ATLANTIS
Journal of the Spanish Association of Anglo-American Studies
45.2 (December 2023): 21-46
e-ISSN 1989-6840
DOI: http://doi.org/10.28914/Atlantis-2023-45.2.02

Advances in the Automatic Lemmatization of Old English: Class IV Strong Verbs (L-Y)

Roberto Torre
Universidad de La Rioja
roberto.torre@unirioja.es

The morphological features of an inflectional language like Old English (OE), which also presents generalized spelling inconsistencies, limit the use of lemmatizing and tagging tools that can be applied to natural languages. Consequently, the development of Natural Language Processing (NLP) models, which crucially depend on lemmatized corpora, is slowed down. Against this background, this article develops a lemmatizer within the framework of Morphological Generation that allows for the type-based automatic lemmatization of OE class IV strong verbs (L–Y). The lemmatizer incorporates a set of algorithms to account for features of inflectional, derivational, morphophonological and diatopic variation. The generated forms are automatically compared with Taylor et al. (2003) and Healey et al. (2004) to confirm their attestation and are assigned a lemma. Overall, the research proves successful in setting up form-lemma associations, while highlighting areas of ambiguity and mismatches. The main conclusion of the article is that taking the route of automatic lemmatization with this methodological framework will contribute to the field of OE lexicography by both lemmatizing attested inflectional forms and by identifying areas for manual revision.

Keywords: Old English; lemmatization; strong verb; Natural Language Generation; Morphological Generation

. . .

Avances en la lematización automática del inglés antiguo. Los verbos fuertes de la clase IV (L-Y)

Las características morfológicas de una lengua flexiva como el inglés antiguo que, además, presenta inconsistencias formales generalizadas, limitan el uso de herramientas de

lematización y etiquetado morfológico que pueden ser aplicadas a los lenguajes naturales. En esta situación, el desarrollo de modelos de Procesamiento del Lenguaje Natural, que dependen necesariamente de corpus lematizados, se ve ralentizado. En este contexto, este artículo diseña un lematizador en el marco de la Generación Morfológica que permite la lematización automática por tipo de los verbos fuertes de la clase IV (L-Y). El lematizador incluye un conjunto de algoritmos que dan cuenta de la variación flexiva, derivativa, morfofonológica y dialectal de estos verbos. Las formas generadas son comparadas de forma automática con los dos corpus de referencia del inglés antiguo (Taylor et al. 2003; Healey et al. 2004) para comprobar su atestiguación y asignarles el lema correspondiente. Los resultados de esta investigación demuestran que se pueden crear asociaciones forma-lema e identificar tanto áreas de ambigüedad formal como asociaciones erróneas. La conclusión principal del artículo es que la exploración de las vías de lematización automática dentro de este marco teórico supone una contribución relevante al campo de la lexicografía del inglés antiguo, tanto al lematizar formas flexivas atestiguadas como al señalar las áreas que deben revisarse.

Palabras clave: inglés antiguo; lematización; verbos fuertes; Procesamiento del Lenguaje Natural; Generación Morfológica

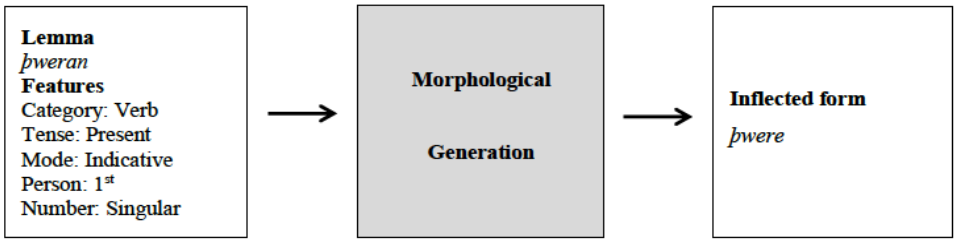
1. Aim and Relevance

This article engages in the automatic lemmatization of Old English (OE) through the Morphological Generation (MG) of inflectional forms of a subset of strong verbs. More precisely, its aims are, on the one hand, to develop an automatic lemmatizer of OE class IV strong verbs (L-Y) based on MG which is able to: (a) generate inflectional forms and account for morphophonological variations at inflectional ending and word stem levels; (b) generate derived counterparts from the simplex inflectional forms; (c) validate the attestation of the generated forms in the OE corpora; and (d) assign lemma to attested forms. On the other hand, the second objective of this article is to test its degree of accuracy.

To do so, MG algorithms must be designed and implemented in the lemmatizer that account for the word formation processes of OE class IV strong verbs (both inflectional and derivational) as well as for diachronic and diatopic variation. Figure 1 shows an example of MG in OE. The lemma *þweran* 'twirl', inflected for person (first), number (singular), tense (present) and mode (indicative), generates the inflected form *þwere* 'I twirl'.

The research is type-based rather than token-based, which implies the assumption of a double layered lemmatizing process: first, the lemmatization of types, the formal abstraction of all the attestations of a word form. Second, the contextual lemmatization of each word form and the disambiguation of competing lemmas. This research deals with the first of these steps.

FIGURE 1. MG in the OE verb *þweran*



In the last decade, significant advances have been made in Natural Language Processing (henceforth NLP) aimed at the understanding and production of natural language by computational means. Such advances include, among others, the development of specific software, like LEMMING (Müller et al. 2015) and GLUE (Wang et al. 2019) and techniques, like automatic text simplification (Saggion 2017) and multitask machine learning (Zang and Yang 2018). Within NLP, *Natural Language Generation* (henceforth NLG) is "the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information" (Reiter and Dale 1997, 1). NLG systems have undergone remarkable advances in different languages. For instance, Forcada et al. (2011) have engaged with Spanish, Khemakhem et al. (2015) with Arabic, Oflazer and Saraçlar (2018) with Turkish and Tapsai et al. (2021) with Thai, just to mention a few. There is also variation in the approach to NLG adopted, ranging from generativist, cognitivist, connectionist or computational perspectives to multilingual approaches.

At the level of the morphological word, the computational generation of language requires the accurate instantiation of the morphological rules that guide word-formation and inflection in different languages. This is known as MG, defined as "the task of producing the appropriate inflected form of a lemma in a given textual context and according to some morphological features" (Ferrés et al. 2017, 110). These systems develop and improve morphological rules by analyzing large, annotated textual corpora and setting up connections between the lemmas, on the one hand, and the inflections and derivatives, on the other.

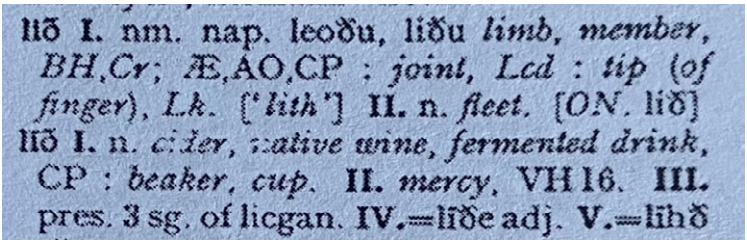
However, several reasons preclude this degree of automation in OE, including the relatively limited amount of textual data, the lack of an orthographic standard and the lack of a lemmatized corpus. The slow pace of progress of the referential lexicographical source, namely the *Dictionary of Old English* (hereafter DOE; Healey et al., 2018), evinces the difficulty of lemmatizing OE. In this article, a method is put forward that will contribute to the automatization of this task.

The remainder of the article is organized as follows. Section two offers an overview of the current panorama of OE lexicographical description. Section three contextualizes this research within the current state of the art. Section four offers an overview of the OE strong verb system and defines the scope of the research. Section five describes the word form generation process. Section six presents the results of the research, while section seven discusses the advances and limits of automatic lemmatization. The main conclusions are dealt with in section eight.

2. OE Lexicographical Description

Classical, glossed textual editions and lexicographical works show variation in their structure and lemma selection, both internally and among different works. This can be observed in the reference dictionaries, namely Bosworth and Toller's (1973) *An Anglo-Saxon Dictionary* and Clark Hall's (1996) *A Concise Anglo-Saxon Dictionary*, whose compilation observes nineteenth century lexicographical practice. Although they provide a wealth of philological information, they are short on terminological precision and analytical systematicity. Divergences can be found in headword spelling, morphological description, internal organization of the lexicographical entries, textual referencing and form attestation. As such, circularity and cross-referencing stand out as the major problems when using these dictionaries. This is illustrated by the entries for *lið* and *līð* in Clark Hall (1996, 220) given in figure 2.

