Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7170–7178
Marseille, 11–16 May 2020

Lemmatising Verbs in Middle English Corpora: The Benefit of Enriching the *Penn-Helsinki Parsed Corpus of Middle English 2* (PPCME2), the *Parsed Corpus of Middle English Poetry* (PCMEP), and *A Parsed Linguistic Atlas of Early Middle English* (PLAEME)

Michael Percillier, Carola Trips
University of Mannheim
B6, 30–32, 68159 Mannheim
percillier@uni-mannheim.de, ctrips@mail.uni-mannheim.de

Abstract

This paper describes the lemmatisation of three annotated corpora of Middle English — the *Penn-Helsinki Parsed Corpus of Middle English 2* (PPCME2), the *Parsed Corpus of Middle English Poetry* (PCMEP), and *A Parsed Linguistic Atlas of Early Middle English* (PLAEME) — which is a prerequisite for systematically investigating the argument structures of verbs of the given time. Creating this tool and enriching existing parsed corpora of Middle English is part of the project *Borrowing of Argument Structure in Contact Situations* (BASICS) which seeks to explain to which extent verbs copied from Old French had an impact on the grammar of Middle English. First, we lemmatised the PPCME2 by (1) creating an inventory of form-lemma correspondences linking forms in the PPCME2 to lemmas in the MED, and (2) inserting this lemma information into the corpus (precision: 94.85%, recall: 98.92%, accuracy: 94%). Second, we enriched the PCMEP and PLAEME, which adopted the annotation format of the PPCME2, with verb lemmas to undertake studies that fill the well-known data gap in the subperiod (1250–1350) of the PPCME2. The case study of reflexives shows that with our method we gain much more reliable results in terms of diachrony, diatopy, and contact-induced change.

Keywords: Lemmatisation, Middle English, verb argument structure

1. Introduction

The project *Borrowing of Argument Structure in Contact Situations*[1] (BASICS) investigates the contact situation between Old French (OF) and Middle English (ME) which set in after the Norman Conquest (1066) and lasted until 1500. More specifically, we hypothesise that the copying of OF verbs with their predicate-argument structures (AS) has favoured and produced grammatical changes in ME. Instead of using the more traditional and more problematic term 'borrowing' we use Johanson's (2002, 287f) term 'copying' as it allows for the non-identicality of original and copied material. We assume that verbs copied from OF to ME are global copies in Johanson's terms, bringing along a block of properties (material, semantic, combinational in syntax and word structure, frequential). So far we have conducted item-oriented, class-oriented, and construction-oriented studies to gain new insights into this contact situation and contact-induced changes in more general terms (see e.g. Percillier (2019), Trips and Stein (2018), Trips and Stein (2019)). To do so, we developed a method that allows us to systematically query verbs of French and non-French origin in the linguistically annotated corpora available for Middle English. In this article we will describe our method in detail, and by means of a corpus study show how it is successfully applied to phenomena relevant to our project.

The outline of the article is as follows: In section 2 we briefly discuss the resources that are currently available for Middle English and that built the basis for our method. Section 3 describes in detail how we developed our method of lemmatising a Middle English corpus (PPCME2), including its evaluation. In section 4 we motivate why lemmatisation should be extended to the other Middle English corpora available and how we applied our method to them. In section 5 we present a quantitative corpus study of the rise of the reflexive system in Middle English to show that with our method we were able to gain much more reliable results in terms of diachrony, diatopy, and contact-induced change. Section 6 summarises our results and concludes.

2. Currently Available Resources

The *Oxford English Dictionary* (Proffitt, 2019), abbreviated as OED, serves as a point of reference for the project, not only because it is an authoritative resource on the English lexicon, but also because it contains a wealth of etymological information. Owing to a cooperation in the project with the OED's principal etymologist Philip Durkin, we were able to obtain a list of 2,026 English verbs copied from French between 1066 and 1500 based on an explicit query. The verbs in this list constitute the starting point of the project, as they are the loan words whose argument structure is thus introduced to English and can thereafter extend to other verbs.

The ways in which these copied verbs were used should be verified empirically in a corpus. For ME, the *Penn-Helsinki-Parsed-Corpus of Middle English* (Kroch and Taylor, 2000, henceforth PPCME2), presents the advantage of being syntactically annotated. The corpus is based on the Middle English section of the Diachronic Part of the Helsinki Corpus of English Texts and consists of

[1] This project is funded by the Deutsche Forschungsgemeinschaft, DFG research grant TRI555/6-1, TRI555/6-2 Sachbeihilfe. For further information see `https://tinyurl.com/dfgbasics/`.

56 texts, totaling ca 1.2 million words. Following the Helsinki corpus it is divided into four periods: M1 (1150–1250), M2 (1250–1350), M3 (1350–1420), and M4 (1420–1500).[2] The annotation format used is *Penn-Treebank*, which can be queried using the specialized software tool *CorpusSearch* (Randall, 2010). The format uses sets of parentheses to represent the clause hierarchy, as illustrated for Modern English in the example below.[3]

```
(1)  ((IP-MAT (ADVP-TMP (ADV Then))
              (NP-SBJ (D the)
                      (N child))
              (VBD became)
              (ADJP (ADJR happier)
                    (CONJ and)
                    (ADJR happier))
              (E-S .)))
```

At the lowest level of the tree hierarchy, each form is assigned a part-of-speech (POS) tag. Consequently, the annotation format, in combination with *CorpusSearch*, makes it possible to search for specific grammatical properties, such as past tense verbs using the *VBD* tag, or specific forms such as *became*. However, due to frequent spelling variation in ME data and the existence of irregular verb paradigms, queries for all forms of a verb, such as *become*, are not readily available by searching for verb stems in ME corpora. To remedy this, all lexical verb forms in the PPCME2 were lemmatised, a process described in Section 3.

