast, *aganum*, *agende* and *aget* are not provided by the DOE, although the dictionary includes other forms such as *agætþ*, *aganan*, *agiode*, *ahgan* (García Fernández, 2015).

With the results of the comparison with other sources, the definition of *Norna* queries is improved. Consider the class 1 weak verb *behȳdan* 'to conceal'. A search for the canonical prefix, stem and ending turns out forms like *behyd, behydan, behydanne, behyde, behydest, behydeð, behydað, behydaþ, behyded, behydon*. However, the feedback from previous searches allows *Norna* to find non-canonical spellings for prefixes, such as *be-* in *bihyd, bihydde, bihyded, bihydest*; instances of the *e/i/y* stem variation as shown by *behed, behedan, behid, behidaþ, behiddan, behidde, behydan, behydaþ, behydd*, etc.; instances of consonantal gemination such as *behydest/ behyddest*; weak endings like *behydda, behyddun*; syncope in ending as in *behydest/behydst*; other variations in endings, as found in *behydanne/ behydenne*, *behyddest/behyddyst*; different degrees of assimilation in the second person, such as *behyddest/behyddes*; different degrees of assimilation in the third person, thus *behydeð, behytt, behyt*. This leaves the unpredictable stem spelling -u- found in the DOE for the forms *behud, bihud, bihuddest*. In other words, although manual revision is necessary, the searches based on the division of the word into prefix, stem and ending on the one hand, and canonical forms as well as predictable

variations, on the other, guarantees the assignment of lemma to most of the inflections of this verb.

4. LEMMATISING THE YCOE

The previous sections of this article have addressed the question of where to find the data necessary to lemmatise an Old English treebank. Put in a few words, the solution proposed above boils down to retrieving the information from a set of related knowledge bases. This section intends to answer the question of what procedure should be adopted to link the lemmatisation provided by the knowledge bases to the treebank that is going to be lemmatised, the YCOE.

To begin with, a general scheme of treebank annotation is needed, so that the lemma tag is related to the other parts of annotation and tagging. A scheme of layers for treebank annotation is presented in Figure 7. These layers are imported automatically from the knowledge bases once a given inflectional form has been attributed to a lemma entry. Overall, the information provided by the treebank is enriched in two aspects, metalinguistically and linguistically. On the metalinguistic side, metadata are added referring to secondary sources. On the linguistic side, a distinction is drawn between three blocks: lemma (comprising headword, alternative spelling, lexical category and meaning), inflectional morphology (including inflectional class and inflectional paradigm) and derivational morphology (which consists of lexical prime and derivational paradigm).

Metalinguistic	**Linguistic**		
	Lemma	Inflectional Morphology	Derivational morphology
Secondary sources	Headword Alternative spelling Lexical category Meaning	Inflectional class Inflectional paradigm	Lexical prime Derivational paradigm

Figure 7. ***A scheme of layers for treebank annotation****.*

As can be seen in the data flow in Figure 10, the assignment of lemma consigns the information in the blocks in Figure 7. It is necessary, therefore, to devise a linking procedure for providing the inflectional forms in the YCOE with the lemmas available from the lemmatiser *Norna*. The format of the linking procedure can be described as is presented in Figure 8.

