Honkapohja, Alpo; Suomela, Jukka

# Lexical and function words or language and text type? Abbreviation consistency in an aligned corpus of Latin and Middle English plague tracts

*Published in:*
Digital Scholarship in the Humanities

*DOI:*
10.1093/llc/fqab007

Published: 01/01/2022

*Document Version*
Publisher's PDF, also known as Version of record

# Lexical and function words or language and text type? Abbreviation consistency in an aligned corpus of Latin and Middle English plague tracts

**Alpo Honkapohja**
School of Philosophy, Psychology, and Language Sciences, and Angus McIntosh Centre for Historical Linguistics, University of Edinburgh, Edinburgh, UK

**Jukka Suomela**
Department of Computer Science, Aalto University, Espoo, Finland

## Abstract

This study examines the consistency of medieval abbreviation practices in a parallel corpus consisting of Latin and Middle English copies of a plague treatise attributed to John of Burgundy (JB). Focusing on different versions of the treatise enables us to maximize textual and lexical overlap, comparing differences caused by text type, word type, and language. We examine how the following variables affect the consistency of abbreviating across manuscript witnesses: A language: Latin versus English; B text type: recipes versus running text; C word type: lexical versus function words; and D: the number of characters in a word. Variables A–D are compared using a parallel corpus of automatically aligned rich TEI P5 XML-tagged transcriptions of six manuscript witnesses to the JB treatise. The alignment process is based on computer–human collaboration and custom-built alignment tool which uses sections tagged in the TEI XML file and word division. The results reveal that abbreviation was overwhelmingly more consistent in Latin than in the Middle English and somewhat more consistent in recipes. High token counts of frequent lexical items had a major effect on the results. Word length worked better than division into lexical and function words.

**Correspondence:**
Alpo Honkapohja, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK.
**E-mail:**
alpo.honkapohja@gmail.com, alpo.honkapohja@ed.ac.uk

## 1 Introduction

Abbreviation was an integral part of medieval manuscript culture. The necessity of having to make each copy of a text by hand created a need for a system for saving time and writing space. Medieval abbreviation practices reached intricate heights especially in the hands of the multiple generations of scribes who copied texts in Latin, the main literary language of Western Europe. The existence of such an elaborate system raises questions about the consistency of scribal behaviour regarding its use. In order to be useful, any abbreviation system needs to be able to encode information into a shorter form and to expand it


doi:10.1093/llc/fqab007

again. To what extent did scribes actually do this? Did they tend to copy abbreviated words as they stood in their exemplar or did they actively expand and abbreviate? Does their behaviour vary between Latin and vernacular English, lexical words and function words, or text types such as recipes? This study seeks to answer these questions by analysing how the scribal behaviour of abbreviating and expanding words varies line by line in an aligned parallel corpus of a textually very close group of texts.

We focus on Latin and Middle English versions of a single text: a plague treatise attributed to John of Burgundy (JB), the most popular such treatise extant in England and elsewhere in northern Europe before 1500. We study how five scribes treat abbreviation in different versions of the treatise, which survive in six closely related manuscripts, the so-called Voigts-Sloane Sibling Group. These manuscripts contain a medical anthology, which, in addition to other texts, includes no fewer than three versions of the JB tract. Focusing on different versions of the treatise enables us to maximize textual and lexical overlap, comparing differences caused by text type, word type, and language. We compare how the following variables affect the consistency of abbreviating across manuscript witnesses:

A language: Latin versus English,
B text type: recipes versus running text,
C word type: lexical versus function words, and
D: the number of characters in a word.

The article has the following structure. Section 2 gives some background information on the quantitative approaches to manuscript abbreviations and explains how the current study differs from Honkapohja (2018), which used a 'bag of words' approach. It also presents the data, the JB plague tract, and the Voigts-Sloane Sibling group of manuscripts, as well as the annotation scheme used. Section 3 describes the alignment process. We compare variables A–D using a parallel corpus of automatically aligned rich Text Encoding Initiative (TEI) P5 XML-tagged transcriptions of six manuscript witnesses to the JB treatise. The alignment process is based on computer–human collaboration and a custom-built alignment tool which uses sections tagged in the TEI XML file and word division. Section 4 elaborates on the variables, discussing them in the context of secondary literature. Section 5 presents and discusses the results. Section 6 presents the conclusions and elaborates further on their importance.

## 2 Background and Data

Manuscript abbreviations have been an important means for localizing and dating scribal hands, using both qualitative and, increasingly, quantitative methodologies. The importance of abbreviation to palaeography shows the number of available handbooks that contain a section on abbreviations (Hector, 1966 [1958]; Petti, 1977; Brown, 1990; Clemens and Graham, 2007) and specialized dictionaries dedicated exclusively to the subject (Chassant, 1970 [1845]; Cappelli, 1990 [1899]). Early quantitative approaches to Latin palaeography include Hälvä-Nyberg (1988), which focused on Greek and Latin epigraphy, and Bozzolo *et al.* (1990), which focused on Latin and French liturgical books from the fifteenth century. Quantitative digital approaches include Cottereau (2005), Thaisen (2011), Camps (2016), Shute (2017), and Smith (2018). These studies have uncovered a number of factors that can condition abbreviation, such as line justification, genre, and chronology. Nevertheless, there remain areas which have, so far, received little attention, such as how much scribal abbreviation and expansion behaviour varied when copying closely related texts. It has so far only been addressed by a handful of studies, including Middle English dialectological work, which considered variation between 'short forms' and 'long forms' in addition to a range of other orthographical phenomena in a corpus consisting of two or three manuscript witnesses (Samuels, 1983; Thaisen and Da Rold, 2009, pp. 287, 289–90; Thaisen, 2011; see Section 4.1).

This study expands on the results of an earlier one (Honkapohja, 2018), which used a slightly different corpus from five of the six manuscripts used in this study. The conclusion of the previous study and starting point of the present one was that the scribes had an individual inventory of function words, especially in Middle English. However, '[w]hen it comes to Latin names of ingredients, apothecaries' weights and some verbs, the scribes reproduce the forms found in the exemplar, and the frequencies are very close in all manuscripts, with only minor variation' (267). The

previous study left some questions open. This study aims to answer and test them using a more robust methodology.

One of the unanswered questions of the previous study was whether the results were more an effect of text type or language. The study used a corpus which only contained Latin recipes, but were compared with all of the Middle English found in the manuscripts. The comparison was between a single text type in Latin and a more miscellaneous sample of Middle English. This left unclear whether the observed differences were primarily because of language or text type. The current study addresses both of these issues by using XML tagging which separates between text types in addition to languages (see Section 2.3). Moreover, Honkapohja (2018) relied on a 'bag of words' approach which compared token frequencies. The analysis concentrated on a few representative word types, which emerged as interesting based on their similar or different token frequencies. Essentially, the selection fell on the high-frequency function words or lexical words that looked similar based on their token frequencies. This study uses automatically aligned XML, which enables comparing how the different manuscript witnesses treat words in the same part of the text, looking at each aligned row of tokens in a corpus consisting of plague treatises in six closely related manuscripts.

## 2.1 The JB corpus

The corpus used in this study consists of digital transcriptions of two Latin versions and one English version of a popular plague treatise. The JB copies come from a group of closely related manuscripts known as the Voigts-Sloane Sibling Group. The group consists of six medical manuscripts that have strong textual similarities. These manuscripts originated in professional book production in London or its vicinity between 1450s and 1490s.

The treatise attributed to JB was extremely common in England and elsewhere in northern Europe. The original was composed by a medical practitioner, known as John, Johannes, or Jean of Burgundy or, in some Middle English copies, of Bordeaux, who was active in Liège in 1365 during the second epidemic (see Sudhoff, 1911; Waley Singer, 1916; Matheson, 2005 on different versions of the treatise; and Honkapohja and Jones, 2020 on identity of the author). The tract survives altogether in more than a hundred manuscripts, adapted into at least ten different versions (Sudhoff, 1911; Matheson, 2005) and translated into French, English, Dutch, and Hebrew. The treatise became especially popular in England and Scotland, in which it was the dominant plague text until the late sixteenth century (Jones, forthcoming).

## 2.2 The Voigts-Sloane sibling Group

The current study includes JB texts from six codices connected to the professional book market in London or its vicinity between 1450 and 1490s. Four of the manuscripts date right from before William Caxton started the first printing press in England; two shortly after. These codices form a group known as the Voigts-Sloane Siblings, first identified by Voigts (1990). The production circumstances, multilingualism, and dialect of the group were subject to an extensive codicological and linguistic analysis by Honkapohja (2017). Table 1 presents the six manuscripts that constitute the group.

The earliest surviving manuscript in the group is London, British Library, MS Sloane 2320, a quarto-sized paper manuscript, which was most likely copied in the 1450s. We will refer to it as Sloane 2320.

The second oldest are three pocket-sized codices, London, British Library, MS Sloane 3566, Cambridge, Trinity College, MS O.1.77, and Boston, Collectanea medica, Boston Medical Library, Francis A. Countway Library of Medicine, MS Ballard 19, which we will refer to collectively as the pocket-sized siblings and individually as Sloane 3566, Trinity and Boston. These manuscripts date to the late 1450s or 1460s. The pocket-sized siblings seem to be slightly later than Sloane 2320. Trinity O.1.77 contains the dates 1459 and 1460 in margins or flyleaves. Boston MS 19 was dated by Doyle (1956) to the summer of 1468.[1] Sloane 3566 does not contain dates, but contains one printed woodcut image (Honkapohja, 2017; Fig. 6), making it a hybrid manuscript.

