UNIVERZITA KARLOVA – FILOZOFICKÁ FAKULTA
ÚSTAV ANGLICKÉHO JAZYKA A DIDAKTIKY

Filologie – Anglický jazyk

Marie Vaňková

ANALÝZA STŘEDNÍ ANGLIČTINY ONLINE: TVORBA A VYUŽITÍ DATABÁZE SPELLINGOVÝCH VARIANT ZALOŽENÉ NA LAEME

ANALYSING EARLY MIDDLE ENGLISH ONLINE: CONSTRUCTION AND USE OF A LAEME BASED SPELLING DATABASE

DISERTAČNÍ PRÁCE

Vedoucí práce: Mgr. Ondřej Tichý, PhD.

2021

Prohlašuji, že jsem disertační práci napsala samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Poděkování

Ondřeji Tichému a Janu Čermákovi za odborné konzultace, všestrannou podporu a důvěru.

To Margaret Laing, Rhona Alcorn, Benjamin Moulineux, Vasilis Karaiskos, Raffaella Baechler and Alpo Honkapohja for a most encouraging and enjoyable time spent in Edinburgh and valuable advice.

Svému muži Jakubovi a celé své rodině za podporu, povzbuzování a nekončící trpělivost.

Abstrakt

Práce se zabývá sestavením a testováním webového nástroje na analýzu textů v rané střední angličtině, vytvořeného z dat dostupných v *Linguistic Atlas of Early Middle English (*LAEME).

Jako základ pro návrh nástroje slouží úvodní teoretický přehled o historicko-lingvistickém výzkumu středoanglických textů, se zaměřením na nářečí a vztahy mezi psaným a mluveným jazykem. Práce dále podrobně vysvětluje metodologii tvorby nástroje, přičemž postupuje od struktury databáze, do níž byla data z LAEME převedena k poloautomatickému procesu zpracování dat a výstupním datům. Zpracování dat spočívalo především v segmentaci jednotlivých variant slov na menší úseky a určení, které segmenty si vzájemně odpovídají. Následně jsou popsány jednotlivé uživatelské funkce nástroje a jejich použití je vyzkoušeno na krátkých analýzách.

Třebaže nástroj vyžaduje rozsáhlejší testování a úpravy, dosavadní testování nebyly objeveny závažnější chyby a nástroj lze označit za použitelný. Podařilo se otevřít nové možnosti (rychlejšího) přístupu k datům z LAEME a nástroj je navíc otevřen možnostem dalšího rozšíření, včetně přidávání zápisových variant slov z dalších období vývoje angličtiny.

Abstract

The present thesis deals with the construction and testing of a web-based tool for analysis of Early Middle English texts, created from the data available in the Linguistic Atlas of Early Middle English (LAEME).

The introductory theoretical overview of research into Middle English texts focuses on dialectology and the relation between spoken & written language and it serves as a springboard for the development of the tool. The thesis further presents a detailed explanation of the methodology behind the tool. It describes the structure of the database containing the transformed data from LAEME and then it moves on to the semi-automatic data processing and types of output data. This processing consists mainly in the segmentation of LAEME spelling variants into smaller units and in determining which segments in a group of variants correspond to one another. The thesis also describes the individual functions available within the tool and tests their use on short sample analyses.

Although more extensive testing and modifications of the tool are required, it has so far revealed no crucial errors and the tool can be described as useable. The project succeeded in opening new possibilities of faster access to LAEME data. Furthermore, the tool is prepared for future upgrades, including the addition of data from other periods of development of the English language.

Table of contents

List of tables

List of figures

Abbreviations and notation

CoNE	Corpus of Narrative Etymologies
DB	database
EME	Early Middle English
LAEME	Linguistic Atlas of Early Middle English
LALME	Linguistic Atlas of Late Middle English
LME	Late Middle English
LP	Linguistic Profile
LSS	Litteral Substitution Set
ME	Middle English
OF	Old French
PDE	Present Day English
PrOE	Proto-Old English
PSS	Potestatic Substitution Set
SWM	South-West Midlands

Notation

LAEME files are referenced by their number preceded by a hash, e.g. (text) #8. The index of files mentioned in the thesis is available in appendix 7.1.

