Old English morphological inflection generation with UniMorph. Assessment with a relational database and training guidelines

Generación de flexión morfológica con UniMorph. Evaluación con base de datos relacional y pautas de entrenamiento

Javier Martín Arista
Universidad de La Rioja
javier.martin@unirioja.es

Abstract: The aim of this article is to assess the morphological inflection generation of Old English of the UniMorph data set. The method of this study is based on McCarthy et al.´s (2020) model of generation of putative morphological paradigms. The assessment includes inflections (morphological features and values), inflectional forms and stems. The question is also addressed of plausibility, understood as the effective attestedness of an inflectional form. The assessment tasks are carried out in a relational database specifically designed for filing and comparing the relevant data sets, including treebanks and databases of Old English lexicographical and textual sources. The overall conclusion is that the Old English UniMorph data set is consistent and robust. On the basis of the assessment, however, training guidelines of the generation model are proposed that include characters, diacritical marks, the prefix ge- in verbs, the superlative grade of adjectives, the adjectivally inflected participle and some local shortcomings.
Keywords: morphological inflection generation, UniMorph, relational database, treebank, Old English.

1 Introduction

This article engages in morphological inflection generation, which, according to Çöltekin (2019), is "the task of generating a word based on its lemma and morphological features. For example, given the German lemma *aufgeben* 'to give up' and the morphological tags {V.PTCP, PST}, the task is to predict the inflected form *aufgegeben*."

The target language of the study is Old English, the diachronic variety of English spoken in England between approximately the 6th and the 11th centuries of the Christian Era. It belongs to the West-Germanic Group of the Indo-European family of languages and is characterised by its explicit generalised morphological inflection and its consistently Germanic lexicon. Around 3,000 texts that approximately comprise 3 words million have been kept, most written in the West-Saxon variety in the 9th and, above all, in the 10th. century. Synchronic and diatopic variation, as well as the lack of a written standard, result in the unpredictiveness of the spelling of a remakable number of textual forms, as has been remarked by authors like Johnson (2009). In this line, an algorithm devised for generating Old English forms, even at the level of the syntactic word, requires a thorough design and an extensive training, as Torre Alonso (2021) shows. Matters are further complicated by the randomness of textual transmission, throughout which the vast majority of the texts might have got substantitally modified with respect to the original version or simply lost. These aspects should be taken into account when the task of generating Old English is undertaken.

For these reasons, the aim of this article is to assess the Old English data set of

morphological inflection generation provided by UniMorph and available from https://raw.githubusercontent.com/unimorph/ang/master/ang. The UniMorph Project (https://unimorph.github.io) has defined a universal schema for morphological annotation with which data sets from 142 languages have been annotated.

More specifically, this study intends to contribute to the training of an inflection model of Old English in two directions: by gauging the accuracy of the inflections, inflectional forms and lemmas of the Old English UniMorph data set and by presenting a number of guidelines for the training of the inflection model.

The scope of the article is restricted to the syntactic word, that is to say, morphologically simplex words, affixed words and compound words that are written as one segment. The sources include treebanks and relational databases from Old English lexicographical and textual sources.

The article is structured as follows. Section 2 presents the data sets and the method of the study, including the design and implementation of the relational database. Section 3 assesses the generation model as to inflections (morphological features and values), inflectional forms and stems. The question of plausibility is also raised in this section. Section 4 discusses some weak points of the generation model, including their relevance for further training. Finally, Section 5 draws the conclusions of the article.

2 Data sets and method

This study relies on two types of data sets, to wit, treebanks and relational databases. While treebanks (Böhmová et al., 2003) and databases can be described as computerised data table collections (Jurafsky and Martin, fc.), they differ from each other in two important respects, at least in the context of this study. Firstly, treebanks are available from the Internet in open access, whereas relational databases are not always public. Secondly, treebanks tend to be more available for linguistic comparison and analysis than relational databases. An important consequence of this is that treebanks usually represent final products whereas relational databases can be updated.

Beginning with the treebanks, two data sets belong to this category: the Old English segment of the UniMorph Project and the York annotated corpora of Old English, both the prose and the poetry segments.