Finally, the group also includes two lavishly decorated parchment manuscripts copied in display hands that may be as late 1480s or 1490s: Cambridge, Gonville and Caius College, MS 336/725 and Tokyo, Takamiya Collection, MS 33. Voigts (1990, p. 27) christened them as the 'second generation of a family that grew prosperous'. Such manuscripts were often

**Table 1.** The Voigts-Sloane Sibling Group manuscripts

| | MS | Size | Dates |
|---|---|---|---|
| Sloane 2320 | London, British Library, MS Sloane 2320 | Medium (quarto) | 1450?[a] |
| **The pocket-sized siblings** | | | |
| Sloane 3566 | London, British Library, MS Sloane 3566 | Small | 1450–70? |
| Trinity | Cambridge, Trinity College, MS O.1.77 | Small (sextodecimo) | 1459–60 (in astrological calculations) |
| Boston | Boston, Collectanea medica, Boston Medical Library, Francis A. Countway Library of Medicine, MS Ballard 19 | Small | 1468 |
| **The second generation** | | | |
| Gonville | Cambridge, Gonville and Caius College, MS 336/725 | Large | 1480–90 |
| Tokyo | Tokyo, Takamiya Collection, MS 33 | Large | 1480–90 |

[a]A marginal note in the manuscript mentions the year 1454. The note is in a different booklet than the JB tracts and may have ended up in the same covers later, possibly bound by Sir Hans Sloane himself in the sixteenth or seventeenth century (Honkapohja, 2017, pp. 50–62). It is, however, likely that the JB texts were also copied in the 1450s and thus predate the other manuscripts in the Group.

commissioned by upper class audiences, who continued to provide a market for luxurious hand-copied books even after the printing press had made books available for the middle classes (Smith, 1996, pp. 30, 76). We will refer to these manuscripts as the second generation and individually as Gonville and Tokyo.

Despite the differences in date and size, the texts in these manuscripts are very close copies, which make them very suitable for the present study, as it is dependent on an aligned parallel corpus. The defining feature of the Voigts-Sloane Siblings is a bilingual anthology of ten to twelve medical texts. Texts in the anthology address a number of areas typical of medieval medicine in general, such as diagnosis by uroscopy and treatment by diet or laxative and purgative remedies, or ones more specifically typical of the fifteenth century such as the emphasis on astrology or alcohol-based medications (for a full discussion, see Honkapohja, 2017, pp. 10–17). The anthology texts, somewhat unusually, occur in a nearly identical order in all six manuscripts with only minor textual variation (Honkapohja, 2017, pp. 10–17). The anthology also contains three versions of the JB tract: the Latin long version, the Middle English short version, and a Latin epistolary version. The three versions of the JB tract are in the same order in all six manuscripts and form a textually very close subtype (Waley Singer, 1916; Sudhoff, 1911; Matheson, 2005). Figure 1 illustrates the length of the different versions as well as abbreviation density in each section.

## 2.3 Annotation and frequencies

To study the treatment of abbreviations, the JB versions were encoded in TEI P5 XML and annotated for abbreviations as well as textual divisions. The encoding differentiates between three text types: recipes, running text, and incipits/explicits.[2] These are all transcribed and separated by XML tags whose position is used to create Fig. 1.

The tagging of abbreviations makes use of the TEI P5 Guidelines tags for encoding abbreviations and their expansions. The system is similar to the one proposed by Matthew Driscoll for Old Norse (Driscoll, 2009; Cummings, 2009; Stutzmann, 2010). The specific tagging choices are presented in Honkapohja (2013), in which the examples come from the Trinity and Boston manuscripts, which are both included in the present study.

```
<w>
<choice>
<abbr><am>&amp;</am></abbr>
<expan><ex>and</ex></expan>
</choice>
</w>
```

For the purposes of textual alignment and analysis, textual divisions were annotated using the  tag, which is used to indicate a subdivision of the text (TEI P5 Guidelines, 4.1). Textual divisions of different types were indicated by the use of the attributes @type and @n. Five different values were used for @type: ‘chapter’, ‘recipe’, ‘incipit’, ‘explicit’, and

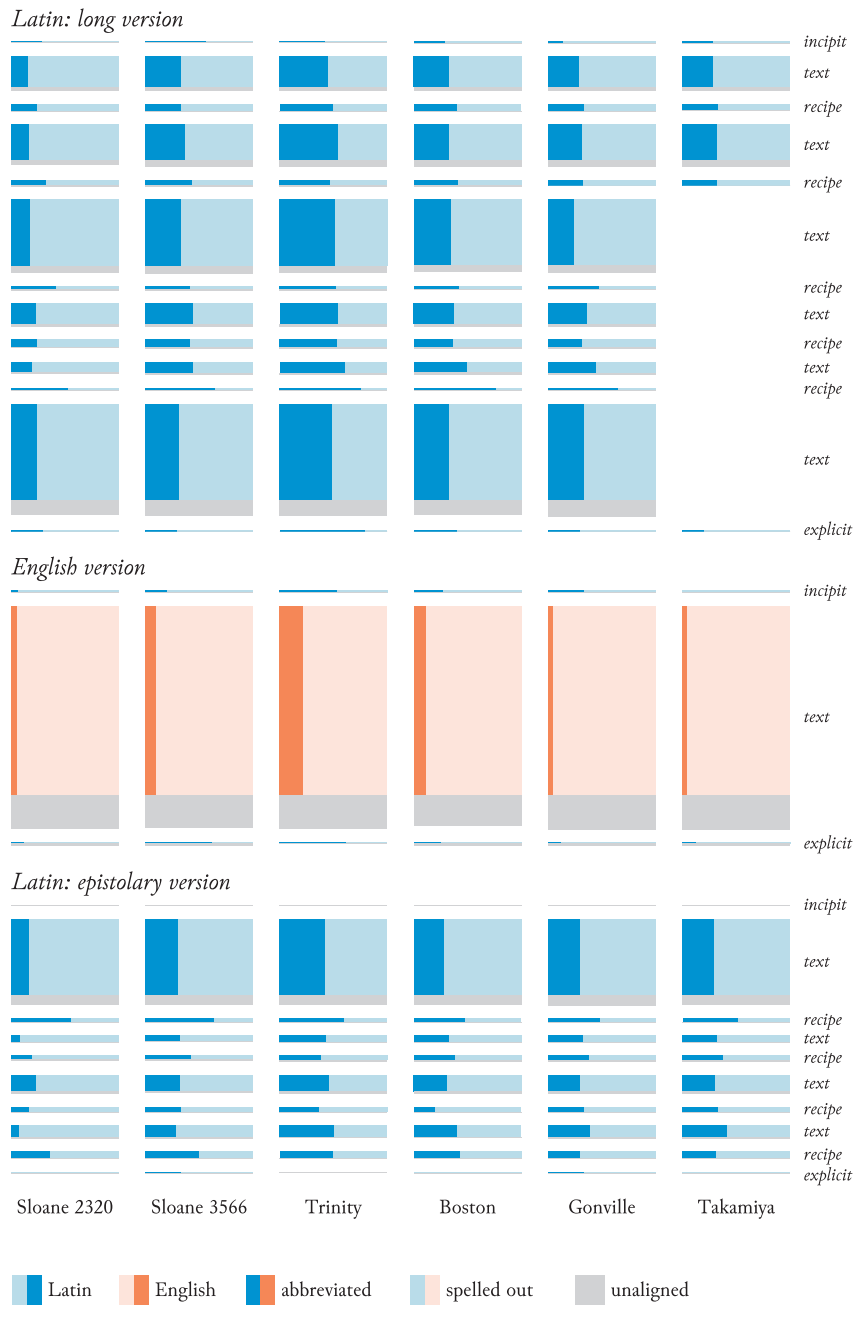


**Fig. 1** Sections, text types, and languages in the data (Tokyo lacks an entire leaf, which had the effect of missing alignment. For this reason, it was left out of the analysis. Both Tokyo and Gonville were copied by the same scribe.)

'running_text'. The three different versions of the JB treatise were separated by using the @n attribute, giving it the values 'JB_Latin_long', 'JB_English', and 'JB_Epistolary'. To facilitate alignment and analysis, recipes were also given individual names: @n='JBlong1', @n='epistolary1', etc. The tags were used for alignment on the level of the textual divisions.

# 3 Alignment

Alignment took place both on the higher level of aligning textual divisions and on the word level when aligning lines in all manuscript witnesses. The alignment of textual divisions was carried out as follows: we developed an alignment tool https://github.com/suomela/medieval-abbreviations that attempts automatic alignment, based on the TEI XML source files and the attributes given for the  tags. The alignment tool was written in Python 3, using the lxml library (for parsing XML input files and for producing human-readable HTML output) and the XlsxWriter module (for generating Excel tables with the alignment results).

## 3.1 Textual alignment

The tool first parses the XML input files, splits them in sections using the  elements, and finds a correct alignment on the level of the sections. This is based on the attributes of the elements (e.g. the element  corresponds to the first recipe in all texts) and the placement of the elements (e.g. we can uniquely identify the section of running text between the first and second recipes). This way we can identify twenty-six sections that are represented in at least five of the six texts.

Then we iterated three following steps, which refined both the XML encoding and the alignment parameters of the tool. First, we ran the alignment tool and inspected the results by spot checking. Secondly, if we noticed misalignments that are due to mistakes in the XML markup (typically, slight differences in where the scribes have marked a recipe as beginning or ending), we changed the textual divisions in the XML documents according to the usage of most scribes. Thirdly, if we noticed systematic misalignments that are due to spelling variation, we refined the alignment rules in the tool and repeated. Finally, once it seemed that the automatic tool was able to produce a near-perfect alignment, we produced Excel tables with the alignment results, manually checked all of them, and annotated any of the remaining misalignments that needed to be corrected.

## 3.2 Word-level alignment: Two normalized forms

For the purpose of word-level alignment, each section was flattened into a list of words. We discarded punctuation, line breaks, and other annotations, and kept only the contents of the <w> tags. For each abbreviated word, we extracted its expanded version (annotated by <ex> tags). A key obstacle for automatic alignment is the large amount of spelling variation and errors; our next step was to normalize the text.

For the purposes of normalization, it turned out to be beneficial to produce two different normalized versions of the text: a 'conservative' normalized form and a 'sloppy' normalized form. Both forms serve the same purpose of eliminating spelling variation and typical mistakes. For example, the words *medicine* and *medycyn* have the same normalized forms, and *evil* and *euylle* have the same normalized forms. However, there is a fine balance between the following objectives: different spellings of the same word have the same normalized form, but different words have different normalized forms.