References to lexical items are given in the format used in CoNE. The attested form written in italics is followed by lemma/lexel in capital letters and optionally by translation given in quotes, e.g. *ȝeld* GIELDAN "yield, pay", *þoruh* THROUGH. If a specific *item* (see subchapter 3.6) is referenced, the *lexel* is followed by a specification of word class (partial *grammel*) separated by a slash, e.g. *will* WILL/N. Additionally, specific position in the item (slot) can by given in brackets, e.g. STONE/N (3).

Sets (see subchapter 3.6) are enclosed in curly brackets, e.g. {*f*, *v*}

Single characters (litterae) or digraphs are written in italics, e.g. *c*, *k*.

Presumed sound values are written in square brackets and standard IPA characters are employed, e.g. [s], [ɣ].

Pieces of code or LAEME source data are marked by `the courier font`.

References to electronic sources

Introduction to LAEME: “Laing & Lass 2013” plus subchapter number, e.g. Laing & Lass 2013: 2.1.

LAEME Grammel commentary: “Laing 2013: Grammel commentary”

Corpus of Narrative Etymologies: “CoNE, change label”, e.g. CoNE, PH

1. Introduction

The purpose of the present project is to use data available in *The Linguistic Atlas of Early Middle English* (LAEME) to develop a research tool for the analysis of Early Middle English spelling and dialects. The tool is conceived as a spelling database and an interface designed specifically to access data therein. The construction of the database consisted in processing LAEME data to allow comparison of spelling variants at the level of segments (slots) rather than morphemes. The interface offers a simple search tool for the database as well as more sophisticated ways of data presentation, mainly a mapping tool adapted to the new database structure.

The design of the tool seeks to respond to the problems of research into Early Middle English, which is notorious for the extreme level of spelling variation (Black, 1999; Laing & Lass 2013). Its aim is to devise new, a possibly faster, ways of exploring LAEME data, which is exceptionally rich and well organised, but not always easy to search.

1.1. The structure of the thesis

The theoretical part of the thesis discusses theoretical and methodological issues connected with research into Middle English as well as methodological concepts and principles considered relevant for the construction of the tool. It also presents several electronic resources and projects which share certain features with the new spelling database.

The methodological chapter explains the process of transforming LAEME data into the new "segmental" database and describes its structure. It also defines the structure of data retrievable from the database and comments on the envisaged approach to querying. The chapter "Results" briefly discusses problematic forms which were difficult to process, and then it moves to the description of the user interface and its various functions and features. The use of the tool is subsequently demonstrated on a series of practical examples and the chapter concludes with a commentary on the strong and weak points of the tool and selected issues of a more theoretical nature.

Similarly to other fields, research into the history of English has reached a stage where the major developments have been described but there is still space for more detailed and comprehensive analyses, which exploit electronic processing of data. This course of research should hopefully lead to a deeper understanding of the development of English as well as our knowledge of language in general.

2. Theoretical background

Descriptions of the development of English in the "Early Middle" period will always be based on a limited amount of surviving material, which seems sparse in comparison with the later stages and highly chaotic in comparison with Old English.

The central question underlying the following theoretical discussion is how to approach the sources in their complexity to maximally exploit LAEME data as well as the possibilities of electronic processing and construct a useful research tool. This question suggests structuring the chapter in a manner that reflects the progression from a maximally realistic description of the linguistic reality to theoretical models and methodological observations which in turn serve as a basis for the construction of electronic resources and tools.

Accordingly, the theoretical chapter is divided into three subchapters. The first one outlines the present state of research into Early Middle English, moving from a general introduction and extralinguistic context to a more focused discussion of specific areas calling for attention of scholars. These include the problem of relations between spoken and written language, our present knowledge of scribal practice, the nature of linguistic change and previously described phonological changes. The aim of the chapter is to summarize the most relevant findings and theories which can inform improvements in research methodology.

