A semiautomatic lemmatisation procedure for treebanks. Old English strong and weak verbs.

Darío Metola Rodríguez
University of La Rioja
Edificio Filologías
26004, Logroño (La Rioja) Spain
dario.metola@unirioja.es

Marta Tío Sáenz
University of La Rioja
Edificio Filologías
26004, Logroño (La Rioja) Spain
marta.tio@unirioja.es

Abstract

The aim of this paper is to present a semiautomatic lemmatisation procedure implemented on database software that is compatible with the morphological tagging of treebanks. The language of analysis is Old English, for which parsed corpora are available but they are not lemmatised. The scope of the paper includes both strong and weak verbs. After describing the lemmatisation procedure, the paper discusses the results of automatic searches and compares them with the lists of inflectional forms by lemma provided by other lexicographical sources. The conclusions insist on the limits of automatic lemmatisation and the compatibility with treebank parsing.

1 Introduction

This paper deals with lemmatisation in a corpus of Old English and focuses on the seven classes of strong verbs, and classes 1 and 2 of weak verbs. The analysis reported here is based on the lemmatiser *Norna*, a building block of the lexical database of Old English *Nerthus* (www.nerthusproject.com). *Norna*, in turn, draws on the information available from the *Dictionary of Old English Corpus* (hearafter DOEC, Healey et al., 2004), the *Helsinki Corpus* (Rissanen et al., 1991), the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al., 2003) and the *York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug, 2001). Of these, only the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al., 2003) and the *York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug, 2001) are parsed. The parsing includes syntactic categories and functions as well as lexical and morphological tagging. For this reason, these corpora are commonly known as *treebanks*. In their current state, these treebanks are unlemmatised. That is to say, the attestations of the inflections of a given lemma are not related to the dictionary word, which results in a lower descriptive power, especially as regards paradigmatic analysis and the quantification of morphological and lexical aspects.

Moreover, the standard dictionaries of Old English, including *An Anglo-Saxon Dictionary, A Concise Anglo-Saxon Dictionary* and *The student's Dictionary of Anglo-Saxon,* constitute valuable sources of philological data, although they are not based on an extensive corpus of the language but on the partial list of texts listed in their prefaces or introductions. On its part, *The Dictionary of Old English* (henceforth DOE) is based on the corpus mentioned above, but is still in progress (the letter H was published in 2016). All in all, neither the textual nor the lexicographical sources of Old English are fully lemmatised. At the same time, treebanks could improve their descriptive power and searchability by incorporating lemma tags. For these reasons, the aim of this paper is to present a semiautomatic lemmatisation procedure implemented on database software that is compatible with the morphological tagging of treebanks. The first step was to carry out a concordance of the texts that the *Dictionary of Old English* provides in its corpus. The concordance by word consists of three million lines, one per word in the corpus. The concordance by fragment, in contrast, contains around two hundred thousand fragments of texts identified with the short title with which they appear in the DOEC, as in *Eala ðu cleric ne wana ðu æfre wexbreda fram sidan* [Abbo 000100 (103.1)]. The target of the analysis is the data retrieved from the word concordance to the DOEC, which turns out an index of approximately one hundred and ninety thousand inflectional forms. Once the data has been identified and extracted from the concordance, the process of lemmatisation starts. The different types of verbs are lemmatised in turn, depending on their formal transparency. That is to say, strong verbs have been lemmatised in the first

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 88–94,
Prague, Czech Republic, January 23–24, 2018.

Keywords: semi-automatic lemmatisation, treebanks, Old English, weak verbs

place and then weak verbs from the second class have been processed. The former can be identified by stem and inflectional ending, the latter by inflectional ending only, but their inflectional paradigm is more transparent that the one of the weak verbs from the first and the third class. Then, weak verbs from class 1 have been lemmatised. Weak class 3, anomalous, contracted and preterite-present verbs will be dealt with in further research.