In addition to the OED, the *Middle English Dictionary* (Schaffner et al., 2018), henceforth MED, constitutes a further dictionary resource that is relevant for the lemmatisation of a ME corpus.

The MED uses unique numerical identifiers (henceforth MED-IDs) for each entry that can serve to disambiguate homonyms. Furthermore, entries in the MED and the OED are linked, so that using both resources in tandem makes it possible to distinguish between native and copied ME verbs by checking them against the list of verbs copied from French provided by the OED.

3. Lemmatisation of the PPCME2

As previously stated, the lemmatisation of a ME corpus, in particular of its verbs, is a crucial step for any study in which queries of specific verbs or semantic verb classes are to be undertaken. Given the absence of lemmatised ME corpora or any gold standard for the lemmatisation of ME data, the lemmatisation process relies on the semi-manual assignment of graphemic verb forms to their respective lemmas.[4] The process is divided into two major steps: (1) the creation of an inventory of form-lemma correspondences linking forms in the PPCME2 to lemmas in the MED, and (2) the insertion of this lemma information into the corpus.

3.1. Assignment of Form-Lemma Correspondences

Verb forms were extracted from the PPCME2, and each verb form was paired with a lemma and the corresponding ID extracted from the MED. This assignment of verb forms to lemmas was undertaken manually by four trained research assistants and one of the authors using a spreadsheet application. They also had the option of specifying multiple lemmas or marking their choices as doubtful. In total, 19,320 graphemic verb forms were assigned to 2,979 lemmas as primary matches, alongside 4,973 lemmas specified as additional possible matches. The resulting form-lemma links were exported to the tabular CSV (Comma Separated Values) format.

3.2. Insertion of Lemma Information into the Corpus

Using the inventory of form-lemma correspondences just mentioned, the insertion of lemma information is performed. For every verb marked with a POS tag beginning with *V* in the corpus,[5] the following instructions are carried out:

The main approach is a lexical lookup in the inventory of form-lemma correspondences. Should this not return any results, two fallbacks are used: (1) Spelling variants are generated and queried for corresponding lemmas. The grapheme substitution rules used are shown in Table 1.[6] Further, forms containing hyphens or tildes are assigned spelling variants without these characters. (2) The form is stemmed and checked against all stemmed forms in the form-lemma inventory. Stemming is achieved by removing the following ME inflectional suffixes: *+d*, *+d+d*, *+t*, *+t+t*, *an*, *ande?*, *dd?*, *den?*, *e*, *e+d*, *e+t*, *ede?*, *enn?*, *e?st*, *et*, *in?d?e?*, *ingg?e?*, *ode*, *odest*, *oden*, *ten?*, *th*, *tt?*, *yde?*, *ynde?*, *ynn?*, *yngg?e?*, and *yst*.[7]

The lemma information is appended directly to the form in the corpus, so as to still comply with the Penn-Treebank format and related software such as *CorpusSearch*. Each piece of inserted information is demarcated by @ characters and specified by an attribute. Verb lemmas are specified by the attribute *l*, and MED-IDs by the attribute *m* (see Example (2)). For verbs occurring in the list of French-based verbs, an additional attribute *e* (for *etymology*) is defined as *french* (see Examples (3)/(7)). For other verbs, the attribute *e* receives the value *nonfrench*. The attribute *w* (for *warning*) indicates that the lemma was matched using either the spelling substitution or the stemming method (see Examples (3)/(6) and (4) respectively), or that the manual form-lemma match was deemed doubtful (see Example (5)). For verbs spelt as multiple words, the information is appended

[2] Information from `http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/description.html`.

[3] Example adapted from `http://corpussearch.sourceforge.net/format.html`.

[4] The tool and resources to apply the lemmatisation process are available via the project's *BASICS Toolkit* website: `http://tinyurl.com/basicstoolkit` under the section *Lemmatizer*.

[5] Lexical *be*, *do*, and *have* need not be lemmatised as their tags (*B**, *D**, and *H** respectively) already reveal their lemma.

[6] In the PCCME2, the character sequences *+d*, *+g*, and *+t* represent the graphemes <ð, ȝ, þ> respectively.

[7] Question marks refer to the regular expression quantifier specifying that the preceding character may or may not occur.

Grapheme	Substitution
i	e/y
e	i
y	i/g/+g
u	v/ou
v/ou	u
th	+t/+d
+t/+d	th
g	+g/y
+g	g/y
ll	l
nn	n
pp	p

Table 1: Grapheme substitution rules for generating Middle English spelling variants

to the final element (see Example (6)). For verb forms that can be assigned to multiple lemmas, all matching lemmas are inserted (see Example (7)). Should no form-lemma correspondence have been found even after the stemming method, the lemma and MED-ID are marked as *NA*. Such cases include forms which are wrongly tagged as verbs in the corpus, as shown for the Roman numeral in Example (8). The lemma insertion process is summarized in Figure 1.