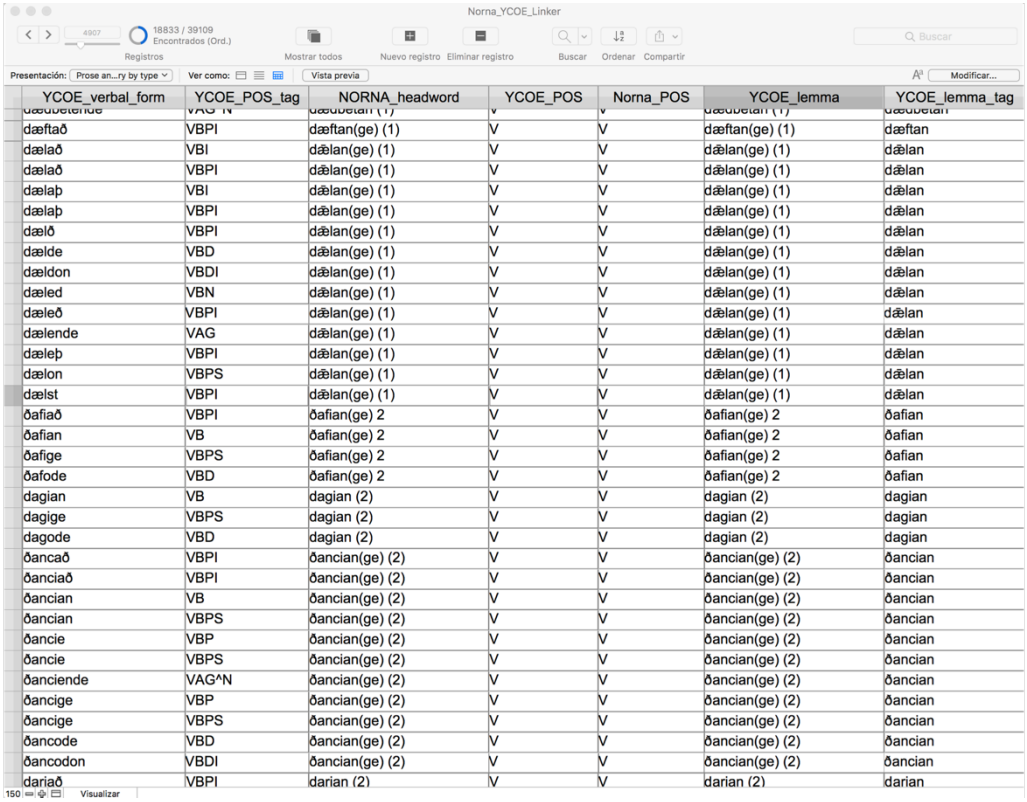
input: [[[inflectional form]LEXICAL CATEGORY]]SYNTACTIC CATEGORY
output: [[[[inflectional form]LEMMA]LEXICAL CATEGORY]]SYNTACTIC CATEGORY

Figure 8. ***Format of the linking procedure****.*

The bracketing in Figure 8 relates *Norna* to the YCOE. The input in this figure represents the format of the YCOE, which is currently unlemmatised, while the output shows the lemmatised version of the YCOE, so that the lemma assigned to the inflectional form is retrieved from *Norna*. The retrieval of the lemmas in *Norna* requires a relational algorithm that is implemented on the grid of databases. This algorithm is displayed in Figure 9. It states that when the form and the lexical category coincide in *Norna* and the YCOE, the lemma in *Norna* is attached to the inflectional form in the YCOE.

IFF
NORNA inflectional form = YCOE form;
AND
NORNA POS = YCOE POS;
THEN
NORNA headword >>> YCOE lemma tag
Figure 9. ***The retrieval algorithm****.*

The implementation of the retrieval algorithm on Filemaker software can be seen in Figure 10. The layout is called the *Linker*.



YCOE_verbal_form	YCOE_POS_tag	NORNA_headword	YCOE_POS	Norna_POS	YCOE_lemma	YCOE_lemma_tag
dæftað	VBPI	dæftan(ge) (1)	V	V	dæftan(ge) (1)	dæftan
dælað	VBI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælað	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælaþ	VBI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælaþ	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælð	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælde	VBD	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dældon	VBDI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dæled	VBN	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dæleð	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælende	VAG	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dæleþ	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælon	VBPS	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
dælst	VBPI	dælan(ge) (1)	V	V	dælan(ge) (1)	dælan
ðafiað	VBPI	ðafian(ge) 2	V	V	ðafian(ge) 2	ðafian
ðafian	VB	ðafian(ge) 2	V	V	ðafian(ge) 2	ðafian
ðafige	VBPS	ðafian(ge) 2	V	V	ðafian(ge) 2	ðafian
ðafode	VBD	ðafian(ge) 2	V	V	ðafian(ge) 2	ðafian
dagian	VB	dagian (2)	V	V	dagian (2)	dagian
dagige	VBPS	dagian (2)	V	V	dagian (2)	dagian
dagode	VBD	dagian (2)	V	V	dagian (2)	dagian
ðancað	VBPI	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðanciað	VBPI	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancian	VB	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancian	VBPS	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancie	VBP	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancie	VBPS	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðanciende	VAG^N	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancige	VBP	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancige	VBPS	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancode	VBD	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
ðancodon	VBDI	ðancian(ge) (2)	V	V	ðancian(ge) (2)	ðancian
dariað	VBPI	darian (2)	V	V	darian (2)	darian

Figure 10. ***The Linker Norna-YCOE****.*

The Linker matches, for instance, the inflectional form *dælon* in the YCOE and the homonymous *dælon* in *Norna*. Since the YCOE POS tag and the Norna POS tag coincide (both select the lexical category verb), the lemma corresponding to *dælon* in *Norna* is attributed to this form in the YCOE. This process is shown in Figure 11.

input: [[[dælon]V]]SYNTACTIC CATEGORY
output: [[[[dælon]DÆLAN]V]]SYNTACTIC CATEGORY
Figure 11. ***Instantiation of the linking procedure****.*

In the remainder of this section, this procedure is applied to the lemmatisation of the fragment in (2).