Figure 2. The entries for *lið* and *līð* in Clark Hall (1996)


līð I. nm. nap. leoðu, liðu *limb, member,*
BH,Cr; Æ,AO,CP : *joint, Lcd* : *tip* (*of*
finger), *Lk.* ['*lith*'] II. n. *fleet.* [*ON.* lið]
līð I. n. *cider, native wine, fermented drink,*
CP : *beaker, cup.* II. *mercy,* VH 16. III.
pres. 3 sg. of licgan. IV.=līðe adj. V.=līhð

The DOE by Healey et al. (2018), for its part, constitutes a great advance in the field of OE lexicography. It is being compiled under strict criteria of exhaustiveness and accuracy, both guaranteed by the electronic implementation and analysis of Healey et al.'s (2004) *Dictionary of Old English Corpus* (henceforth DOEC), which comprises 3,000,000 words. However, while constituting the leading project in OE studies, the DOE has, to date, only reached the letter I and is not expected to be completed in the near future.

The second major corpus of OE is the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (hereafter YCOE) which includes up to 1,500,000 words. Other partial

corpora are Rissannen et al.'s (1991) *Helsinki Corpus of English texts* (300,000 words) and Pintzuk and Plug's (2001) *York-Helsinki Parsed Corpus of Old English Poetry* (70,000 words). However, none of these corpora are lemmatized and only the YCOE and the York-Helsinki corpus incorporate linguistic metadata, including morphological tagging (that is, part of speech category and inflectional features) and syntactic parsing.

This panorama evinces the need for advances in the methodological and analytical procedure of OE lexicography if the discipline is to enter the field of digital humanities. In this vein, Martín Arista et al.'s (2021) ParCorOEv2: *An Open Access Annotated Parallel Corpus Old English-English* anticipates new avenues of research. ParCorOEv2 currently files 110,000 word tokens and although in terms of size there is no comparison with the YCOE, it does have several advantages as it is enriched with linguistic metadata not found in the YCOE, including gloss translation and word lemmatization.

The strengths of a lemmatized corpus are its search power and the establishment of paradigmatic form-lemma associations on a principled basis. Lemmatized corpora allow searches for a single lemma to be conducted, as well as the retrieval of all the inflectional forms of the lemma attested in the corpora in the result. The identification of form-lemma associations allows for, among other possibilities, the exhaustive analysis of complementation patterns, the semantic analysis of lexical classes, the paradigmatic organization of the lexicon, the development of automatic translation tools and natural language generation.

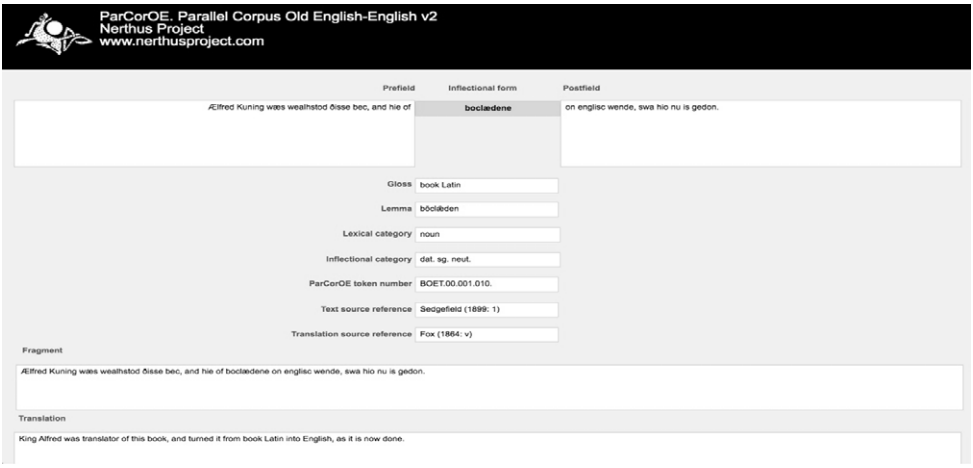
Consequently, the lemmatization of the OE lexical stock is a crucial task to be undertaken in order to close the gap between classical philological studies and the application of up-to-date computational procedures for OE linguistic analysis. That said, the present state of compilation, the size of the reviewed corpora and the textual data they store preclude statistical analysis and limit the potential machine training needed to develop automatized lemmatizers or taggers. Steps need to be taken to enrich the extant corpora with the required information to automatize lemma assignment processes. Such an enhancement will contribute to accelerate the development of a lexicographical product that meets twenty-first century standards.

3. Previous Approaches to OE Automatic Lemmatization

This research is framed within the *Nerthus* project (www.nerthusproject.com). The goals of this project are to provide a comprehensive analysis of the OE lexical stock and turn the findings into modern, electronic, searchable lexicographical products (Martín Arista 2012; 2013). Martín Arista et al.'s (2021) ParCorOEv2 is the project's most recent contribution to academia. It is a searchable, lemmatized corpus comprising around 110,000 records. Its compilation is still in progress and will reach 300,000 records in its forthcoming version. Each record files an inflectional form (token) along with its tagging (file, number, lemma, lexical category, inflectional category and gloss).

Furthermore, each token displays a concordance with the prefield and postfield of the concorded word, as well as the OE text, the Present-Day English (PDE) translation and the sources for both texts. Figure 3 shows the record for *bodclæden* 'Latin book' in Boethius.

FIGURE 3. A sample view of ParCorOEv2 (Martín Arista et al. 2021)



ParCorOEv2 is being developed in database format, which has several advantages over dictionaries and textual corpora. One is that lexical databases are designed under fixed sets of rules, units and relations, which guarantee the formal unicity, exhaustiveness and systematicity of the final product. Another is that lexical databases are flexible, they can be easily modified or expanded and they can be combined and linked.

ParCorOEv2 is fed with the lexicographical, morphological, etymological and contextual information stored in the Knowledge Base of OE (KBOE; Martín Arista 2017). The KBOE stores data from Clark Hall's (1996) dictionary as well as from the dictionaries of Bosworth and Toller (1973), Sweet (1976) and Healy et al. (2018). Seebold (1970), Heidermanns (1993) and Orel (2003) account for most of the etymological data, while the DOEC provides the textual background. However, lemma assignment in ParCorOEv2 is yet to be implemented manually.

Some recent attempts have been made to automatize lemma assignment, including the works by Novo Urraca and Ojanguren López (2018), who have successfully incorporated lemma assignment to the YCOE syntactic treebanks; Metola Rodríguez (2015; 2017), who has tackled the lemmatization of strong verbs; Tío Sáenz (2019), who has dealt with weak verbs; and García Fernández (2020), who has engaged in the identification of lemmas for preterit-present, anomalous and contracted verbs.

These works adopt different approaches and scopes which depend on the features of the different verb classes.

Metola Rodríguez (2015; 2017) and Tío Sáenz (2019) have developed sets of search codes, named query strings (QS). Each QS aims to search for a particular feature of the verbal inflection, including *ablaut* or gemination of consonants, among others. These particular QSs are based on the morphological properties of strong and weak verbs, respectively. Specifically, Metola Rodríguez (2015; 2017) focuses on the stem while Tío Sáenz (2019) focuses on the inflectional endings in an attempt to maximize data retrieval.

In (1) below, the fourth query string (QS4) proposed by Metola Rodríguez (2017, 70) for the verb *beodan* 'to command' is shown. QS4 aims to identify specific stems which may be preceded or followed by any prefix and/or inflectional ending.[1]

(1) ==*beod*, ==*bead*, ==*bud*, ==*bod*, ==*bied*, ==*biet*, ==*biest*.

The use of the wildcard (*), however, makes these searches virtually unrestricted, resulting in the return of a remarkable number of undesired results, including *beada* 'counsellor' and *beadas* 'tables' which belong in the nominal class. Consequently, filtering and manual revision of the data is essential. As it stands, Metola Rodríguez (2017, 73) quantifies the accuracy of his approach at around 80% before manual revision when comparing his lemmatized forms with those of the DOE.

While Tío Sáenz's (2019) approach faces the same problems, she modifies the methodology of her research in two ways. By including participial inflectional forms in the targets of her QS and by comparing her findings not only with the DOE, but also with the YCOE, Tío Saénz (2019) is able to lemmatize 6,300 forms from weak verbs beginning I-Y, while acknowledging (Tío Sáenz 2019, 544) that validation of this data in the textual fragments of the DOEC is still pending.

For her part, García Fernández (2020) departs from QS-based studies and moves closer to an MG approach. This author compiles a list of inflectional forms to which prefixes are attached in order to develop complex counterparts. The attestation of these generated complex forms is checked both in the DOEC and the YCOE. The simplex forms are obtained from a selection of grammars, including Brunner (1965), Campbell (1987) and Hogg and Fulk (2011), among others. The prefixes are those provided in Kastovsky (1992), along with their spelling variants. For the sake of illustration, out of the inflectional forms of *swapan* 'sweep', García Fernández (2020) identifies the following complex forms in the YCOE that are not included in the DOEC—with lemmas given in brackets—*tosweop* (***toswapan***); *emswapen*, *ymbswapen*, *ymbsweop*, *ymbsweopan*, *ymbsweopon* (***ymbswapan***).[2]

[1] The wildcard (*) stands for any number of characters at either side of the stem.

