A Parsed Linguistic Atlas of Early Middle English*

Robert Truswell, Rhona Alcorn, James Donaldson (University of Edinburgh), Joel Wallenberg (Newcastle University)

1. Introduction

We describe a new parsed corpus which enriches material from the Linguistic Atlas of Early Middle English (LAEME, Laing 2013–) with explicit annotation of syntactic structure in the format of the Penn Parsed Corpora of Historical English (PPCHE, Kroch and Taylor 2000; Kroch, Santorini, and Delfs 2004; Kroch, Santorini, and Diertani 2016). This corpus is known as the Parsed Linguistic Atlas of Early Middle English (PLAEME).

In the paper, we will introduce the PPCHE format and identify a data gap in the Middle English portion of the corpora (Section 2), and then demonstrate how LAEME can fill the data gap (Section 3). Section 4 describes the process of constructing PLAEME from LAEME materials, and Section 5 extends recent studies of Middle English syntax based on the PPCHE to PLAEME data.

2. The Penn Parsed Corpora of Historical English

The Penn Parsed Corpora of Historical English (PPCHE) represent written English from the period c.1150–1914 by means of some 5.8 million words of running prose (with a small amount of verse) in more than 600 texts. They include the Penn–Helsinki Parsed Corpus of Middle English, 2nd Edition (Kroch and Taylor 2000), the Penn–Helsinki Parsed Corpus of Early Modern English (Kroch et al. 2004), and the Penn Parsed Corpus of Modern British English, 2nd Edition (Kroch et al. 2016). The York–Toronto–Helsinki Corpus of Old English Prose (Taylor, Warner, Pintzuk, and Beths 2003), though not technically one of the PPCHE, follows almost identical annotation guidelines and provides the same type of data for the entirety of extant Old English prose, with manuscript dates from the 9th through early 12th centuries.

Each word in the PPCHE is annotated with a part-of-speech tag, sometimes indicating basic morphological information as well (e.g. `BEP` for a present-tense form of *be*). Phrases at various levels of syntactic constituency are grouped with brackets, which are labelled for their syntactic category (e.g. `NP`). These labels are often extended to indicate grammatical function as well (e.g. `NP-SBJ` for a subject noun phrase), and movement (i.e. displacement phenomena) and other nonlocal dependencies are indicated with various 'trace' and empty category tags. Each sentence tree in these corpora is also associated with a unique ID node, so that it is clear where each piece of data comes from in the output of a given search.[1] As an

* Thanks to Meg Laing and two anonymous reviewers for comments on an earlier draft. The construction of PLAEME has been supported by a British Academy/Leverhulme Small Research Grant, while the collaboration between Alcorn, Truswell, and Wallenberg emerged from a networking event in Campinas, Brazil, funded by the British Council and organized by Susan Pintzuk and Charlotte Galves. Thanks to Susan Pintzuk and Aaron Ecay for helping Truswell get started with the automatic parsing procedure, to Akiva Bacovcin, Aaron Ecay, and Meredith Tamminga for making their scripts publicly available, and to Meg Laing for constant encouragement and support.

[1] Tree IDs can also be useful in controlling for text-specific idiosyncrasies in mixed-effects statistical models.

example, the full sentence token (i.e. matrix clause) in (1a) is shown in its PPCHE format in (1b). This illustrates the labelled bracketing format that is used in the corpus itself. Figure 1 gives the same information in a tree representation.

(1) a. *All thinges were made by it* 'all things were made by it'

b.
```
( (IP-MAT (NP-SBJ (Q All) (NS thinges))
          (BED were)
          (VAN made)
          (PP (P by)
              (NP (PRO it)))
          (. ,))
  (ID TYNDNEW-E1-H,I,1J.10))
```

***** Insert Figure 1 about here *****

The PPCHE format is designed for practicality rather than theoretical accuracy. Thus, the syntactic parses in these corpora, for example, do not identify VP nodes: instead verbs and their associated arguments and adjuncts are grouped under a single IP node with no indication of internal sub-structure (as shown in (1b) and Figure 1). This decision allows the corpora to remain agnostic on the difficult question of the boundaries of VP in sentences from various stages of historical English. It also makes querying the corpora more straightforward, with no loss in search accuracy. This is characteristic of the kinds of decisions made in the PPCHE: the great advantage of this format is that it can be consistently implemented and easily queried.

Moreover, the stability of the PPCHE format is such that it can be generalized across different languages and genres, making possible controlled studies of syntactic variation across time, genre, and language. For example, the Icelandic Parsed Historical Corpus (Wallenberg et al. 2011) uses essentially the same notation with, as far as is possible given the constraints of a different language, identical parsing decisions. There exist similar parsed diachronic corpora for historical French (Martineau et al. 2010) and historical Portuguese (Galves and Faria 2010); these, along with the YCOE mentioned above, the Parsed Corpus of Early English Correspondence (Taylor et al. 2006), the HeliPaD for Old Saxon (Walkden 2016), the Parsed Corpus of Early New High German (Light 2011), and the Parsed Corpus of Middle English Poetry (Zimmerman 2014–) form sister corpora to the PPCHE. The approach to parsing is similar enough in all of these corpora that queries (especially using the CorpusSearch query language; Randall 2000/2013) can be run on all of them with only minor modifications, yielding quantitatively comparable output. Because of its great flexibility and portability, the PPCHE format has become the pre-eminent format for diachronic syntax research.

The PPCHE do, however, have a number of limitations. They are built from published editions rather than manuscript texts, which means there is always at least one extra layer between analyst and primary data. Further, many of the source manuscripts are of unknown provenance, so information about time and place of composition is often limited in scope and precision—in particular, most texts are localized only to coarse-grained dialect areas. A third limiting factor is the absence of lemmatization. A significant portion of PPCHE materials—especially those which pre-date the emergence of a national written standard—exhibit a great deal of non-semantic textual variation, due partly to differences in pronunciation, e.g. ME *mon, hond, lond* (with Southern rounding of Old English /ɑ:/) vs. *man, hand, land* (without), and partly to different spelling practices, e.g. *ðu*, *þou*, *þu*, *you*, *yow*, *yu*, *thu* 'thou'.

Controlling for such extensive spelling variation is a laborious process and one which is especially prone to error.

A major shortcoming of these corpora is the paucity of data for the period c.1250–1350. Only one text (the 3534-word *Kentish Sermons*) is included from the late 13th century, while the earliest 14th-century text is the *Ayenbite of Inwyt*, from 1340. This data gap is no accident: with very few exceptions, the PPCHE are composed of prose texts, whereas the surviving English material from 1250–1340 is overwhelmingly verse. The gap is, however, extremely limiting for research on this period of English, and especially its syntax: recent corpus studies have emphasized that 1250–1340 is a transitional period for many syntactic phenomena, including the establishment of *not* as the expression of sentential negation (Ecay and Tamminga 2017), the fixing of the syntax of the dative alternation in ditransitive constructions (Bacovcin 2017), and the introduction of argumental headed *wh*-relatives (Gisborne and Truswell 2017).[2] Any further parsed material from the late 13th and early 14th centuries would naturally help piece together the details of this period of wild change in English grammar.