The second subchapter briefly introduces the field of historical dialectology and explains the main methodological challenges of research into Middle English texts. It links the theoretical observations from the preceding chapters with typical problems faced by a historical dialectologists. The chapter also surveys specific principles and methods proposed by researches to address the issues. The aim of the chapter is to establish continuity of the methodology of the present thesis with previous research and identify methodological problems to which the thesis should or could respond.

The third subchapter continues the methodological strand and it focuses specifically on the use of electronic resources in research. It shows how the methods and models from the previous chapter are applied to create specific atlases and corpora. Its concluding part assesses the potential of computers in research into Middle English and it identifies the most useful procedures and techniques, which deserve to be incorporated in electronic tools in the future.

2.1. Theoretical problems of research into Early Middle English

Early Middle English (EME, *ca* 1150-1300) as a stage of development stands between Old English and Late Middle English and some sources describe it merely as a "transitional phase" (Black, 1999: 155). The accounts which try to identify the distinctive characteristics of EME typically mention its extreme dialectal diversity (Corrie, 2006: 86) or irregular disorganised spelling (Black, 1999; Laing & Lass, 2013; Faulkner, 2019; Smith, 2020 etc.).

The Linguistic Atlas of Early Middle English (LAEME) as well as the spelling database, which is the subject of this thesis, can be regarded as attempts at counterbalancing the irregularity of Early Middle English material with well-designed research tools, exploiting the potential of electronic processing. Obviously, such attempts would be impossible without a sound understanding of how the diversity originated and where to look for regularity behind the chaos on the surface, which is the topic of this subchapter.

2.1.1. General introduction and extralinguistic context

The term *Early Middle English* as employed in LAEME is applied to the English language in the period ca 1150-1325 (Laing & Lass, 2013: 1.5.1). Although a hundred years separate this period from the Norman Conquest, the effects of this crucial historical event "played an important part in some of the developments which shaped the form of Middle English. The most prominent one is a rupture in the writing tradition resulting in marked differences between *written* English of the 11th century and that of the 13th" (Vaňková, 2016: 9).

The new rulers, who spoke Norman French, had littles interest in using English (Kohnen, 2014: 72; Upward & Davidson 2011: 66) and the loss of institutional support resulted in the decline of the "West Saxon spelling standard" (Upward & Davidson, 2011: 68). This means that scribes were not longer trained to use a relatively fixed spelling system and they "naturally used their own dialects. On top of that, they had no standard spelling which they could follow and had, for that reason, to rely on their own intuition when transforming spoken English into the written form" (Vaňková, 2016: 10). As a result, we often find numerous and diverse EME forms of a single word in place of comparatively fewer and more homogenous OE forms.

The "bright side" of this situation from the linguistic perspective is that spelling became a closer representation of the spoken language. In fact, some of the phonological changes which must have affected English a long time before became detectable in EME extant texts (Smith, 2007: 34). Horobin & Smith (1999) explicitly speak of the Middle Ages as a period of relatively "close correlation between spoken and written language" (Horobin & Smith, 1999: 362).

Hopeful, as this might sound, this brief characteristic of written EME essentially states that scribes were likely to rely on their ears and native dialect when writing. What it cannot tell us is what exactly were the resources available to each scribe, for instance, whether s/he was used to reading Old English texts or whether s/he was trained to write Anglo-Norman, Latin or both. The latter is especially relevant as it has been shown that Anglo-Norman scribes introduced some innovations into the spelling system, which triggered changes in its structure (see Upward & Davidson, 2011: 68; Fisiak, 1986: 15).

To summarize, Middle English could be characterised as a stage of "a striking lack of uniformity in the employed spelling systems" (Vaňková, 2016: 10). As for dialectal diversity, it may be more precise to state that compared to OE, EME provides *better material* for the study of dialects, while actual differences in speech were perhaps less marked that the surface variation in spelling might suggest. It has been noted that the distinction between spoken and written language is vital in research into EME dialects, because we need to carefully distinguish between dialectal differences (sound changes) and differences in the spelling systems. While detailed analyses of the spelling systems are necessary for sound reconstruction, knowledge of the previously described sound changes and dialectal differences (albeit based on incomplete data) are necessary for a correct understanding of an individual spelling system (see e.g. McIntosh et al., 1989; Laing & Lass, 2013).