[2] Italics and boldface as in the original.

Notwithstanding the advance towards the lemmatization of OE of this approach, it overlooks two major features of corpus studies, to wit, formal variation and textual attestation. With respect to the former, by assuming a direct formal link between the stems and inflectional endings of the simplex and complex forms, García Fernández (2020) disregards several well-attested morphophonological changes and ignores spelling variation. As for the latter, García Fernández (2020) connects the attestation of complex forms to the occurrence of simplex forms in the corpora. Consider the case of *beþearfst* 'you have need' as an illustration. The inflectional form *þearfst*, belonging in the paradigm of þurfan 'to need', is not attested in the DOEC (García Fernández 2020, 133). This prevents the author from generating the form *beþearfst* in the paradigm of *beþurfan* 'to have need', although this is, however, attested in the DOEC, as (2) shows.

(2) [PsGlC (Wildhagen) 015600 (15.2)]
Ic sæde drihtne god min eart þu forþon goda minra þu ne ***beþearfst***
'I have said to the Lord, thou art my God, for thou **hast** no **need** of my goods' (Douay-Rheims 1971, 586).

The review of these methods shows that there are several areas for improvement as regards technical procedures, levels of accuracy and textual attestation in the development of a process of automatic lemmatization of OE. This research designs a method that offers improvements in all these areas.

4. An Overview of the OE Strong Verb System

Against the background described above, this research pursues the automatic lemmatization of OE class IV strong verbs (L-Y). The scope of the research has been limited to the letters that have not yet been published by the DOE. In the remainder of this section, I offer an overview of the OE strong verb system.

The classification of OE strong verbs into seven classes based on the different *ablaut* patterns of the verbal paradigms constitutes, according to von Mengden (2011, 123-24), an undisputed fact in the description of OE. This classification is inherited from the original Proto-Germanic grades (Mailhammer 2007, 58), given in figure 4.

Figure 4. Proto-Germanic *ablaut* patterns

	ablaut pattern	root	vowel 1	vowel 2	vowel	vowel 4
I	e-a-Ø-Ø	CViC	CeiC	CaiC	CØiC	CØiC
II	e-a-Ø-Ø	CVuC	CeuC	CauC	CØuC	CØuC
III	e-a-Ø-Ø	CVCC	CeCC	CaCC	CØCC	CØCC

IV	e-a-e:-Ø	CVR	CeR	CaR	Ce:R	CuR
V	e-a-e:-e	CVC	CeC	CaC	Ce:C	CuC
VI	a-o:-o:-a	CVC	CaC	Co:C	Co:C	CaC

The historical evolution of these phonological syllabic distributions leads to the formation of the OE strong verb classes. A seventh class was analogically created in OE, consisting of originally reduplicating verbs, which did not show *ablaut* patterns in Proto-Germanic. This is summarized in figure 5.

Figure 5. The seven classes of OE strong verbs (adapted from Campbell 1987)

	Infinitive	**Preterit 1**	**Preterit 2**	**Past Participle**
I	*rīdan* 'ride'	*rād*	*ridon*	*geriden*
II	*bēodan* 'command'	*bēad*	*budon*	*geboden*
III	*bindan* 'bind'	*band*, *bond*	*bundon*	*gebunden*
IV	*beran* 'bear'	*bær*	*bǣron*	*geboren*
V	*giefan* 'give'	*geaf*	*gēafon*	*gegiefen*
VI	*faran* 'go'	*fōr*	*fōron*	*gefaren*
VII	hātan 'command'	hēt	*hēton*	*gehāten*

The system is organized around the gradation patterns seen in the stem vowel of the verb forms. Four groups are distinguished. The infinitive grade, for the present tense; the preterit 1 grade for the 1st and 3rd singular forms; the preterit 2 grade, for the other forms of the preterit, and the past participle grade. Regarding class IV, Krygier (1994, 49) summarizes its *ablaut* in the series eLV – æLV – ǣLV – oLV, where L stands for a liquid sound (l, r) and V for a vowel sound, thus *stelan – stæl – stǣlon – stolen*. Three verbs resist this distribution and present a different *ablaut* pattern while retaining the original stem consonant, i.e. *cuman* 'come', *niman* 'take' and *striman* 'resist'.[3]

Only Levin (1964) and von Mengden (2011) argue against this traditional representation of the OE strong verb paradigm. Levin (1964) opts for subsuming and rearranging classes IV and V, displacing *cuman*, *niman* and *striman* to class V and incorporating *metan* 'measure', *seon* 'see' and *biddan* 'pray' into class IV. For its part, von Mengden (2011) defends the incorporation of a fifth vowel grade resulting from the *i*-mutation of the stem vowel in the second and third persons of the present indicative. Von Mengden's (2011) approach is justified on the basis that apophonic changes

[3] Liquid and nasal consonants form the group of the resonant sounds which characterized the Proto-Germanic *ablaut* series for this category (see figure 4; cf. Levin 1964, 157).

originally conveyed morphological value. Figure 6 shows von Mengden's (2011) five-vowel series.

Figure 6. Von Mengden's (2011) class IV strong verbs with apophonic variants

	1 (Inf.)	1' or 5	2 (Pret. 1)	3 (Pret. 2)	Past Part.
IV	*stelan* 'steal'	*stilþ*	*stæl*	*stǣlon*	*Gestǣlen*

For the current research, I shall stick to the traditional classification in the terms described by Krygier (1994). Therefore, following Clark Hall's *A Concise Dictionary of Anglo-Saxon*, the following underived verbs (L-Y) have been analysed: *niman*, *sceran*, *stelan*, *stenan*, *striman*, *swelan*, *teran* and *þweran*.[4]

5. The Morphological Generation Process

This section describes the methodological lines guiding this research. The choice of the strong verb class to conduct type-based automatic research is justified for the following reasons. First, strong verbs lie at the origin of lexical creation in OE. In Kastovsky's (1992, 297) words:

> As these examples show, strong verbs, or, rather, the various stem allomorphs of strong verbs with their different *ablaut* grades form the basis for both suffixal and suffixless derivatives, which in turn may act as the starting-point for further derivational series, as in drincan drunc{en) ->-drunc + n + ian->-drunc + n + ing, or faran 'travel'->for f. 'journey'-+fer + an (< */or+j + an-) 'go on a journey, travel, set out'-*fer + end m. 'sailor' jer+nessi. 'passage, transition, passing away.' Hinderling's (1967, 2) claim that a description of word-formation in the Germanic languages has to take the strong verbs as its starting-point is thus fully justified.

Second, the *ablaut* patterns of this class shown in figure 5 constitute a solid base for MG algorithm instantiation, even though "in Old English the productivity value of *ablaut* disappeared almost without a trace" (Krygier 1994, 17).

The upcoming sections describe the implementation of MG algorithms regarding inflectional endings (5.1), word internal mutations (5.2), the generation of complex forms (5.3) and the automation of the attestation process (5.4).

[4] There is no consensus in the sources consulted regarding paradigm and class adscription for *stenan* 'groan, roar'. Krygier (1994, 50) lists it among the class IV strong verbs; Sweet (1976, 161) also considers it a strong verb, although he includes it in his fifth class; Bosworth and Toller (1973, 915), however, assign this lemma to the weak paradigm.

5.1. Generating Class IV Strong Verb Forms: Inflection

The selected verbs are inflected for infinitive, present indicative (person and number), preterit indicative (person and number), present subjunctive (number), preterit subjunctive (number), inflected infinitive, present participle, past participle and imperative (number). Example (3) illustrates the inflection of *stelan* 'steal'.

(3)

Infinitive	*Stelan*	Pres. subj. (sg.)	*Stele*
Inflected Infinitive	*Stelenne*	Pres. subj. (pl.)	*Stelen*
Pres. ind. (1st sg.)	*Stele*	Pret. subj. (sg.)	*Stæle*
Pres. ind. (2nd sg.)	*Stelest*	Pret. subj. (pl.)	*Stælen*
Pres. ind. (3rd sg.)	*Steleþ/steleð/steleth*	Pres. part.	*Stelend*
Pres. ind. (pl.)	*Stelaþ/stelað/stelath*	Past part.	*Stolen*
Pret. ind. (1st/3rd sg.)	*Stæl*	Imperative (sg.)	*Stel*
Pret. ind. (2nd sg.)	*Stæle*	Imperative (pl.)	*Stelaþ/stelað/stealth*
Pret. ind. (pl.)	*Stælon*		

As can be seen in (3), formally ambiguous forms arise, as in the first singular present indicative and the singular present subjunctive. Given that the research is not token-based, duplicated forms within the same verbal paradigm will be reduced to one single type at a later stage (see section 5.3). The graphemes <ð>/<þ>/<th> are duplicated, giving rise to couplets as in *stelaþ/stelað/stelath*.