We made two different choices here: the conservative form is chosen so that it errs rather on the side of false negatives (failing to eliminate spelling variation), while the sloppy form is chosen so that it errs rather on the side of false positives. For example, the conservative form is sufficient to detect that *medicine* and *medycyn* are spelling variations of the same word, while we will need the sloppy form to guess that *venomys* and *venymous* refer to the same word.

To construct the conservative normalized form, we first normalized the capitalization, eliminated punctuation and spaces, applied special rules for some common short words (*ye*, *he*, *hit*, *hyt*, *his*, *hem*, *fro*, *froo*, *yerof*), and replaced & with *et*. Then we replaced þ with *th* and ȝ with *gh*, *z*, or *y*, depending on the context. Similarly, we replaced *ph* with *f*. At this point, the text is still a rather faithful representation of the original text and easily human-readable.

Then we proceeded to apply more aggressive rules that, in essence, eliminate variation inside each of the following groups of letters:

- th, d, t
- m, n
- z, s, c, k
- i, j, y, e, a
- u, v, w

The process was iterative. The selection of letters was based on trying, observing the result, and adding or removing rules based on them. In addition to ones mentioned above, we also eliminated repeated letters (e.g. *evill* versus *evil*), applied special rules related to *cio* versus *tio* and *hour* versus *our*, and eliminated certain vowels around an *r* (e.g. *suffir* versus *suffre*) and near the end of the word (e.g. *medycyne* versus *medycyn*). The text at this point is no longer human-readable, but based on our experiments, it preserves enough of both Latin and English text so that there are very few false positives in our alignment.

The sloppy form applies much more aggressive rules, preserving only very little of the text. For example, most of the vowels are completely eliminated. We emphasize that the normalized forms are not intended for any kind of human consumption but to give the reader some idea of how the computer sees the text. Here are short examples:

- **original**: Here begynnyth a noble tretys made of a gode phisician Iohnn of Burdeux (...)
- **conservative**: hr bigimit i mobl trtic mit of i got ficicim iohm of burtiux (...)
- **sloppy**: r bgmt i mbl trt mt of i gt ftm im of brt (...)

Many of the sloppy forms are no longer unique and hence an alignment solely based on them would result in numerous false positives, while the conservative form is not sufficient to eliminate all spelling variation. We addressed this issue by performing the alignment primarily based on the conservative forms, and then filling in the gaps with the help of the sloppy forms.

In our experience, the following strategy worked very well with our data: We first set the bar very high and tried to find high-quality alignments. Such alignments divide the section into shorter parts that are still unaligned. Then we gradually lowered the bar and processed the unaligned parts again. For our purposes, a high-quality alignment is a place where we have a long segment of text in which the conservative normalized forms agree. We started at a word boundary, accumulated complete words until the total length of the normalized words is at least some number $k$, and then we checked if the normalized forms agreed between different texts. Initially, our threshold was $k = 39$, and we gradually lowered it until we had $k = 1$. Then we repeated the same process for sloppy normalized forms, starting again at $k = 39$ and stopping at $k = 2$. Note that we require the aligned part starts and ends at a word boundary (indicated by the <w> tags in XML), but we are happy to accept different numbers of words; for example, in to matches with into. Overall, there were more than 4,500 lines in the alignment tables produced by the tool. The end result was then checked and annotated manually. In total, eleven lines were false positives, and six lines were false negatives. The aligned lines were subject to further corpus analysis.

# 4 Variables

The alignment tool produces a table in which aligned parts and unaligned parts show in different colours. Furthermore, each alignment is labelled with an indication of the quality of the alignment, and alignments that were done based on only short parts or only sloppy forms are highlighted; this makes it easier to see which parts need to be verified carefully. Output from the automatically aligned texts was classified based on a number of criteria, three of which can be taken automatically from the XML encoding, but one was annotated manually: a division into function and lexical words. These were subject to a comparative analysis of scribal consistency across manuscript witnesses.

## 4.1 Consistency and constrained selection

The notion of consistency requires some clarification. It seeks to model the fact that the production of manuscript books was an act of encoding and decoding. 'If a scribe [...] is a copyist of someone else's work, [...] he has first to decode the language of the

author and then re-encode it into his own system' (Laing 1999, p. 251). A number of studies note that, when re-encoding texts, scribes copied some words more *literatim* than others (Benskin and Laing, 1981, pp. 5.1–5.7; McIntosh *et al.*, 1986, pp. 1.6–1.8; Thaisen and Da Rold, 2009). Benskin and Laing also use the term 'constrained selection'—referring to 'the tendency of a scribe to reproduce from his copytext forms that he would rarely produce spontaneously' (Thaisen and Da Rold, 2009, p. 287). They mention a difference between a scribe producing an abbreviated *w*$^{t}$ as opposed to the full form with as one possible area in which one can expect to find variation. The difference between abbreviation and expansion has thus been identified as one where constrained variation may operate.

The present aligned parallel corpus provides an excellent testbed for investigating scribal behaviour. If an aligned word is abbreviated in all five manuscripts, this shows that the scribes copied these words *literatim*. The word is abbreviated consistently by all manuscript witnesses. If a word is abbreviated by four, it is still fairly consistent and leads to asking why one manuscript witness deviates from the other four. In contrast, an overlap in abbreviation practices of two or three words is more difficult to interpret. It could be caused by chance or it could have an explanation in some shared practices of these scribes—perhaps the two copies are closely related. If a word is abbreviated only by a single scribe, we can conclude that this manuscript stands out for some reason. The data were encoded to test whether this is constrained by factors such as the language, text type, type of the word, or length of the word.

## 4.2 Latin and vernacular

The first variable is related to Latin and vernacular, which can be tested by comparing the consistency between passages tagged as English or Latin with the <xml: lang> tags (see Section 2.3). Our working hypothesis is that abbreviation is going to be not only more frequent but also more consistent in Latin. There are a few reasons for this assumption. The abbreviation and suspension system used in Latin emerged over centuries of copying texts and became famously complicated. Vernacular abbreviation systems, such as in Middle English, were for the most part based on the Latin system, but both the number of different abbreviations and their density remained lower than in Latin (Hector(1958) 1966, pp. 36–37; Hasenohr, 2002, pp. 82–83; LAEME, 2013, 3.4.5.1). Essentially, abbreviations were just more central to Latin than to Middle English. Moreover, at least the Boston scribe, William Ebesham (Doyle, 1956), was not a medical practitioner, and codicological analysis suggests that the copyists of the other manuscripts were likely to be commercial book artisans in London (see Honkapohja, 2017, ch 2, 3, and 6). If these scribes were copying specialized subject matter in a second language, it makes sense to assume they would copy Latin more faithfully.

## 4.3 Text type: Recipes

Another variable which one might expect to constrain selection is the text type of recipes, of which there are several (see Fig. 1 and Section 2.3). A text type can be defined as 'a specific linguistic pattern in which formal/structural characteristics have been conventionalized in a specific culture for certain well-defined and standardized uses of language' (Görlach, 2004, p. 105). Though some of its characteristics can vary across time, the recipe is one of the clearest examples of a text type and it is safe to assume that it was easily recognizable to any literate member of fifteenth -century society. The recipe as a text type can be defined, on the one hand, by its purpose, giving instructions, and, on the other, by the presence of certain 'obligatory linguistic features and formulas' (Carroll, 1999, p. 27). Recipes have been subject to several recent studies in historical linguistics (Görlach, 1992; Carroll, 1999; Grund, 2003). 'In the Middle English period recipes could appear singly, scribbled into the margins of other texts, or collected with charms, medicinal recipes or household information' (Carroll, 1999, p. 28). It is easy to see how selection could also be constrained by genre or text type.

The working hypothesis that abbreviation might be more consistent in recipes derives from the need for precise instruction. For example, Alonso-Almeida (2003, p. 14) argues regarding recipes (and their punctuation): 'If you think for instance of the case medieval medicinal recipes, the final therapeutic solution suggested in those recipes depends entirely in the correct reading of the text'. Consequently, research question B tests whether abbreviation consistency is higher in sections of the text tagged as recipes than elsewhere.

### 4.4 Lexical and function words

Another set of factors that could constrain which words were abbreviated and which expanded is the type of the word itself. Hasenohr (2002) notes that the extensive Latin system of abbreviation covered words that 'come often under the pen' (80). These include adverbs, pronouns, and prepositions, essentially things that can be grouped under the umbrella category: function words.[3]

To investigate this, we classified the aligned output into lexical and function words. The definition of lexical word was mainly from the Longman Grammar (Biber *et al.*, 2000). Lexical words are 'members of open classes' and 'the main carriers of meaning in a text' (Longman Grammar, 2.2.3.1). They include nouns, verbs, adjectives, and declinable adverbs. Function words, on the other hand, 'provide the mortar which binds the text together' and 'indicate relationships between lexical words or larger units'. They are members of closed classes (Longman Grammar, 2.2.3.2). They include prepositions, pronouns, and closed-class adverbs. We did, however, deviate from this definition to include forms of the verb 'to be'—following Hasenohr (2002) and Cottereau (2005). Otherwise, we followed Longman definitions while encoding things as lexical and function words, which worked for the most part, but also led to a handful of difficulties.

The encoding required decisions about a number of difficult cases. Adverbs in particular required case by case decisions. Latin *vero*, *modo* and *nihil* were counted as adverbs and function words. In addition, we counted a number of adverbs appearing in medical recipes as function words. These include the indeclinable adverb *semel* 'once' as a medical function word, and also *paulatim* 'by little and little, by degrees, gradually', *cito* 'quickly', *equaliter* 'equally', and *quousque* 'until that time'. The abbreviation *ana* is also counted as a function word. The abbreviation *.i.* 'id est' is counted as a function word as it has a common function of introducing synonyms. In English *much*, an intensifier, both an adjective and an adverb, along with its inflectional forms *more* and *most*, were classified as function words. The prepositions in phrasal/prepositional verbs were counted as function words (e.g. *put out*). These caveats aside, it should be easy to test whether the division between lexical and function words works as a variable that explains what gets abbreviated *literatim*.