(2)
[cotempo,ÆTemp:0.2.3_ID, cotempo,ÆTemp:0.2.3_ID]
Her æfter fyligð an lytel cwyde be gearlicum tidum þæt nis to spelle geteald ac elles to rædenne þam ðe hit licað. Ic wolde eac gif ic dorste gadrian sum gehwæde andgit of ðære bec þe BEDA se snotera lareow gesette gegaderode of manegra wisra lareowa bocum be ðæs geares ymbrenum fram anginne middaneardes.

Here after follows a short treatise on the seasons of the year. It is not to be told as a homily but to be read by whoever likes it. I would also gather, if I dared, some slight knowledge from the book that the very wise teacher Bede compiled and gathered from the books of many wise teachers about the course of the year from the beginning of the World.

In the corresponding POS file, given in (3), the sequences +a, +d, +t have been replaced, respectively, by æ, ð, þ, both small and capital.

(3)
<T03990000100,0.1>_CODE DE_FW TEMPORIBUS_FW ANNI_FW
cotempo,ÆTemp:0.1.2_ID
Her_ADV^L æfter_P fyligð_VBPI an_NUM^N lytel_Q^N cwyde_N^N be_P
gearlicum_ADJ^D tidum_N^D ._, <T03990000200,0.2>_CODE Þæt_D^N
nis_NEG+BEPI to_P spelle_N^D geteald_VBN ac_CONJ elles_ADV to_TO
rædenne_VB^D þam_D^D ðe_C hit_PRO^N licað_VBPI ._.
cotempo,ÆTemp:0.2.3_ID
<T03990000300,1.0>_CODE DE_FW $DIE_FW ._. cotempo,ÆTemp:1.0.4_ID
<T03990000400,1.1>_CODE Ic_PRO^N wolde_MDD eac_ADV gif_P ic_PRO^N
dorste_MDD gadrian_VB sum_Q^A gehwæde_ADJ^A andgit_N^A of_P ðære_D^D
bec_N^D þe_C BEDA_NR^N ,_, se_D^N snotera_ADJ^N lareow_N^N gesette_VBD
,_, &_CONJ gegaderode_VBD of_P manegra_Q^G wisra_ADJ^G lareowa_N^G
bocum_N^D be_P ðæs_D^G geares_N^G ymbrenum_N^D ,_, fram_P anginne_N^D
middaneardes_N^G ._. cotempo,ÆTemp:1.1.5_ID

The PSD file fragment corresponding to the POS file fragment in (3) follows in (4).

(4)
((CODE <T03990_ÆTemp_B1.9.4>))
((CODE <T03990000100,0.1>)
 (LATIN (FW DE) (FW TEMPORIBUS) (FW ANNI)) (ID cotempo,ÆTemp:0.1.2))
((IP-MAT (PP (ADVP-LOC (ADV^L Her))
      (P æfter))
   (VBPI fyligð)
   (NP-NOM (NUM^N an) (Q^N lytel) (N^N cwyde)
         (CP-REL *ICH*-1))
   (PP (P be)
      (NP-DAT (ADJ^D gearlicum) (N^D tidum)))
   (, .)
   (CODE <T03990000200,0.2>)
   (CP-REL-1 (WNP-NOM-2 (D^N Þæt))
         (C 0)
         (IP-SUB-0 (NP-NOM *T*-2)
               (NEG+BEPI nis)

```
                    (PP (P to)
                        (NP-DAT (N^D spelle)))
                    (VBN geteald))
                (IP-SUB (CONJP (CONJ ac)
                            (IPX-SUB-CON=0 (ADVP (ADV elles))
                                (IP-INF (TO to)
                                    (VB^D rædenne)
                                    (NP-DAT (D^D þam)
                                        (CP-REL (WNP-3 0)
                                            (C ðe)
                                            (IP-SUB (NP *T*-3)
                                                (NP-NOM
(PRO^N hit))
                                                (VBPI
licað)))))))))
   (. .)) (ID cotempo,ÆTemp:0.2.3))
( (CODE <T03990000300,1.0>)
  (LATIN (FW DE) (FW $DIE)
      (. .)) (ID cotempo,ÆTemp:1.0.4))
( (CODE <T03990000400,1.1>)
  (IP-MAT-0 (NP-NOM (PRO^N Ic))
     (MDD wolde)
     (ADVP (ADV eac))
     (CP-ADV (P gif)
          (C 0)
          (IPX-SUB=0 (NP-NOM (PRO^N ic))
               (MDD dorste)))
     (VB gadrian)
     (NP-ACC (Q^A sum) (ADJ^A gehwæde) (N^A andgit))
     (PP (P of)
         (NP-DAT (D^D ðære) (N^D bec)
              (CP-REL (WNP-1 0)
                   (C þe)
                   (IP-SUB (NP *T*-1)
                        (NP-NOM (NR^N BEDA)
                             (, ,)
                             (NP-NOM-PRN (D^N se) (ADJ^N snotera)
(N^N lareow)))
                        (VBD (VBD gesette) (, ,) (CONJ &) (VBD
gegaderode))
                        (PP (P of)
                          (NP-DAT (NP-GEN (Q^G manegra) (ADJ^G
wisra) (N^G lareowa))
                               (N^D bocum)))))))
     (PP (P be)
         (NP-DAT (NP-GEN (D^G ðæs) (N^G geares))
              (N^D ymbrenum)))
     (, ,)
     (PP (P fram)
         (NP-DAT (N^D anginne)
```