The paradigm given in (3) does not correspond to the classical West Saxon description traditionally represented in OE grammars. It is, however, an artificially standardized paradigm that serves as the starting point for the design of the programming algorithms that underlie the MG of inflectional forms. In other words, the paradigm contains a set of reconstructed forms upon which the lemmatizer operates to generate forms that show well-attested morphophonological changes. Such changes may occur both in the inflectional endings—namely assimilation of consonants, simplification of consonant clusters and weakening—as described in (4) below or in the stem, as happens with the replacement of dental fricatives with dental plosives predicted by Verner's rule. The underlying reasoning behind this methodological decision is that these reconstructed forms allow the derivation of all other assimilated and syncopated forms in the various verbal paradigms.

I draw on Campbell (1987, 299-300) and Hogg and Faulk (2011) to develop the algorithms guiding the MG of alternating inflectional endings. In this respect, (4a-c) show alternations in the present indicative first, second and third person singular, (4d) in the preterit indicative plural form, (4e) in the inflected infinitive, (4f) in the present

participle and (4g) in the past participle. Duplications of <ð>/<th> have been omitted for the sake of clarity.[5]

(4)

a. -e > -æ
-e > -o
-e > -u

b. -est > -ist
-est > -st
-dst > tst > -st
-þst > -sst > -st
-þs > -ts
-ngst > -ncst
-gst > -hst > -xt

c. -eþ > -iþ
-eþ > -þ
-eþ/-iþ > -et/-it
-tþ/-dþ > tt
-sþ > -*st*
-gþ > hþ
-ngþ > -ncþ
-tt > -t
-þþ > t

d. -on >-an
-on > -un

e. -enne > -anne
-enne > -onne
-enne > -ene
-enne > -ane

f. -end > -and
-end > -ind
-end > -ende
-end > -ande
-end > -onde
-end > -ænde
-end > -endi
-end > -ændi

g. -en > -in
-en > -æn

[5] I draw strictly on Campbell's (1987) account of inflectional endings.

These algorithms lead the lemmatizer to rewrite the paradigm in (3) and generate forms with the alternative endings that respond to dialectal variations or phonological processes of assimilation and elision. These algorithms operate on a recursive basis. Once the operation is completed, the system searches through the selected forms again to apply the next modification. In a nutshell, the algorithm behind (4a) *-e* > *-a* commands the system to look for those forms in the paradigm ending in *-e* and to rewrite them as *-a*. Once this process is completed, the system searches again for the forms ending in *-e* and rewrites them as *-o*. In a third stage, the same forms are generated with a *-u* ending. Thus, the lemmatizer generates several plausible forms for the first person singular of the present indicative which respond to different diatopic variants. Considering the paradigm in (3), the forms *stele*, *stelæ*, *stelo* and *stelu* are generated.

Finally, the present and past participles are inflected following the weak and strong adjectival paradigms. Following Campbell (1987, 266-72), the endings ø; *-ne*; *-es*; *-um*; *-e*; *-ra*; *-u*; *-re*; *-a*; *-an*; *-ena* are attached to the participial forms.

5.2. Generating Class IV Strong Verb Forms: Mutation

Taking the forms in (3) as a starting point, this section describes the algorithms designed to generate mutated forms, including phenomena like *i*-mutation of second and third present indicative singular forms, Verner's law or diatopic variation. The linguistic and theoretical evidence that underlie these algorithms comes from Campbell (1987, 312-13) and Krygier (1994, 50). The set of stem rewriting algorithms is described in figure 7.

FIGURE 7. MG of mutated forms

	Algorithm Description	**Linguistic motivation**
Algorithm #M1	*-e-* > *-i-* in 2nd and 3rd person singular	*i*-mutation
Algorithm #M2	*-e-* > *-ie-* > *-i-* after initial *sc-*	Palatal diphthongization
Algorithm #M3	*-e-* > *-eo-* in present indicative	Diphthongization
Algorithm #M4	*-eo-* > *-ea-* - in present indicative	*u*-mutation
Algorithm #M5	Extension of *-eo-* to the present system	Analogical extension
Algorithm #M6	*-i-* > *-io-* in present indicative	*u*-mutation
Algorithm #M7	Extension of *-io-* to the present system	Analogical extension
Algorithm #M8	*-ie-* > *-i-* > *-y-* in present indicative	Dialectal monophthongization in lWS
Algorithm #M9	*-æ-* > *-e-* in preterit forms	Diatopic variation
Algorithm #M10	*-æ-* > *-ea-* in preterit forms	*u*-mutation
Algorithm #M11	*-o-* > *-a-* in preterit forms	Diatopic variation
Algorithm #M12	*-u-* > *-y-* in past participle	*i*-mutation
Algorithm #M13	-þ- > -d-	Verner's law

5.3. Generating Class IV Strong Verb Forms: Derivation

The algorithms described in sections 4.1 and 4.2 account for the generation of simplex verbs. The complex counterparts are generated by attaching preverbal items to the simple forms. To do so, the items L-Y as described by Metola Rodríguez (2015) and García Fernández (2020) have been selected. Although outside of this alphabetical selection, the prefix *ge-* has also been included in the set of preverbal items given its participation in the formation of past participles. Since the scope of this article is limited to the type-based lemmatization of OE, the spelling variations of these preverbal items has been accounted for and a distinction has been made between canonical (lemma) and non-canonical forms. In this respect, the grapheme <þ>, rather than < ð> or <th>, has been selected as canonical. Example (5) shows the canonical form (in bold) and non-canonical forms of the selected preverbs.

(5) **ge-**(*cg-*, *g-*, *ga-*, gæ-, gæn-, gær-, *gad-*, *gan-*, *gar-*, *ged-*, *gen-*, *gem-*, *ger-*, *gi-*, *gif-*, *gim-*, *gy-*); **med-**(*me-*, *met-*, *mi-*, *mid-*, *mið-*, *miþ- mith-*, *mod-*); **mis-**(*miss-*, *mus-*); **niþer-**(*neoþer-*, *nioþer-*, *nyþer-*, *nieþer-*, *niþor-*, *niðer*, *neoðer-*, *nioðer-*, *nyðer-*, *nieðer-*, *niðor-*, *nither*, *neother-*, *niother-*, *nyther-*, *niether-*, *nithor-*); **o-**; **of-**(æf-, *af-*, *off-*); **ofer-**(*eofer-*, *eofor-*, *ofær-*, *ofern-*, *ofor-*, *of'-*, *ofyr-*, *ouer-*, *ouyr-*); **on-**; **or-**; **oþ-**(*oeþ-*, *oð-*, *oeð- oth-*, *oeth-*); **onweg-**(*anweg-*, *aweg-*, *unweg-*); **riht-**(*reht-*, *reoht-*, *rieht-*, *ryht-*); **sam-**; **sin-**; **sub-**; **to-**; **twi-**(*twig-*, *twy-*); þri-(þry-, *þrie-*, ðri-, *ðry-*, *ðrie-*, *thri-*, *thry-*, *thrie*); þurh-(þorh-, *ðurh-*, ðorh-, *thurh-*, *thorh-*); **un-**; **under-**(*und-*, *undern-*, *ynder-*); **up-**(*upp-*); **ut-**(*utt-*, *vt-*); **uþ-**(*uð-*, *uth-*); **wan-**; **wiþ-**(*wið-*, *with-*); **wiþer-**(*wiþere-*, *wiþyr-*, *wiðer-*, *wiðere-*, *wiðyr-*, *wi*ther-, *wi*there-, *withyr*); **ymb-**(*ym-*, *ymbe-*, *emb-*, *embe-*, *eme-*, *imb-*).

The implementation of the algorithms to generate complex forms brings to an end the MG of inflectional forms. Before conducting the search in the corpora, duplicated forms need to be deleted when they are generated in the same verb paradigm. Example (6) illustrates this point.

(6) *gestenan* > *gesteanaþ* (pres. ind. pl.); *gestenan* > *gesteanaþ* (imp. pl.)

As seen in (6), the form *gesteanaþ* occurs twice in the paradigm of *gestenan*, both as a present indicative plural and as an imperative plural. This being a type-based study, morphological tagging is not relevant at this stage. To avoid the undesired identification of duplicated forms in a paradigm, homographic couplets or triplets are reduced to just one single occurrence. It may, however, be the case that homographic forms are generated in two different verbal paradigms. In such cases, both forms would be maintained to guarantee the assignment of the two lemmas. Lemma disambiguation would come from future contextual analysis, although there are in fact no such instances in this study.