UniMorph consists of a schema and a set of databases for cross-linguistic morphological annotation. Morphological inflection generation in UniMorph is based on the UniMorph Schema (Sylak-Glassman, fc.), which comprises 23 dimensions of meaning (morphological categories) and 212 features. For Old English, the UniMorph Schema has been applied to the major lexical categories noun, adjective and verb. The relevant features include: ACC (accusative case), ADJ (adjective), DAT (dative case), FEM (feminine gender), GEN (genitive case), IMP (imperative mode), IND (indicative mode), INS (instrumental case), LGSPEC1 (weak declension of the adjective), LGSPEC2 (strong declension of the adjective), MASC (masculine gender), N (noun), NEUT (neuter gender), NFIN (non-finite form of verb (infinitive and inflective infinitive)), NOM (nominative case), PL (plural number), PRS (present tense), PST (past tense), SBJV (subjunctive mode), SG (singular number), V (verb), V.PTCP (verbal participle). In the case of the adjective, the most inflective lexical class (as it can be declined according to a weak and a strong declension and can be graded for the comparative and the superlative), the generation with UniMorph turns out 67 inflectional forms, some of which are presented for illustration in Figure 1 with the corresponding morphological tags.

aberendlic (ADJ;NEUT;SG;NOM;LGSPEC2)
…
aberendlicena (ADJ;FEM;PL;GEN;LGSPEC1)
…
aberendlicu (ADJ;NEUT;PL;ACC;LGSPEC2)
…
aberendlicum (ADJ;FEM;PL;DAT;LGSPEC2)

Figure 1. UniMorph inflectional forms and tags of *aberendlic* (extract).

The second treebank used as data set for this study comprises the prose and the poetry parts of the York corpora of Old English (hereafter YCOE): *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (1,500,000 words; Taylor et al., 2003) and *The York-Helsinki Parsed Corpus of Old English Poetry* (50,000 words; Pintzuk and Plug, 2001). The YCOE is morphologically tagged and syntactically annotated. It comprises a POS (part of speech)

file and a PSD (syntactic parsing) file for each text.

Turning to the relational databases, this study draws on unlemmatised and lemmatised data sets. *The Dictionary of Old English web corpus* (Healey et al., 2004), henceforth DOEC, contains 3,000,000 words. It is not lemmatised, neither does it provide morphological tagging or syntactic annotation, but it is generally considered to gather all the written records of the language. The DOEC was compiled as the corpus of the *Dictionary of Old English* (DOE; Healey, 2018), which has published the letters A-I so far. This electronic dictionary can be accessed online and offers, along with meaning definitions and citations, attestations per lemma that can be searched by headword, attested spelling, part of speech and occurrence, among other criteria.

The other lemmatised data set is *ParCorOEv2. An open access annotated parallel corpus Old English-English* (ParCorOEv2; Martín Arista et al. 2021). ParCorOEv2 is a deeply annotated parallel corpus that is aligned at word level. It currently holds 110,000 records and another 140,000 are expected by the end of 2022. These lemmatised data sets are complementary as to headword spelling. While the DOE opts for a late spelling of headwords (10th-11th century), ParCorOEv2 renders a classical spelling (9th-10th century), in such a way that the combined used of the two sources provides a wider inventory for comparison.