### 4.5 Word length

Finally, the results were analysed simply based on the number of characters. Abbreviation as constrained variation is also likely to be constrained by what Petti (1977) called economy of space and economy of time. A language user may abbreviate when writing with great speed, or abbreviation may be necessitated by spatial constraints, such as the edge of a page or the end of a quire.

The importance of spatial constraints for abbreviation has been established by a number of previous studies. For example, Thaisen (2011, p. 79) found that abbreviation was more frequent in those quires in which space was most limited. Shute (2017) discovered that William Caxton's typesetters used abbreviation as a means to achieve right margin justification. Two comprehensive studies were carried out by Cottereau (2005) and Camps (2016, pp. cclii–cclix), who investigated how a number of variables affected abbreviation, including the number of syllables, position in the line and position in the quire. Among other things, Camps found a significant positive correlation between the number of syllables and the number of abbreviations (2016, p. cclii).

The current study does not focus on economy of space as such. The design of the current study with its aligned parallel corpus makes taking the end of the line, page, or quire into account impracticable. However, spatial constraint/one purely space related constraint is encoded as a variable: word length. We use simply the number of characters. The way the current study is set up allows determining whether language, text type, and word type affect abbreviation consistency more than economy of space. It also allows comparison between word length and the division into lexical and function words (described in Section 4.4).

## 5 Results and analysis

### 5.1 Latin and English: What gets abbreviated verbatim

The first comparison of consistency is between Latin and English. Figure 2 illustrates the number of words

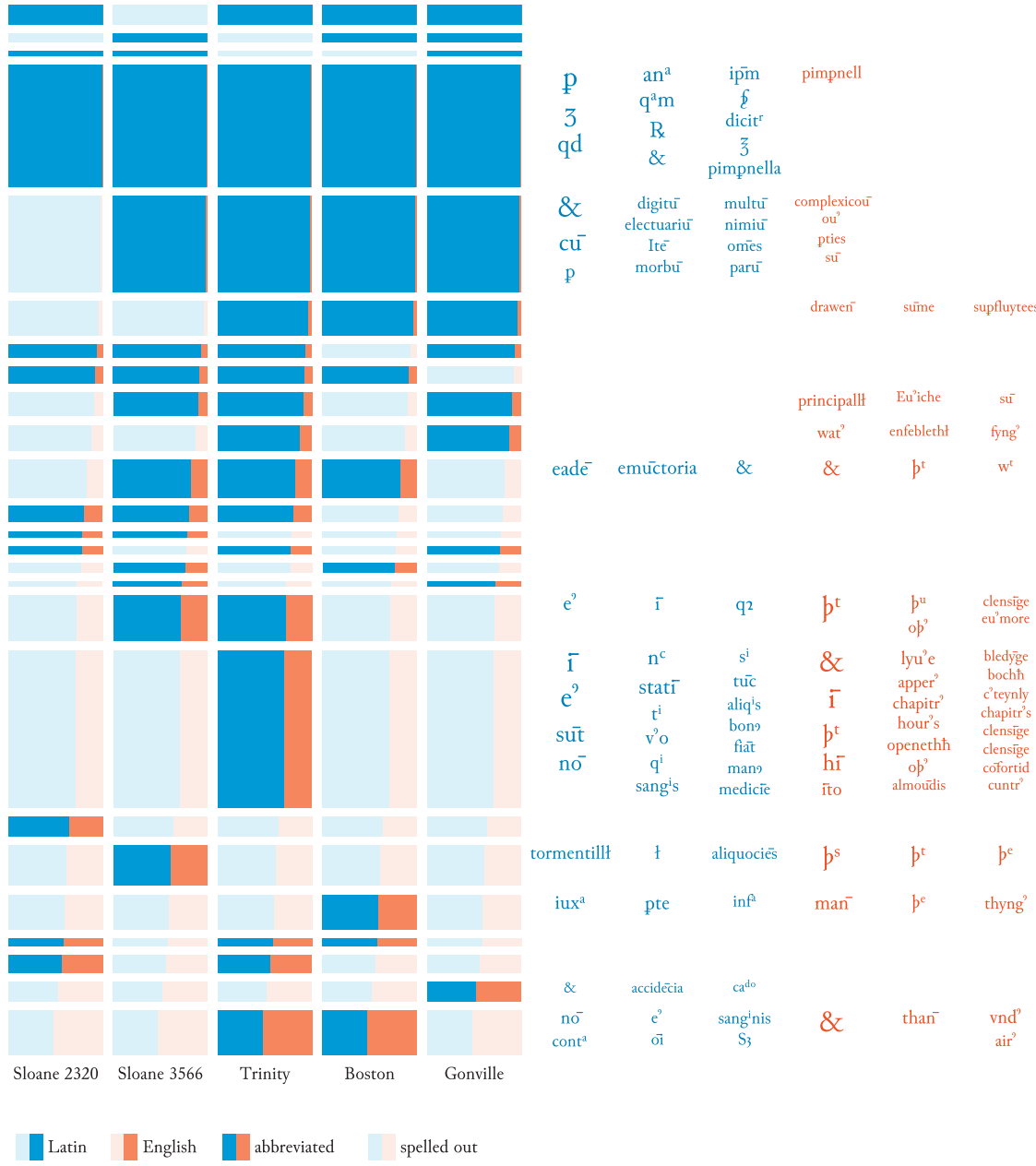


**Fig. 2** The consistency of abbreviating Latin and English

abbreviated consistently across the manuscripts. Abbreviated Latin words are shown as blue. Abbreviated English words are shown as orange. Expanded words are shown as lighter/paler blue/orange. The section on the right displays the most common words that are shared by these manuscripts. The words have primarily been listed by the number of occurrences and secondarily in alphabetical order. Font size corresponds with the token count: a larger font corresponds with a higher token count.

The results indicate that Latin is abbreviated overwhelmingly more consistently than English. There are altogether 275 aligned words consistently abbreviated by all five scribes. Out of these, 273 are Latin and only two English.[4] The counts are similarly overwhelming when considering aligned words abbreviated by four out of the five witnesses or three out of five. Table 2 shows the exact counts of aligned lines. According to these counts, 99.27% of words abbreviated by all five are Latin, as are 97.95% of words abbreviated by four out of five, and 87.19% of words abbreviated by three out of five witnesses.

In contrast, English abbreviation is less likely to be aligned across manuscripts, which shows as overlapping pairs or individual manuscripts clustering near the bottom of Fig. 2. The heavily abbreviating Trinity has altogether 356 abbreviated word tokens which the others do not abbreviate, out of which 103 are English (28%). The figures are similar for the others as well. The greatest number of overlapping English abbreviations is between Trinity and Boston, shown in Table 3. Just over half of the abbreviations shared by both are in English, 53/102 = 51.96%. A closer examination reveals that this is caused by the high token count of a single word, *&* 'and'.

A high token count of this frequent function word causes >70% of the overlap. As Table 3 shows, there are twenty-eight aligned lines on which both scribes abbreviate 'and' whereas the other three expand it. Both scribes use it overlappingly twenty-eight times. In addition, there are three other lexical items that are abbreviated by the two and occur more than once. Two of these contain an abbreviation type which was one of the most popular ones in English, the 'hook' abbreviation: *vnd*$^{9}$ 'vnder' (three tokens) and *ay*$^{9}$ 'air' (two tokens). Moreover, both scribes are in the habit of sometimes making a horizontal bar above 'n' in *thaⁿ* 'than, then', which may or may not indicate an abbreviation (De la Cruz-Cabanillas and Diego-Rodriquez, 2018, p. 172). Function words and abbreviation types can thus have a major effect in shared abbreviation practices across two witnesses. When it comes to recipes, however, there are certain formulaic elements which are abbreviated consistently across all five manuscript witnesses.

## 5.2 Text types: Recipes, running text, and meta-text

The recipe is one of the most universally recognizable text types (see Section 4.3). Some of the features that

**Table 3.** English abbreviations shared by Trinity and Boston

| | |
|---|---|
| 28 | & 'and' |
| 3 | vnd$^{9}$ 'vnder' |
| 3 | thaⁿ 'than, then' |
| 2 | ay$^{9}$ 'air' |
| 1 | bledyng$^{9}$ 'bleeding' |
| 1 | clēsing 'cleansing' |

**Table 2.** The percentage of Latin and English abbreviation

| **Consistently abbreviated by five MSS** | **Latin** | **Percentage** (%) | **English** | **Percentage** (%) |
|---|---|---|---|---|
| | 273/275 | 99.27 | 2/275 | 0.73 |
| **Consistently abbreviated by four MSS** | **Latin** | | **English** | |
| 2320 + Trinity + Boston + Gonville | 44/44 | 100 | 0/44 | 0 |
| 3566 + Trinity + Boston + Gonville | 214/218 | 98.17 | 4/218 | 1.83 |
| 2320 + 3566 + Trinity + Gonville | 29/31 | 93.55 | 2/31 | 6.45 |
| 2320 + 3566 + Trinity + Boston | 35/38 | 92.11 | 3/38 | 7.89 |
| TOTAL | 322/331 | 97.28 | 9/331 | 2.72 |
| **Consistently abbreviated by three MSS** | **Latin** | | **English** | |
| 3566 + Boston + Gonville | 20/20 | 100 | 0/20 | 0 |
| 2320 + 3566 + Gonville | 11/11 | 100 | 0/11 | 0 |
| Trinity + Boston + Gonville | 75/78 | 96.15 | 3/78 | 3.85 |
| 3566 + Trinity + Gonville | 49/54 | 90.74 | 5/54 | 9.26 |
| 3566 + Trinity + Boston | 71/86 | 82.56 | 15/86 | 17.44 |
| 2320 + 3566 + Trinity | 29/36 | 80.56 | 7/36 | 19.44 |
| 2320 + Trinity + Gonville | 14/18 | 77.78 | 4/18 | 22.22 |
| 2320 + Trinity + Boston | 10/17 | 58.82 | 7/17 | 41.18 |
| Total | 279/320 | 87.19 | 41/320 | 12.81 |

identify a recipe had become so formulaic and easy to recognize that there were common abbreviations used for them. Many words specific to recipes show up among the most consistently used abbreviations. These include logograms used for measurements such as ʒ 'drachm', ℞ 'recipe', *β* 'semis [half]', as well as slightly longer formulaic abbreviations such as *an*$^{a}$ 'ana', instructing one to add the same amount of an ingredient as previously. Indeed, Rogos-Hebda (2018, p. 54) suggests that characteristic abbreviations could function as a visible pragmatic marker to help readers differentiate between text types. Figure 3 shows the things shared by the various manuscripts.