3. LAEME: A solution to the PPCHE data gap

Since long before the advent of modern corpus linguistics, many different types of corpora have been compiled to feed and support all kinds of historical linguistic inquiry. Many of these corpora are tailored to the requirements of specific research communities. As a result, an efficient alternative to creating a corpus from scratch is often to adapt an existing corpus for other purposes. This approach has been pursued successfully in many cases: the PPCHE initially adapted the unparsed Helsinki Corpus of English Texts (Rissanen et al. 1991) by adding syntactic annotation, while the Penn Discourse Treebank (Prasad et al. 2008) consists of stand-off annotation of discourse relations designed to supplement the Penn Treebank (Marcus, Santorini, Marcinkiewicz, and Taylor 1999).

In a similar vein, the PLAEME project consists of adapting an existing corpus, namely *A Linguistic Atlas of Early Middle English* (LAEME), to remedy the PPCHE's data gap. First published online in 2008, LAEME was compiled from 167 samples of Early Middle English amounting to some 650,000 words in all. As a corpus, it contains all the necessary ingredients for adaptation to PPCHE format.

First, all of the LAEME materials were written between c.1150 and c.1325, with a sizeable portion falling within the period for which PPCHE materials are lacking. The samples—a mixture of official records, prose, poetry, and lyrics—are admittedly more diverse than is usually found in syntactic corpora. Moreover, they are weighted significantly towards verse texts, a genre dispreferred by syntacticians since authors may manipulate word order for reasons related to metre and rhyme.[3] Nonetheless, verse texts can help address many types of question relating to syntactic variation and change. It is possible to assess the magnitude of the effect of verse on the syntactic phenomenon of interest, and we sketch one method for

[2] In addition to these recent studies, overviews of these phenomena are included in standard reference works on Middle English syntax, including Mustanoja (1960), Visser (1963–73), and Fischer (1992). For background on Early Middle English syntax, see Moessner (1989).
[3] The PPCHE contain very few verse texts, but do include the *Ormulum*.

addressing this issue in the conclusion. Moreover, in the case of English c.1300, there is little choice, as there are not enough surviving prose texts.[4]

Second, most of LAEME's samples are rich in metadata: 120 (72%) are localized to a particular county—many to a particular town[5]—and most are dated to within 10–25 years of their likely production date. Third, all of the samples were diplomatically transcribed from original manuscripts rather than printed editions, thus minimizing the number of layers between analyst and primary evidence.

One further, enormous advantage of LAEME lies in its system of tagging. Although LAEME was constructed primarily for the benefit of historical dialectologists, its compiler had an eye on the possibility of future parsing from the outset and annotated the samples accordingly. As a result, LAEME tags have a number of features that make this corpus eminently suitable for automatic parsing.

Each tagged object in LAEME has the structure `$lexel/grammel_FORM`, with the lexel element identifying the word lexically (in much the same way as a lemma but often with additional information about word sense),[6] and the grammel identifying it grammatically. Grammels minimally specify word category but typically add morphological information such as number, person, tense, and mood. The tag is attached to a form which represents the unit of analysis, i.e. the word or morpheme (inflectional and derivational morphemes are tagged separately in LAEME to facilitate the study of their histories). Words and morphemes are transcribed using upper case for plain text manuscript letters; capitalization is indicated by a leading asterisk, e.g. manuscript *man* and *Man* ('man') are transcribed `MAN` and `*MAN` respectively. Lower-case letters are reserved for special characters: `y` for thorn ('þ'); `d` for edh ('ð'); `z` for yogh ('ȝ'); `g` for insular 'g' ('ᵹ'); `m` and `n` for abbreviated forms of these letters, etc. Word-internal morpheme boundaries are generally indicated by `+`.

Our ability to parse LAEME efficiently and accurately depended primarily on LAEME's grammels. A basic requirement for a syntactically parsed corpus is, of course, part-of-speech tagging. However, although many corpora incorporate this feature, LAEME's grammels are an order of magnitude more informative than usual. This can be seen by comparing the number of distinctions made: the PPCHE POS tagset contains 92 distinct tags, while LAEME (even disregarding distinct tags for affixes) uses over 2,000 distinct grammels. This is largely because the grammels indicate a word's grammatical function as well as basic part-of-speech information.

This makes it possible to reconstruct automatically a great deal of information about constituency and other syntactic relations implied by LAEMEs grammels, and to automatically annotate the texts with that information in the PPCHE format. For example, the

[4] Perhaps for this reason, no attempt is made in PPCHE to eliminate certain other potential sources of noise. For example, the *Kentish Sermons* (Laud Misc. 471), the only late 13th-century text in the PPCHE, is a collection of five sermons translated from French. According to Hall (1963: 669), the English translations are markedly literal and show French interference in their word order and idiom.

[5] The unlocalized samples exhibit regionally colourless types of language or are not consistent with a single variety.

[6] LAEME only provides lexels for content words, not function words like `yE` in (2b) below.

LAEME representation of (2a), given in (2b), can be identified as a noun phrase on the basis of constituent adjacency and information contained in the constituents' grammels: `TN` identifies `yE` as a determiner; `aj` identifies `IUELE` as an adjective; and `n` identifies `MAN` as a singular noun.

(2) a. *þe iuele man* 'the evil man'
b.
```
$/TN_yE
$evil/aj_IUELE
$man/n_MAN
```

LAEME's grammels also identify grammatical function. In (3b), for example, the `Od` element identifies `yE RIzTE LAWE` as a direct object NP. (Indirect objects are identified by `Oi`, and the absence of any functional element in the grammels in (2b) identifies it as a subject NP.)

(3) a. *þe riȝte lawe* 'the right law'
b.
```
$/TOd_yE
$right/ajOd_RIzTE
$law/nOd_LAWE
```

In (4b), `NER yE SE` can be recognized as a prepositional phrase on the basis of `pr` in its constituents' grammels: `pr` alone identifies `NER` as the preposition, `<pr` identifies `yE` and `SE` as its dependents, and `<` indicates the relative position of the preposition to its dependents.

(4) a. *ner þe se* 'near the sea'
b.
```
$near/pr_NER
$/T<pr_yE
$sea/n<pr_SE
```

LAEME's grammels additionally project certain nonlocal dependencies. In (5b), for example, `HEOM` is tagged as a 3pl personal pronoun (`P23`) governed by a following preposition (`>pr`), and `TO` is tagged as a preposition (`pr`) that governs an object in a marked position, i.e. to its left (`<`).

(5) a. *& heom co(m) to þe halȝa gast* 'and the holy ghost came to them'
b.
```
$&/cj_&
$/P23>pr_HEOM
$come/vSpt13_COm
$to/pr<_TO
$/TN_yE
$holy/aj_HAL+gA $-ig/xs-aj_+gA
$ghost/n_GAST
```

The discontinuous PP in (5) can be reconstructed by matching the elements tagged `pr<` (the preposition) and `>pr` (its object).

In example (6) `RTIOd` identifies `dAT` as a relative pronoun (`RT`) with inanimate reference (indicated by `I`) in direct object function (`Od`), but there is nothing to indicate that it is an argument of `BRING+EN`. We have discovered, however, in the course of constructing PLAEME (see next section) that almost every relativized object is an argument of the next main verb encountered in the narrative. On that basis we have automatically annotated these dependencies with a high degree of accuracy.