The method of this study is based on McCarthy et al.´s (2020) model of generation of putative morphological paradigms, which comprises two steps, training and generation. At the step of training, the aim is to relate the existing lemmas to the existing paradigms through an inflection model; while at the generation step, the aim is to relate the extracted lemmas to the putative paradigms via a trained model. This study contributes to the training of an inflection model for Old English and, ultimately, to the congruence of the existing and the putative paradigms of the language. An assessment is carried out and training guidelines are defined with a view to improving the training of the model, so that it generalises well to new data of Old English. The assessment and the training guidelines revolve around two main aspects, namely, the morphological paradigms and the lemma set that is inputted to them. Since the paradigms consists of the morphological features and values (inflections) as well as their exponents (inflectional forms), a distiction must be drawn between the assessment of inflections, on the one hand, and the assessment of exponents, on the other hand. The quality of the lemma set determines the plausibility of the outcome of the generation of morphological inflection. The tasks that this method require include (1) the assessment of inflections: are there counterparts of the morphological features and values of the UniMorph Schema as applied to the Old English data set in other tagged data sets? (2) the assessment of inflectional forms: do the morphological exponents of the UniMorph Schema as applied to the Old English data set include the inflectional forms tagged in other data sets? And (3) the assessment of the lemma set: are there substantial differences between the lemma set of the UniMorph Old English data set and the lemma lists of other lemmatised sources? With this assessment, it will be possible to address the question of plausibility: (4) how many putative forms are attested in the written records? Tasks 1 and 2 require a data set with POS tagging, while task 3 calls for a lemmatised data set. Task 4 needs to rely on an extensive unlemmatised inventory. For these reasons, the YCOE is used for the assessment of inflections and inflectional forms, while lemmas are assessed with respect to the DOE and PacCorOEv2. Task 4 should necessarily draw on a full inventory of the forms attested in the written records of Old English. The DOEC has been selected for this task.

Not only the amount of data but also the need for falsifiability advise the automation of the four tasks described above. To this end, a relational database has been specifically designed for the undertaking. It has been implemented in Claris FileMaker Pro software (version 19.3.2.206). The database consists of six layouts, five of which correspond to the data sets reviewed above (UniMorph, DOEC, DOE, YCOE and ParCorOEv2). The sixth is a summary layout that combines all the data sets.

The Old English data set of UniMorph has been downloaded in txt format and imported into the database. It comprises 42,068 inflectional forms with the corresponding lemmas (1,867).

The DOEC has been concorded and indexed with AntConc 3.5.9 (Anthony, 2020). The concordance to the DOEC has 3,075,444 lines, while there are 194,327 types in the index.

The DOE has been searched by headword and attested spelling. A total of 15,907 lemmas and 83,477 inflectional forms have been found.

The inflectional forms and morphological tags of the YCOE have been extracted with BBEdit (version 14.0.2). A total of 106,202 types, corresponding to 1,595,674 tokens, have been extracted. The resulting types have been edited with the characters <æ>, <ð> and <þ> and tagged for lexical category, on the basis of the YCOE POS labels given in the corpus manuals (https://www-users.york.ac.uk/~lang22/YCOE/doc/annotation/YcoeLite.htm#pos_labels). Figure 2 illustrates this process.

POS file

<T06940000100,1>_CODE De_FW scientia_FW ._.
coalcuin,Alc_[Warn_35]:1.2_ID +arest_ADVS^T ealre_Q^G +tingen_N^G +aighwylce_Q^I m+an_N^D is_BEPI to_TO secene_VB^D ,_, hw+at_WPRO^N seo_BEPS se_D^N so+de_ADJ^N

inflectional form	morphological tag	lexical category
ærest	ADVS^T	Adverb
ealre	Q^G	Adjective
þingen	N^G	Noun
æighwylce	Q^I	Adjective
mæn	N^D	Noun
is	BEPI	Verb
to	TO	Preposition
secene	VB^D	Verb
Hwæt	WPRO^N	Pronoun
Seo	BEPS	Verb
Se	D^N	Demonstrative
Soðe	ADJ^N	Adjective

Figure 2. Extraction of types from the YCOE and lexical category tagging.

When designing the relational database, the types from the YCOE represented the field of reference. The data from the other layouts have been imported as an update for the reference field. With these premises, the total amount of files in the relational database is 106,202. The summary layout consists of the following fields: YCOE inflectional form, YCOE morphological tag, YCOE lexical category, DOE lemma, DOE attestation, UniMorph morphological tag, UniMorph lemma, and ParCorOEv2 lemma. A file with these fields and their values is presented in Figure 3.

Field	Value
YCOE_inflectional_form	bit
YCOE_morphological_tag	BEPI
YCOE_lexical_category	verb
DOE_lemma	bītan
DOE_attestation	44
UniMorph_morphological_tag	V;IMP;SG
UniMorph_lemma	bitan
ParCorOEv2_lemma	bītan

Figure 3. The summary layout in the relational database.