(2) `(VAG settyng@l=setten@m=39654@e=nonfrench@)`

(3) `(VAG consyderyng@l=consideren@m=9387@e=french@w=substitution@)`

(4) `(VB tellyn@l=tellen@m=44693@e=nonfrench@w=stemming@)`

(5) `(VBI wilne@l=wilnen@m=52815@e=nonfrench@w=doubt@)`

(6) `(VBP21 vnder)(VBP22 stont@l=understonden@m=48362@e=nonfrench@w=substitution@)`

(7) `sesyd@l=seisen@m=39243@e=french@l=cesen@m=7155@e=french@`

(8) `(VAN iii@l=NA@m=NA@)`

With this additional annotation, the PPCME2 can be queried for syntactic structures as before, but also for specific verbs. Using *CorpusSearch*, this is achieved by specifying the lemma with the *exists* function, e.g. `(*l=setten@* exists)`. To distinguish between homonyms, the MED-ID can also be used for unambiguous queries, e.g. `(*m=39654@* exists)`.

3.3. Evaluation

The lemmatisation of verbs in the PPCME2 treated 128,523 verbs in total. 110,827 verbs (86.23%) were directly assigned matching lemmas. Additionally, 6,469 verbs (5.03%) were assigned a lemma using spelling substitution, and 10,359 verbs (8.06%) using stem comparison. The total of lemmatised verbs is thus 127,655 (99.32%), whereas 868 verbs (0.68%) could not be assigned any lemma.

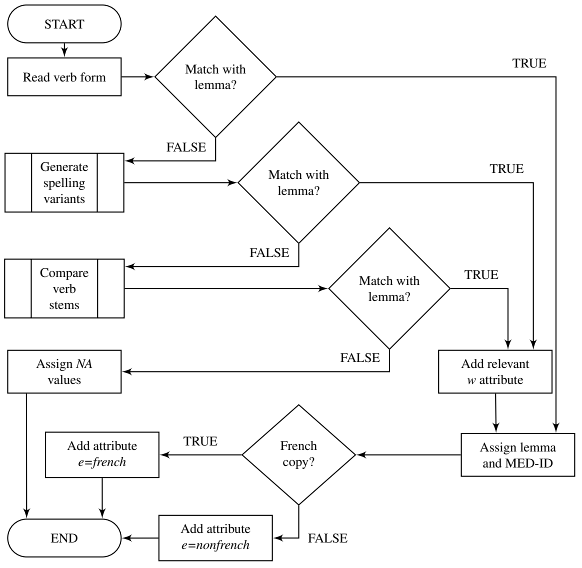

Figure 1: Lemma Insertion Process

The lemmatisation process was evaluated based on a manually verified random sample of 100 tokens, using the following categorisation: (1) true positives for forms with a correct lemma suggestion, (2) false positives for forms with only incorrect lemma suggestions, (3) false negatives for forms marked with *NA* that should have been matched with a correct lemma, and (4) true negatives for forms with an appropriate *NA* marking, e.g. a lemma unidentifiable even upon manual verification, or a form that is actually not a verb. Precision was thus determined to be 94.85%, recall 98.92%, and accuracy 94%.

4. Extending the Lemmatisation to Further ME Corpora

The present section describes the application of the verb lemmatisation method devised for the PPCME2 to other ME corpora, namely the *Parsed Corpus of Middle English Poetry* (Zimmermann, 2018), abbreviated as PCMEP, and the *Parsed Linguistic Atlas of Early Middle English* (Truswell et al., 2018), abbreviated as PLAEME.

4.1. Motivation for the Lemmatisation of Additional Corpora

Due to its focus on prose texts, the PPCME2 contains a large data gap in the period M2 (1250–1350). This gap is problematic in terms of documenting diachronic developments, as M2 contains only a fourth to a third of the word counts in other periods. Further, there is a lack of representative diatopic variation, as only texts from the southeast of England are to be found in M2. The addition of texts from the PCMEP and PLAEME fills this gap both in terms of diachrony (see Figure 2) and diatopy (see Figure 3).

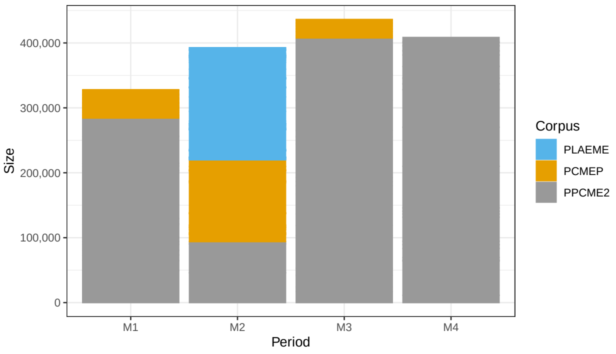

Figure 2: Word counts by ME sub-period for PPCME2, PLAEME, and PCMEP

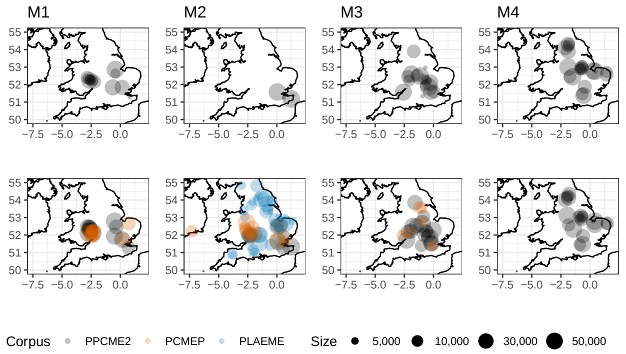

Figure 3: Localised texts by ME sub-period for PPCME2 only (top), and with additions from PLAEME and PCMEP (bottom), text size shown in $\log_{10}$