The interactions between spoken and written language have attracted the attention of scholars for decades and a considerable body of knowledge has been accumulated. The rest of this chapter presents the main findings and theories considered relevant for the present thesis. Some of these findings are going to be referred to further on in connection with methodological issues. The relation between written and spoken language and scribal approaches to copying will be discussed first. The following section will cover the problem of the spread of changes as well as selected phonological and orthographic changes.

2.1.2. Written language and scribal practice

This subchapter addresses problems connected with the nature of written evidence in ME and the relationship between spoken and written language. The issues are examined mainly from the theoretical perspective. Methodological implications connected with the actual use of the written data as evidence are going to be discussed in the subchapter about historical dialectology.

The subchapter opens with a general theory about written language. Its second part presents more specific theoretical models proposed for analyses of spelling systems. The final part summarizes useful findings in the field of textual transmission and scribal practice, which provide important contexts for the interpretation of written ME sources and their linguistic systems.

2.1.2.1. Written language

It may appear natural to perceive written language as "a mere veil blurring the actual constitution of language facts" (Vachek & Luelsdorff, 1989: 103). This might not be a major obstacle to synchronic research, however, historical enquiries in which written texts are our only source require a greater attention to the role and use of the written medium.

The treatment of written language as a partly autonomous system has been present in the work of historical dialectologists based in Edinburgh as well as in the writings of the Prague school. The earliest papers on the topic were written by Artymovyč in 1932 and his ideas were further developed by Vachek in several papers from 1939-1973. Vachek discussed the problem from the functionalist perspective (Vachek & Luelsdorff, 1989: 92-93).

Angus McIntosh published a paper on the topic as early as the mid 20th century (McIntosh et al., 1989). He highlighted the importance of studying the written form of English as a system in its own right and he criticised the treatment of written language as something "inferior" to spoken language (e.g. Bloomfield 1933 as cited in Linell 2019: 3). Such disregard for writing used to be the weakness of older studies of ME dialects which focused primarily on reconstructing the sound of older English and regarded variation in orthography as unimportant.

McIntosh introduced two useful terms to clarify the relationships between spoken and written language: *correlation at the phonetic level* and *systemic correlation* (McIntosh et al., 1989: 2-3). The first expression refers to a correspondence between a specific sound and a specific symbol and it is described by statements like "*s* stands for [s]". *Systemic correlation* occurs if a contrast between symbols corresponds to a contrast between sounds. For instance, if the medial vowel in *libbe* (*live*) is different from *habbe* (*have*), it is reasonable to assume that the pronunciation of the two differed as well.

It is important to note, however, that parallel variation in the two systems does not always imply correlation. This can be exemplified by the varying pronunciations of /p/ as opposed to varying shapes of *p*. The spoken and the written system behave *analogically* in that there is variation but none of the different shapes of *p* can be related to a specific pronunciation, i.e. we cannot speak of correlation in this case (McIntosh et al., 1989: 11).

The natural but unjustified tendency to always expect correlation between spoken and written language can lead to statements like "the text fails to reflect the difference between long and short vowels", but such statements are "misleading" according to McIntosh who claims that the fact that texts do not reflect all the features of spoken language is perfectly natural (McIntosh et al., 1989: 11).

McIntosh does not forget to stress that greater attention to orthography should eventually lead to better results in the field of historical phonology, stating that "written texts will always be ransacked for information about spoken language and they can be the more fully exploited to this end the more carefully we explore the nature of their relationship to their spoken equivalent" (McIntosh et al., 1989: 7). Although the approach advocated by McIntosh, but also other scholars like Vachek is now widely accepted, Smith (2020) has recently evaluated the study of writing as "surprisingly under-researched" (Smith, 2020: 14).