(6) a. *ðat ghe ne migte hi(m) bringen on* '… that she may not prove against him'
b.
```
$/RTIOd_dAT
$/P13NF_GHE
$/neg-v_NE
$may/vpt13_MIGTE
$/P13>prM_HIm
$bring/vi_BRING+EN $/vi_+EN
$on{p}/pr<{rh}_ON
```

Despite the large amount of syntactic information contained in LAEME's grammels, some useful syntactic distinctions are not represented. The `cj` grammel, for example, is used for both subordinating and coordinating conjunctions, in contrast to the standard syntactic treatment of the former as complementizers. On the whole, however, the grammels' rich morphosyntactic information makes this particular corpus of tagged texts a near-perfect quarry for constructing a new syntactically parsed corpus of Early Middle English.

4. The construction of PLAEME

4.1. Text selection

We intend PLAEME as a bridge between the bodies of research carried out by historical dialectologists (using LAEME) and diachronic syntacticians (using the PPCHE). Historical dialectologists can use PLAEME to investigate syntactic variation and change, in addition to the lexical, phonological, morphological, and orthographic variation that has typically occupied users of LAEME (e.g. Laing and Lass 2009, Gardner 2011, Studer-Joho 2014, Alcorn 2015, Lass and Laing 2016, Stenbrenden 2016). Diachronic syntacticians can use PLAEME to fill the data gap in existing parsed corpora, and can begin to investigate geographical variation, in addition to variation over time.

Parsing all of LAEME would allow us to include a wide range of Early Middle English texts, including multiple parallel versions of a number of individual texts for focused analysis of dialectal and idiolectal variation. Unfortunately, we are presently unable to parse the whole of LAEME for lack of resources. We have therefore initially chosen to maximize the diversity of the texts included in PLAEME, and we hope in future to be able to expand PLAEME to include parallel versions of texts.

PLAEME currently consists of 69 texts (189,713 words). This is an exhaustive sample of LAEME texts which meet the following criteria:

1. The text is from 1250–1325;
2. no parsed version of the text currently exists;
3. the text is longer than 100 words.

Where multiple versions of the same text meet those criteria, we have chosen a single version with a view to balancing material across dialects as far as possible; all else being equal we parsed the longest version.

LAEME files do not correspond to conventional texts, but to samples of individual text languages, in the sense of Laing (2010: 237, fn.1). Accordingly, there is a many-to-many relationship between LAEME files and conventional texts. Because of this, we have split LAEME files into their component texts and only included those texts in PLAEME which

meet the above criteria. For instance, the LAEME file digby86mapt contains 19 verse texts from the manuscript Digby 86 that are in unmixed South West Midlands language. Several of these verse texts (for instance, *The Fox and the Wolf*) have already been parsed in at least one version (in this case, in the Parsed Corpus of Middle English Poetry). We have not reparsed those texts, but have split digby86mapt into 19 files corresponding to those 19 texts, and included those (e.g. digby86bede, the *Sayings of Bede*) of 100 words or more which have not previously been parsed. Again, this makes PLAEME more accessible to researchers familiar with the PPCHE, where files correspond to texts rather than text languages. The documentation for PLAEME details the correspondences between PLAEME files and LAEME files.

Figure 2 shows how material from PLAEME fills the PPCHE data gap. The combination of material from the PPCHE, the York–Toronto–Helsinki Parsed Corpus of Old English Prose (YCOE, Taylor et al. 2003), and PLAEME now offers unbroken coverage of the history of English, from the start of the written record to 1914.

***** Insert Figure 2 around here *****

The most common reason why texts included in PLAEME were omitted from the PPCHE is because PLAEME is composed of verse texts, with the exception of a few small legal documents. We have included three sample analyses in Section 5 to show the prospects for use of PLAEME in syntactic analysis and to indicate that use of verse texts does not produce outlandish results.

4.2. Annotation

We used an automatic annotation process to project a first-pass representation of syntactic structure based as far as possible on the information contained in LAEME's grammels (see Section 3), supplemented by inferences based on basic English grammar (for example, sentences do not have two subjects; or if a sentence has a direct object but no subject, the subject is probably either null or displaced).

The automatic annotation process, which uses the corpus revision function of CorpusSearch (Randall 2000/2013), inserts labelled brackets around likely constituents, as well as many empty categories. For example, the information in (4) tells us that *ner þe se* is a prepositional phrase, but also indirectly implies that *þe se* is a noun phrase, because *þe* and *se* are tagged as determiner and noun respectively, both dependent on the immediately preceding preposition. Accordingly, we automatically project the following labelled bracket representation of (4).[7]

```
(7)   (PP (P ner-near)
          (NP (D +te-the)
              (N se-sea)))
```

Similarly, we automatically project (8a) on the basis of (6): the `*T*` ('trace') indicates that the relative clause appears to contain a direct object gap, while `*ICH*` ('insert constituent here') indicates that the NP *him* is a nonlocal complement of *on*. During manual correction of

[7] In addition to projecting the syntactic structure implicit in (4), (7) shows the LAEME transcription converted to PPCHE norms (where `+t`, rather than `y`, represents thorn, for example), the preservation of LAEME lexels as lemmata, and the postulation of lemmata for function words like *the*, which were not annotated with lexels in LAEME.

this automatically generated parse, indices are added to indicate that `(WNP 0)` is associated with the direct object trace, `(NP-OB1 *T*)`, and the pronoun *him* with the complement of *on*, `(NP *ICH*)`. Moreover, the context surrounding this example, in the Middle English *Genesis and Exodus*, makes it clear that this is actually a free relative: *And seið ioseph hire pulde don ðat ghe ne migte hi(m) bringen on* 'And (she) says that Joseph would do to her what she could not prove against him'. The LAEME annotation does not systematically distinguish between headed and free relatives, so the distinction has to be made manually in this, and most other, cases. This manual editing leads to the final representation in (8b).[8]

(8) a.
```
(CP-REL (WNP 0)
        (C +dat-that)
        (IP-SUB (NP-OB1 *T*)
                (NP-SBJ (PRO ghe-she))
                (NEG ne-ne)
                (MD migte-may)
                (NP him-him)
                (VB bring+en-bring)
                (PP (P-RH on-on)
                    (NP *ICH*))))
```

b.
```
(CP-FRL (WNP-1 0)
        (C +dat-that)
        (IP-SUB (NP-OB1 *T*-1)
                (NP-SBJ (PRO ghe-she))
                (NEG ne-ne)
                (MD migte-may)
                (NP-2 him-him)
                (VB bring+en-bring)
                (PP (P-RH on-on)
                    (NP *ICH*-2))))
```

Almost all of the information in (7) and (8a) is implicit in the LAEME tags, but the automatic conversion process makes that information visible in a format which can be queried by standard software like CorpusSearch, and more readily assimilated by syntacticians used to thinking in terms of constituent structure.

This initial annotation process is error-prone. Some errors can be corrected automatically with adjustment rules (for example, LAEME tags both *that* and *and* as conjunctions, as mentioned in Section 3, so we automatically retag *that* as a complementizer, while leaving *and* as a conjunction), but a substantial residue of errors remains, in part because of the lack of explicit indication of sentence boundaries in early Middle English manuscript material. We have therefore hand-corrected all automatically annotated texts using Annotald (Beck, Ecay, and Ingason 2011), bespoke software for rapid and accurate correction of PPCHE-format corpora. The virtues of this two-stage process are a combination of speed and accuracy: correction of automatically annotated text is both faster and more accurate than manual annotation typically is.