4.2. Application and Evaluation of the Lemmatisation

Given that the PCMEP and PLAEME adopted the annotation format of the PPCME2, applying the lemmatisation method can be performed with only minor adjustments. Specifically, the PLAEME contains additional formatting conventions inherited from the *Linguistic Atlas of Early Middle English* (Laing, 2013) which require slight modifications in the lemmatisation process, e.g. character sequences such as *+w* to represent the grapheme <p> ('wynn'), morpheme boundaries marked with the = character, and the presence of so-called *lexels* appended to word forms. Lexels are "elements identifying the word lexically (in much the same way as a lemma but often with additional information about word sense)" (Truswell et al., 2019, 23), so that the lemmatisation of verbs in the PLAEME could be argued to be redundant. However, applying the process to the PLAEME allows a common scheme for the three corpora, as well as the insertion of additional information concerning verb origin. Due to the aforementioned differences in PLAEME, the insertion of verb lemmas therein, as shown in Example (9), differs slightly from the format previously shown in Examples (2–8), as additional information such as lexels should be maintained.

(9) `(VBP sing=ez-sing@l=singen@m=40448 @e=nonfrench@)`

The lemmatisation of verbs in the PCMEP and PLAEME is applied in several rounds, whose respective evaluations are based on random samples of 100 verb forms. Round 1 consists in applying the lemmatisation process to the PCMEP and PLAEME without any modifications, with the exception of the aforementioned adjustments required due to differences in the PLAEME format. Given that the process was tailored to the PPCME2, any PCMEP/PLAEME form belonging to a lemma not encountered in the PPCME2 is expected to be marked as *NA*. This is indeed reflected in the comparatively low recall values observed for round 1 (see Table 3). For round 2, all *NA* forms in the PLAEME were assigned a matching lemma, i.e. by repeating the process undertaken for the PPCME2 described in section 3.1. Furthermore, errors encountered in the random samples used for the evaluation of the preceding round were corrected. The reason for excluding the PCMEP from the verification of *NA* forms lies in the presence of lexels in the PLAEME, which eases the verification process. Any new lemmas identified in the PLAEME may also occur in the PCMEP and should therefore be lemmatised when the process is run again. This is visible with the clear improvement in recall from round 1 to round 2 for both the PCMEP and PLAEME. The improvement for round 3 then consists in manually assigning lemmas for any remaining *NA* forms in the PCMEP, which raises recall for this corpus even further. One observation made in Tables 2, 3, and 4 is that values do not consistently improve as one would expect, e.g. precision for the PPCME2 and PCMEP or recall for the PLAEME. Upon closer inspection, inconsistent precision values are caused by certain forms not including the entire set of homonyms to which they should be assigned, and receding recall values for PLAEME are due to inadequate handling of forms containing ^ (circumflex/caret) characters to mark superscripts. The correction of these errors results in the overall improvements observed in round 4.

Corpus	Round 1	Round 2	Round 3	Round 4
PPCME2	94.95%	97.00%	95.88%	97.00%
PCMEP	94.38%	92.22%	91.92%	94.90%
PLAEME	81.01%	89.80%	92.22%	91.84%

Table 2: Precision for the lemmatisation process on PPCME2, PCMEP, and PLAEME, based on random samples of 100 forms per corpus

5. Case Study: the Rise of Reflexives in ME

In this section we discuss empirical findings gained by applying the method and tools presented above which shed new light on the development of reflexive strategies in ME. We will show that our method not only provides new insights into contact-induced change on the basis of our etymological distinction but also traces this development in

Corpus	Round 1	Round 2	Round 3	Round 4
PPCME2	98.92%	100.00%	100.00%	100.00%
PCMEP	88.42%	96.70%	100.00%	100.00%
PLAEME	75.29%	98.88%	95.60%	100.00%

Table 3: Recall for the lemmatisation process on PPCME2, PCMEP, and PLAEME, based on random samples of 100 forms per corpus

Corpus	Round 1	Round 2	Round 3	Round 4
PPCME2	94%	97%	95%	97%
PCMEP	84%	85%	92%	95%
PLAEME	64%	89%	89%	92%

Table 4: Accuracy for the lemmatisation process on PPCME2, PCMEP, and PLAEME, based on random samples of 100 forms per corpus

a more reliable and valid way by filling the diachronic and diatopic gaps described above.

It is well known that languages differ in the way they formally mark reflexivity. In a very general sense a predicate can be called ‘reflexive’ whenever two of its arguments refer to the same person, i.e. when they are co-referent. Present-Day English (PDE) is a language which uses the set of *self*-pronouns to mark all three persons in reflexive situations as well as the adnominal intensifier, see the examples in (10) (Kemmer, 1993; König and Siemund, 2000, 48)). These forms inflect for person, number, and in the third form, for gender.

(10) a. I was talking to **myself**.
b. You were talking to **yourself**.
c. Pierre was talking to **himself**.
d. Mary was talking to **herself**.
e. Pierre **himself** wanted to become president.

In contrast, the Modern French system uses the first and second person pronouns to mark co-reference and has an invariable third person reflexive marker (*se*), see the examples in (11).

(11) a. Je **me** plaisais.
b. Tu **te** plaisais.
c. Pierre **se** plaisait.
‘I/you/Pierre was/were pleased.’