(NP-GEN (N^G middaneardes))))
(. .)) (ID cotempo,ÆTemp:1.1.5))

The lemmatised POS file fragment can be seen in (5). The lemmatisation of verbal forms has been done automatically and revised manually. The lemmas of non-verbal categories have been assigned manually on the basis of the information found in the dictionary knowledge bases discussed above.

(5)
<T03990000100,0.1>_CODE DE_FW TEMPORIBUS_FW ANNI_FW
cotempo,ÆTemp:0.1.2_ID
Her_**HĒR**_ADV^L æfter_**ÆFTER**_P fyligð_**FOLGIAN**_VBPI an_**AN**_NUM^N
lytel_**LȳTEL**_Q^N cwyde_**CWIDE**_N^N be_**BE**_P
gearlicum_**GĒARLIC**_ADJ^D tidum_**TĪD**_N^D ._,
<T03990000200,0.2>_CODE Þæt_**ÐÆT**_D^N
nis_**NEWESAN**_NEG+BEPI to_**TŌ**_P spelle_**SPELL**_N^D
geteald_**GETELLAN**_VBN ac_**AC**_CONJ elles_**ELLES**_ADV to_**TŌ**_TO
rædenne_**RÆDAN**_VB^D þam_**SE**_D^D ðe_**ÐE**_C hit_**HE**_PRO^N
licað_**LĪCIAN**_VBPI ._.
cotempo,ÆTemp:0.2.3_ID
<T03990000300,1.0>_CODE DE_FW $DIE_FW ._. cotempo,ÆTemp:1.0.4_ID
<T03990000400,1.1>_CODE Ic_**IC**_PRO^N wolde_**WILLAN**_MDD
eac_**ĒAC**_ADV gif_**GIF**_P ic_**IC**_PRO^N
dorste_**DURRAN**_MDD gadrian_**GADERIAN**_VB sum_**SUM**_Q^A
gehwæde_**GEHWÆÐE**_ADJ^A andgit_**ANDGIET**_N^A of_**OF**_P
ðære_**SE**_D^D
bec_**BŌC**_N^D þe_**ÐE**_C BEDA_**BEDA**_NR^N ,_, se_**SE**_D^N
snotera_**SNOTOR**_ADJ^N lareow_**LĀRĒOW**_N^N gesette_**GESETTAN**_VBD
,_, &_**AND**_CONJ gegaderode_**GEGADERIAN**_VBD of_**OF**_P
manegra_**MANIG**_Q^G wisra_**WĪS**_ADJ^G lareowa_**LĀRĒOW**_N^G
bocum_**BŌC**_N^D be_**BE**_P ðæs_**SE**_D^G geares_**GĒAR**_N^G
ymbrenum_**YMBRENE**_N^D ,_, fram_**FRAM**_P anginne_**ANGINN**_N^D
middaneardes_**MIDDANGEARD**_N^G ._. cotempo,ÆTemp:1.1.5_ID

Finally, the lemmatised PSD file fragment is presented in (6). The lemmas have been imported from the POS file fragment.