2 Lemmatisation procedure for strong verbs

On the lemmatiser *Norna*, inflectional forms of strong verbs are assigned a lemma on the basis of a reference list of verbs from each strong class that has been retrieved from the lexical database Nerthus and supplemented with information from Krygier (1994) and Hogg and Fulk (2011).
The second step in the process of lemmatisation in *Norna* has already been implemented for the seven classes of strong verbs of Old English. Query strings are defined and launched in the database, so that the results are compared with the existing sources, and the conclusions of such comparison are used to refine the query strings.
After these query strings have been inputed to the lemmatiser, the assignments of lemmas to attestations are filed and compared with the lexicographical and philological sources, so that the feedback of previous searches is used to improve subsequent query strings.
With this method, the design of the search algorithm is stepwise. The target of the first step is the simplex word. The underived verbs in the reference list of the seven strong classes have been inflected for the infinitive, inflected infinitive, present participle and past participle; present indicative singular and plural, present subjunctive singular and plural, preterite indicative singular and plural, preterite subjunctive singular and plural, imperative singular and plural.
The second step in the creation of the algorithm is focused on the complex word. It consists on the compilation of a list of elements that may be attached to simplex strong verbs to form derived or compound verbs. Originally, the inventory of preverbal elements, retrieved from the lexical database of Old English *Nerthus*, includes affixes with a very specific meaning, such as the negative prefix *un-*, the pejorative prefix *mis-* as well as the aspectual prefixes *eft-* and *ed-*; the Germanic pure prefixes *ā-*, *be-*, *for-*, *ge-*, *of-*, *on-*, *tō-* (de la Cruz 1975); the spatial and temporal adverbs and prepositions that are going through grammaticalisation resulting in a telic marker (Brinton and Traugott 2005; Martín Arista and Cortés Rodríguez 2014), including *adūn-*, *æfter-*, *æt-*, *āweg-*, *beforan-*, *betwux-*, *ðurh-*, *forð-*, *fore-*, *fram-*, *geond-*, *in-*, *niðer-*, *oð-*, *ofer-*, *onweg-*, *under-*, *ūp-*, *ūt-*, *wið-*, *wiðer-*, and *ymb-*; and fully free forms that appear in compound verbs such as *āgēn-*, *and-*, *ðri-*, *dyrn-*, *efen-*, *ful-*, *hearm-*, *mǣg-*, *mān-*, *nyd-*, *riht-*, *twi-*, *wyrg-*. Having the preverbal elements, the roots and the set of inflections presented above, the third step in the design of the search algorithm is the definition of query strings that can be applied on Filemaker. Four query strings (QS) have been defined. QS1 is devoted to the stems and inflections by using the operator (wild card) for exact matches in Filemaker (==). The part of QS1 that searches the corpus for the inflections of *bēodan* can be seen in (1).

(1)

==beodan, ==bead, ==budon, ==beode, ==bead, ==biedest, ==biedst, ==bietst, ==biest, ==bude, ==beodeð, ==beodeþ, ==biett, ==bietð, ==bietþ, ==bead, ==beodaþ, ==beodað, ==budon, ==beode, ==bude, ==beoden, ==buden, ==beod, ==beodað, ==beodaþ

The target of the second QS (QS2) is prefix *ge-*, the most frequent in Old English (Martín Arista 2012b), to such an extent that most strong verbs present a simplex and complex form prefixes with *ge-*. QS2 for *gebēodan* is shown in (2).

(2)

==gecimban, ==gecamb, ==gecumbon, ==gecumben, ==gecimbe, ==gecamb, ==gecimbst, ==gecimbest, ==gecumbe, ==gecimbð, ==gecimbþ, ==gecimbeð, ==gecimbeþ, ==gecamb, ==gecimbaþ, ==gecimbað, ==gecumbon, ==gecimbe, ==gecambe, ==gecimben, ==gecumben, ==gecimb, ==gecimbaþ, ==gecimbað, ==gecimbanne, ==gecimbenne, ==gecimbende

QS3 has been created for accounting the existence of complex strong verbs with preverbs different from *ge-*. The wild card (*) in (3) represents any preverbal elements attached to the base and its inflections.