The clustering of words in recipes near the bottom in Fig. 3 shows that words in recipes are abbreviated more consistently than in running text. The numbers tell us a similar story. There are altogether 287 aligned words in recipes. Out of these, 62 are abbreviated the same way across all five manuscript witnesses, that is, 21.6%. This contrasts starkly with the 213 aligned words out of a total of 3,441 elsewhere, which is 6.2%. It is thus possible to conclude that recipes are abbreviated more consistently. There are, however, differences in which words are likely to cause this effect.

Table 4 shows the words with highest token frequencies aligned in all five manuscripts. It reveals that all of the aligned words that occur three times or more have to do with the formulaic elements of recipes ʒ 'drachm' (10), *an*$^{a}$ 'ana' (7), ℞ 'recipe' (6), *β* 'semis' (4), and ȝ 'ounce' (3)—some of the most straightforward generically defining words for the text type. In contrast, the most frequently abbreviated words that occur in non-recipe sections are predominantly function words such as *p* 'per', *qd* 'quod', *q*$^{a}$*m* 'quam', followed by the conjunction & 'et/and'.

Function words can have a major effect on separating one manuscript witness from all of the others. In particular, the high frequency of a single function word makes one manuscript witness diverge from all the others: Sloane 2320. The scribe of this manuscript has a tendency to expand the function word *et* where others use an ampersand, and not specific to the text type of recipes (see discussion in Section 5.3 and Table 5 for word counts). Furthermore, the Sloane 2320 scribe's frequencies of *et* cut across both counts and represent about 5% of the scribe's abbreviation (recipe 15 = 5.2%, non-recipe 189 = 5.49%). The reason this manuscript witness stands out from all of the others, more than anything else, due to a single high-frequency item.

The results are reasonably clear in proving that the text type of recipes is very consistently abbreviated with regard to the formulaic elements, which are specific in identifying them as belonging to the text type. They are also clear with respect to function words, as the scribal practice of the Sloane 2320 scribe of expanding *et* makes this manuscript witness differ from the rest. However, the working hypothesis influenced by Alonso-Almeida (2003, p. 14) was also that we could see constrained selection, in which ingredients would be abbreviated more consistently as the outcome of the recipe depended on the correct reading of the text. The results are not quite as clear with respect to lexical words such as ingredients, as there is a steady proportion of such words that are abbreviated by all five, abbreviated by four, three, two out of five, or possibly just one. Figure 4 shows the percentages of words abbreviated in different manuscript witnesses categorized into three groups: content words, function words, and the five recipe-specific formulaic words ('drachm', 'recipe', 'semis', 'ana', and 'ounce').

The consistency of abbreviating lexical words varies more than function words or recipe *formulae*. There are thirty lexical words in recipes abbreviated by five scribes, but there are also fourteen abbreviated by four witnesses, eleven abbreviated by three witnesses, fifteen abbreviated by two witnesses, and twenty-one which are only abbreviated by a single one. Although the effects are fairly clear regarding the five recipe-specific abbreviations, which are commonly abbreviated by all five, and function words, which show the tendency of the Sloane 2320 scribe to expand *et*, they are less clear regarding lexical words, which include ingredients.[5] Examples of these range from names of ingredients such as *pimpnella* (abbreviated by all five), to *diptanū* 'diptanum [dittany]' (abbreviated by four MS witnesses) to *radic*$^{9}$ 'radicem [root]' (abbreviated by two MS witnesses to *sandał* 'sandalis [sandalwood]' (only abbreviated by Trinity)). Consequently, the greater consistency of abbreviating recipes is mainly caused by the formulaic elements specific to them, which has implications for the working hypothesis (see Section 4.3) that scribes might have copied recipes more consistently than running text to transmit precise information. The scribes abbreviate the formulaic frame very consistently, but


**Fig. 3** The consistency of abbreviating recipes and running text

their individual preferences for using certain types of abbreviations also have an effect across recipes, and their profiles regarding function words remain intact.

## 5.3 Word types: Lexical and function words

But how much can we say when the results are divided into words classified as 'function' or 'lexical' words? Do these support the idea that function words would be most variable and lexical words most consistent? As was already apparent, this division has an effect, but it is partly one of an overlap. Figure 5 shows the output broken down using this division arranged in such a way that the highest degree of overlap in lexical words shows at the top of the diagram as blue, and the highest number of overlaps in function words shows at the bottom as black.

**Table 4.** Words abbreviated by all five scribes

| Non-recipe | Recipe |
|---|---|
| 16: ꝑ ‘per’ | 10: ʒ ‘drachm [dram]’ |
| 9: qd ‘quod [that]’ | 7: an[a] ‘ana [the same amount]’ |
| 6: q[a]m ‘quam [than]’ | 6: ℞ ‘recipe [take]’ |
| 4: & ‘et/and’ | 4: $\beta$ ‘semis [half]’ |
| 4: im ‘ipsum [self]’ | 3: ȝ ‘ounce’ |
| 3: dicit[r] ‘dicitur [is said]’ | 2: ꝑ t[9] ‘parte [part]’ |
| 3: pimꝑnella ‘pimpernel’ | 2: p[9]seruari ‘preservari [preserve]’ |
| 3: quousqȝ ‘quousque’ [until when] | 1: accidentibȝ ‘accidentibus [accidents]’ |
| 3: venenū ‘venenum [poison]’ | 1: armoniac[9] ‘Armenian bole, ammoniac’ |
| 3: vinū ‘vinum [wine]’ | 1: calament[9] ‘calamente [calaminth]’ |
| 3: vniusquisqȝ ‘unusquisque [one/single]’ | 1: cameact[9] ‘Dwarf Elder’[a] |

[a]The translation is based on Horrox (1994, p. 190).

**Table 5.** The most frequent words abbreviated by all except Sloane 2320

| | | | |
|---|---|---|---|
| 111: & ‘et’ | func | 2: multūm ‘multum [many]’ | lex |
| 9: cū ‘cum’ | func | 2: nimiū ‘nimium [too much]’ | lex |
| 4: ꝑ ‘per’ | func | 2: oēs ‘omnes [all]’ | lex |
| 2: digitū ‘digitum’ | lex | 2: quibȝ ‘quibus [which (dat./abl. pl.)]’ | func |
| 2: electuariū ‘electuarium [electuary]’ | lex | 2: sanguinē ‘sanguinem [blood]’ | lex |
| 2: it[9] ‘item [likewise]’ | func | 2: vinū ‘vinum [wine]’ | lex |
| 2: morbū ‘morbum [sickness]’ | lex | | |

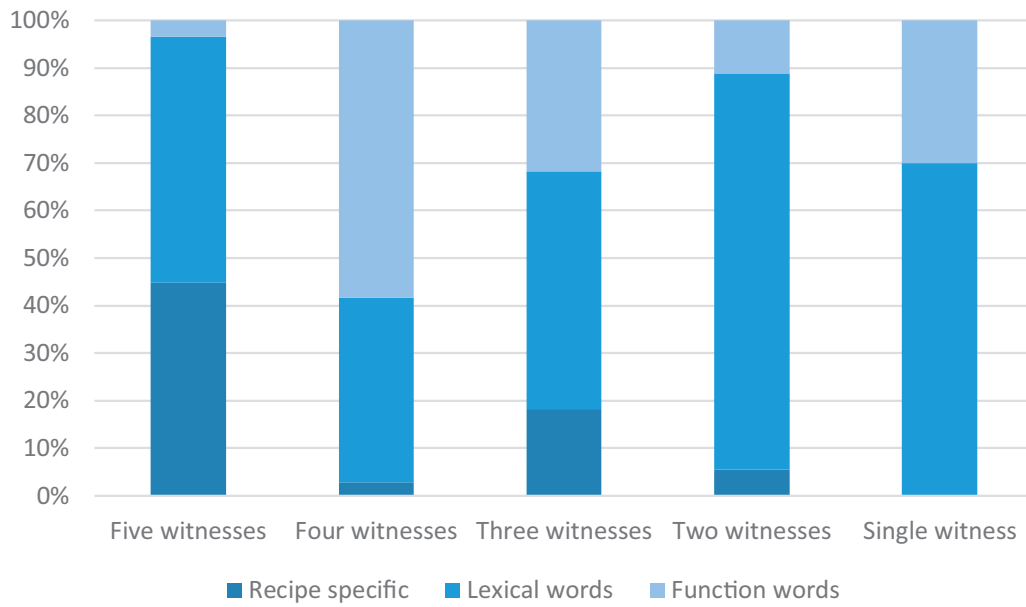


**Fig. 4** The percentage of abbreviation types in recipes


**Fig. 5** The consistency of abbreviating lexical and function words

The results make two manuscripts stand out: Trinity and Sloane 2320. Trinity stands out because 203 out of the 334 words, 60.8%, of the words abbreviated only in this manuscript are function words. Sloane 2320, on the other hand, stands out due to the scribal tendency to abbreviate fewer function words, which shows up as an alignment of the other four manuscripts: 121 out of the 186 aligned words, 65.1%, are function words. These two deserve closer examination.

Trinity's is the result that fits best with the working hypothesis that an individual manuscript might stand out from the others due to its scribe's tendency to use more abbreviations for a whole range of function words (see Section 4.4). The fact that ∼60% of the words are abbreviated only by Trinity is the result that fits most clearly with variable C, that scribes would have individual repertoires of function words. The manuscript contains the highest proportion of abbreviations and 60% of the cause are function words. The scribe uses frequencies for *ī* 'in' (71), *&* 'and' (20), *e*$^{9}$ 'est' (18), *sūt* 'sunt' (8), *þ*$^{t}$ 'that' (8). Although other scribes also use these, the numbers are not as high, which makes the manuscript stand out from the rest. Nevertheless, the fact that these function word abbreviations are not unique may also cause overlap.