3 Assessment with the relational database

This section gauges the accuracy of the Old English UniMorph data set from two perspectives. In the first place, the accuracy of the morphological paradigms is discussed. This includes the morphological features and values (inflections) as well as their exponents (inflectional forms). The assessment of morphological features and values is extensive, whereas the one of their exponents is restricted to the main lexical and morphological classes. This part of the assessment depends on the YCOE layout of the relational database. In the second place, the question of the quality of the lemma inventory of UniMorph is raised. This part of the assessment is carried out with the DOE and the PacCorOEv2 layouts of the relational database.

The first question addressed in this section is whether or not there are counterparts of the morphological features and values of the UniMorph Schema as applied to the Old English data set in other tagged data sets. A total of 6,762 counterparts of UniMorph morphological tags have been found in the YCOE. They correspond to 94 different tags in the YCOE layout, which comprises a total of 94 tags (recall ratio 1). Apart from the different annotation formats, there is no complete coincidence between the two sets of tags for reasons of homography across categories or due to different criteria for category assignment between noun and adjectives or adjectives and verbs (regarding the participle). This kind of mismatch is illustrated in Figure 4.

YCOE tag	UniMorph tag
ADJ	N;DAT;SG
ADJR^N	N;NOM;SG

N	ADJ;FEM;PL;NOM;LGSPEC2
N	ADJ;NEUT;SG; ACC;LGSPEC1
N	V;IMP;SG
N	V;IND;PRS;1;SG
VBN^D	ADJ;NEUT; PL;DAT;LGSPEC1
VBN^N	ADJ;FEM;SG; ACC;LGSPEC2

Figure 4. Morphological tags in the YCOE and UniMorph.

The second question raised in this section is whether or not the morphological exponents of the UniMorph Schema as applied to the Old English data set include the inflectional forms tagged in other data sets. For this purpose, the forms tagged in the YCOE and in the UniMorph data set are compared. From the quantitative point of view, 6,762 UniMorph inflectional forms are filed and tagged in the YCOE, which represents a 0.16 recall ratio. On the qualitative side, the most representative morphological classes of the lexical categories represented in UniMorph (the adjective, the noun and the verb) are considered. Such morphological classes include the weak and the strong forms of the adjective, strong masculine nouns, strong verbs and weak verbs. Weak verbs with strong forms, strong verbs with weak forms, preterite-present verbs and irregular verbs (which are not tagged in UniMorph), as well as the minor declensions of the noun have been put aside. The adjective, the noun and the verb are discussed in turn.

The inflectional forms of the adjective *beald* 'bold' tagged in the YCOE can be seen in Figure 5. The corresponding tags in UniMorph, when available, are given in the right column. The unpredictable spelling *bald* (ADJ^N) is missing in the UniMorph tagging and, more importantly, the comparative *bealdran* (ADJR^N) and the superlative *baldeste* (ADJS^N) are not tagged in UniMorph.

YCOE Inflectional form	**YCOE tag**	**UniMorph tag**
bealdum	ADJ^D	ADJ;FEM;PL; DAT;LGSPEC 2
beald	ADJ^N	ADJ;NEUT;PL ACC;LGSPEC 2
bealda	ADJ^N	ADJ;MASC; SG; NOM; LGSPEC1
bealde	ADJ^N	ADJ;NEUT; SG;ACC; LGSPEC1
bealdne	ADJ^A	ADJ;MASC; SG;ACC; LGSPEC2
baldra	ADJR^N	-
bald	ADJ^N	-
baldeste	ADJS^N	-
bealdran	ADJR^N	-

Figure 5. Adjective in YCOE and UniMorph.