(3)

==*beodan, ==*bead, ==*budon, ==*beode, ==*bead, ==*biedest, ==*biedst, ==*bietst, ==*biest, ==*bude, ==*beodeð, ==*beodeþ, ==*biett, ==*bietð, ==*bietþ, ==*bead, ==*beodaþ, ==*beodað, ==*budon, ==*beode, ==*bude, ==*beoden, ==*buden, ==*beod, ==*beodað, ==*beodaþ

QS4 is the least specific of all the query strings. It looks in the corpus for the stems of strong verbs with any preverbal element and inflectional ending, hence the addittion of the wild card (*) to both sides of the stem, as in (4).

(4)

==*beod*, ==*bead*, ==*bud*, ==*bod*, ==*bied*, ==*biet*, ==*biest*

The previous query strings have been launched in a sequential manner: QS1, QS2, QS3, QS4. After the submission of each query, the resulting hits have been tagged on the lemmatiser *Norna*, with the result that the tags from previous queries could aid the tagging of the hits resulting from subsequent queries. This brings about a simplification of the overall task because, in spite of being likely to find some unexpected spellings, QS4 is redundant with respect to QS1 (endings), QS2 and QS3 (preverbs). Moreover, due to its wide scope, it is predictable that this query strings turns out a remarkably high number of results. For this, the final step in the design of this algorithm is the definition of filters to put aside at least part of the undesired results of QS4, so that manual revision can be diminished dramatically.

Four filters have been designed for this purpose. Filter (F) 1 is intended to isolate verbal forms. It cuts down the hits of QS4 to inflectional forms that end with -odon-, -ast, -est, -ost, -ð, -þ, -iað and-iaþ, thus the operators == and *. The application of F1 to the 17,138 hits of SQ4 reduces this figure to 1,939. F1 is presented in (5).

(5)

==*-on, ==*-odon-, ==*-ast, ==*-est, ==*-ost, ==*-ð, ==*-þ, ==*-iað, ==*-iaþ

F2 is aimed at finding spelling variations in the consonantal endings of verbal forms. It is applied in two steps. The first selects the inflectional forms that end in a consonant, as can be seen in (6).

(6)

==*b, ==*c, ==*d, ==*f, ==*g, ==*h, ==*l, ==*m, ==*n, ==*p, ==*r, ==*s, ==*t, ==*w, ==*x, ==*y, ==*ð, ==*þ

The second step of F2 targets members of the non-verbal classes as well as weak verbs by deleting inflectional forms that end in -on, -en, -an, -es, -um, -end, -as, -est, -ost, -ed, -od, -ig, -ic, -ing, -ung, -un, -us, -nes, -er, -or, -ur, -iað, -iaþ. It must noted that F2 also puts aside the endings -iað, -iaþ, which are selected by F1. When applied to the outcome of SQ4, the first step of F2 reduces its hits from 17,138 to 10,305, which, after the application of the second step of F2, result in 3,533 hits. The second step of F2 is displayed in (7).

(7)

==*on, ==*en, ==*an, ==*es, ==*um, ==*end, ==*as, ==*est, ==*ost, ==*ed, ==*od, ==*ig, ==*ic, ==*ing, ==*ung, ==*un, ==*us, ==*nes, ==*er, ==*or, ==*ur, ==*iað, ==*iaþ

Turning to the comparison with lexicographical sources, the comparison with the inflectional forms provided by the DOE (A-H) has shown that the accuracy of the search algorithm is around 80%.

As regards the comparison with textual sources, the lemmatiser *Norna* has been modified so that it gives access to the inflectional forms that appear both in the DOEC and the YCOE.

A comparison of YCOE and the lemmatiser *Norna* has been carried out. As an illustration, for the letter L the forms in (8a) can be retrieved from the YCOE (165), while those in (8b) appear in *Norna* (148).