The system that we find in PDE was not found in Old English — at that time personal pronouns did double duty both as markers of disjoint reference and as markers of co-reference as shown with the examples in (12) (König and Siemund, 2000; Keenan, 2009, 44):

(12) a. ða **gegyrede** **heo** **hy** mid
then dressed she$_i$.NOM her$_i$.ACC with
hærenre tunecan
of-hair tunic
‘then she dressed herself in a tunic of hair.’
(Mart 190 c.875 in Keenan (2009, 20))

b. forðæm **hi** **him** **ondrædað** ða
because they$_i$.NOM them$_i$.DAT fear the
frecenesse ðe hi ne gesioð
danger that they NEG see
‘because they fear the danger that they do not see.
(CP 433 c.880 in Keenan (2009, 21))

Self functioned as an intensifier and could be combined with a personal pronoun in object position (co-reference with preceding subject):

(13) se **Hælende** sealde **hine**
the lord$_i$.NOM.SG.M gave him$_i$.ACC.SG.M
sylfne for us
self.ACC.M.SG for us
‘The Saviour gave himself for us.’
([ELet 4 1129] in König and Siemund (2000, 45))

In ME, the intensifier occurs more and more often in compound forms (*himself, herself* etc.; cf. König and Siemund (2000, 46)), which leads to a situation where two alternative ways to express reflexivity compete with each other (cf. Peitsara (1997, 280)). We will call them the ‘simple strategy’ (use of personal pronoun) and the ‘*self*-strategy’ (use of *self*-compounds) here. In the course of the ME period, the paradigm of reflexive pronouns developed, disambiguating the reflexive use of verbs (cf. also Keenan (2009), van Gelderen (2000), McWhorter (2004)). Some authors attribute this development to French influence: Einenkel (1916, 50), Mustanoja (1960, 502–3) and Visser (1963, 328) assume that the *self*-compounds are calques of French expressions like *(soi)-même* or built in analogy to French reflexive verbs which used first and second person pronouns to mark co-reference and had an invariable third person reflexive marker *se* just like in Modern French.

In a similar vein, Peitsara (1997) comments on one of her findings in her study of the reflexive strategies in ME and Early Modern English (EModE) based on data from the *Helsinki Corpus* (Matti Rissanen and others (1991)). She finds an increase of overtly reflexive constructions in ME with a peak in the subperiod of ME3 (1350–1420). She notes: “It could be supposed that the peak of frequency in ME3 might have to do with the grammaticalization of the reflexive construction in English but, interestingly enough, it also coincides with the period of most profuse introduction of French (Peitsara, 1997, 287). Her results of comparing the ‘simple strategy’ and the ‘*self*-strategy’ are shown in Figure 4.

So far the rise of the reflexive system that we find in PDE in the light of the contact hypothesis has not been explained in

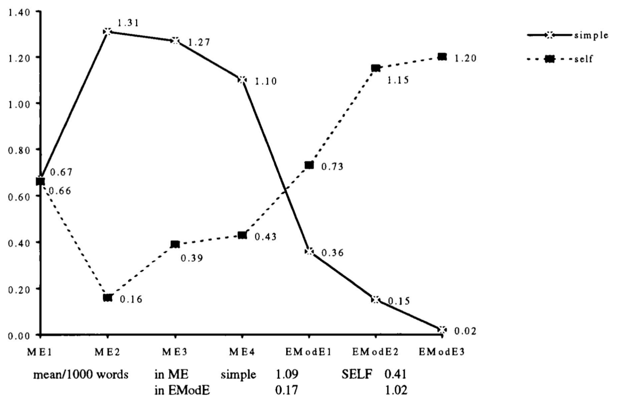

Figure 4: Reflexive strategies in Middle English and Early Modern English per 1000 words (Peitsara, 1997, 289)

a satisfying way, neither empirically nor theoretically. Although Peitsara's article is a valuable contribution to this topic, her empirical study is based on the *Helsinki corpus* and thus her results are as problematic as the results gained from the PPCME2 due to the empirical gap in M2 concerning diachronic and diatopic aspects (see below). The addition of further data (corpora), however, fills this gap and verb lemmatisation, including etymological information, facilitates the investigation of the contact-induced hypothesis. On this basis we seek to answer the following questions: (1) How often does the 'simple strategy' occur with native verbs and verbs copied from OF? (2) How often does the '*self*-strategy' occur with native verbs and verbs copied from OF? The following subsection provides details on the queries used to search for reflexives in the PPCME2, PLAEME, and PCMEP.

5.1. Methodology: Querying Reflexives in Three Corpora

The annotation scheme used in the PPCME2, and by extension the PLAEME and PCMEP as well, contains a `NP-RFL` tag, which however applies only to non-argument reflexives. Any other reflexives are therefore not explicitly marked in the corpora and can occur in the annotation patterns shown in Table 5.

Strategy	Type	Annotation (of 2$^{\text{nd}}$ argument)
Simple	Argument	`NP-OB1`, `NP-OB2`
Simple	Non-argument	`NP-RFL`
Self	Joined	`PRO+N`, `PRO$+N`
Self	Split	`PRO N`, `PRO$ N`

Table 5: Overview of possible annotation for reflexives in the PPCME2, PLAEME, and PCMEP

An example of each strategy shown in Table 5 is given in (14).

(14)
a. I wash me.
b. I fear me.
c. I wash myself.
d. I wash me self.