Figure 6 tabulates the inflectional forms of the strong noun with weak forms *ancor* 'anchor'. UniMorph gives *ancras* (N;ACC;PL) *ancre* (N;DAT;SG), *ancrum* (N;DAT;PL), *ancra* (N;GEN;PL), *ancor* (N;NOM;SG) and *ancras* (N;ACC;PL) but misses the unpredicatable spellings of the strong singular nominative *ancer* (*ancor*), singular accusative *ankor* (*ancor*), singular dative *ancræ* (*ancre*), plural nominative *onceras* and *oncras* (*ancras*) and plural accusative *oncras* (*ancras*). More significantly, UniMorph misses the forms from the weak declension of the noun, including the weak singular genitive *ancran*, singular dative *ancran*, as well as the plural nominative and plural accusative *ancran*. It is also worth commenting that syncopated forms like *ancras* and unsyncopated forms such as *anceras* co-occur in the inflectional paradigm.

YCOE Inflectional form	**YCOE tag**	**UniMorph tag**
ancran	N^A	-
ancras	N^A	N;ACC;PL
ancræ	N^D	-
ancran	N^D	-
ancre	N^D	N;DAT;SG
ancrum	N^D	N;DAT;PL
ancra	N^G	N;GEN;PL
ancran	N^G	-
ancer	N^N	-
anceras	N^N	-
ancra	N^N	N;GEN;PL
ancran	N^N	-
oncras	N^A	-
ancor	N^N	N;NOM;SG
ancras	N^N	N;ACC;PL
ankor	N^A	-
oncras	N^N	-

Figure 6. Masculine noun in YCOE and UniMorph.

The YCOE and UniMorph inflectional forms and tags of the strong verb (class III) *belgan* 'to become angry' can be seen in Figure 7. Out of 13 attested inflectional forms, UniMorph gives 3, *belgaþ* (V;IND;PRS;PL), *belge* (V;IND;PRS;1;SG) and *gebolgen* (V.PTCP;PST). Of the missing forms, *bealh* and *belh* show alternation <h/g> with respect to *bealg* and *belg*. The alternation, as such, is relatively predictable. With respect to *gebolgene*, the adjectival part of the inflection of the present and the past participle has not been distinguished in the UniMorph Old English data set, which does not seem consistent with the choice of the verbal and the adjectival lexical classes. It is also worth pointing out that UniMorph must have considered the prefix *ge-* as derivational, thus distinguishing *belgan* from *gebelgan*. It must be noted in this respect that differences in meaning between the simplex and the *ge*-prefixed verb are scarce and the separation between the simplex and the complex verb is more often a lexicographical decision than a linguistic fact. From the strictly linguistic point of view, the prefix *ge-* plays a central role in inflection as it canonically forms past participles (Cambell, 1987; Hogg and Fulk, 2011) as well as in inflectionally motivated derivation (Kastovsky, 1992; Martín Arista, 2012). If we put aside the *ge-* prefixed forms of *belgan*, UniMorph correctly generates 3 out of 5 attested inflectional forms and misses the relatively unpredictable <ea> spelling.

YCOE Inflectional form	**YCOE tag**	**UniMorph tag**
bealg	VBDI	-
bealh	VBDI	-
belgaþ	VBPI	V;IND;PRS;PL
belge	VBPS	V;IND; PRS;1;SG
belh	VBI	-
gebealg	VBDI	-
gebealh	VBDI	-
gebelg	VBI	-
gebelgan	VB	-
gebelge	VBPS	-
gebolgen	VBN^N	V.PTCP;PST
gebolgene	VBN^N	-
gebulgon	VBDI	-

Figure 7. Strong verb in YCOE and UniMorph.

Figure 8 carries out this analysis with respect to the class 1 wead verb *rǣdan* 'to advise'. There are 23 tagged forms in the YCOE, in contradistinction to the 12 found in UniMorph. Surprisingly, UniMorph generates forms with consonant gemination like *rædde* (V;IND;PST;3;SG) but other geminated form such as the preterite plural *ræddan* are missing. This one, however, could be missing on the basis of the unpredictable spelling *ræddan* for *ræddon*, which has been generated by UniMorph. It is worth pointing out that the interchangeability of the eth and the thorn spelling to represent the voiceless and voiced dental allophones has not been taken into account in UniMorph, given that forms with thorn like *rædaþ* have been generated but the corresponding form with eth (*rædað*) has been missed. On the other hand, the inflection of the infinitive (*to rædanne*) has been generated in UniMorph but has not been tagged in the YCOE because it is not attested in the DOE. It also deserves a word of comment that the inflected present participles *rædene, rædendne*, *rædanne* and *rædendan* are missing in the UniMorph generation. The spelling of *rædd*, *rætst, redst* and *ret* is unpredictable.