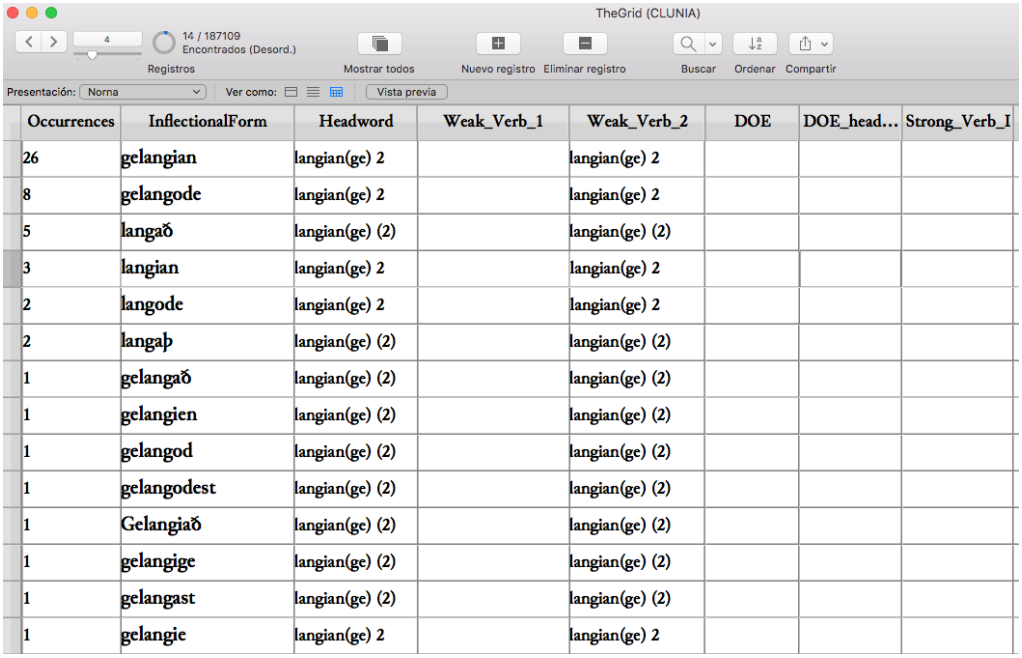
(8)
a. lacað, lacan, laceð, lacende, laðaþ, ladigan, laðod, laðode, lædað, lædan, Læddan, lædde, læddest, læddon, læde, læded, lædeð, lædeþ, Læf, læfan, læfde, læg, lægon, lægun, lær, læran, lærde, lærde, lærdes, lære, læreð, lærest, læs, læst, læstan, læste, læston, læt, læt, læt, Lætað, lætað, lætan, læte, læte, læteð, lætst, lagan, lagon, lah, lata, laþaþ, leag, leanað, leanast, leanige, leccað, leccaþ, lecgað, lecgan, lecge, legde, legdun, legeð, lengde, lengeð, leofa, leofað, leofaþ, leogan, leoge, leohte, leolc, leordan, leorde, Leort, leoþode, lepeþ, Let, Letan, lete, lete, Leton, letton, liban, libban, licað, licgað, licgan, licge, licgean, , lician, licode, licodon, liðan, liðan, liden, liden, lifað, Lifde, lifdon, lifdon, lifgað, lifgan, lifgaþ, lifge, lifiaþ, lifige, ligeð, ligeð, lihteð, limpeð, linnan, linneð, linnið, lixað, lixan, lixeð, lixtan, lixte, lixton, liþ, liþan, locað, locade, locast, locen, lociað, Lofiað, lofian, log, logon, lomp, Longað, losað, losade, losaþ, losian, lucan, ludon, lufast, lufaþ, lufiað, lufian, lufiaþ, lufie, lufien, lufige, lufu, lunnon, lycð, Lyfað, lyfað, lyfde, lyhð, lyhte, lyhteð, lysan, lyst, lyste, lyste, lysteð, lysteð, lysteþ.
b. lac, lace, lece, lec, lacað, lacan, leolc, lecc, laceð, læceð, Læt, let, læte, lætan, Lætað, Leton, lete, læst, læten, læteð, letan, lætst, lett, lætt, læton, lætest, lætaþ, leten, læteþ, leto, leteð, lætenne, Leort, lettes, leteþ, lætæð, lætoð, lætæst, lætin, lætene, Læt.þ, leode, lead, leod, leodan, leoden, lude, ludon, liet, leogað, leoge, lugon, leogan, leag, leoh, luge, leogaþ, lugen, liehð, leogeð, leogendan, leogð, liegeð, lieht, , leah, leore, leoreð, leoran, leor, leorað, lure, leoren, leord, leorest, lorene, leorad, lierest, liereð, loren, les, List, lese, lesan, lisð, lesað, lest, lisseð, lað, liðe, liþe, liðan, liþan, laþ, liðon, liþon, liþ, lið, liðað, lif, life, laf, lifað, lifeð, lifæs, lifæþ, lifen, lifaþ, lifð, lifast, lifeþ, lift, limpð, limpeð, lamp, limpe, limpað, limpes, limpa, limpeþ, lumpe, limpan, lin, lan, linnan, linnen, lunnon, linne, linnið, linneð, leac, Lucan, Luce, locen, lucon, lucað, luc, lycð, leat, luton, lutan, luteð, lute, lut, loten, lutaþ, lutað.