Given the various annotation patterns for reflexives, multiple queries need to be formulated to account for all types of reflexives. Non-argument reflexives marked with the simple strategy can be queried with the `NP-RFL` tag. A distinction by verb etymology can be drawn by specifying the desired value, either `french` or `nonfrench`, in the slot marked by `{ETYMOLOGY}` in (15). Given the common annotation scheme and the extension of the verb lemmatisation process, this query can be applied to the three corpora.

(15)
```
(IP* iDoms V*)
AND (V* iDoms *@e={ETYMOLOGY}@*)
AND (IP* iDoms NP-RFL)
```

The queries for the other reflexive types are not as straightforward given the lack of dedicated marking. Argument reflexives marked with the 'simple strategy' are only annotated with their syntactic function as either `NP-OB1` for direct objects or `NP-OB2` for indirect objects, which makes them indistinguishable from other (non-reflexive) direct and indirect objects. In order to mitigate this problem, the query was restricted to cases in which the subject and object of the clause match in terms of grammatical person, number, and gender. Such cases involving first and second person pronouns (*I VERB me, we VERB us, thou VERB thee, ye VERB you*) are overwhelmingly reflexive, whereas cases involving third persons (*he VERB him, she VERB her, they VERB them*) are characterised by ambiguity as to whether subject and object are co-referential. For this reason, the output of the query shown in (16) was verified manually when applied to third person pronouns. Unlike the previous query, distinct versions have to be used for the PPCME2 and PCMEP on the one hand and the PLAEME on the other hand. For the former, forms of a given pronouns as attested in the MED have to be inserted in the slot marked by `{FORMS}`, e.g. `me|meo|mi|Me|Meo|Mi` for *me*, whereas the presence of lexels in the PLAEME simplifies this process whereby inserting the corresponding lexel, e.g. `*-me`, covers all spelling variants.[8]

(16)
```
(IP* iDoms V*)
AND (V* iDoms *@e={ETYMOLOGY}@*)
AND (IP* iDoms NP-SBJ*)
AND (NP-SBJ* iDoms [1]PRO)
AND ([1]PRO iDoms {FORMS})
AND (IP* iDoms NP-OB*)
```

[8] The query for argument reflexives using the 'simple strategy' is limited to pronoun subjects and is therefore not exhaustive, as an exhaustive query would require the verification of any full noun phrase subject occurring with a third person pronoun object. As the other queries, i.e. for non-argument reflexives using the 'simple strategy', and for any reflexives using the '*self*-strategy', are exhaustive, we refrain from making statements on the contrasts between the two reflexive marking strategies, and focus on the contrasts between French-based and non-French-based verbs, for which the parameters of exhaustive and non-exhaustive queries are identical.

```
        AND (NP-OB* iDoms [2]PRO)
        AND ([2]PRO iDoms {FORMS})
```

Querying reflexives marked with the ‘*self*-strategy’ requires the exclusion of adnominal intensifiers. As such cases are marked by the `-PRN` (‘appositive or parenthetical’) tag, limiting the query to *self*-compounds within object constituents effectively excludes them. In a similar vein to (16), distinct versions of the query shown in (17) have to be prepared for the PPCME2 and PCMEP on the one hand, where spelling variants of *self* have to be listed, i.e. `*self|*selfe|*selph|*selphe|*selef` and 60 further variants listed in the MED, and for the PLAEME on the other hand, for which only the lexel `*-self` needs to be specified. A modification of the query, shown in (18), covers cases in which *self*-compounds are spelled as separate words.

(17)
```
(IP* iDoms V*)
AND (V* iDoms *@e={ETYMOLOGY}@*)
AND (IP* iDoms *OB*)
AND (*OB* iDoms PRO*+N*)
AND (PRO*+N* iDoms {FORMS})
```

(18)
```
(IP* iDoms V*)
AND (V* iDoms *@e={ETYMOLOGY}@*)
AND (IP* iDoms *OB*)
AND (*OB* iDoms PRO*)
AND (*OB* iDoms N*)
AND (N* iDoms {FORMS})
```

The output from the queries just described was annotated for grammatical person so as to be able to investigate whether reflexives developed differently for first, second, and third persons in the ME period. The following subsection presents and discusses the results.

5.2. Results: Changes in the Middle English Reflexive System

The queries described in the preceding subsection were applied to the PPCME2, PCMEP, and PLAEME. The results display great discrepancies when only the PPCME2 is used as opposed to when the PCMEP and PLAEME are added. The left-hand pane of Figure 5 suggests that the ‘*self*-strategy’ for marking reflexives virtually disappears in M2, only represented by 2 tokens, only to surge in M3. This development is also outlined in Peitsara (1997, 289) as shown in Figure 4. Adding data from the PCMEP and PLAEME overcomes the paucity of data for M2 and results in a more plausible development, shown in the right-hand pane of Figure 5, whereby the ‘*self*-strategy’ increases steadily, even if only moderately at first. The new picture that emerges for both reflexive strategies is therefore one of stability from M1 to M2, followed by notable changes from M2 to M3, rather than two subsequent changes in opposite directions as data taken only from the PPCME2 would suggest.

In order to investigate the possibility of French influence on the development of reflexives in the ME period, our queries distinguished French-based verbs (FBVs) and non-French-based verbs (nFBVs). As FBVs are rare in M1 but are copied to English in increasing numbers from M2 onwards (Percillier, 2016, 212), reporting frequencies of FBVs used reflexively may reflect changes in the overall frequencies of FBVs rather than changing characteristics of reflexive use with these verbs. For this reason, the use of reflexives is reported as ratios per 1000 verbs of the same etymological group, as shown in Figure 6.