[8] (8) also shows the `-RH` dash tag, an addition to the PPCHE format. This tag marks rhymes, as indicated by `{rh}` in LAEME's grammels.

5. Case studies

As mentioned earlier, the Early Middle English period saw rapid syntactic change, and in many cases the quantitative investigation of those changes has been severely hampered by the paucity of parsed material from 1250–1350, which we will refer to below as the 'PLAEME window'. In many respects, despite dramatic changes in inflectional morphology and basic word order, the morphosyntax of the *Ormulum* or *Ancrene Wisse* hadn't moved far beyond Old English. By the mid-to-late 14th century, the time of the *Ayenbite of Inwyt* or Chaucer, a recognizably modern syntax had emerged. How this happened in the space of a few generations is quite unclear. We hope that PLAEME can help to cast light on this period of change.

We illustrate the use of PLAEME in relation to an opportunistic selection of three studies of Middle English syntax recently published in Mathieu and Truswell (2017). In none of these cases do we intend to delve too deeply into our findings, or engage with the broader theoretical points made by the authors; rather, we aim to show that, in these PPCHE-based studies, much of the action happened while our back was turned, so to speak. In many cases, investigation of the early stages of a grammatical change has revealed unseen details (see, for instance, Ecay 2015 on properties of affirmative *do* in the early stages of the rise of *do*-support). We hope that investigating the early stages of these rapid changes using PLAEME will be similarly revealing.

5.1. The expression of negation

In Old English, sentential negation was typically expressed by the preverbal particle *ne* (9). Since c.1450, postverbal *not* has been universally used instead (10). Middle English is the transitional period between these two systems.

(9) … *ac hie ne dorston þær on cuman*
but they not dared there in come
'… But they did not dare enter there' (Traugott 2008: 267)

(10) *It is not neccessari to declare what it was* (Capgrave, *Chronicle*, a1464)

The details of the transition have been debated in a recent string of papers (Frisch 1997, Wallage 2008, Ecay and Tamminga 2017). Much of the debate concerns the status of a hybrid system, in which both preverbal *ne* and postverbal *not* appear, as in (11).

(11) *he ne shall nouȝt deceive him*
(*Earliest Prose Psalter*, c.1350; Frisch 1997, cited in Ecay and Tamminga 2017: 206)

Ecay and Tamminga give a graphical representation of the diachrony of these three competing variants; Figure 3 is based on their Figure 13.2, p.207.[9]

[9] Figure 3 is not identical to Ecay and Tamminga's figure in one salient respect: they have a very large point representing a large amount of text from the year 1300. Inspection of their corpus queries (freely available at https://github.com/aecay/digs15-negative-priming) indicates that this is a result of an error in their query, which results in several 14th-century texts being dated at precisely 1300, and several 15th-century texts at precisely 1400. Correcting this error also leads to markedly different regression curves for the diachronic trajectories of the different variants, compared to Ecay and Tamminga's figure. We hasten to

***** Insert Figure 3 around here *****

Figure 3 shows that PPCHE data indicates a stable Early Middle English system that held until c.1250, with *ne* as the majority variant, c.25% of tokens using *ne ... not*, and near-complete absence of *not*. By 1350, *ne* had largely vanished, with most texts using simple *ne* less than 10% of the time. In the late 14th century, both *ne ... not* and *not* are found in significant proportions, with *not* approaching 100% usage by 1400.

What happened between 1250 and 1350 is not clear from the PPCHE data. The only substantial text in this period is the *Ayenbite of Inwyt* (1340), which looks largely similar to texts from 100 years earlier in terms of negation, with majority *ne* use, some use of *ne ... not*, and no simple *not*.[10] However, the *Earliest Prose Psalter*, just ten years later, favours *ne ... not* and *not*, and almost never uses simple *ne*. Lowess regression curves based on PPCHE data alone (represented with dashed lines in Figure 3) suggest that the frequency of *ne* declines from c.1275, following an S-shaped trajectory as it loses ground initially to *ne ... not*, which follows a 'failed change' trajectory peaking around 1300, and then to *not*, which increases steeply in frequency after 1325.

Because there is almost no PPCHE data in the PLAEME window, the dashed regression lines in Figure 3 during that period are based on interpolation between pre-1250 and post-1340 data, plus the minimal amount of data gathered from the 3,534-word *Kentish Sermons*, c.1275. The PLAEME data suggests several refinements to this picture, as represented by the solid lines in Figure 3.[11] The trajectory for simplex *ne* is almost unchanged by the addition of the PLAEME data, but the 'failed change' trajectory for *ne ... not* becomes much less apparent, as the PLAEME data suggests a more or less stable rate of c.25% *ne ... not* use throughout the 13th and early 14th centuries, followed by a decline in tandem with simplex *ne* in the late 14th century. Meanwhile, the emergence of simplex *not* is taken to be more gradual (the solid regression line for *not* moves above 0 about a generation earlier once PLAEME data is added to the analysis). Although *not* remains a marginal variant throughout the PLAEME window, there are some examples as early as the late 13th century. (12) is one of the first.

(12) *Suyc richesse p(re)yse ic nout* ('I do not praise such riches')
(tr323bt, Homily for the anniversary of St Nicholas, c13b1)

All of these adjustments to the diachronic picture that emerges can be seen as consequences of a reduced emphasis on the *Ayenbite* and *Earliest Prose Psalter*. Both of these are outliers in certain ways: the *Ayenbite* is the last major text with a majority of *ne* and almost no *not*,

---

add that this parochial error does not affect any of Ecay and Tamminga's theoretical arguments.

[10] As Laing (2013–) observes, the *Ayenbite* was written by a 70-year old scribe and so its language may be taken as representative of the late 13th rather than the mid 14th century.

[11] See also Laing (2002) for an examination of the distribution of different forms of negation on the basis of an early version of LAEME. Laing's research is partly a response to research by Jack (1978a,b) describing constraints on the syntactic contexts in which *ne ... not* occurs. Jack's work, on the basis of Middle English prose texts, informed Wallage's response to Frisch and was extended by Iyeiri (1992), who examined verse texts including several that feature in LAEME. Thanks to Meg Laing for pointing us to this body of literature.

while the *Earliest Prose Psalter* is the only major text to use >50% *ne ... not*. Adding the PLAEME data reduces the influence of these two texts over the regression lines.

5.2. Case and word order in ditransitives

Bacovcin (2017) describes a complex series of changes in the syntax of recipient–theme ditransitives like *give*, leading ultimately to the emergence of the Present-Day English grammar, which allows an alternation between *give* $NP_{recipient}$ NP_{theme}, and *give* NP_{theme} *to* $NP_{recipient}$. Bacovcin is particularly interested in the diachrony of a 'failed change', in which a third variant, *give to* $NP_{recipient}$ NP_{theme}, as in (13), initially gains ground before gradually disappearing over the course of Middle and Early Modern English.