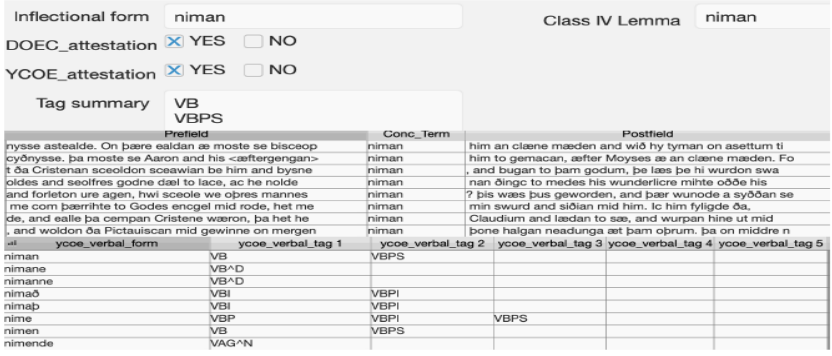
5.4. Automatic Attestation of Forms

Following the principles of MG, the designed algorithms attempt to maximize the generation of plausible inflectional forms, addressing all the potential issues that may lead to formal variation, namely dialectal and morphophonological variation. Given the limited amount of surviving textual material, it is necessary to check which of the generated forms are indeed attested in the selected corpora. To do so, an index needs to be generated for each of the corpora. The index for the DOEC (Healey et al. 2004) has been obtained from the concorded version of the corpus filed in the knowledge base the *Grid* developed by the Nerthus Project. As for the YCOE (Taylor et al. 2003), the index has been obtained by extracting all the forms containing a verbal tag (see appendix 1 for a full list of verbal POS—part of speech—tags and their meaning). A sample view of some of the available forms and their tags is offered in example (7).

(7) *sceare* (VBDS); *terendan* (VAG^N); *undernam* (VBDI)

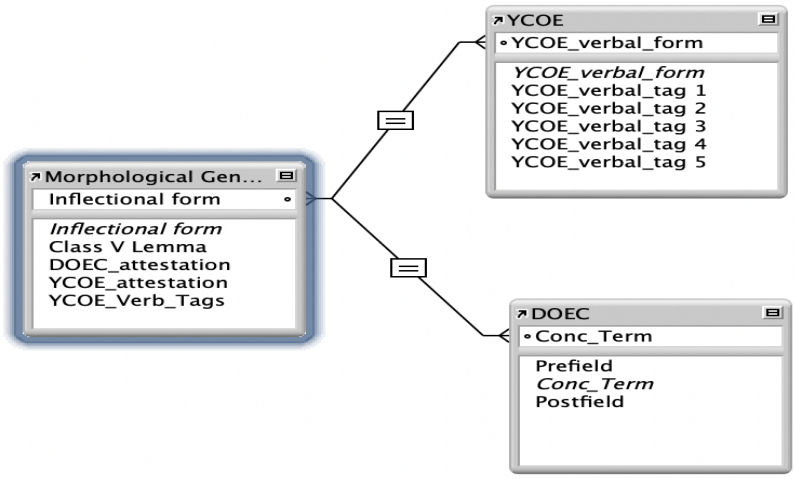
Recapitulating, three lists of forms have been compiled at this stage, namely the MG forms, the index from the DOEC and the verb-tagged forms from the YCOE. These lists are filed in separate databases containing the following data. The MG database includes a field for the generated form (*Inflectional form*), a field for the lemma from which the form has been generated (*Class III Lemma*), a field to check attestation in the DOEC (*DOEC attestation*), a field to check attestation in the YCOE (*YCOE attestation*) and a field for the YCOE POS (*YCOE verb tag*) if the form is attested in the YCOE. The DOEC database has a field for the indexed form in the DOEC (*ConcTerm*), a field for the text before the concorded term (Prefield) and a field for the text following the concorded term (*Postfield*). The YCOE database contains a field for the inflectional form in the YCOE (*YCOE_verbal_form*) and a field for the POS tag (*YCOE_verbal_tag*). Figure 8 shows the information displayed in the three databases.

FIGURE 8. The form *niman* in the MG, DOEC and YCOE databases



The three databases are related to one another through the fields *Inflectional form*, *ConcTerm* and *YCOE_verbal_form* in such a way that if there is a spelling coincidence between the form in *Inflectional form* and *ConcTerm* and/or the *YCOE_verbal_form*, the corresponding *DOEC_attestation* and/or *YCOE_attestation* fields (YES) are activated. Figure 9 shows the grid of relations on the relational database.

FIGURE 9. Relational interface of the MG, DOEC and YCOE databases



6. RESULTS

With the implementation of the inflectional and derivational rules along with those accounting for word internal mutations, the MG lemmatizer generated 328,118 inflectional forms of which 315 were attested in the corpora. These forms corresponded to 45 lemmas, which may or may not themselves be attested. All in all, 174 inflectional forms were attested in the DOEC, belonging in 40 different lemmas, 139 forms from 22 distinct lemmas were attested in both corpora and 2 inflectional forms belonging to the lemmas *geþweran* and *wiþniman* were attested only in the YCOE. The forms attested in the DOEC, organized by lemma, are shown in (8).

(8) **geniman:** *genimanne*, *genimaþ*, *genimað*, *genimende*, *genimest*, *geniomað*, *geniomæ*, *genioman*, *geniomanne*, *geniomende*, *genome*, *genomen*, *genomun*, *genumenan*, *genumenne*, *genymað*, *genyme*, *genymest*, *genymst*, *ginim*, *ginime*, *ginimeð*, *giniomað*, *giniomanne*, *ginom*, *ginome*, *ginomon*, *ginomun*, *ginumen*; **gesceran:** *gescert*, *gescir*, *gescoren*, *gescorene*; **gestelan:** *gesteal*, *gestele*, *gistele*; **gestenan:** *gestæne*, *gestint*, *gystene*; **geswelan:** *geswel*, *geswelinde*; **geteran:** *getearende*, *geteorað*, *geteoraþ*, *geter*; **geþweran:** *geðwære*, *geþwære*, *geþwearan*, *geþworen*; **medniman:**

minime; **medstelan:** *mistel*; **medstenan:** *mesten*; **medteran:** *metere*, *meteren*; **medþweran:** *modþwære*; **niman:** *nimend*, *nimo*, *niomað*, *nioman*, *niomande*, *niomaþ*, *niomende*, *niomendra*, *niomu*, *nomun*, *numenne*, *nymanne*, *nymend*, *nymendan*, *nymendum*, *nymeþ*; **oferniman:** *ofernimð*; **oferstelan:** *oferstælon*; **ofniman:** *ofnimað*, *ofnimð*, *ofnumen*; **ofteran:** æfter, *æfteran*, *æftere*, *after*, *aftere*, *ofter*; **oniman:** *onam*, *oniman*; **onstenan:** *onstenst*, *onstent*; **oteran:** *oter*; **rihtniman:** *rihtnaman*; **rihtsceran:** *rihtscire*; **sceran:** *scære*, *scæron*, *scearan*, *scearen*, *sceor*, *sceort*, *scer*, *sceran*, *scere*, *sceren*, *scerð*, *scir*, *scirað*, *scire*, *scireþ*, *sciru*, *scirþ*, *scorenum*, *scyre*; **stelan:** *stælen*, *steal*, *stelað*, *stelende*, *stelendes*, *steleþ*, *stelþ*, *stilith*; **stenan:** *stæne*, *stænen*, *stean*, *sten*, *stene*; **striman:** *strame*, *strim*, *strym*; **subteran:** *subter*; **swelan:** *swel*; **teran:** *tear*, *tearan*, *teare*, *tearen*, *teart*, *teoran*, *teore*, *teorende*, *teorendum*, *teorenne*, *teoreþ*, *ter*, *terað*, *teraþ*, *terende*, *terendum*, *toren*; **toniman:** *tonaman*; **tosceran:** *toscereð*; **tostenan:** *tostenð*; **toteran:** *toteran*, *totere*, *toteren*, *toterð*, *toterende*, *totoren*, *totorenne*; þurhstenan: þurhstinð; þweran: ðwære, *ðweoran*, *ðweore*, *ðweran*, *þweor*, *þweoran*, *þweore*, *þweoren*, *þwere*, þwert; **underniman:** *undernimað*, *undernimð*, *undernumen*; **unsceran:** *unscoren*; **upniman:** *upnimende*; **utniman:** *utnam*, *utniman*, *utnimð*; **utteran:** *utter*, *utteran*.