YCOE Inflectional form	YCOE tag	UniMorph tag
rædende	VAG^A	V.PTCP;PRS
rædan	VB	V;NFIN
rædene	VB^D	-
rædde	VBD	V;IND; PST;3;SG
ræddan	VBDI	-
ræddon	VBDI	V;IND;PST;PL
redon	VBDI	V;IND;PST;PL
rædd	VBN	-
ræded	VBN	V.PTCP;PST
ræden	VBN	N;NOM;SG
rædað	VBPI	-
rædaþ	VBPI	V;IND;PRS;PL
rædeþ	VBPI	V;IND; PRS;3;SG
redst	VBPI	-
rædan	VBPS	V;NFIN
rædendne	VAG^A	-
rædanne	VB^D	-
rædeð	VBPI	-
ræden	VBPS	N;NOM;SG
rædendan	VAG^G	-

rædende	VAG	V.PTCP;PRS
rætst	VBPI	-
ret	VBI	-

Figure 8. Weak verb in YCOE and UniMorph.

The third question raised in this section has to do with the lemma inventory of UniMorph. That is to say, are there substantial differences between the lemma set of the UniMorph Old English data set and the lemma lists of other lemmatised sources? Since lemmas are used as reference forms (the singular nominative of nouns and adjectives and the infinitive of verbs), this part constitutes, above all, an assessment of the quality of the stems geared to the plausibility of the putative language. To answer this question, the UniMorph data set and the ones from the DOE and PacCorOEv2 have been compared. The comparison of the stems throws the following results. The UniMorph data set inflects 1,867 different stems, from the adjectival, nominal and verbal classes. Of these, 1,069 correspond to the letters A-I (which have already been published by the DOE). Within the letters A-I, 827 stems of UniMorph have been found in the DOE and another 227 have a correlate in ParCorOEv2. This adds up to a total of 1,054 stems out of 1,069, which throws a recall ratio of 0.98.

Finally, this section addresses the question of plausibility. Once the stems, the inflectional features and the values inflections have been it is necessary to relate the generated inflectional forms to the ones attested in the unlemmatised data sets. The concept of plausibility is relevant at this point. Plausibility is the degree of convergence of the putative language generated with the UniMorph Schema and the attested language extracted from the DOEC. The analysis shows that 18,820 out of 42,067 of the generated UniMorph inflections are attested in the DOEC (recall ratio 0.44). By categories, these totals can be broken down as presented in Table 1 (adjectives), Table 2 (nouns) and Table 3 (verbs). Tables 1-3 tabulate the amount of forms present in both data sets.

	UniMorph	DOEC
Gender		
ADJ;FEM	509	534
ADJ;MASC	412	455
ADJ;NEUT	590	636
Declension		
ADJ;SPEC1	901	965
ADJ;SPEC2	651	704

Table 1. Attestedness of UniMorph inflectional forms (adjectives).

	UniMorph	DOEC
Case		
N;ACC	321	380
N;DAT	1,074	1,150
N;GEN	368	394
N;NOM	696	753
Number		
N;SG	1,702	1,829
N;PL	757	848

Table 2. Attestedness of UniMorph inflectional forms (nouns).

	UniMorph	DOEC
Mode		
V;IMP	570	633
V;IND	1,609	1,947
V;SBJV	1,237	1,318
Number		
V;SG	2,453	2,775
V;PL	963	1,123
Tense		
V;PRS	875	1,534
V;PST		
Non-finite forms		
V;NFIN	679	722
V.PTCP;PRS	423	446
V.PTCP;PST	366	383

Table 3. Attestedness of UniMorph inflectional forms (verbs).