The discrepancies in the number of forms can be attributed to the compilations of the DOEC and the YCOE. An avenue of future research in this respect is the identification of forms in the YCOE that has been deleted in the different editions of the DOEC. Apart from this question, the search algorithm has to be modified to include, at least, three aspects: <g> followed by a front vowel in the same syllable, as is the case with ladigan, legde, legdun, legeð, lifige and lufige; <y> for <i> in accented syllables, as in lyfað and lyfde; -an/-un for -on in unaccented (inflectional) syllables, as in lægun and legdun.

3 Lemmatisation procedure for weak verbs

Weak verbs correspond to the modern 'regular' verbs. The changes in their inflection take place in the suffixal part of the word rather than in the stem, as is the case with strong verbs, counterparts of the modern 'irregular verbs'. In order to find the inflectional forms of weak verbs in the database, it is necessary to list a set of inflectional endings of class 1 and class 2 weak verbs that includes the endings of finite forms (indicative, subjunctive and imperative) and non-finite forms (infinitive, inflected infinitive, present participle, past participle and past participle forms inflected as adjective). The choice of forms has been made by comparing Old English verbal paradigms of the three subclasses of weak verbs in four Old English grammars (Campbell, 1987; Hogg and Fulk, 2011; Sievers, 1903; Stark, 1992). The compilation of these inflectional endings will guide the automatic searches in subsequent steps of the analysis. The result of this task is an inventory of 24 different endings for the paradigm of class 1 and 29 different endings for the paradigm of class 2.
On the lemmatiser, the query strings for each of the endings selected are defined as follows: two equal signs followed by a wild card (an asterisk) and the spelling of the ending. Thus, we will obtain all inflectional forms in the database with the requested ending. For example, we use the query string ==*ianne, the canonical inflectional ending of the inflected infinitive of class 2 weak verbs, and obtain 160 inflectional forms included in the database, most of which are likely to be lemmatised under a lemma from the second weak class. We launch 24 different queries for the canonical endings of class 1 weak verbs and another 29 for the counterparts of the second class of weak verbs. Then, all the hits are checked with the lists of weak verbs available in *Nerthus*. These lists of reference contain more than 2,000 verbs each. To achieve the maximum degree of accuracy in the initial stages of the project, the

criterion was to assign lemma only to combinations of stems of the weak verbs as they appear in the reference lists and canonical endings.
The results of this lemmatisation process consist of more than 4,680 inflectional forms lemmatised with class 1 weak verbs and more than 6,600 inflectional forms for class 2 weak verbs. The lemmatisation of these forms provides paradigmatic information on weak verbs that can contribute to the development of treebanks, since lemmatisation relates the syntactic analysis of the inflectional forms that are lemmatised under the same headword, thus allowing for extensive descriptive analysis as well as quantification. For instance, the inflectional forms lemmatised in the database for class 2 weak verb *lufian* are the following: *lufað, lufiað, lufode, lufige, lufast, lufie, lufaþ, lufodon, lufodest, lufiað, lufien, lufigen, gelufiað, gelufode, gelufie, lufian, gelufod, lufianne, lufiende, lufienne, lufodes, lufiendum, lufodan, gelufoda, lufod, gelufian, gelufodes, gelufodne, lufoden*. All these inflectional forms belong to different parts of the paradigm of the same lemma; therefore, lemmatisation is proved to be an effective tool for linking together syntactic analysis of a given lemma. In turn, we are able to take a look at all the inflectional forms for a given lemma with the click of a button, as shown in Figure 1. The two leftmost columns list the number of occurrences and the inflectional forms, and the column *Headword* shows the lemma assigned to them.