(13) *he ne shal nouȝt ȝeven to God his quemeyng* 'he shall not give God his appeasement'
(*Earliest Prose Psalter,* c.1350)

Two approaches to the modelling of failed changes, and their relationship to successful changes, have recently been proposed. Postma (2010, 2017) models the U-shaped diachronic trajectory of a failed change as the first derivative of the S-shaped trajectory of a successful change, implying that each failed change is directly linked to a successful change. Bacovcin proposes instead, based on the diachrony of the order in (13), that failed changes result from interactions among multiple successful changes, in this case a first change introducing *to* as a marker of dative case on recipients in both word orders, and a second change introducing the modern split system, in which *to* only surfaces in theme–recipient orders. The first change feeds the rise of examples like (13); the second change bleeds it.[12]

Bacovcin's Probability Multiplication Model predicts a trajectory for examples like (13) as shown in Figure 4 (based on Bacovin's Figure 7.2, p.98), where the proportion of V $NP_{recipient}$ NP_{theme} orders that contain *to* before the recipient rises throughout Middle English to a peak at approximately 1350.[13]

***** Insert Figure 4 around here *****

Unfortunately, data concerning the relevant variants in Early Middle English is very scarce: Bacovcin's analysis of the failed change applies specifically to ditransitives with transfer-of-possession meanings (canonically *give*), where the recipient precedes the theme, both arguments are realized as full NPs, and neither argument is topicalized. Although this strict specification is empirically justified, it means that a very large amount of data is discarded, and only a handful of examples retained. Bacovcin has only four such examples that he dates to within the PLAEME window (1 with *to*) and 51 from the century prior to the PLAEME

[12] The hybrid *ne ... not* negator described in Section 5.1 is also a failed change, as an Early Middle English innovation which disappeared in the 15th century. To our knowledge it has not yet been studied from the modelling perspective of Postma or Bacovcin. In principle, PLAEME data could also be used to ground a model of the interactions between the failed change in *ne ... not*, and the successful replacement of *ne* by *not*.

[13] Bacovcin's original scripts and datasets for the analysis of Middle English ditransitives are available as online supplementary materials accompanying Mathieu and Truswell (2017), at http://www.oup.co.uk/companion/DiGS15. Although Bacovcin's data covers both recipient–theme and theme–recipient orders, here we only show data concerning recipient–theme orders, as the spread of *to* in theme–recipient orders was clearly well on its way to completion by the start of the PLAEME window.

window (9 with *to*). Eight of these examples with *to* occur in the same text, namely the *Brut Chronicle*, normally dated to the end of the 14th century but dated in Bacovcin's data at 1215. Disregarding this possibly misleading data point, the PPCHE data amounts to 40 tokens pre-1250 (1 with *to*) and 4 tokens during the PLAEME window (1 with *to*). After the PLAEME window, data is plentiful, supporting the gradual decline of the failed change, but the abruptness of the rise of V *to* $NP_{recipient}$ NP_{theme} is very much open to question.

Because of the scarcity of data, Figure 4 shows the addition of data from both PLAEME and the Parsed Corpus of Middle English Poetry (PCMEP, Zimmermann 2014–), which contains 38 poems (107,299 words) from c.1150–1420. This adds 128 further data points (16 from PCMEP, of which 9 are within the PLAEME window, and 112 from PLAEME, although 64 of the these are from a single file, namely buryFft, the English texts from the Sacrist's Register of Bury St Edmunds, dated to c.1300). In Figure 4, we have also adjusted the date of the *Brut Chronicle* to 1400.

The effect of adding these extra data points can be seen by comparing the dashed regression line (PPCHE data only) with the solid line (data from PPCHE, PCMEP, and PLAEME). The revised estimates for the diachrony of V *to* $NP_{recipient}$ NP_{theme} show a more gradual increase in frequency, and a lower and later peak frequency. We have not investigated the implications of these findings for Bacovcin's theoretical claims about the modelling of failed changes. In particular, the asymmetry of Bacovcin's curve (based on only PPCHE data) is crucial to his argument for the Probability Multiplication Model. The revised curve including PLAEME and PCMEP data appears visually to be more symmetrical, and so to offer less strong support for Bacovcin's model, but we have not attempted any quantitative model comparison.

A reviewer comments that the choice of ditransitive construction in PLAEME texts could be affected by factors such as weight, animacy, information structure, metre, and rhyme. Bacovcin investigated the first three of these for his data, but we have made no effort to do so in this brief report. In principle, though, the method described in Section 6 below could be used to investigate whether there is any systematic effect of metre and rhyme on choice of ditransitive construction.[14]

5.3. Argumental and adverbial wh-*relatives*

Truswell and Gisborne (2017) describe the introduction of headed *wh*-relatives in Middle English, and the division of labour between interrogative and demonstrative forms as relativizers. One of their key claims is that there is no straightforward replacement of demonstrative with interrogative forms. In other words, it is not the case that interrogative forms are co-opted into the system of relative pronouns as direct replacements for the demonstrative forms that were radically levelled at the start of the Middle English period. Rather, interrogative *where* and demonstrative *there* coexisted as relativizers throughout Middle English, while there was a 100-year gap between the disappearance of the *se*-series of argumental demonstrative relatives and the first uses of argumental *which* (and later *whom* and *who*) as relativizers. That 100-year gap is the period labelled as M2 in the PPCHE, spanning 1250–1350, almost perfectly coextensive with the PLAEME window. Of course, given the scarcity of data from that period, any claims of absolute absence are crying out for re-evaluation in the light of further data.

[14] Note that most relevant examples in PLAEME come from wills and other legal documents, rather than verse texts, particularly buryFft.

Further analysis of PPCHE data concerning *where* and *there* as relativizers, beyond that carried out by Truswell and Gisborne, suggests that little of interest happened during the PLAEME window. Although the two forms did indeed coexist for centuries, they were functionally specialized: *where* was an R-pronoun in the sense of van Riemsdijk (1978): it could precede a preposition in relativizers such as *whereby* or *wherethrough* (14), but there are no instances of *where* as a strictly locative relativizer in the PPCHE until the *Ayenbite* in 1340. *There* remains the locative relativizer throughout Early Middle English, including the PLAEME window, in examples such as (15), and is rapidly replaced by *where* in the late 14th century. Moreover, although *there* clearly is used as an R-pronoun in Old and Early Middle English, it is apparently not an R-pronoun when used as a relativizer. That is, there are no examples like (14) with *there* replacing *where.*

(14) *For þe eareste Pilunge, hwer of al þis uuel is, nis buten of pride* 'For the first peeling, where all this evil comes from, is just of pride' (*Ancrene Riwle*, c13b)

(15) *bi hald inwart þer ich am & ne seh þu me naut wið uten þin heorte* 'Look inward, where I am, and you don't see me without your heart' (*Ancrene Riwle*, c13b)

This refines the empirical picture presented by Truswell and Gisborne, but supports the basic claim of largely stable coexistence of *where* and *there* as relativizers throughout Middle English, including the PLAEME window. PLAEME data indeed supports this claim: *where* is used in relative clauses only as an R-pronoun, and *there* as a locative relative.

Of more interest is the diachrony of argument-gap *wh*-relatives in PLAEME. Figure 5 graphs the percentage of argument-gap relatives with a *wh*-relativizer in Middle English.