Remarkable differences arise when class totals are considered. Whereas 4,269 UniMorph generated nominal forms are attested from a total of 7,280 (recall ratio 0.58), the corresponding percentages in adjectives and verbs are much lower: 8,205 out of 18,712 adjectives (recall ratio 0.43) and 6,346 of 16,075 (recall ratio 0.39).

4 Discussion

Two main lessons can be learned from the assessment of UniMorph morphological inflection generation presented in Section 3. The first has to do with the concept of putative language. While the putative language generated with the UniMorph schema is adequate given the quality of the stems and the inflections, both features and values, its

plausibility is relatively low. The comparison of the UniMorph and the DOEC data sets suggests that the grammatically canonical inflectional paradigms are scarcely attested in the written records. This is particularly the case with the lexical class of the verb.

This leads us to the next lesson that can be learned from the data presented in Section 3. Plausibility, defined as the effective attestation of the grammatically canonical inflectional paradigms, is a fully random consequence of the process of textual transmission and, consequently, cannot be considered a weak point of the Old English UniMorph data set. This data set has proved robust in terms of the choice of stems, morphological features and values, but it also presents some weak points which are summarised in the remainder of this section and discussed as to their relevance for further training of the generation model.

To begin with, a number of unpredictable spellings have been mentioned that the UniMorph data set misses. However, such local irregularities do not seem compatible with a framework geared to cross-linguistic comparison, which seeks regularities rather than highly language-specific phenomena.

Other local shortcomings, which may be revised, affect the singular masculine accusative from the strong adjectival declension (ADJ;MASC;SG;ACC;LGSPEC2), which is mistaken for the plural feminine (ADJ;FEM;PL;NOM;LGSPEC2) in instances like *gyldenne* (*gylden* 'golden'), *mihtigne* (*mihtig* 'mighty'), *eadigne* (*ēadig* 'wealthy), *elþeodigne* (*elþēodig* 'foreign') and *hefigne* (*hefig* 'heavy'). Furthermore, nouns are not classified by gender, which might result in the wrong categorial tagging of at least thirty adjectival inflectional forms, such as *middan* (*midde* 'middle), *fyrene* (*fȳren* 'of fire'), *neowe* (*niwe* 'new'), *woge* (*woh* 'perverse'), etc., which are tagged as dative nouns.

More general questions include, in the first place, characters and diacritical marks. As for characters, problematic choices of the UniMorph data set include the character <ƿ> (wynn) to represent the grapheme <w>, which appears in 145 forms; and the letter <þ> (thorn) to represent the interchangeable pair <þ/ð>, which affects 815 forms. Editors of Old English texts do not use the wynn and tend to prefer the eth over the thorn, in such a way that the letter eth, as a general rule, subsumes <þ> and <ð>. Regarding diacritics, marking vocalic length and palatalisation in inflectional forms, as UniMorph does, is completely unprecedented, with the exception of some teaching materials. Finally, it is necessary to indicate the vocalic length of lemmas because vowel length is meaningful in Old English. While the changes of characters and diacritics would certainly contribute to the standarisation of the data set, the marking of vocalic length in lemmas would improve the applicability of the data set to subsequent analysis.

Also of general import is the question of the verbal prefix *ge-*. Its degree of generalisation suggests that both the simplex and the *ge-* affixed forms should be conjugated for all verbs. In this respect, it must be borne in mind that the prefix *ge-* is attached to 8,337 forms of verbs in the YCOE, out of a total of 33,986 verbal tokens.

The adjectival inflection of present and past participles is ignored in the current state of the UniMorph data set. Even if we put aside proto-auxiliary verbs like *bēon* 'to be' and *habban* 'to have', there are 1,469 present participles with agreement traceable to the nominal head in the YCOE, and 3,082 past participles.

Something similar happens to adjective gradation, which has been generated only partially. There are 895 adjectives graded for the superlative in the YCOE, of which 78 only have been generated in the UniMorph data set; and another 956 comparatives, of which 78 only have been processed in the data set at stake. This represents a recall ratio of 0.08 in adjective gradation. It must be remarked in this respect that the size of the YCOE is approximately one half of the DOEC, which contains all the written records of the language. This ratio makes the figures just given even more significative.