(6)
((CODE <T03990000100,0.1>)
(LATIN (FW DE) (FW TEMPORIBUS) (FW ANNI)) (ID
cotempo,ÆTemp:0.1.2))
((IP-MAT (PP (ADVP-LOC (ADV^L Her_HĒR))
(P æfter_**ÆFTER**))
(VBPI fyligð_**FOLGIAN**)
(NP-NOM (NUM^N an_**AN**) (Q^N lytel_**LȳTEL**) (N^N cwyde_**CWIDE**)
(CP-REL *ICH*-1))
(PP (P be_**BE**)
(NP-DAT (ADJ^D gearlicum_**GĒARLIC**) (N^D tidum_**TĪD**)))
(, .)
(CODE <T03990000200,0.2>)

(CP-REL-1 (WNP-NOM-2 (D^N Þæt_**ÐÆT**))
(C 0)
(IP-SUB-0 (NP-NOM *T*-2)
(NEG+BEPI nis_**NEWESAN**)
(PP (P to_**TŌ**)
(NP-DAT (N^D spelle_**SPELL**)))
(VBN geteald_**GETELLAN**))
(IP-SUB (CONJP (CONJ ac_**AC**)
(IPX-SUB-CON=0 (ADVP (ADV elles_**ELLES**))
(IP-INF (TO to_**TŌ**)
(VB^D rædenne_**RÆDAN**)
(NP-DAT (D^D þam_**SE**)
(CP-REL (WNP-3 0)
(C ðe_**ÐE**)
(IP-SUB (NP *T*-3)
(NP-NOM
(PRO^N hit_**HE**))
(VBPI
licað_**LĪCIAN**))))))))) ((CODE <T03990000300,1.0>)
(LATIN (FW DE) (FW $DIE)
(. .)) (ID cotempo,ÆTemp:1.0.4))
((CODE <T03990000400,1.1>)
(IP-MAT-0 (NP-NOM (PRO^N Ic_**IC**))
(MDD wolde_**WILLAN**)
(ADVP (ADV eac_**ĒAC**))
(CP-ADV (P gif_**GIF**)
(C 0)
(IPX-SUB=0 (NP-NOM (PRO^N ic_**IC**))
(MDD dorste_**DURRAN**)))
(VB gadrian_**GADERIAN**)
(NP-ACC (Q^A sum_**SUM**) (ADJ^A gehwæde_**GEHWÆDE**) (N^A andgit_**ANDGIET**))
(PP (P of_**OF**)
(NP-DAT (D^D ðære_**SE**) (N^D bec_**BŌC**)
(CP-REL (WNP-1 0)
(C þe_**ÐE**)
(IP-SUB (NP *T*-1)
(NP-NOM (NR^N BEDA_**BEDA**)
(, ,)
(NP-NOM-PRN (D^N se_**SE**) (ADJ^N snotera_**SNOTOR**) (N^N lareow_**LĀRĒOW**)))
(VBD (VBD gesette_**GESETTAN**) (, ,) (CONJ &) (VBD gegaderode_**GEGADERIAN**))
(PP (P of_**OF**)
(NP-DAT (NP-GEN (Q^G manegra_**MANIG**) (ADJ^G wisra_**WĪS**) (N^G lareowa_**LĀRĒOW**))
(N^D bocum_**BŌC**)))))))
(PP (P be_**BE**)
(NP-DAT (NP-GEN (D^G ðæs_**SE**) (N^G geares_**GĒAR**))
(N^D ymbrenum_**YMBRENE**)))

(, ,)
(PP (P fram_**FRAM**)
(NP-DAT (N^D anginne_**ANGINN**)
(NP-GEN (N^G middaneardes_**MIDDANGEARD**))))
(. .)) (ID cotempo,ÆTemp:1.1.5))

To finish up this section, the lemmatisation procedure, which has been devised and applied above, is assessed. The main aspect of the assessment is that it is possible to lemmatise a treebank automatically with the information available from knowledge bases. This procedure guarantees a lemmatisation that bridges the gap between the infomation split in type and token analysis and permits several types of paradigmatic generalisation. For instance, all the inflectional forms in the paradigm can be gathered under the lemma, including predictable and unpredictable paradigmatic forms. Furthermore, textual frequency can be gauged, including all the occurrences of the inflectional form (token) and the lemma (type). It is also possible to assess the occurrence of empty morphs, as in verbs in which the presence or absence of the prefix *ge-* does not seem to cause a change of meaning. On the side of spelling, the variant spellings of the inflections of a lemma can be compared, including the alternative spellings with eth and thorn. All these aspects can be used as feedback to refine the lists of inflectional forms and lemmas, which ultimately improves the quality of lemmatisation. Overall, the lemmatisation of treebanks can contribute to research venues in the linguistics of Old Engish that combine morphology and semantics (such as the analysis of empty morphs, verbal tense and aspect) or syntax and semantics (like collocations and complementation).