High frequencies of abbreviated function words are also the cause for overlap with the other four manuscripts. Many of the overlapping pairs with higher counts involve Trinity (Trinity—Gonville 7/57; Trinity—Sloane 3566 29/104; Trinity—Sloane 2320 18/41). The likely reason is therefore as the Trinity scribe abbreviates considerably more, there is a greater likelihood of overlap when the others abbreviate as well. What is more, overlapping pairs and threes, such as Trinity, Boston, and Gonville (43 func/72 words = 59.7%), Trinity and Boston (58/97 = 59.8%), Sloane 3566, Trinity and Boston (42/77 = 54.5%), and Sloane 3566 and Trinity (53/95 = 55/8%), are also caused by function words. There is, however, also a difference caused by the absence of abbreviated function words.

One of the most striking differences caused is that the majority of words abbreviated by all five scribes are classified as lexical, whereas ones abbreviated by all except Sloane 2320 are classified as function words. Table 5 lists words that are expanded by Sloane 2320, but abbreviated by others. The number on the left shows token count. The column on the right shows whether the word is a lexical or a function word.

The table reveals a huge difference caused by the scribe's tendency for a single high-frequency item. By far, the highest token count is provided by hits (111 out 199) in which the Sloane 2320 scribe is the only one who writes an expanded *et*, where the others use an ampersand. The results are thus influenced by the high token counts of a single high-frequency word.

However, the Sloane 2320 scribe's tendency to expand where others abbreviate cuts deeper. As Fig. 5 and Table 5 reveal, there are also certain other function words (*cū* 'cum', *ꝑ* 'per', *it*$^{9}$ 'item) as well as a number of lexical words which the Sloane 2320 expands, but others abbreviate. Ones which have a token count of two or higher include *cum*, *digitum*, *electuarium*, *morbum*, *multum*, *nimium*, *omnes*, *sanguinem*, and *vinum*. In other words, the list also includes lexical words related to medicine, which one could perhaps assume are words that would be copied most literatim: *electuarium* 'a medicinal conserve or paste' (OED), *morbum* 'sickness', and *sanguinem* 'blood'. All of these words contain a nasal, which is very commonly abbreviated by the horizontal bar, also known as 'tittle' or 'macron'. Figure 6 shows the raw counts of this abbreviation type. The results revealed by Fig. 6 show clearly that the scribe of Sloane 2320 also uses fewer macrons. The scribe's tendency to use fewer abbreviations applies across the board.

The Sloane 2320 scribe's practice of using fewer abbreviations also shows in his tendency to use the most consistent abbreviation less: the crossed-p abbreviation *ꝑ* for 'per/por/par', as well as words containing it, such as English *pimpnel*, or Latin *pimpnalla* and *ꝑiculo* 'periculo [danger]'. It would appear that some abbreviation types and commonly abbreviated syllables are likely to be abbreviated similarly or differently by the various scribes.

To test the commonality, we checked for the consistency of abbreviating the syllable 'per', whose abbreviation could also stand for 'par' or 'por'.[6] Figure 7 displays the results. They reveal a number of things. First, the 'per' syllable is abbreviated at least once in every single manuscript. There are nineteen aligned words which are not abbreviated in any, but they all contain either 'por' or 'par': *appareat* (15), *epar* (15), *apparet* (13), *porros* (12), *porys* (12). Secondly, once again Sloane 2320 differs from the rest, as the scribe is also somewhat more likely to expand the 'per' syllable. There are eleven cases in which the practice of the Sloane 2320 scribe differs from the others. Thirdly, the abbreviation is used more consistently for 'per' than for 'por' or 'par', as words such as *departyd*, *corpore*, and *tempore* are among ones which are

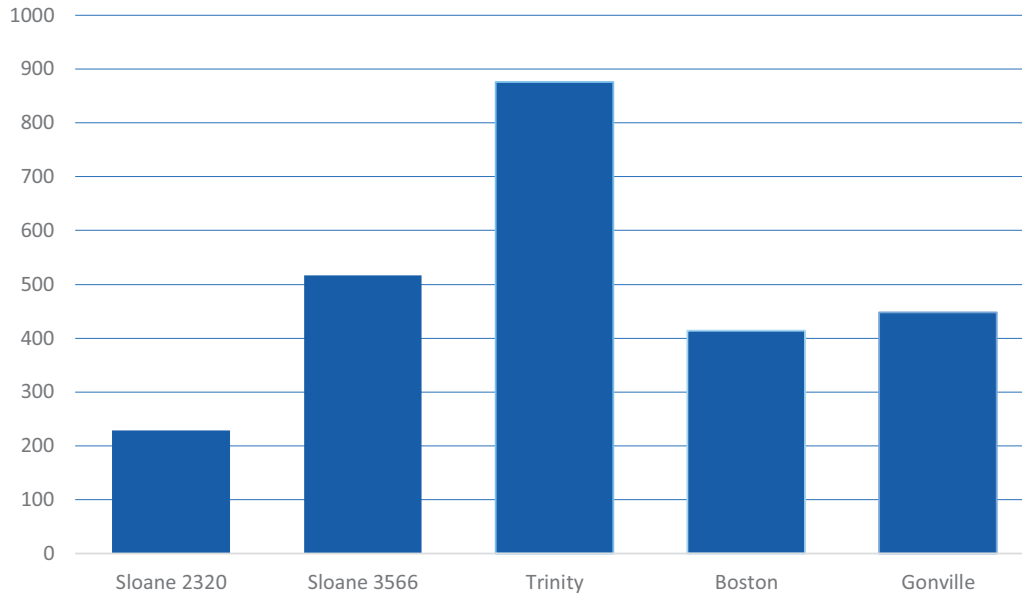


**Fig. 6** The number of horizontal bars

abbreviated only by few manuscripts rather than all, whereas ones containing 'per' are commonly abbreviated across the whole line/streak.

## 5.4 Word length

Finally, we examined whether a division based on word length (original, not normalized) will produce different results to ones based on text type and the division into function/lexical. For this part of the analysis, we divided words automatically into five groups based on the number of characters: 4 or less, 5–6, 7–8, 9–10, and 11 or more. The results are displayed in Fig. 8.

The results prove a useful way of differentiating between some of the scribal practices and allow seeing how different thresholds could yield different ways of dividing the material. The general tendency displayed by Fig. 8 is that longer words are more likely to be abbreviated consistently by all five manuscripts or sometimes expanded by the Boston or Gonville scribes, but abbreviated by the other four. One likely cause for this is that longer words are more likely to be abbreviated, a tendency also noted by Camps (2016) and Cottereau (2005).

Shorter words, four characters or less, make the two manuscripts stand out which already did in the lexical/function division. Trinity stands out for the high number of short words. Similarly, the single most influential factor is the tendency of the Sloane 2320 scribe to expand the ampersand 111 times, where all others abbreviate it.

However, examining these groupings also reveals some surprising divisions. For example, Boston stands out from the others predominantly because its scribe abbreviates words that are five or six characters long: these include *iux*$^{a}$ 'iuxta' (4) and *part*$^{9}$ 'parte' (4) as well as a single token for abbreviation *inf*$^{a}$ 'infra' (2). Similarly, there are words shared by Trinity and Boston which fall into the five to six-character word category—some of them English (*vnder*, *contra*, and *than*).

An important question that needs to be asked is whether the division into function and lexical words works better than the one based simply on the character length. On the whole, it looks like a threshold of four characters and below or five characters and above would serve to make a useful distinction here. We will compare three: all, only Trinity, and all but Sloane 2320. The results are shown in Table 6 below.

The results show that in all cases, a division based on character length actually proves to be