***** Insert Figure 5 around here *****

The PPCHE indicate that complementizer *that* is by far the most common relativizer prior to 1250, despite sporadic use of *wh*-forms. After the PLAEME window, the *Ayenbite* (1340) shows almost no argument-gap *wh*-relatives (less than 1% of the 1,117 argument-gap relatives use *which*, and other *wh*-relativizers are not used at all for argument-gap relatives). Around 10 years later, the *Earliest Prose Psalter* uses *which* in 12.3% of the 413 headed relatives with an NP argument gap, and argument-gap *wh*-relatives are clearly beginning to be established by the late 14th century, for instance in Chaucer and the Wycliffite Bible.

As for the PLAEME window, most texts behave in line with Truswell and Gisborne's claims, and categorically use *that* (or zero) for argument-gap relatives. However, there are some occasional examples of *which* or *who*, such as (16), in line with the sporadic earlier argument-gap *wh*-relatives noted above.

(16) *al erue and prim and pilde der. // Qpel man mai sen on perlde her.* 'all cattle and serpents and wild animals which one may see in the world here' (genexodt, *Genesis and Exodus*, c14a1)

This may suggest that argument-gap *wh*-relatives were available at a very low frequency for some time prior to their late 14th-century diffusion. Alternatively, the emergence of argument-gap *wh*-relatives may reflect repeated sporadic innovations throughout Middle English, that only began to diffuse in the second half of the 14th century.

6. Conclusion

The Parsed Linguistic Atlas of Early Middle English is a fairly small parsed corpus, but it can be of significant use as a supplement to the Penn Parsed Corpora of Historical English in tracking the multiple overlapping rapid changes that take place in the sparsely documented period 1250–1325. The case studies in Section 5 show how PLAEME can inform our understanding of changes in the morphosyntax of negation, ditransitives, and headed relative clauses. We expect that PLAEME will be of similar use in investigating other grammatical changes in Middle English.

In many respects discussed in the case studies in Section 5, PLAEME data is conservative, in the sense that syntactic variables observed in texts from 1250–1325 have more in common with texts from before 1250 than with texts from after 1325. This suggests that the grammatical changes in question were often more abrupt than previously thought. Alternatively, it is possible that the results we have presented reflect some systematic difference between PLAEME and PPCHE texts, such as a difference between verse and prose.

A possible method for investigating the influence of verse on the syntax of a text may be to use the Parsed Corpus of Middle English Poetry (PCMEP). PCMEP overlaps temporally with PLAEME to an extent, but also overlaps significantly with the prose texts of the PPCHE. By comparing results from PCMEP and the PPCHE, it is possible to assess whether verse texts behave differently from prose texts with respect to the variable of interest. If they do not, then properties of verse texts in PLAEME are probably also not attributable to an effect of verse.

PLAEME currently covers approximately one third of the material included in the unparsed Linguistic Atlas of Early Middle English. We hope in future to expand PLAEME to cover the period 1150–1250, and parallel versions of a single text, to open up these Penn-format resources to canonical historical dialectology methods. Also for the future is the possibility of parsing the Linguistic Atlas of Older Scots (Williamson 2008–), which is in a similar format to LAEME, to facilitate the quantitative investigation of the much-neglected diachrony of Scots syntax.

References

Alcorn, Rhona (2015). Pronoun innovation in Middle English. *Folia Linguistica Historica* 36: 1–17.

Bacovcin, Hezekiah Akiva (2017). Modelling interactions between morphosyntactic changes. In Mathieu and Truswell (2017), pp.94–103.

Beck, Jana, Aaron Ecay, and Anton Karl Ingason (2011). Annotald, version 1.3.8. https://annotald.github.io/

Ecay, Aaron (2015). *A Multi-step Analysis of the Evolution of English Do-support*. PhD thesis, University of Pennsylvania.

Ecay, Aaron, and Meredith Tamminga (2017). Persistence as a diagnostic of grammatical status: The case of Middle English negation. In Mathieu and Truswell (2017), pp.202–215.

Fischer, Olga (1992). Syntax. In Norman Blake (ed.) *The Cambridge History of the English Language*, vol. II: 1066–1476, pp.207–408. Cambridge: Cambridge University Press.

Frisch, Stefan (1997). The change in negation in Middle English: A NEGP licensing account. *Lingua* 101: 21–64.

Galves, Charlotte, and Pablo Faria (2010). The Tycho Brahe Corpus of Historical Portuguese. Department of Linguistics, University of Campinas (UNICAMP). http://www.tycho.iel.unicamp.br/~tycho/

Gardner, Anne-Christine (2012). Word formation in Early Middle English: Abstract nouns in the *Linguistic Atlas of Early Middle English*. In Paul Rayson, Sebastian Hoffmann, and Geoffrey Leech (eds.) *Studies in Variation, Contacts, and Change in English 6: Methodological and Historical Dimensions of Corpus Linguistics*. www.helsinki.fi/varieng/journal/volumes/06/gardner.

Gisborne, Nikolas, and Robert Truswell (2017). Where do relative specifiers come from? In Mathieu and Truswell (2017), pp.25–42.

Hall, Joseph (1963). *Selections from Early Middle English 1130–1250. Part I*. 2nd edition. Oxford: Clarendon Press.

Iyeiri, Yoko (1992). *Negative Constructions in Selected Middle English Verse Texts*. PhD thesis, University of St Andrews.

Jack, George (1978a). Negative adverbs in early Middle English. *English Studies* 59: 295–309.

Jack, George (1978b). Negation in later Middle English prose. *Archivum Linguisticum* n.s. 9: 58–72.

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2004). Penn–Helsinki Parsed Corpus of Early Modern English, release 3. University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-3/index.html

Kroch, Anthony, Beatrice Santorini, and Ariel Diertani (2016). Penn Parsed Corpus of Modern British English, 2nd edition, release 1. University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/PPCMBE2-RELEASE-1/index.html

Kroch, Anthony, and Ann Taylor (2000). Penn–Helsinki Parsed Corpus of Middle English, 2nd edition, release 4. University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html

Laing, Margaret (2002). Corpus-provoked questions about negation in early Middle English. *Language Sciences* 24: 297–321.

Laing, Margaret (2010). The reflexes of OE *beon* as a marker of futurity in Early Middle English. In Ursula Lenker, Judtih Huber, and Robert Mailhammer (eds.) *English Historical Linguistics 2008, vol. 1: The History of English Verbal and Nominal Constructions*, pp.237–254. Amsterdam: John Benjamins.

Laing, Margaret (2013–). A Linguistic Atlas of Early Middle English, 1150–1325, version 3.2. University of Edinburgh. http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html

Laing, Margaret, and Roger Lass (2009). Shape-shifting, sound-change and the genesis of prodigal writing systems. *English Language and Linguistics* 13: 1–31.

Lass, Roger, and Margaret Laing (2016, in press). Q is for what, when, where? The 'q' spellings for OE hw-. *Folia Linguistica Historica* 37.

Light, Caitlin (2011). Parsed corpus of Early New High German. University of Pennsylvania. http://enhgcorpus.wikispaces.com/

Martineau, France, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin (2010). Modéliser le changement : les voies du français. Université d'Ottawa. http://www.arts.uottawa.ca/voies/voies_fr.html.

Mathieu, Eric, and Robert Truswell (2017). *Macro-change and Micro-change in Diachronic Syntax*. Oxford: Oxford University Press.

Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor (1999). Treebank-3 LDC99T42. Philadeplhia, PA: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC99T42

Moessner, Lilo (1989). *Early Middle English Syntax*. Tübingen: Niemeyer.

Mustanoja, Tauno (1960). *A Middle English syntax*, Part I. Helsinki: Société néophilologique.
Postma, Gertjan (2010). The impact of failed changes. In Anne Breitbarth, Christopher Lucas, Sheila Watts, and David Willis (eds.) *Continuity and Change in Grammar*, pp.269–302. Amsterdam: John Benjamins.
Postma, Gertjan (2017). Modelling transient states in language change. In Mathieu and Truswell (2017), pp.75–93.
Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Mitsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech.
Randall, Beth (2000/2013). CorpusSearch 2: A tool for linguistics research. http://corpussearch.sourceforge.net/CS.html
Riemsdijk, Henk van (1978). *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases*. Dordrecht: Foris.
Rissanen, Matti, Merja Kytö, Leena Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Terttu Nevalainen, and Helena Raumolin-Brunberg (1991). The Helsinki Corpus of English Texts. University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
Stenbrenden, Gjertrud Flermoen (2016) *Long-Vowel Shifts in English, c. 1050–1700: Evidence from Spelling*. Cambridge: Cambridge University Press.
Studer-Joho, Nicole (2014). *Diffusion and Change in Early Middle English: Methodological and Theoretical Implications from the LAEME Corpus of Tagged Texts*. Tübingen: Francke Verlag.
Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen (2006). York–Helsinki Parsed Corpus of Early English Correspondence. University of York and University of Helsinki. Distributed through the Oxford Text Archive. http://ota.ox.ac.uk/desc/2510
Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths (2003). The York–Toronto–Helsinki Parsed Corpus of Old English Prose. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm
Traugott, Elizabeth Closs (2008). Syntax. In Richard Hogg (ed.) *The Cambridge History of the English Language. Volume 1: Beginnings to 1066*, pp. 168–289. Cambridge: Cambridge University Press.
Visser, Fredericus Th. (1963–73). *An Historical Syntax of the English Language.* Leiden: Brill.
Walkden, George (2016). The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21: 559–571.
Wallage, Phillip (2008). Jespersen's Cycle in Middle English: Parametric variation and grammatical competition. *Lingua* 118: 643–674.
Wallenberg, Joel C., Anton K. Ingason, Einar F. Sigurðsson, and Eiríkur Rögnvaldsson (2011). Icelandic Parsed Historical Corpus (IcePaHC), version 0.9. http://www.linguist.is/icelandic_treebank
Williamson, Keith (2008–). A Linguistic Atlas of Older Scots, Phase 1: 1380–1500, version 1.2. http://www.lel.ed.ac.uk/ihd/laos1/laos1.html
Zimmermann, Richard (2014–). Parsed Corpus of Middle English Poetry. http://pcmep.net/

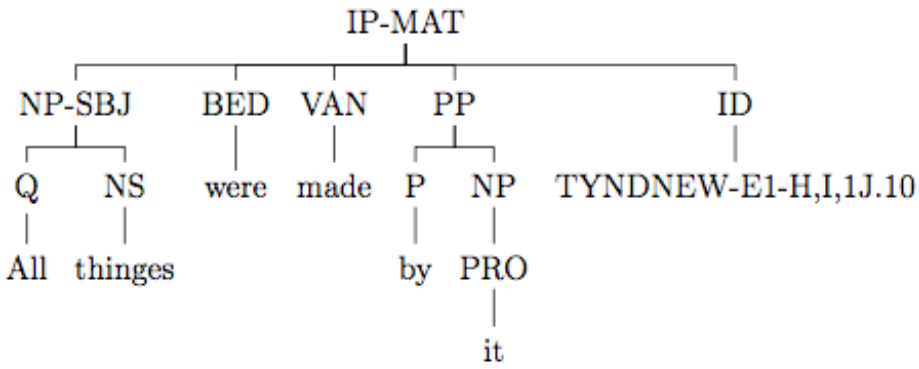

Figure 1: An example of a full sentence token in PPCHE format, represented as a tree diagram

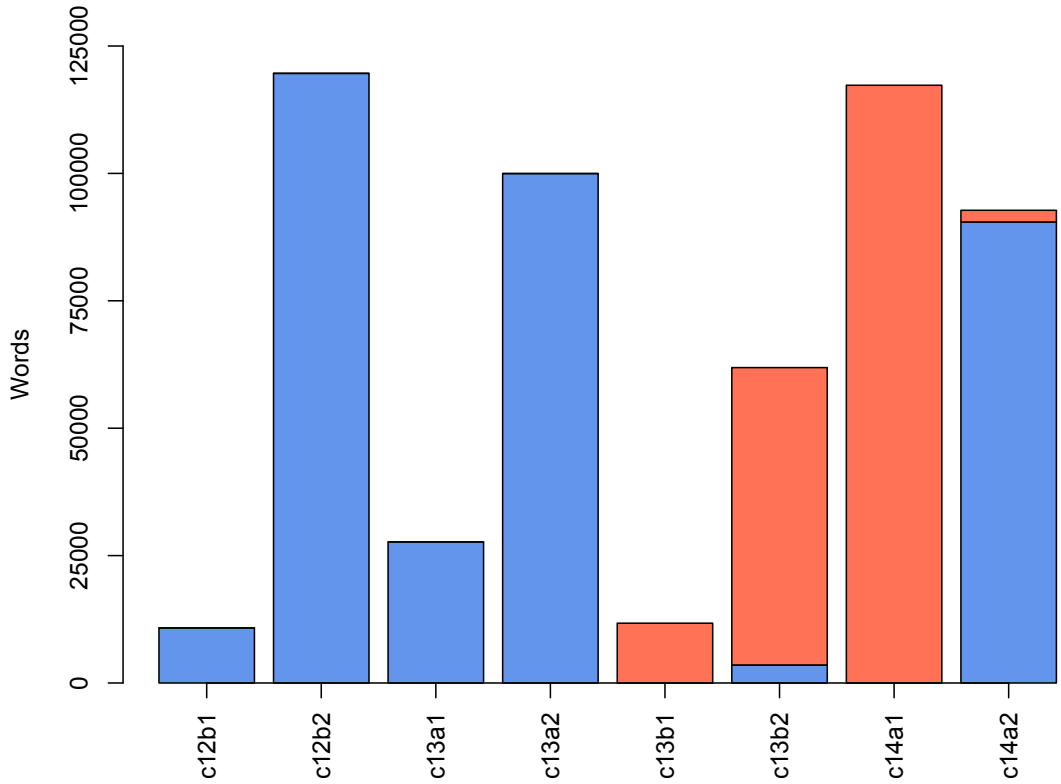

Figure 2: Material from the PPCHE (in blue) is complemented by material from PLAEME (in red).