The forms attested both in the DOEC and the YCOE are listed in (9). Forms are given along with the POS tag provided by the YCOE. Whenever a given form is assigned different verbal tags in the corpus, all the tags are stated.

(9) **geniman:** *geniman*-(VB), *genyman*-(VB), *genimenne*-(VB^D), *genime*-(VBP)-(VBPS), *genimst*-(VBPI), *genimeþ*-(VBPI), *genimeð*-(VBPI), *genimþ*-(VBPI), *genimð*-(VBPI), *genom*-(VBDI), *genam*-(VBDI), *gename*-(VBD)-(VBDS), *genomon*-(VBDI), *genamon*-(VBDI), *genoman*-(VBDI), *genaman*-(VBDI), *genamun*-(VBDI), *genimen*-(VB), *genamen*-(VBD)-(VBDS), *genim*-(VBI), *genym*-(VBI), *genumen*-(VBN)-(VBN^A), *genumenum*-(VBN^D), *genumene*-(VBN^A)-(VBN^N); **gesceran:** *gescyrt*-(VBI)-(VBN), *gescer*-(VBDI); **gestelan:** *gestæle*-(VBPS), *gestilst*-(VBPI), *gestilð*-(VBPI); **gestenan:** *gesteon*-(VB), *gestent*-(VBPI); **geþweran:** *geþwere*-(VBPS), *geþwerað*-(VBPI), *geþwer*-(VBI), *geðwere*-(VBPS); **niman:** *niman*-(VB)-(VBPS), *nyman*-(VB)-(VBPS), *nimenne*-(VB^D), *nymenne*-(VB^D), *nimanne*-(VB^D), *niomanne*-(VB^D), *nime*-(VBP)-(VBPI), *nyme*-(VBP)-(VBPS), *nimest*-(VBPI), *nimst*-(VBPI), *nymst*-(VBPI), *nimeþ*-(VBPI), *nimeð*-(VBPI), *nymeð*-(VBPI), *nimþ*-(VBPI), *nymþ*-(VBPI), *nimð*-(VBPI), *nymð*-(VBPI), *nimaþ*-(VBI)-(VBPI), *nymaþ*-(VBPI), *nimað*-(VBI)-(VBPI), *nom*-(VBDI), *nam*-(VBDI), *nome*-(VBDS), *name*-(VBD)-(VBDS), *nomon*-(VBDI), *namon*-(VBDI), *noman*-(VBDI), *naman*-(VBDI), *nimen*-(VB)-(VBPS), *nomen*-(VBD), *namen*-(VBD)-(VBDS), *nim*-(VBI), *nym*-(VBI), *nymað*-(VBI)-(VBPI), *nimende*-(VAG^N), *nymende*-(VAG), *nymenda*-(VAG), *numen*-(VBN), *numene*-(VBN^N), *nymen*-(VB), *nymene*-(VB^D); **oferniman:** *ofernime*-(VBPS), *ofername*-(VBDS)-(VBPS), *ofernumen*-(VBN); **oferstelan:** *oferstæle*-(VBPS); **onniman:** *onniman*-(VB), *onnime*-(VBPS); **sceran:** *scieran*-(VB), *sciran*-(VB), *scyran*-(VB), *sceoran*-(VB), *scyreþ*-(VBPI), *scireð*-(VBPI), *scerað*-(VBI), *sceare*-(VBDS), *sceron*-(VBPS), *sciren*-(VBPS), *scorene*-(VBN^A)-(VBN^N), *scirð*-(VBPI), *scyrð*-(VBPI); **stelan:** *stelan*-(VB), *stelenne*-(VB^D), *stele*-(VBPS), *steleð*-(VBPI), *stel*-(VBI), *stæle*-(VBDS), *stælon*-(VBDI), *stilð*-

(VBPI); **stenan:** *stenst*-(VBPI), *stent*-(VBPI), *stint*-(VBPI); **swelan:** *swelan*-(VB), *swelað*-(VBPI), *swelt*-(VBI)-(VBPI), *sweolt*-(VBDI), *swealt*-(VBDI)-(VBPI), *swilt*-(VBPI); **teran:** *teran*-(VB), *tere*-(VBP), *teorað*-(VBPI), *tæron*-(VBDI), *terendan*-(VAG^N), *torinne*-(VBPS), *tirð*-(VBPI); **toniman:** *tonimað*-(VBI), *tonymað*-(VBI)-(VBPI), *tonumen*-(VBN); **tostenan:** *tostent*-(VBPI); **toteran:** *toterenne*-(VB^D)-(VBN^A), *toterað*-(VBPI), *totære*-(VBDS), *totæron*-(VBDI), *totorene*-(VBN^N); **underniman:** *underniman*-(VB), *undernyman*-(VB)-(VBPS), *undernam*-(VBDI), *undernim*-(VBI), *undernymað*-(VBI); **understenan:** *understenst*-(VBPI), *understent*-(VBPI); **unsceran:** *unscorene*-(VBN^N); **wiþsceran:** *wiðsceorað*-(VBPI), *wiðstent*-(VBDI)-(VBPI).

Finally, the forms attested in the YCOE are given in (10).

(10) **geþweran:** *geðwer*-(VBI); **wiþniman:** *wiðnam*-(VBDI)

On the qualitative side, each of the 315 forms attested in the corpora are assigned a distinct lemma. This means that the MG lemmatizer is 100% efficient as regards form-lemma association for class IV strong verbs in that there is no competition between the generated lemmas. However, this data might be influenced by the scope of the research and the variability of OE spelling. The inflectional system of OE, along with its dialectal and diachronic spelling variations, causes different lexical paradigms to develop homographic forms. Myers (1966, 153) explains that changes occur for phonological influence and paradigmatic analogy and concludes that "when too many conflicting analogies are possible, some confusion is bound to result." The following section discusses the degree of accuracy of the MG lemmatizer in this regard.

7. Advances and Limitations of Automatic MG Lemmatization

Spelling variation in OE may lead to the generation of formally ambiguous forms. Homographs may develop within the same lexical class (11a) or across categories (11b, 11c) and in (11) the selected samples are given with the lemma assigned by the MG lemmatizer.

(11)

a. *swealt* (VBDI) (VBPI) ~ **swelan**; *wiðstent* (VBDI) ~ **wiþstenan**
b. *naman* (VBDI) ~ **niman**
c. *oter* ~ **oteran**; *rihtscire* – **rihtsciran**

Example (11a) displays inflectional forms generated in the paradigms of *swelan* and *wiþstenan*. However, lemma assignment is not undisputed, as these forms could also belong in verbal paradigms of *sweltan* 'die, perish' and *wiþstandan* 'withstand, resist', respectively. Example (11b) shows an ambiguous form belonging in the paradigms

of *niman* 'take' and *nama* 'name'. The morphological tagging provided by the YCOE confirms the attestation of *naman* as a verbal element at least once and validates lemma assignment. Likewise, the forms shown in (11c) are homographs of inflectional forms belonging in the nominal paradigms of *nama* 'name', *oter* 'an otter' and *rihtscir* 'parish'. Their attestation as verbal forms, however, cannot be proved without textual analysis. A search in the DOEC confirms that there is just one occurrence of the form *rihtscire*, which is given in example (12).

(12) [LawVIAtr 003300 (21)]
& gif man ænig lic of ***rihtscire*** *elles hwar lecge, gelæste man þone saulsceat swa þeh into þam mynstre, þe hit to hyrde.*
'And if any body is buried elsewhere than in the **parish** to which it properly belongs, the payment shall nevertheless be made to the church to which the deceased belonged' (Robertson 1925, 97).

This example confirms that although the MG lemmatizer has developed a plausible verbal form, the form-lemma association does not stand up to contextual analysis and so it has to be dismissed. The case with the form *oter* is similar. There are several fragments in the DOEC which show occurrences related to this form. They can be classified into three major groups, namely glosses, proper names and inflections of the noun *oter* 'otter'. As a gloss, *oter* appears in fragments [ÆGl 061100 (309.9)] and [CollGl 25 012600 (126)], with the meaning 'otter'. As a proper name, it occurs in fragments [Rec 10.6.2 001100 (3.1)], [Rec 10.6.2 006100 (15.3)] and [Rec 10.6.2 008700 (21.3)], the first of which is given in (13).

(13) [Rec 10.6.2 001100 (3.1)]
Her kyð on þissere boc þæt ***Oter*** *& his cild cwede saccles Aluric þane Reda, & his ofspring.*
'Here is announced in this book that **Oter** and his child declared Aluric the Red and his offspring free of charge' (own translation, based on Förster 1933, 47-48).

Finally, there are occurrences of the noun *oter* 'otter' in the genitive case in the constructions *oteres hol* 'otter's hole', *oteres ham* 'otter's home' and *oteres pol* 'otter's pool', used in topographic descriptions, usually in land grants registered in charters. Consider the samples in (14) as illustrative.