On the other hand, the quality of the lemmatisation crucially depends on the exhaustiveness and accuracy of the information provided by the knowledge bases. In the present state, the knowledge bases of the *Nerthus* Project permit the lemmatisation of the verbal lexicon exclusively. This project has opted for lemmatising the verbal class in the first place because the inflections of non-verbal classes are less transparent, given that many inflectional endings are shared by the declensions of nouns and adjectives. Nevertheless, the realisation of the arguments of the sentence as well as the relations between clauses in the complex clause are determined by verbs. In other words, the information on verbs is central to the morphological and syntactic interpretation of the sentence and any advance in this line represents a significant contribution.

Another aspect that deserves discussion is automatisation. Verbal forms have been lemmatised automatically with a categorial filter. That is to say, two homonymous forms get the same lemma if their POS tag is V in both Norna and the YCOE. This guarantees the accuracy of most verbal forms, except pairs of verbal homonyms from two different verbs. For example, an inflectional form like *seo* may correspond to both *bēon* 'to be' or *sēon* 'to see'. This issue calls for manual revision.

Finally, the correspondence between *Norna* and the YCOE has not been fully attained when it comes to lemmatising verbs beginning with *ge-*. Norna unifies simplex verbs and the corresponding verbs with the prefix ge-. Thus, gaðerian and gegaðerian appear in *Norna* under *gaðerian(ge)*. For verbs not beginning with *ge-*, the simplex verb lemma can be assigned straightforwardly, but the complex verb lemma cannot be directly attributed to verbs beginning with *ge-* because this sequence is not always a prefix, thus *gēatan* 'to say yeah', or involves other preverbs, as in *geandwyrdan* 'to present'. This question requires further research.

5. CONCLUSION

This article has devised a lemmatisation procedure for Old English treebanks, as represented by *The York-Toronto-Helsinki Parsed Corpus of Old English Prose.* It has addressed two research questions, to wit, where to find the data for lemmatisation and how to link the information on lemmatisation available from the sources to the treebank. *The solution adopted in this article is to resort to the knowledge bases compiled by the Nerthus* Project. On the grounds of these knowledge bases, a semi-automatic procedure has been implemented and assessed. The assessment insists on the possibility of lemmatising the verbal lexicon on a semi-automatic basis as well as on the different paradigmatic generalisations for which the lemmatisation of the treebank allows. In this respect, linking inflections and lemmas makes it possible to calculate the textual frequency of the lemmas, and to analyse the patterns of spelling variation and morphological variation of the inflectional forms of the lemmas in the corpus. These aspects, in turn, bring the possibility of conducting studies that combine morphology and semantics (such as the analysis of empty morphs, verbal tense and aspect) or syntax and semantics (like collocations and complementation).

REFERENCES

Abeillé, A. (2003). Introduction. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer. xiii xxvi.

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

Brunner, K. (1965). *Altenglische Grammatik nach der Angelsächsischen Grammatik von Eduard Sievers* (3rd ed.). Tübingen: Max Niemeyer Verlag.

Campbell, A. 1987 (1959). *Old English Grammar*. Oxford: Oxford University Press.

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

Fowler, R. (1972). *Wulfstan´s Canons of Edgar*. Oxford: Oxford University Press.

García Fernández, L. (2015). The *Lemmatisation of Derived Preterite-Present and Irregular Verbs on a Lexical Database of Old English*. Master´s Thesis. University of La Rioja.

García Fernández, L. Preterite-present verb lemmas from a corpus of Old English. In P. Guerrero Medina, R. Torre Alonso and R. Vea Escarza (Eds.), *Verbs, Clauses and Constructions: Functional and Typological Approaches*. Newcastle: Cambridge Scholars Publishing. Forthcoming.

García García, L. (2012). Morphological causatives in Old English: the quest for a vanishing formation. *Transactions of the Philological Society,* 110(1), 112-148. doi: 10.1111/j.1467-968X.2012.01287.x

García García, L. (2013). Lexicalization and morphological simplification in Old English jan-causatives: some open questions. *Sprachwissenschaft,* 38(2), 245-264.

González Torres, E. (2010a). The Continuum Inflection-Derivation and the Old English *suffixes -a, -e, -o, -u*. *ATLANTIS,* 32.1, 103-122.

González Torres, E. (2010b). The bases of derivation of Old English affixed nouns: status and category. *Studia Anglica Posnaniensia,* 46(2), 21-43.