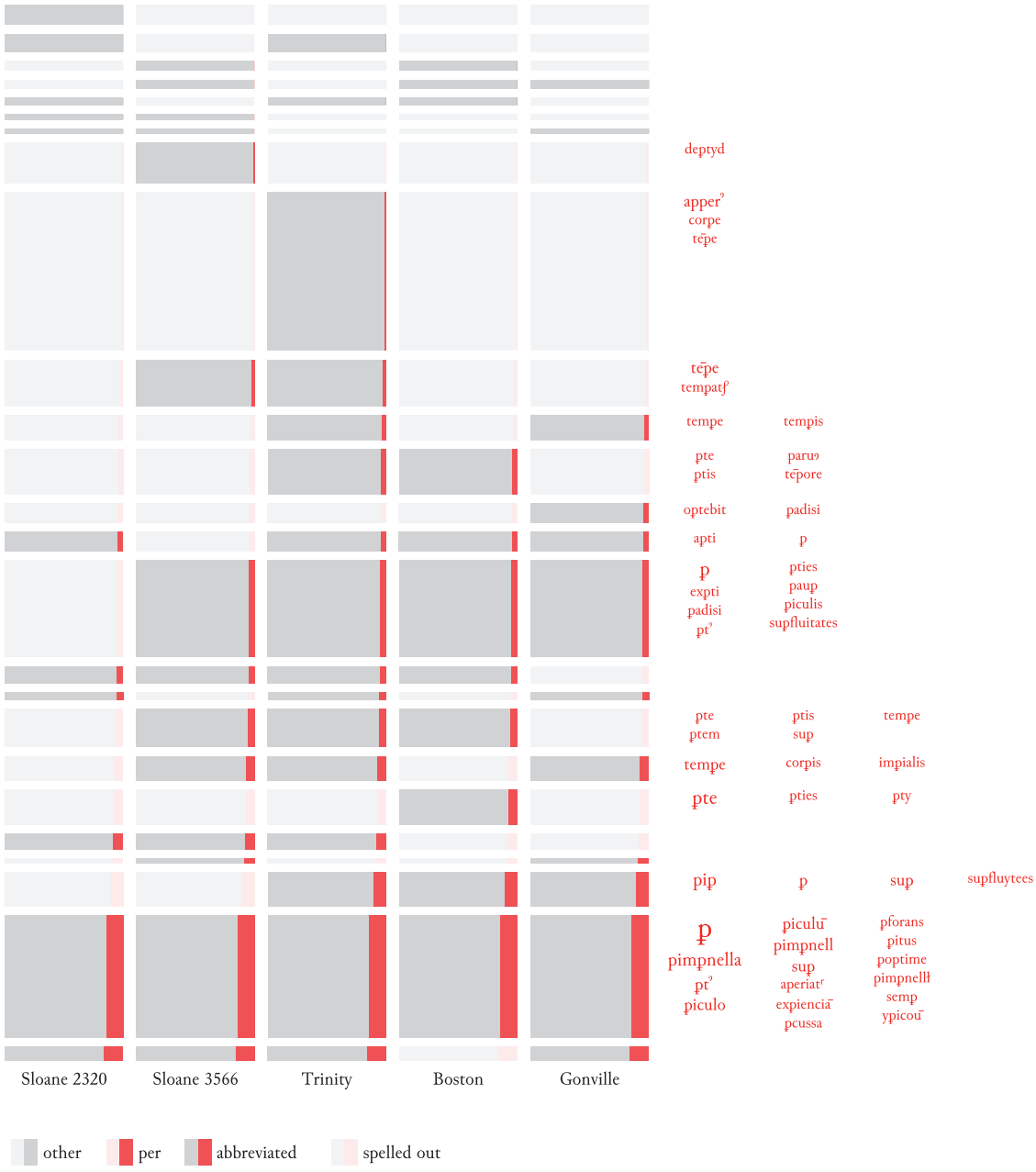


**Fig. 7** The consistency of abbreviating per/por/par syllables

the more efficient in forming a dividing line. The likely reason for this is, no doubt, the number of longer function words that were classified as such—including adverbials. All in all, forming a table which divides the words based on the number of characters could be useful for exploratory data analysis—if there was a user interface which would give the user a figure such as the above.

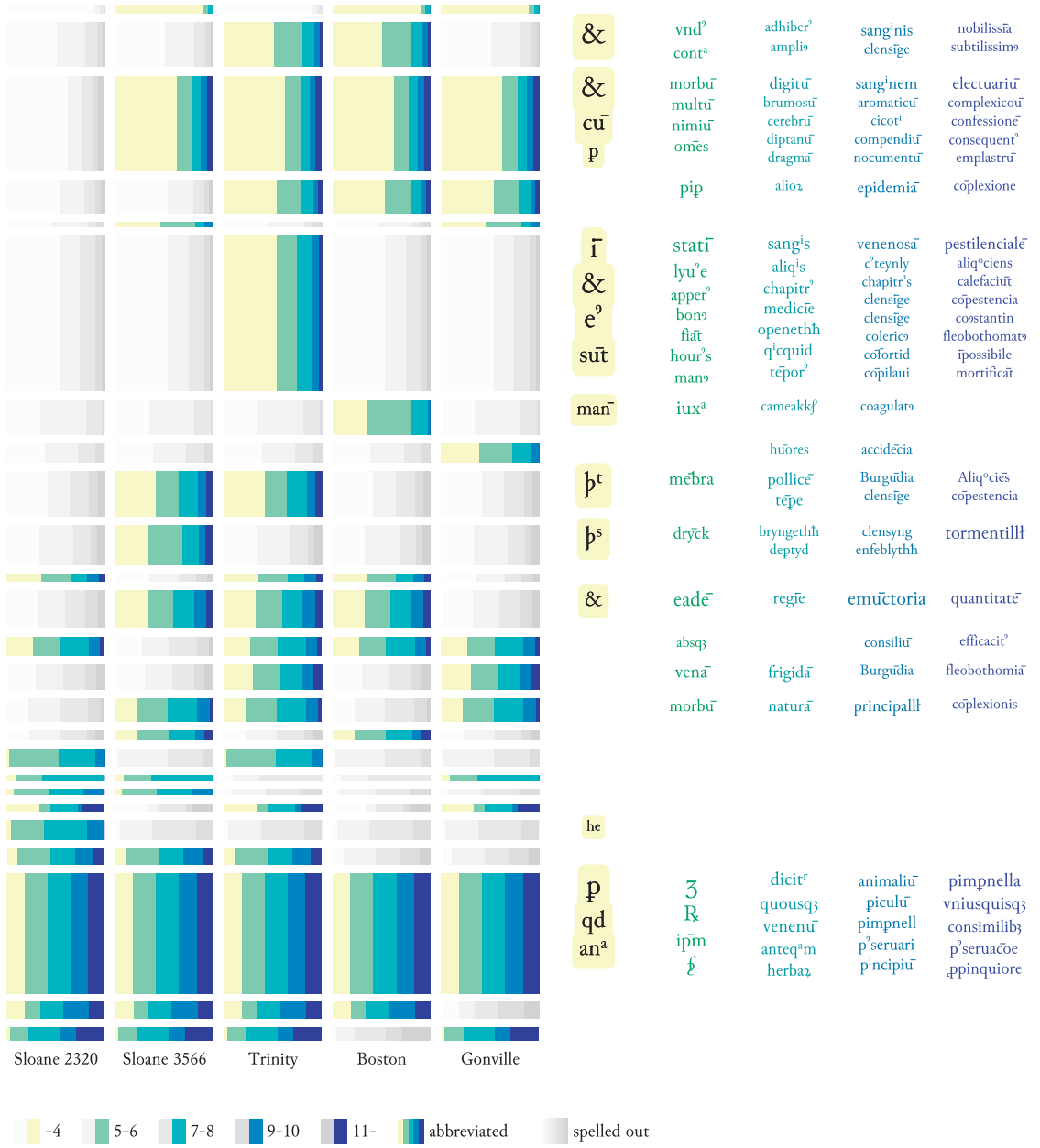


**Fig. 8** The consistency of abbreviating sorted by word length

## 6 Conclusions

This study examined how four sets of variables affect the consistency of abbreviation: A, language (Latin versus English); B, text type (recipes versus running text); C, word type (lexical versus function words); and D, word length.

The results are overwhelmingly clear regarding variable A, language, and fairly clear regarding variable B, text type. Latin is abbreviated far more consistently than the English.

**Table 6.** Comparison of function words and character length

| | func | 1–4 characters |
|---|---|---|
| All five | 78/275 | 50/275 |
| Only Trinity | 143/356 | 75/356 |
| All but Sloane 2320 | 140/218 | 136/218 |

The scribes also copied recipes more consistently than running text and incipits/explicits least consistently. However, this was shown mainly to result from the consistent use of abbreviations specific to recipes (r(ecipe), measurements), whereas lexical words such as ingredients do not show greater consistency.

The results also showed a division into function words and lexical words to work, but in a somewhat different manner than expected. It turned out that high token counts of a single word can affect consistency. The main example of this is the tendency of Sloane 2320 to expand *et* where others use an ampersand. This makes the manuscript witness stand out from the others. High token frequencies of common function words are also likely to lead to overlap, which shows as alignment pairs and threes. In particular, when it comes to the most abbreviating scribe, that of Trinity, there will be aligned words in which his practice will overlap with one or two other scribes, which shows as similarity between these.

In general, the working hypothesis that content and function words would work well as a diagnostic turned out to have somewhat mixed results. Manual encoding into lexical and function words was compared with an automatic one based on word length. Word length turned out to work even better as a distinguishing variable. A couple of explanations suggest themselves. On the one hand, longer words are more likely to contain an abbreviation. On the other, the high and consistent frequencies of short high-frequency words and commonly used endings are likely to have an effect, simply by virtue of their high token frequency. Moreover, abbreviation type also turned out to be influential. The ones investigated were the macron and the crossed-p for 'per/por/par'. Sloane 2320, which does not abbreviate as much, turned out not to differ also in terms of macron and the crossed-p.

All in all, this study revealed that abbreviations have definite potential as a diagnostic for scribal practice. The results show there is much potential for going further. It can be concluded, though, that there is definite potential in using a quantitative profile of abbreviation practices in identifying scribes—it might be possible to develop a tool in which the abbreviation scores of two scribes are compared. If convenient parameters can be found, differences such as the one that sets Sloane 2320 apart do suggest that a convenient sample of abbreviation forms can yield distinctive profiles, which could potentially be used to differentiate between scribes. Ideally, the tool would be able to use TEI XML or other datasets in which the abbreviations are annotated, but some of these results such as the tendency of Sloane 2320 scribe to expand *et* and *and* instead of the ampersand could also be studied by non-tag-based tools like Stylo. Although saving space (abbreviating long words, more abbreviations in small MSS) does play its part, it is by no means the whole story and there are scribal differences. The scribal level of language variation is often identified as problematic for automated authorship attribution. The results here show there is much potential in this type of quantitative approach and it might yet form an important part of digital palaeographers' toolkit.

## Funding

This work was supported by the Swiss National Science Foundation (SNFS) [grant number 174409]. We are grateful to Sara Norja for copyediting and language check.

## References

**Alonso-Almeida, F.** (2003). An assessment of punctuation in some middle English Gilbertus Anglicus's versions of 'the Sekenesse of Wymmen. *Philologica Canariensia*, **8**: 13–42.

**Benskin, M. and Laing, M.** (1981). Translations and *Mischsprachen* in Middle English manuscripts. In Benskin, M. and samuels, M. L. (eds), *So Many People Longages and Tonges: Philological Essays in Scots and Mediaeval English presented to Angus McIntosh.* University of Edinburgh, Edinburgh, pp. 55–106.

**Benskin M. and Laing M. and Karaiskos V., and Williamson K.** (2013) *An Electronic Version of a*

*Linguistic Atlas of Late Mediaeval English*. Edinburgh: The Authors and The University of Edinburgh.

**Biber, D., Johansson S., Leech, G., Conrad, S., and Finegan, E.** (2000). *Longman Grammar of Spoken and Written English, 3rd Impression*. Harlow: Longman/Pearson Education.