(14)
a. [Ch 492 (Birch 782) 000400 (4)]
Aerest on suþewearddan of oteres hole up andlang wiliges oþþa lace.
'First at the south end from **Otter's** Hole up along the River Wylye as far as the stream' (Grundy 1919, 271).

b. [Ch 896 (Kem 703) 000500 (4)]

Of	*moter-a*	*ford-e*	*andlang*	*moter-a*	*lac-e*
From	speaker-GEN.PL	ford-DAT.SG	along	speaker-GEN.PL	watercourse-DAT.SG
ðæt	on	oter-es	ham		
until	to	otter-GEN.SG	hemmed.land[ACC.SG]		

'From speakers' ford along speakers' watercourse until **Otter's** Home' (own translation).[6]

Examples (14a) and (14b) show that the form-lemma association established by the MG lemmatizer cannot be confirmed and that the attested inflectional forms cannot be assigned the corresponding verbal lemma. Nevertheless, the fact that these associations are not valid does not diminish the possibilities of the lemmatizer, which is able to highlight points of conflict and underscore the areas where revision might be needed. Let us consider the case of the lemma *geniman*. An exhaustive revision of the collection of secondary sources filed in KBOE retrieves the attested forms *genim*, *geniman*, *genimð*, *genime*, *genimeð*, *genimes*, *genimeþ*, *genimþ*, *geniomað*, *genioman*, *genom*, *genome*, *genomon*, *genoom*, *genumen*, *genumene*, *genumini* and *genumni*. Against this background, the lemmatizer generates and validates the attestation of *genam*, *genaman*, *genamen*, *genamon*, *genim*, *geniman*, *genimð*, *genime*, *genimeð*, *genimeþ*, *genimþ*, *genom*, *genomon*, *genumen* and *genumene* in both the DOEC and the YCOE and the forms *geniomað*, *genioman* and *genome* in the DOEC. The lemmatizer fails to generate the unexpected—although attested—spellings *genimes*, *genoom*, *genumini* and *genumni*. Finally, the lemmatizer assesses the attestation of *gename*, *genamun*, *genimað*, *genimaþ*, *genimen*, *genimenne*, *genimst*, *genoman*, *genumenum*, *genym* and *genyman* in both the DOEC and the YCOE corpora and of *genimanne*, *genimende*, *genimest*, *geniomæ*, *geniomanne*, *geniomende*, *genomen*, *genomun*, *genumenan*, *genumenne*, *genymað*, *genyme*, *genymest*, *genymst*, *ginim*, *ginime*, *ginimeð*, *giniomað*, *giniomanne*, *ginom*, *ginome*, *ginomon*, *ginomun* and *ginumen* in the DOEC. Despite their attestations, these forms were not included in the reviewed grammars. The validity of these associations is shown in (15).

(15)

a. [JnGl (Ru) 046700 (10.31)]

ginom-on	ł	onhof-on	stan-as	iude-as	þætte
Take-PRET.PL	and	take.up-PRET.PL	stone-ACC.PL	Jew-NOM.PL	that
hię	*gistend-un*	*hine*.			
3PL.NOM	stone-PRET.PL	3SG.ACC			

'The Jews **caught** and took up stones to stone him' (own translation).

[6] The translation is based on the word-by-word gloss provided by The LangScape Project (2008). The morphological tagging of the interlinear gloss is my own.

b. [ChronC (O'Brien O'Keeffe) 038700 (894.42)]
Þa foron hi to & geflymdon þone here & þæt geweorc abræcon & ***genamun*** *eall þæt ðær þinnan wæs ge on feo*, *ge on wifum*, *ge eac on bearnum & brohton eall inn to Lundenbyrig & ða scypu ealle oþþe tobræcon oððe forbærndon oþþe to Lundenbyrig brohton oððe to Hrofesceastre*.
'They then marched up and put the army to flight and stormed the work, and **took** all that there was within, as well money, as women and children, and brought all to London; and all the ships they either broke in pieces, or burned, or brought to London, or to Rochester' (Thorpe 1861, 71).

All things considered, the lemmatizer presents a high degree of accuracy as regards the generation of inflectional forms and their attestation in the corpora. It only fails to identify non-predictable spellings such as the ones discussed above and the non-standard development *strima(e)ndi* (Brunner 1965, 300). In spite of this limitation, these non-predictable spellings contribute to the revision of the MG rules and to the gradual improvement of the overall MG process.

8. Conclusions

Although a vast amount of research on OE morphosyntax grounded in the above-described corpora is currently being published, OE in its current state of description and lexicographical analysis is not yet an appropriate language with which to develop NLP research on an extensive basis. The relatively limited number of written records and the lack of lemmatized corpora are the main handicaps for this goal. The slow progress of the DOE, which incorporates attested forms for each headword, highlights the need to develop tools that allow for the identification of forms in the DOEC.

In a similar vein, this article has put forward a way to overcome these issues by exploring and checking the potential development of an automatized lemmatization process of the OE corpora. For this purpose, an automatic lemmatizing tool based on MG has been developed that has been tested on three dimensions: (a) automation; (b) validation; and (c) accuracy. The conclusion is drawn that the designed method presents multiple advantages over previous approaches. It (a) systematizes the attestation process; (b) reduces the amount of manual work and the revision of results; (c) accounts for a wider scope of inflectional and derivational possibilities; and (d) increases the degree of accuracy in form-lemma association.

As for automation, despite the formal variations of OE spelling, this paper has shown that a systematic procedure can be put forward to automatize type-based lemmatization. Regarding validation, the MG lemmatizer developed has been able not only to generate inflectional forms, but also to assess their attestation in the selected corpora. This paper has also shown that several inflectional forms that are not provided in the literature are validated as inflectional forms of the verbs analyzed. Finally, regarding accuracy, each

of the generated forms attested in the corpora has been assigned a single verbal lemma. Nevertheless, the research has also shown that not all associations hold and that token-based research is necessary for disambiguation.

This means that the limits of automation have been reached given that, even when type-based lemmatization can be automatized to a substantial extent, the spelling and morphological properties of OE prevent the MG analysis from systematizing and implementing rules to account for all the spelling variants found in the corpora. However, this points to some possibilities for the refinement of the MG lemmatizer.

This article, then, opens new lines of research. First, the completion of the analysis of class IV verbs. Two major groups of verbs have been left out of the scope of this analysis since only verbs and prefixes beginning with letters L-Y have been studied. This leaves L-Y verbs with A-I prefixes still to be investigated. Furthermore, only non-recursive derivation has been considered. Thus, double-prefixed verbs like *ongeniman* 'to take away' have been disregarded. The second line of research involves extending this method to other strong verb classes and later, to other variable lexical categories.[7]

Works Cited

Bauer, Renate and Ulrike Krischke, eds. 2011. *More than Words: English Lexicography and Lexicology Past and Present*. Bern: Peter Lang.

Bosworth, Joseph and Thomas N. Toller. (1898) 1973. *An Anglo-Saxon Dictionary*. Oxford: Oxford UP.

Brunner, Karl. 1965. *Altenglische Grammatik, nach der Angelsachsischen Grammatik von Eduard Sievers*. Tübingen: Max Niemeyer.

Campbell, Alistair. (1959) 1987. *Old English Grammar*. Oxford: Oxford UP.

Chambers, Raymond Wilson, Max Förster and Robin Flower, eds. 1933. The *Exeter Book of Old English Poetry: Facsimile*. London: Humphries.

Clark Hall, John Richard. (1894) 1996. *A Concise Anglo-Saxon Dictionary. Supplement by Herbert D. Merritt*. Cambridge: Cambridge UP.

Ferrés Daniel, Ahmed AbuRa'ed and Horacio Saggion. 2017. "Spanish Morphological Generation with Wide-Coverage Lexicons and Decision Trees." *Procesamiento del Lenguaje Natural* 58: 109-16.

Forcada, Mikel et al. 2011. "Apertium: a Free/Open-Source Platform for Translation." *Machine Translation* 25 (2): 127-44.

Förster, Max. 1933. "The Preliminary Matter of the Exeter Book." In Chambers, Förster and Flower 1933, 44-54.

[7] This research has been conducted with the support of Grant PID2020-119200GB-100 funded by MCIN/AEI/10.13039/501100011033.

García Fernández, Laura. 2020. *Lemmatising Old English on a Relational Database. Preterite-Present, Contracted, Anomalous and Strong VII Verbs*. Munich: Utzverlag.

Grundy, George B. 1919. "The Saxon Land Charters of Wiltshire." The *Archaelogical Journal* 76: 143-301.