To close this section, a comparison is presented that subsumes stem and inflection quality. The stems and inflections of UniMorph can be found in 9,795 inflectional forms of the YCOE (recall ratio 0.09). The stems of ParCorOEv2 enhanced with the training suggestions made in this section have 49,264 correlate inflectional forms in the YCOE (recall ratio 0.46). This comparison must be taken with caution because the lemma set of UniMorph is different from ParCorOEv2. As has been shown above, the morphological generation and the choice of unlemmatised forms in UniMorph are consistent when considered independently. On the other hand, this overall assessment of stem

plus inflection in terms of plausibility indicates that the Old English UniMorph data set should address, at least, the issues raised in this discussion. Finally, further research is needed in the relevance for other historical languages of the method for gauging plausibility put forward in this article.

5 Conclusion

This article has assessed the morphological inflection generation of Old English of the UniMorph data set, including inflections (morphological features and values), inflectional forms, stems and plausability. Although this data set is consistent and robust, training guidelines of the generation model have been proposed that include characters, diacritical marks, the verbal prefix *ge-*, the superlative grade of adjectives, the participle with adjectival inflection and some local shortcomings.

Acknowledgements

Grant PRX19/00389 and grant PID2020-119200GB-100, funded by Ministerio de Ciencia, Innovación y Universidades.

Bibliographical references

Anthony, L. 2020. AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

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

Çöltekin, Çağrı. 2019. Cross-lingual morphological inflection with explicit alignment. *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 71–79, Association for Computational Linguistics.

Cotterell, R., C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection,* pages 1-27, Association for Computational Linguistics.

Healey, A. (ed.), J. 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.

Healey, A. (ed.). 2018. *The Dictionary of Old English in electronic form A-I.* Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

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

Johnson, B. 2009. *Using the Levenshtein algorithm for automatic lemmatization in Old English*. MA Thesis, The University of Georgia.

Jurafsky, D., and J. H. Martin. *Speech and Language Processing* (3rd. edition). Forthcoming.

Kastovsky, D. 1992. Semantics and vocabulary. In R. Hogg (ed.) *The Cambridge history of the English language I: The beginnings to 1066*, pages 290-408, Cambridge University Press, Cambridge.

Martín Arista, J. 2012. The Old English prefix *ge-*: A panchronic reappraisal. *Australian Journal of Linguistics,* 32(4):411–433.

Martín Arista, J., S. Domínguez Barragán, L. García Fernández, E. Ruíz Narbona, R. Torre Alonso, R., and R. Vea Escarza. 2021. *ParCorOEv2. An open access annotated parallel corpus Old English-English*. Nerthus Project, Universidad de La Rioja, www.nerthusproject.com.

McCarthy, A. D., C. Kirov, M. Grella, A. Nidhi, P. Xia, K. Gorman, E. Vylomova, S. J. Mielke, G. Nicolai, M. Silfverberg, T. Arkhangelskij, N. Krizhanovsky, A. Krizhanovsky, E. Klyachko, A. Sorokin, J. Mansfield, V. Ernštreits, Y. Pinter, C. L. Jacobs, R. Cotterell, M. Hulden, and D. Yarowsky. 2020. UniMorph 3.0: Universal Morphology. *Proceedings of the 12th Conference on Language*

Resources and Evaluation (LREC 2020), pages 3922–3931, European LanguageResources Association.

Sylak-Glassman, J. 2016. *The Composition and Use of the Universal Morphological Feature Schema* (UniMorph Schema). Working draft, v. 2. Forthcoming.

Taylor, A., A. Warner, S. Pintzuk, and F. Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* [https://www-users.york.ac.uk/~lang22/YcoeHome1.htm].

Torre Alonso, R. 2021. Old English Class I Strong Verbs Lemmatization: A Morphological Generation Approach. *Studia Neophilologica.* To appear. DOI: 10.1080/00393274.2021.2010128.