Occurrences	InflectionalForm	Headword	Weak_Verb_1	Weak_Verb_2	DOE	DOE_head…	Strong_Verb_I
26	gelangian	langian(ge) 2		langian(ge) 2			
8	gelangode	langian(ge) 2		langian(ge) 2			
5	langað	langian(ge) (2)		langian(ge) (2)			
3	langian	langian(ge) 2		langian(ge) 2			
2	langode	langian(ge) 2		langian(ge) 2			
2	langaþ	langian(ge) (2)		langian(ge) (2)			
1	gelangað	langian(ge) (2)		langian(ge) (2)			
1	gelangien	langian(ge) (2)		langian(ge) (2)			
1	gelangod	langian(ge) (2)		langian(ge) (2)			
1	gelangodest	langian(ge) (2)		langian(ge) (2)			
1	Gelangiað	langian(ge) (2)		langian(ge) (2)			
1	gelangige	langian(ge) (2)		langian(ge) (2)			
1	gelangast	langian(ge) (2)		langian(ge) (2)			
1	gelangie	langian(ge) 2		langian(ge) 2			

Figure 1: Inflectional forms for the lemma *langian(ge)* (2) 'to grieve' in *Norna*.

The lemmatisation of Old English weak verbs with this procedure needs checking. The DOE (A-H) is the lexicographical source used for this task. As stated above, this dictionary is still in progress but the information listed for verbs from A-H is detailed and central to this investigation. Through the online version of this dictionary, it is possible to search inflectional forms by lemma. For this reason, the lemmatiser *Norna* permits to compare the attested spellings for weak verbs from A to H with those that appear in the DOE. Many unpredictable spellings of weak verbs are available in the DOE that cannot be found by means of automatic lemmatisation, since most of them include spelling variations in prefixes, stems and endings. For instance, for the lemma *bǣdan*(*ge*) (1) 'to force', seven inflectional forms can be obtained by automatic lemmatisation: *bædað, baedde, bædde, bæddon, bæde, bædeð, bæden*. However, the DOE includes eight extra attested spellings in the entry to this verb: *bæddan, baedendrae, baedendre, bædendre, bædendum, bædt, beadætþ* and *bedændræ*. The feedback of previous searches are input into the database and used to refine subsequent searches, so that the amount of manual revision can be reduced.
This task of comparison results in the identification of 12,000 extra inflectional forms for weak verbs, on the grounds of the information found in the DOE verb entries in the letters A-H. The analysis of these extra forms will allow to compile a list of the recurrent variants of canonical forms of weak verbs and, eventually, normalisations patterns (by prefix, stem or ending). This correspondence will be applied to

the search for the inflectional forms of the verbs beginning with the letters I-Y. As illustration, the inflectional form *aflemde* is lemmatised under the headword *āflyman* (1) 'to put to fly' in the DOE. This change of vowel < e > ≈ < y > will be included in the normalisation patterns for the stem of the verbal form. Similarly, the DOE includes the inflectional form *aredad* in the paradigm of the lemma *āredian* (2) 'to arrange', which is a variant of *aredod*, the past participle form. This pattern <o> ≈ <a> is consequently added to the normalisation patterns taking place in the endings of the inflectional forms. In this context, the prefix of the inflectional form *gifered* from the lemma *ferian(ge)* (1) 'to carry' shows the recurrent pattern of variation <gi> ≈ <ge>, which is added to the list of normalisation patterns for the prefixal segment of the attested form. The application of these three types of patterns in the study of inflectional forms I-Y, with the same queries as in the search for the canonical endings of weak verbs, will undoubtedly lead to the lemmatisation of a larger number of inflectional forms for class 1 and class 2 weak verbs and, more importantly, will result in the reduction of manual revision.