2.1.2.1.1. Spelling systems

It hardly seems surprising that the treatment of written language as an independent system has a prominent place in the work of the authors of LAEME Laing & Lass (2013: 1.4). The theoretical bases of their approach are described in the Introduction to LAEME. The core concept employed to characterize written language is *spelling system*, which is defined as "mapping of some chosen set (or sets) of linguistic units into a set of visual signs" (Laing & Lass, 2013: 2.2.1). The definition deliberately avoids speaking about "correspondences between written and spoken language", which would be imprecise because written symbols do not always correspond to sounds. They can also correspond to other linguistic units like morphemes or words. It is possible to distinguish between several types of correspondences between written language and linguistic units, which can coexist in a single system. In Smith's words, "the two levels of language do, even if in a complex way, map back onto the 'same' language" (Smith, 2007: 35). The differing "levels of representation" (Laing & Lass, 2013: 2.2.1) allow us to distinguish between "phonographic" systems, representing at the level of phoneme, and "logographic" systems, representing at the level of words (Smith, 2007: 31).

It might seem that an ideal spelling system would represent at the level of sounds and the representation would be bi-unique, i.e. each grapheme would correspond to one phoneme (Laing & Lass, 2013: 2.2.1.; Smith, 2007: 33), but this is virtually never the case, even in languages with high level of correlation like Czech (cf. also Vachek & Luelsdorff, 1989: 96). Moreover, *logographic* systems are not without their virtues, as they can enable communication between speakers of language varieties whose spoken forms are mutually unintelligible. The frequently quoted example of this is the writing system of Chinese, but in rare cases also PDE (Sebba, 2007: 110).

2.1.2.1.2. Laing & Lass' classification of writing systems

Laing & Lass prefer to speak about *logography* as a principle rather than "logographic systems" and they offer a somewhat finer classification of "supra-phonemic levels of representation" (Laing & Lass, 2013: 2.2.1):

Logography refers to correspondence at the level of the word, i.e. a sequence of characters represents a word but the sounds in the words cannot be easily linked to the individual characters one by one. *Logography* is abundant in PDE.

"*Morphography* refers to representation on the morphemic level; in other words, the string of characters does not represent a specific sequence of phonemes but a morpheme, which can be pronounced differently in dependence on its position (Laing & Lass I, 2013: 2.2.1)." (Vaňková, 2016: 21). Vachek (Vachek & Luelsdorff, 1989) points out that easier identification of morphemes resulting from this kind of representation in fact makes the system more efficient (Vachek & Luelsdorff, 1989: 97).

Yet another kind of representation is found in abbreviations and icons, which are devices commonly employed by medieval scribes. Laing & Lass (2013) point out that if we view the levels of representation as a cline with "pure" sound-to-spelling mapping at one end, abbreviations and icons would stand at the other end, i.e. the offer very little, if not no "phonological clues" (Laing & Lass I, 2013: 2.2.1), which would allow to grasp the sound. It might be said that icons point directly to concepts, just as spoken words do.

Diacritics and doubling of letters

Practices like the use of diacritics or doubling of letters to indicate length are also strategies which do not represent at the level of the phoneme. The most common phenomena from this category in EME are: " (a) doubling of consonants to indicate that the preceding vowel is short;

(b) doubling of vowels to indicate length; (c) the use of accents on vowels to indicate their quantity.“ (Laing & Lass, 2013: 2.1). The common practice of scribes, whereby bi-unique correspondences between units are not maintained is termed *literal substitution*. This concept is going to be described shortly.