González Torres, E. (2011). Morphological complexity, recursiveness and templates in the formation of Old English nouns. *Estudios Ingleses de la Universidad Complutense* 19, 45-70.

Hajičová, E., Abeillé, A., Hajič, J., Mírovský, J. & Urešová, Z. (2010). Treebank annotation. In N. Indurkhaya & F. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 167-188). Boca Raton, FL: Chapman & Hall/CRC.

Hargrove, H. L. (1902). *King Alfred´s Old English Version of St. Augustine´s Soliloquies*. New York: Henry Holt and Company.

Haug, D. (2015). Treebanks in historical linguistic research. In C. Viti (Ed.), *Perspectives on Historical Syntax,* (pp. 188-202). Amsterdam: John Benjamins.

Healey, A. diPaolo (Ed.). (2016). *The Dictionary of Old English: A to H.* Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

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

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.

Hogg, R. M. & Fulk, R. D. (2011). *A Grammar of Old English*. Oxford, Wiley-Blackwell.

Kastovsky, D. (1992). Semantics and vocabulary. In R. M. Hogg (Ed.), *The Cambridge History of the English Language I*: *The Beginnings to 1066* (pp. 290-408). Cambridge: Cambridge University Press. doi: 10.1017/CHOL9780521264747.006

Marcus, M., Marcinkiewicz, M, & Santorini, B. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics,* 19(2), 313-330.

Marsden, R. (2004). *The Cambridge Old English Reader*. Cambridge: Cambridge University Press.

Martín Arista, J. (2012a). Lexical database, derivational map and 3D representation. *RESLA-Revista Española de Lingüística Aplicada,* (Extra 1), 119-144.

Martín Arista, J. (2012b). The Old English Prefix *Ge-*: A Panchronic Reappraisal. *Australian Journal of Linguistics,* 32(4), 411-433. doi: 10.1080/07268602.2012.744264

Martín Arista, J. (2013a). Recursivity, derivational depth and the search for Old English lexical primes. *Studia Neophilologica,* 85(1), 1-21. doi: 10.1080/00393274.2013.771829

Martín Arista, J. (2013b). *Nerthus. Lexical Database of Old English: From word-formation to meaning construction*. Research Seminar, School of English, University of Sheffield.

Martín Arista, J. (2014). Noun layers in Old English. Asymmetry and mismatches in lexical derivation. *Nordic Journal of English Studies,* 13(3), 160-187.

Martín Arista, J. (2017a). El paradigma derivativo del inglés antiguo. *Onomazeín,* 37, 144-169.

Martín Arista, J. (2017b). The Design and Implementation of a Pilot Parallel Corpus of Old English. Paper presented at the SHELL Session of the 2017 International Medieval Conference. Leeds, University of Leeds, United Kingdom. July, 4.

Martín Arista, J. (2017c). The Nerthus Project at the crossroads. From lexical database to parallel corpus of Old English. Lecture delivered at the 2017 International Conference of SELIM. Málaga, University of Málaga, Spain.

Martín Arista, J. (2017d). The Semantic Poles of Old English. Towards the 3D Representation of Complex Polysemy. *Digital Scholarship in the Humanities*. Forthcoming. doi: 10.1093/llc/fqx004.

Martín Arista, J. (coord.). *Parallel Corpus of Old English Prose*. Nerthus Project. Universidad de La Rioja. In preparation.

Martín Arista, J. & Cortés Rodríguez, F. (2014). From directionals to telics: meaning construction, word-formation and grammaticalization in Role and Reference

Grammar. In M. A. Gómez González, F. Ruiz de Mendoza Ibáñez & F. Gonzálvez García (Eds.), *Theory and Practice in Functional-Cognitive Space,* (pp. 229-250). Amsterdam: John Benjamins. doi: 10.1075/sfsl.69.10mar

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

Martín Arista, J. & Metola Rodríguez, D. The lemmatisation of Old English strong verbs on a corpus-based lexical database. Forthcoming.

Martín Arista, J. & Vea Escarza, R. (2016). Assessing the semantic transparency of Old English affixation: adjective and noun formation. *English Studies,* 97(1-2), 61-77.

Mateo Mendaza, R. (2013). The Old English exponent for the semantic prime TOUCH. Descriptive and methodological questions. *Australian Journal of Linguistics,* 33(4), 449-466. doi: 10.1080/0726.8602.2013.

Mateo Mendaza, R. (2014). The Old English adjectival affixes ful- and –ful: a text-based account on productivity. *NOWELE-North-Western European Language Evolution,* 67.1, 77-94. doi: 10.1075/nowele.67.1.