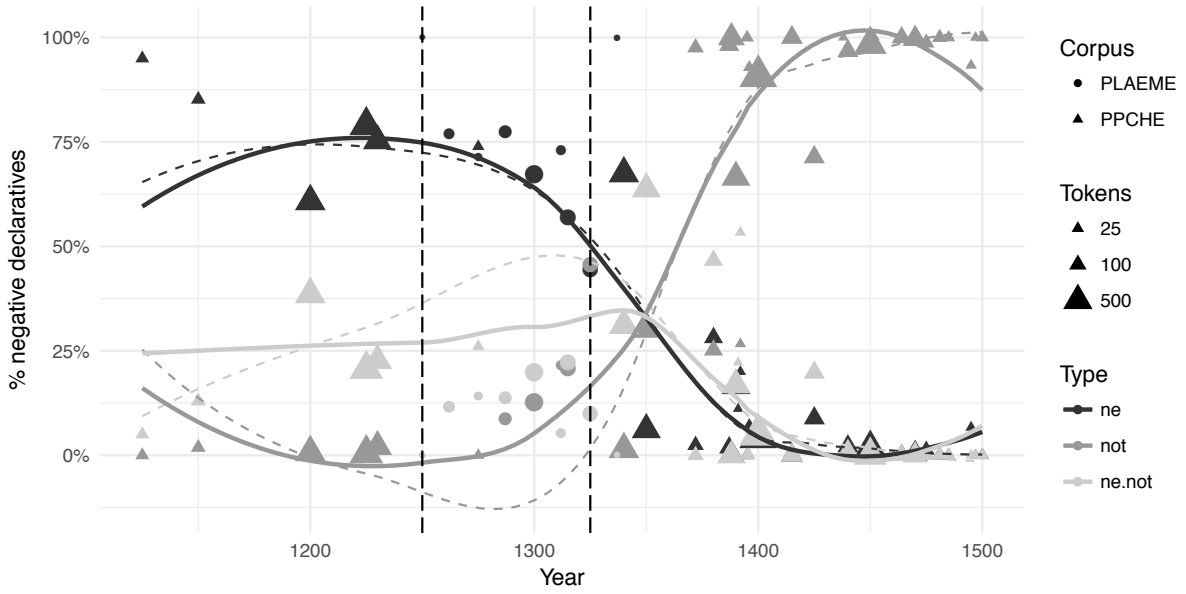

Figure 3: Competing expressions of sentential negation in Middle English. Vertical lines indicate the PLAEME window. The dashed lines represent lowess regression lines taking only PPCHE data into account. The solid lines are based on both PPCHE and PLAEME data.

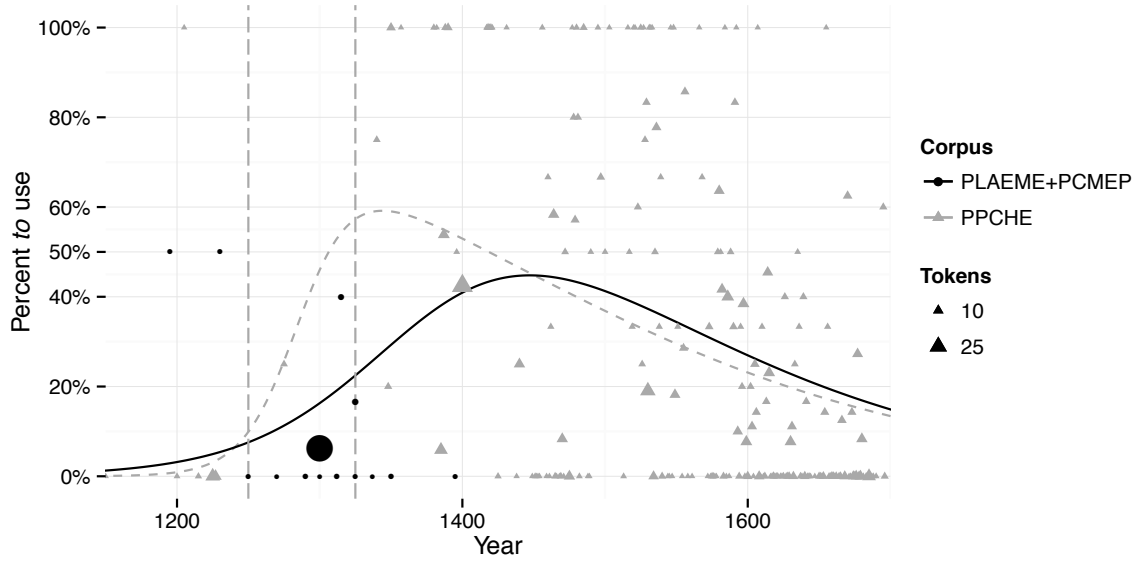

Figure 4: Competition between recipient–theme ditransitives with and without *to* in Middle and Early Modern English. The vertical lines indicate the PLAEME window. The dashed curve represents the fit from the Probability Multiplication Model adopted by Bacovcin (2017). The solid line represents the fit from the same model to PPCHE, PCMEP, and PLAEME data combined.

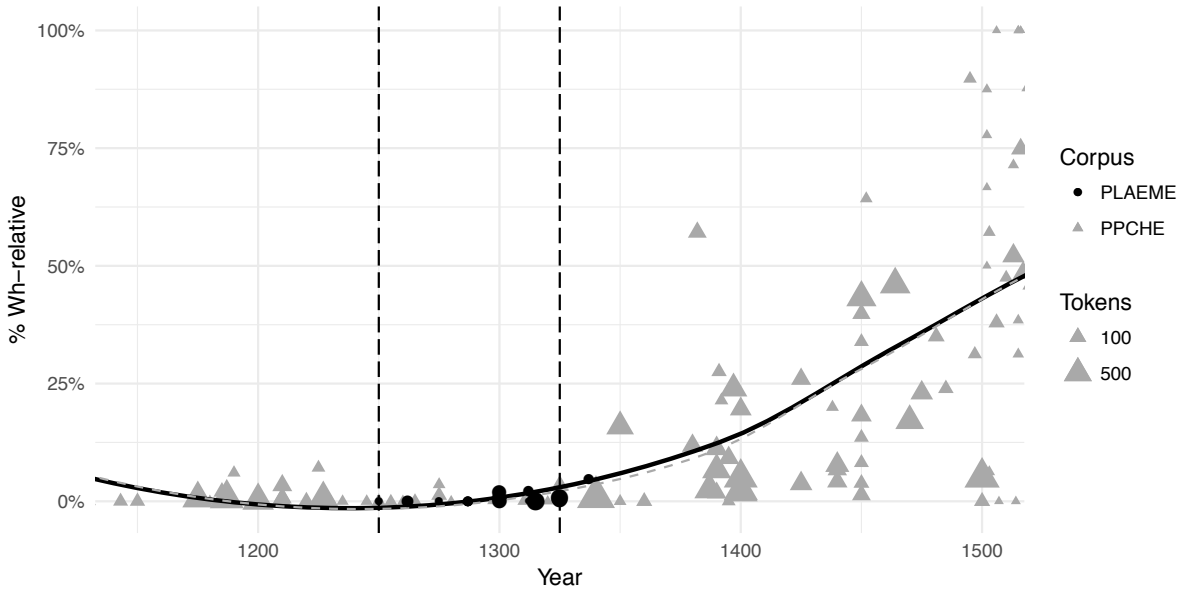

Figure 5: Use of *wh*-forms in argument-gap headed relative clauses in Middle and Early Modern English. Vertical lines indicate the PLAEME window. The dashed and solid lowess regression lines (PPCHE data only and PPCHE + PLAEME data combined) are virtually identical.