Healey, Antonette di Paolo, ed. 2018. The *Dictionary of Old English: A to I*. Toronto: Centre for Medieval Studies, University of Toronto.

—, John Price Wilkin and Xin Xiang, eds. 2004. The *Dictionary of Old English Web Corpus*. Toronto: Centre for Medieval Studies, University of Toronto.

Heidermanns, Frank. 1993. *Etymologisches Wörterbuch der germanischen Priäradjective*. Berlin: De Gruyter.

Hogg, Richard, ed. 1992. The *Cambridge History of* the *English Language I:* The *Beginnings to 1066*. Cambridge: Cambridge UP.

— and Robert D. Fulk. 2011. *A Grammar of Old English*. Chichester: Wiley-Blackwell.

Kastovsky, Dieter. 1992. "Semantics and Vocabulary." In Hogg 1992, 290-408.

Khemakhem, Aida et al. 2015. "ISO Standard Modeling of a Large Arabic Dictionary." *Journal of Natural Language Engineering* 22 (6): 849-79.

Krygier, Marcin. 1994. The *Disintegration of* the *English Strong Verb System*. Bern: Peter Lang.

LangScape Project. 2008. "The Language of Landscape: Reading the Anglo-Saxon Countryside." Version 0.9. [Accessed online on 2 August, 2021].

Levin, Samuel. 1964. "A Reclassification of the Old English Strong Verbs." *Language* 40 (2): 156-61.

Mailhammer, Robert. 2007. The *Germanic Strong Verbs: Foundation and Development of a New System*. Trends in Linguistics. Studies and Monographs 183. Berlin: De Gruyter.

Márquez, Lluis, Chris Callison-Burch and Jian Su, eds. 2015. *Proceedings of* the *2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics.

Martín Arista, Javier. 2012. "Lexical Database, Derivational Map and 3D Representation." *RESLA-Revista Española de Lingüística Aplicada* (Extra 1): 119-44.

—. "Nerthus. Lexical Database of Old English: From Word-Formation to Meaning Construction." Research Seminar given at the School of English, Sheffield, 2013.

Martín Arista, Javier. "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, September 2017.

— et al., comp. 2021. *ParCorOEv2. An Open Access Annotated Parallel Corpus Old English-English*. Nerthus Project, Universidad de La Rioja.

Metola Rodríguez, Darío. 2015. "Lemmatisation of Old English Strong Verbs on a Lexical Database." PhD diss., University of La Rioja.

—. 2017. "Strong Verb Lemmas from a Corpus of Old English: Advances and Issues." *Revista de Lingüística y Lenguas Aplicadas* 12: 65-76.

Müller, Thomas et al. 2015. "Joint Lemmatization and Morphological Tagging with Lemming." In Márquez, Callison-Burch and Su 2015, 2264-74.

Myers, Louis M. 1966. The *Roots of Modern English*. Boston: Little Brown.

Novo Urraca, Carmen and Ana Elvira Ojanguren López. 2018. "Lemmatising Treebanks: Corpus Annotation with Knowledge Bases." *RAEL-Revista Electrónica de Lingüística Aplicada* 17 (1): 99-120.

Oflazer, Kemal and Murat Saraçlar, eds. 2018. *Turkish Natural language Processing*. Berlin: Springer.

Orel, Vladimir. 2003. *A Handbook of Germanic Etymology*. Leiden: Brill.

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

Reiter, Ehud and Robert Dale. 1997. "Building Applied Natural Language Generation Systems." *Natural Language Engineering* 3 (1): 57-87.

Rissanen, Matti et al., comp. 1991. The *Helsinki Corpus of English Texts*. Department of English, University of Helsinki.

Robertson, Agnes J. 1925. The *Laws of* the *Kings of England from Edmund to Henry I*. Cambridge: Cambridge UP.

Saggion, Horacio. 2017. *Automatic Text Simplification*. Berlin: Springer.

Seebold, Elmar. 1970. *Vergleichendes und etymologisches Wörterbuch der germanischen starken Verben*. Berlin: De Gruyter.

Sweet, Henry. (1896) 1976. The *Student's Dictionary of Anglo-Saxon*. Cambridge: Cambridge UP.

Tapsai, Chalermpol, Herwig Unger and Phayung Meesad. 2021. *Thai Natural Language Processing*. Berlin: Springer.

Taylor, Ann et al. 2003. The *York-Toronto-Helsinki Parsed Corpus of Old English Prose*. York: U of York.

The Holy Bible. (1899) 1971. Translated from the Latin Vulgate (Douay Rheims Version). Tan Books.

Thorpe, Benjamin. 1861. The *Anglo-Saxon Chronicle*. Vol 2. London: Longman, Green and Roberts.

Tío Sáenz, Marta. 2019. "The Lemmatisation of Old English Weak Verbs of a Relational Database." PhD diss., University of La Rioja.

Von Mengden, Ferdinand. 2011. "Ablaut or Transfixation? On the Old English Strong Verbs." In Bauer and Krischke 2011, 123-39.

Wang, Alex et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." Paper delivered at the 2019 ICLR Conference, New Orleans, May 2019.

Zhang, Yu and Qiang Yang. 2018. "An Overview of Multi-Task Learning." *National Science Review* 5: 30-43.

Received 27 August 2021 Revised version accepted 3 October 2022

Roberto Torre Alonso is a Lecturer at the University of La Rioja, Spain. He is a member of the Functional Grammars Research Group at the same university and is currently participating in the development of *ParCorOEv2. An Open Access Annotated Parallel Corpus Old English-English*. His research interests include Old English morphology, the semantic-syntactic interrelation of Old English verbs and Old English verb lemmatization.

Appendix 1. YCOE POS verb tags

POS TAG	CATEGORY	POS TAG	CATEGORY
AX	Infinitive	HVI	Have, imperative
AXD	Past, ambiguous form	HVN	Have, past participle (vb. or adj.)
AXDI	Past, unambiguous indicative	HVN^N	Have, past participle (vb. or adj.), nominative
AXDS	Past, unambiguous subjunctive	HVP	Have, present, ambiguous form
AXG	Present participle	HVPI	Have, present, unambiguous indicative
AXI	Imperative	HVPS	Have, present, unambiguous subjunctive
AXN	Past participle (verbal or adjectival)	MD	Modal, infinitive
AXP	Present, ambiguous form	MD^D	Modal, infinitive, inflected
AXPI	Present, unambiguous indicative	MDD	Modal, past, ambiguous form
AXPS	Present, unambiguous subjunctive	MDDI	Modal, past, unambiguous indicative
BAG	Present participle	MDDS	Modal, past, unambiguous subjunctive
BAG^N	Present participle, nominative	MDI	Modal, imperative
BE	Be, infinitive	MDP	Modal, present, ambiguous form
BE^D	Be, infinitive, dative	MDPI	Modal, present, unambiguous indicative
BED	Be, past, ambiguous form	MDPS	Modal, present, unambiguous subjunctive
BEDI	Be, past, unambiguous indicative	VAG	Present participle
BEDS	Be, past, unambiguous subjunctive	VAG^A	Present participle, accusative
BEI	Be, imperative	VAG^D	Present participle, dative
BEN	Be, past participle	VAG^G	Present participle, genitive
BEN^A	Be, past participle, accusative	VAG^I	Present participle, instrumental
BEN^D	Be, past participle, dative	VAG^N	Present participle, nominative
BEN^G	Be, past participle, genitive	VB	Infinitive
BEN^N	Be, past participle, nominative	VB^D	Infinitive, inflected
BEP	Be, present, ambiguous form	VBD	Past, ambiguous form
BEPH	Be, present, ambiguous imp./subj.	VBDI	Past, unambiguous indicative
BEPI	Be, present, unambiguous indicative	VBDS	Past, unambiguous subjunctive
BEPS	Be, present, unambiguous subjunctive	VBI	Imperative
HAG	Have, present participle	VBN	Past participle (verbal or adjectival)
HAG^A	Have, present participle, accusative	VBN^A	Past participle (verbal or adjectival), accusative
HAG^D	Have, present participle, dative	VBN^D	Past participle (verbal or adjectival), dative
HAG^N	Have, present participle, nominative	VBN^G	Past participle (verbal or adjectival), genitive
HAG^G	Have, present participle, genitive	VBN^I	Past participle (verbal or adjectival), instrumental
HV	Have, infinitive	VBN^N	Past participle (verbal or adjectival), nominative
HV^D	Have, infinitive, inflected	VBP	Present, ambiguous form
HVD	Have, past, ambiguous form	VBPH	Ambiguous imperative/subjunctive
HVDI	Have, past, unambiguous indicative	VBPI	Present, unambiguous indicative
HVDS	Have, past, unambiguous subjunctive	VBPS	Present, unambiguous subjunctive