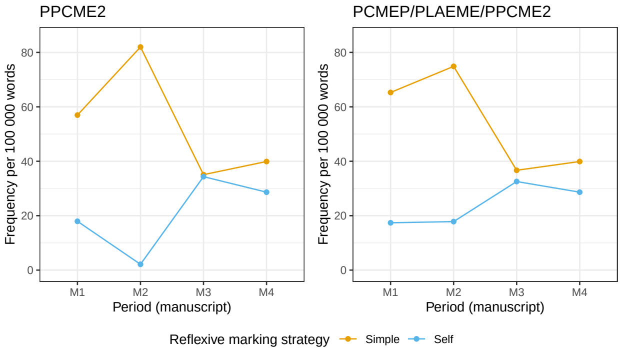

Figure 5: Normalised frequencies of ‘simple’ and ‘*self*’ reflexive marking strategies in the PPCME2 only (left), and the PCMEP/PLAEME/PPCME2 combined (right)

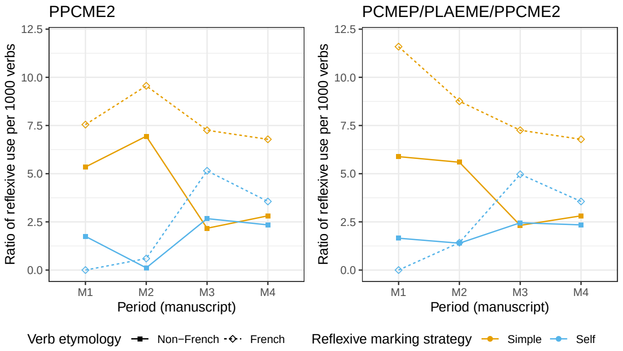

Figure 6: Ratios of reflexive strategy use per 1000 verbs of the same etymological group in the PPCME2 (left) and combined data from the PCMEP/PLAEME/PPCME2 (right)

The right-hand pane of Figure 6 suggests several contrasts between the developments of reflexives with FBVs and nFBVs. The ‘*self*-strategy’ steadily increases with FBVs from M1 to M3 with FBVs as the ‘simple strategy’ steadily decreases. At the same time, reflexives with nFBVs exhibit comparative stagnation from M1 to M2, with noticeable changes occurring only from M2 to M3. An analysis relying solely on data from the PPCME2, as shown in the left-hand pane of Figure 6, would not have revealed these differences, as FBVs and nFBVs appear to develop in a more parallel manner. A further contrast relates to the proportion of reflexive contexts per number of verbs, as FBVs display a higher ratio. This contrast is also visible when only relying on PPCME2 data, but in a less pronounced form.

Taking grammatical person into account, as shown in Figure 7, reveals that changes occur more rapidly for the third person than for the first and second persons. This is apparent from both the PPCME2 data and the combined data PCMEP/PLAEME/PPCME2 data. However, the timing and extent of these changes are different when the M2 data gap is filled. As the right-hand pane of Figure 7 shows, the more important changes with regard to FBVs in the third person occur in M2, when FBVs first enter the English language in larger numbers, with the 'simple strategy' strongly decreasing and the '*self*-strategy' strongly increasing, as opposed to this happening only in M3 as the view in the left-hand pane would suggest. As regards nFBVs, we observe strong fluctuations from M1 to M2 for the PPCME2 data which should in fact be viewed as stagnation, as was already established in Figure 5.

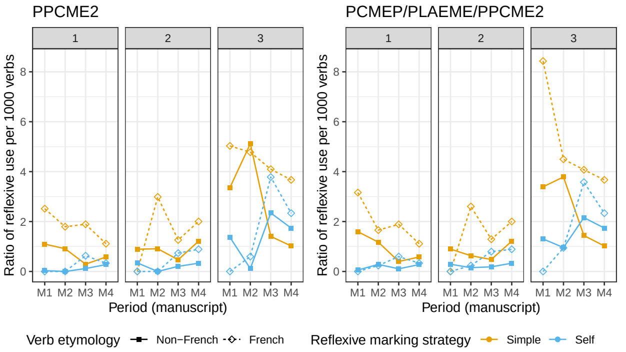

Figure 7: Ratios of reflexive strategy use per 1000 verbs of the same etymological group in the PPCME2 (left) and combined data from the PCMEP/PLAEME/PPCME2 (right), accounting for grammatical person

The consequences of the contrasts between the PPCME2 data and the combined PCMEP/PLAEME/PPCME2, and the questions that they raise for a more in-depth analysis of the developments of reflexives in ME, are addressed in the concluding section.

6. Conclusions

In this article we presented a method to lemmatise the linguistically annotated corpora of Middle English (ME) that are presently available. The *Penn-Helsinki Parsed Corpus of Middle English 2* (PPCME2), which is the major annotated corpus for ME, had long been affected by a data gap in the M2 period (1250–1350) due to its focus on prose texts. Recently, this gap has been filled by corpora such as the *Parsed Corpus of Middle English Poetry* (PCMEP) and *A Parsed Linguistic Atlas of Early Middle English* (PLAEME), which adopt the annotation format of the PPCME2. We described the extension of our verb lemmatisation method, which we originally developed for the PPCME2, to the PCMEP and PLAEME.