Mateo Mendaza, R. (2015a). Matching productivity indexes and diachronic evolution. The Old English affixes ful-, -isc, -cund and -ful. *Canadian Journal of Linguistics,* 60(1), 1-24.

Mateo Mendaza, R. (2015b). The search for Old English semantic primes: the case of HAPPEN. *Nordic Journal of English Studies,* 15, 71-99.

Mateo Mendaza, R. (2016). The Old English exponent for the semantic prime MOVE. *Australian Journal of Linguistics,* 34(4), 542-559. doi: 10.1080/07268602.2016.1169976

Metola Rodríguez, D. (2015). *Lemmatisation of Old English Strong Verbs on a Lexical Database*. PhD Dissertation, University of La Rioja, Spain.

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.

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

Nivre, J. (2008). Treebanks. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (Volume 1). Berlin: Mouton de Gruyter, 225-241.

Novo Urraca, C. 2015. Old English Deadjectival Paradigms. Productivity and Recursivity. *NOWELE-North-Western European Language Evolution,* 68 (1):61-80.

Novo Urraca, C. 2016a. Old English suffixation. Content and transposition. *English Studies,* 97(6). Forthcoming.

Novo Urraca, C. 2016b. Morphological relatedness and the typology of adjectival formation in Old English. *Studia Neophilologica,* 88(1). Forthcoming.

O'Neill, P. P. (2001). *King Alfred's Old English Prose Translation of the First Fifty Psalms*. Cambridge, Massachusetts: The Medieval Academy of America.

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

Rissanen M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T. & Raumolin-Brunberg, H. (1991). *The Helsinki Corpus of English Texts*. Department of Modern Languages, University of Helsinki.

Rosén, V., Meurer, P. & de Smedt, K. (2005). Constructing a parsed corpus with a large LFG grammar. In M. Butt & T. H. King (Eds.), *Proceedings of the LFG'05 Conference University of Bergen*. Stanford, CA.: CSLI Publications. Retrieved from:

http://web.stanford.edu/group/cslipublications/cslipublications/LFG/10/lfg05.html
Sedgefield, W. J. (1899). *King Alfred's Old English Version of Boethius De Consolatione Philosophiae*. Oxford: Clarendon Press.
Sievers, E. (1903). *Old English Grammar* (A. S. Cook, Trans). Boston: The Athenaeum Press. (Original work published in 1885).
Sweet, H. 1976 (1896). *The Student's Dictionary of Anglo-Saxon*. Cambridge: Cambridge University Press.
Taylor, A., Warner, A., Pintzuk, S. & Beths, F. (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Department of Language and Linguistic Science, University of York.
Taylor, A., Marcus, M. & Santorini, B. (2003). The Penn Treebank: An Overview. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora,* (pp. 5-22*)*. Dordrecht: Kluwer.
Tío Sáenz, M. (2015). Regularization of Old English Weak Verbs. *Revista de lingüística y lenguas aplicadas,* 10, 78-89.
Torre Alonso, R. (2011a). The Morphological Structure of Old English Complex Nouns. *ATLANTIS* 33(1), 127-146.
Torre Alonso, R. (2011b). Affix Combination in Old English Noun Formation: Distribution and Constraints. *RESLA-Revista Española de Linguística Aplicada,* 24, 257-279.
Vea Escarza, R. (2012). Structural and Functional Aspects of Morphological Recursivity. *NOWELE-North-Western European Language Evolution,* 64/65, 155-179. doi: 10.1075/nowele.64-65.09esc
Vea Escarza, R. (2013). Old English Adjectival Affixation. Structure and Function. *Studia Anglica Posnaniensia*, 48(2)-3, 1-21.
Vea Escarza, R. (2014). Split and unified functions in the formation of Old English nouns and adjectives. *Revista de Lingüística y Lenguas Aplicadas,* 9, 110-116. doi: 10.4995/rlyla.2014.2086
Vea Escarza, R. (2016a). Recursivity and inheritance in the formation of Old English nouns and adjectives. *Studia Neophilologica,* 88, 1-23. doi: 10.1080/00393274.2015. 1049830
Vea Escarza, R.(2016b). Old English affixation. A structural-functional analysis. *Nordic Journal of English Studies,* 15(1), 101-119.
Wilcox, J. (1994). *Ælfric´s Prefaces*. Durham: Durham Medieval Texts.
Wright, J. & Wright, M. (1925). *Old English Grammar*. United States: Oxford University Press.