The lemmatisation procedure has limitations that need to be taken into account. There are many unexpected spellings within the paradigms of weak verbs, not foreseeable abbreviations, as well as many ambiguities that can only be lemmatised with the help of dictionaries. Gemination and simplification of consonants are also recurrent in verbal forms, thus excluding the automatic lemmatisation of several inflectional forms. This is the case with *cunedon*, from *cunnian*(*ge*) (2) 'to try', which gets consonant simplification. In addition, and according to the DOE, some inflectional forms lemmatised with this method belong to different categories and to other verbs classes. The explanation for this phenomenon is the overlapping of endings among verbs, nouns, adjectives and adverbs, like the ending -e. The inflectional form *dere* was initially lemmatised under the lemma *derian(ge)* (1) 'to hurt', given that it includes the stem of a weak verb of the first class and one of the canonical endings of this class; but the comparison with the DOE showed that, in fact, *dere* is a nominal rather than a verbal form. Finally, the unpredictable forms *auandod*, from *āfandian* (2) 'to test', and *hergendne*, from *erian* (1) 'to plough', which are also available from the DOE, are hard to find by automatic means, at least in this stage of the project.

4 Conclusion

The lemmatisation procedure for the strong verbs of Old English presented in this paper has an accuracy of around 80% before manual revision. After checking the results with lexicographical and textual sources, search strings can be refined and manual revision reduced. The compatibility with treebanks has been addressed with respect to the YCOE. As for weak verbs, the lemmatisation procedure has allowed for the identification of more than 20,000 inflectional forms of weak verbs. However, the comparison with the DOE, as the main lexicographic source, has also proved a crucial step of the lemmatisation procedure because it turns out more inflectional forms but, above all, because it identifies recurrent spelling variants with which normalisation patterns can be defined and subsequent searches can be refined. The inclusion of additional patterns of normalisation and the gradual improvement of searches are likely not only to find more lemmas and inflectional forms but also to reduce the necessary amount of manual revision. In spite of the limitations of semiautomatic lemmatisation, this procedure has allowed us to find a large amount of inflectional forms from weak verbs in Old English.

It will be necessary, therefore, to check the inflectional forms of the DOEC and the YCOE, but further guidelines for search strings have been obtained. To conclude, a step has been taken towards the inclusion of lemma tags into treebanks, which could reinforce the paradigmatic dimension of these parsed corpora and contribute to the retrievability of the information that they contain, including the important aspect of the quantification of the occurrences of a given lemma.

References

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

Brinton, L. and E. Closs Traugott. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

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

Clark Hall, J. R. 1996 (1896). *A Concise Anglo-Saxon Dictionary*. Toronto: University of Toronto Press.

de la Cruz, J. 1075. Old English Pure Prefixes: Structure and Function. *Linguistics* 13: 47-82.

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.

Healey, A. diPaolo (ed.) with J. Price Wilkin and X.Xiang. 2004. *The Dictionary of Old English Web Corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto. Available online at http://www.doe.utoronto.ca/pages/pub/web-corpus.html.

Hogg, R. M. and R. D. Fulk 1992 (2011). *A Grammar of Old English*. Oxford, Wiley-Blackwell.

Krygier, M. 1994. *The Disintegration of the English Strong Verb System*. Frankfurt am Main: Peter Lang.

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

Martín Arista, J. and F. Cortés Rodríguez. 2014. From directional to telics: meaning construction, word-formation and grammaticalization in Role and Reference Grammar. In M. A. Gómez González, F. Ruiz de Mendoza Ibáñez and F. Gonzálvez García (eds.), *Theory and Practice in Functional-Cognitive Space*. Amsterdam: John Benjamins. 229-250.

Martín Arista, Javier (ed.), Laura García Fernández, Miguel Lacalle Palacios, Ana Elvira Ojanguren López and Esaúl Ruiz Narbona. 2016. *NerthusV3. Online Lexical Database of Old English*. Nerthus Project. Universidad de La Rioja. [www.nerthusproject.com]

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

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

Sievers, E. 1903 (1885). *An Old English Grammar*. Boston: The Athenaeum 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.