**Bozzolo, C., Coq, D., Muzerelle, D., and Ornato, E.** (1990) Les abréviations dans les livres liturgiques du XVe siècle: pratique et théorie. In *Actas del VIII coloquio del Comité internatiocional de paleografía Latina*, Madrid-Toledo, pp. 17–27.

**Brown, M.** (1990). *A Guide to Western Historical Scripts from Antiquity to 1600*. London: British Library.

**Camps, J.-B.** (2016). *La`Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique*. Paris-Sorbonne. <https://doi.org/10.5281/zenodo.1116735> (accessed 8 April 2020).

**Cappelli, A.** (1990 [1899]). *Lexicon Abbreviaturarum Dizionario Di Abbreviature Latine Ed Italiane*. Milano: Hoepli.

**Carroll, R.** (1999). The Middle English recipe as a text-type. *Neuphilologische Mitteilungen* **100:** 27–42.

**Chassant, L.-A.** (1970 [1845]). *Dictionnaire des abréviations latines et francaises usitées dans les inscriptions lapidaires et métalliques, les manuscrits et les chartes du moyen âge*. Hildesheim: Georg Olms Verlag.

**Clemens, R. and Graham, T.** (2007). *Introduction to Manuscript Studies*. Ithaca: Cornell University Press.

**Cottereau, E.** (2005) La copie et les copistes français de manuscrits aux XIVe et XVe siècles. Etude sociologique et codicologique. PhD Thesis, Université Paris Panthéon-Sorbonne.

**Cummings, J.** (2009). Converting Saint Paul: a new Tei P5 edition of the conversion of Saint Paul using stand-off methodology, *Literary and Linguistic Computing*, **3**: 307–17.

**De la Cruz-Cabanillas, I. and Diego-Rodriquez, I.** (2018). Abbreviations in medieval medical manuscripts. *Selim*, **23**: 163–83.

**Doyle, A. I.** (1956). The work of a late-fifteenth-century English scribe, William Ebesham. *Bulletin of the John Rylands Library Manchester*, **39**: 298–325.

**Driscoll, M.J.** (2009). Marking up abbreviations in Old Norse-Icelandic manuscripts. In Saibene, M.G. and Buzzoni, M. (eds), *Medieval Texts—Contemporary Media: The Art and Science of Editing in the Digital Age*. Pavia: Ibis, pp. 13–34.

**Görlach, M.** (1992). Text-types and language history: the cookery recipe. In Rissanen, M., Ihalainen, O., Nevalainen, T. and Taavitsainen, I. (eds), *History of Englishes: New Methods and Interpretations in Historical Linguistics*. Berlin & New York: Mouton de Gruyter, pp. 736–61.

**Görlach, M.** (2004). *Text Types and the History of English, Trends in Linguistics (series)*. Berlin and New York: Mouton de Gruyter.

**Grund, P.** (2003). The golden formulas: genre conventions of alchemical recipes in the Middle English period. *Neuphilologische Mitteilungen*, **104**(4): 455–75.

**Hälvä-Nyberg, U.** (1988). Die Kontraktionen auf den Lateinischen Inschriften Roms und Afrikas: bis zum 8. Jh. N. Chr. Helsinki: Suomalainen tiedeakatemia.

**Hasenohr, G.** (2002). Écrire en latin, écrire en roman: réflexions sur la pratique des abréviations dans les manuscrits français des XIIe et XIIIe siècles. In Banniard, M. (ed), *Langages et Peuples d'Europe: Cristallisation des Identités Romanes et Germaniques (VIIe-XIe siècle)*. Toulouse, Conques: CNRS-Université de Toulouse Le Mirail, pp. 79–110.

**Hector, L. J.** (1958). *The Handwriting of English Documents*. Ilkley: Scolar Press.

**Honkapohja, A.** (2018). ''Latin in Recipes?' A corpus approach to scribal abbreviations in 15th-century medical manuscripts', in *A Multilingual Approach to Language History: New Perspectives on Language Mixing. (Papers from the symposium Historical Code-switching: The Next Step held in Tampere, 11–13 June 2014)*. Mouton de Gruyter.

**Honkapohja, A.** (2017). *Alchemy, Medicine, and Commercial Book Production: A Codicological and Linguistic Study of the Voigts-Sloane Manuscript Group*. Turnhout: Brepols.

**Honkapohja, A.** (2013). Manuscript abbreviations in Latin and English: History, typologies and how to tackle them in encoding. *Studies in Variation, Contacts and Change in English Volume 14: Principles and Practices for the Digital Editing and Annotation of Diachronic Data*. <http://www.helsinki.fi/varieng/series/volumes/index.html> (accessed 18 March 2021).

**Honkapohja, A. and Jones, L.** (2020). From Practica Phisicalia to Mandeville's Travels: Untangling the Misattributed Identities and Writings of John of Burgundy. *Notes and Queries*. 10.1093/notesj/gjz161. 0029-3970.

**Horrox, R.** (1994). *The Black Death*. Manchester: Manchester University Press.

**Jones, L.** (forthcoming). *Time, Space, and the Plague: Rereading English and French Plague Tracts, 1348-1750*.

Dissertation, University of Ottawa (a book based on this dissertation is under contract with MQUP).

**Kestemont, M.** (2014). 'Function words in authorship attribution from black magic to theory?', In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 59–66.

**Laing, M.** (1999). Confusion wrs confounded—litteral substitution sets in early Middle English writing systems. *Neuphilologische Mitteilungen*, **100**(3): 251–70.

**LAEME = Laing, M.** (2013). *A Linguistic Atlas of Early Middle English, 1150–1325*, Version 3.2. Edinburgh: The University of Edinburgh. <http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html>

**Matheson, L. M.** (2005). *Médecin sans Frontières?*: The European Dissemination of John of Burgundy's Plague Treatise. *ANQ*, **18**(1): 19–30.

**Petti, A. G.** (1977). *English Literary Hands from Chaucer to Dryden*. London: E. Arnold.

**Rogos-Hebda, J.** (2018). Text and Image: Revisiting Traube's Halbgraphische objekte in a paleographic-pragmatic approach to scribal abbreviation. *Kwartalnik Neofilologiczny*, Lxv, 1/2018.

**Samuels, M.** (1983). The scribe of the Hengwrt and Ellesmere manuscripts of the Canterbury Tales. *Studies in the Age of Chaucer*, **5**: 49–65.

**Shute, R.** (2017). Pressed for space: the effects of justification and the printing process on fifteenth-century orthography. *English Studies*, **98**(3): 262–82.

**Smith, D.** (2018). The predictability of {-S} abbreviation in Older Scots manuscripts according to stem-final Littera. In Alcorn, R., Kopaczyk, J., Los, B. and Molineaux, B. (eds), *Historical Dialectology in the Digital Age*. Edinburgh: Edinburgh University Press, pp. 187–211.

**Smith, J.** (1996). *An Historical Study of English: Function, Form and Change*. London and New York: Routledge.

**Stutzmann, D.** (2010). Paléographie statistique pour décrire, identifier, dater. Normaliser pour coopérer et aller plus loin? In Fischer, F., Fritze, C. and Vogeler, G. (eds), *Kodikologie und Paläographie im digitalen Zeitalter 2—Codicology and Palaeography in the Digital Age 2*. Norderstedt: BoD, pp. 247–77.

**Sudhoff, K.** (1911). Pestschriften aus des ersten 150 Jahren nach der Epidemie des "schwarzen Todes" 1348. III. Aus Niederdeutschland, Frankreich und England. *Archiv für Geschichte der Medizin*, **1**(2): 58–70.

**TEI P5 Guidelines** <https://tei-c.org/guidelines/p5/>.

**Thaisen, J.** (2011). Adam Pinkhurst's short and long forms. In Thaisen, J. And Rutkowska, H. (eds), *Scribes, Printers, and the Accidentals of their texts*. Bern: Peter Lang, pp. 73–90.

**Thaisen, J. and Da Rold, O.** (2009). The linguistic stratification in the Cambridge Dd copy of Chaucer's 'Canterbury Tales'. *Neuphilologische Mitteilungen*, **110**(3): 283–97.

**Voigts, L.** (1990). The 'Sloane Group': related scientific and medical manuscripts from the fifteenth century in the Sloane Collection. *The British Library Journal*, **16**: 26–57.

**Waley Singer, D.** (1916). 'Some Plague Tractates (Fourteenth and Fifteenth Centuries)', in *Proceedings of the Royal Society of Medicine ix*.

## Notes

1 Doyle (1956) dates the manuscript with unusual precision to summer 1468. He identified the manuscript as 'a little book of physick', copied by the Westminster-based freelance scribe William Ebesham for John Paston (II). Two letters from Ebesham to Paston survive in the collection known as the Paston Letters. Ebesham's letters reveal that he carried out the commission when John Paston was visiting the continent as a part of the wedding retinue of Princess Margaret of York in 1468.

2 Incipits and explicits were excluded from the analysis because of their low word count.

3 Interestingly, the division between lexical and function words has parallels with stylometry, the computational study of writing style. Approaches based on word-level units, the most important of which was the division between content and function words, dominated the field in the 1990s (for an overview, see Kestemont, 2014, pp. 59–61). In stylometry, the function-content division was eventually superseded by *n*-grams (Kestemont, 2014, p. 62).

4 They also include ingredients such as the aforementioned 'pimpernel'. Notably, the English word by all manuscripts, *pimp̨nel* 'pimpernel', is actually attested with two tokens and is abbreviated by all scribes on both of the occasions (the Latin *pimp̨nella* is also abbreviated by all scribes three times). The next section examines how much more common abbreviation is in the text type of recipes.

5 The instance where only four witness abbreviate is caused by the Sloane 2320 scribe writing *ana* without a superscript. The instance where Trinity and Boston expand *recipe* twice and *semis* once may reflect these two being textually close, which is also apparent from shared mistakes not present in others.

6 This involved some further manual tagging. All words which contained this sequence of letters were tagged as yes. If another abbreviation was used (such as *macron* for the final *m* in parum, or abbreviating tempore with a macron—or *aperiatur* with sup-r '-ur', which was typically used to abbreviate the Latin passive suffix *-tur*), the word was tagged as other. Words which did not contain this word sequence were left blank.