In our case study of the development of reflexives in ME, we showcase how the common verb lemmatisation scheme facilitates queries across the three corpora with only minimal adjustments. Contrasting results one would obtain when only consulting the PPCME2 with results from a combination of data from the three corpora, we show that the combined data reveal patterns that would have been more difficult to discern or missed entirely when using the PPCME2 only: (1) distinct developmental patterns for reflexives with French-based verbs where changes already occur from M1 to M2, as opposed to (2) reflexives with other verbs which remain comparatively stable from M1 to M2, and undergo changes towards M3, and (3) the changes with regard to reflexives with French-based verbs occur most rapidly in the third person.

A more thorough investigation into the developments of reflexives in the ME period would have to offer explanations for these observations. For instance, the rise of the '*self*-strategy' occurring first with French-based verbs in the third person may be linked to the fact that French possessed an unambiguous way of marking reflexivity in the third person which English hitherto lacked, as either *se* or *soi/elle/eux même(s)*, with the latter being replicable in English as *him/her/it/them self/selves*. A further investigation of this possibility entails looking at the other locus of contact-induced change besides copied verbs and their argument structures, namely the effect of translation. In order to do so, the corpora need to be harmonised not only in terms of syntactic annotation, which is inherent to their design, and their verb lemmatisation, which we have achieved with the method described in the present paper, but also in terms of their metadata. Specifically, a determination of which texts are English originals and which texts are translations from French, Latin, or other languages, is crucial for investigating the role of translation effects. The harmonisation of metadata across the three corpora for other variables is important for the study of ME in general, as this should enable further types of studies such as genre effects.

7. Bibliographical References

Einenkel, E. (1916). *Geschichte der englischen Sprache*, volume II: Historische Syntax. Trübner, Strasbourg.

Johanson, L. (2002). Contact-induced change in a code-copying framework. In Mari C. Jones et al., editors, *Language change: the interplay of internal, external and extra-linguistic factors*, pages 285–313. de Gruyter, Berlin.

Keenan, E. (2009). Linguistic theory and the historical creation of English reflexives. In Paola Crisma et al., editors, *Historical Syntax and Linguistic Theory*, chapter 2, pages 17–40. Oxford University Press, Oxford.

Kemmer, S. (1993). *The Middle Voice*. Number 23 in Typological Studies in Language. Benjamins Publishing, Amsterdam and Philadelphia.

Kroch, A. and Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2), release 3*. University of Pennsylvania, Philadelphia.

König, E. and Siemund, P. (2000). The development of complex reflexives and intensifiers in English. *Diachronica*, XVII(1):39–84.

Laing, M. (2013). *A Linguistic Atlas of Early Middle English, 1150–1325, Version 3.2*. The University of Edinburgh, Edinburgh.

Matti Rissanen, M. K. et al. (1991). *The Helsinki Corpus of English Texts*. Department of Modern Languages, University of Helsinki, Helsinki.

McWhorter, J. (2004). What happened to English? In Hartmut Czepluch et al., editors, *Focus on Germanic Typology*, number 6 in Studia typologica, pages 19–60. Akademie Verlag, Berlin.

Mustanoja, T. F. (1960). *A Middle English Syntax Part I*. Number 23 in Mémoires de la Société Néophilologique de Helsinki. Société Néophilologique, Helsinki.

Peitsara, K. (1997). The development of reflexive strategies in english. In Matti Rissanen, et al., editors, *Grammaticalization at Work: Studies of Long-Term Developments in English*, pages 277–370. Mouton de Gruyter, Berlin.

Percillier, M. (2016). Verb lemmatization and semantic verb classes in a Middle English corpus. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 209–214.

Percillier, M. (2019). Dynamic modelling of medieval language contact: The case of Anglo-Norman and Middle English. In Roger Schöntag et al., editors, *Diachrone Migrationslinguistik: Mehrsprachigkeit in historischen Sprachkontaktsituationen*, pages 79–99. Peter Lang, Berlin.

Michael Proffitt, editor. (2019). *Oxford English Dictionary*. Oxford University Press, Oxford, 3 edition.

Randall, B. (2010). CorpusSearch (Version 2.003.00)[Computer Software].

Paul Schaffner, et al., editors. (2018). *Middle English Dictionary*. University of Michigan, Ann Arbor.

Trips, C. and Stein, A. (2018). A comparison of multigenre and single-genre corpora in the context of contact-induced change. In Richard Whitt, editor, *Diachronic corpora, genre and language change*, Studies in Corpus Linguistics, pages 241–260. Benjamins, Amsterdam and Philadelphia.

Trips, C. and Stein, A. (2019). Contact-induced changes in the argument structure of Middle English verbs on the model of Old French. In Eitan Grossman, et al., editors, *Journal of Language Contact. Special Issue on Valency and Transitivity in Contact*, pages 232–267. Brill, Leiden.

Truswell, R., Alcorn, R., Donaldson, J., and Wallenberg, J. (2018). *A Parsed Linguistic Atlas of Early Middle English*. University of Edinburgh.

Truswell, R., Alcorn, R., Donaldson, J., and Wallenberg, J. (2019). A Parsed Linguistic Atlas of Early Middle English. In Rhona Alcorn, et al., editors, *Historical Dialectology in the Digital Age*, pages 19–38. Edinburgh University Press, Edinburgh.

van Gelderen, E. (2000). *A history of the English reflexive pronouns: person, self, interpretability*. Linguistik Aktuell/Linguistics Today 39. John Benjamins, Amsterdam and Philadelphia.

Visser, F. T. (1963). *An Historical Syntax of the English Language*. E. J. Brill, Leiden.

Zimmermann, R. (2018). *The Parsed Corpus of Middle English Poetry (PCMEP)*.