2.1.2.1.3. Development of written language

The characteristics of spelling systems presented above are clearly connected with the problem of writing tradition. The nature of correspondences between writing and linguistic units change as the system of written language develops. In the words of Vachek, “in its very first beginnings written utterances were hardly more than signs of the second order” and “they constituted very primitive quasi-transcriptions of the phonic make-up of the corresponding spoken utterances” (Vachek & Luelsdorff, 1989: 95). Writing systems of this sort require only a relatively small inventory of symbols and more or less common idea of their “value”. This is of course advantageous for the establishment of a new writing system. Vachek (Vachek & Luelsdorff, 1989) further claims that there is a natural tendency for “written utterances” to become “symbols (…) of the first, not just of the second order” (Vachek & Luelsdorff, 1989: 98). Thus, *logography* arises only with a certain continuity which allows the members of the community to “learn” the extra correspondences between visual signs and higher linguistic units and perpetuate them. In fact, the acquisition of writing skills in our time often consists in learning to write something different from what we hear. The association of apparent “mismatches” between the written and spoken systems with history and past stages of the language is actually reflected in the terms employed to describe such phenomena, e.g. “historical residues and conventionalisations” (Smith, 2007: 32), “fossil distinctions” (Lass, 1997: 57 as cited in Smith, 2007:34) or “ghost contrasts” (Laing & Lass, 2013: 2.2.1).

The usual pattern of divergence of writing from pronunciation is that writing remains stable, while pronunciation changes. The opposite is attested for the Germanic runic alphabet Futhark, where a change in pronunciation motivated a modification of the runes (Smith, 2007: 33). The transfer of runic *wynn* or characters like *edh* into a different script might be regarded as similar in principle.

2.1.2.1.4. Commentary

The claim that written language is not a mere “reflection” of spoken language is justified by the fact that strings of written symbols may represent linguistic units directly, even though the system usually involves sound-to-symbol mapping. Moreover, written language can develop independently. The natural tendency in the development of written language, evidenced also in English, is to move from a system close to transcription to representation on the level of higher units.

The principle in ME seems to be predominantly alphabetic spelling which has since evolved into a system with a strong logographic component. Early Middle English is definitely one of the periods of development which Vachek (1989) considered especially difficult to research because of the "notoriously smaller stability of the written norm (…) with all its numerous differentiations, regional as well as individual" (Vachek & Luelsdorff, 1989: 119). On the other hand, the process of "re-establishment" of written English in Early Middle period appears to be unique in the history of the language and its special character should hopefully be worth overcoming the difficulties in research.

Both written and spoken languages are to some extent independent systems and they share a number of traits like variation and the importance of oppositions. The symbols used for representation are arbitrary in both cases, which Smith (2007) aptly expressed by comparing letters to currency. He states that people "have simply agreed, as they do when assigning values to money (coins, paper), to assign sound values to particular symbols" (Smith, 2007: 31). The parallelism of spoken a written language invites considerations of what our analyses of written language may contribute to our understanding of language in general (Vachek & Luelsdorff, 1989: 100). All of this entails that the study of written language deserves its own framework. The next section discusses specific models proposed to describe spelling systems.

2.1.2.2. Models of the writing system

Michael Benskin (1997) and his colleagues responsible for the creation of A Linguistic Atlas of Late Medieval English (LALME) propose to use the model of *litterae* as a framework for dealing with ME spelling systems. Laing (2013) uses the same framework, avoiding the use of structuralist concepts like *grapheme* and *phoneme*. The main reason for this decision explained in Laing & Lass (2013) is that "such concepts do not always characterise what our scribes appear to be doing", which is why the authors prefer "to use a theoretical framework and notation that cohere more closely with what scribes would have experienced in their education" (Laing & Lass, 2013: 2.3.1).

The terminology is based on the 5th century Latin work of Aelius Donatus *Ars maior*, which was presumably used in training of the scribes. Donatus defines the term *littera* as follows:

> Littera is the smallest unit of articulated sound ... littera is (a) sound which is capable of being written alone ... littera has three properties: name, shape, power [= sound value]. For one must ask what the littera is called, what its shape is, and what its power is. (Laing & Lass, 2013: 2.2.1)

Within the framework, *littera* is an "abstract object" and "the stream of litterae in writing is represented by a sequence of figurae" (Laing & Lass, 2013: 2.3.1), i.e. letter shapes. According to Donatus, each *littera* may have one *potestas* but more *potestates* are allowed in the proposed framework. *Potestas* refers to the sound. For instance, lowercase *f* and uppercase *F* were two different *figurae* of the same *littera* having a few possible *potestates* including [f] and [v]. The possibility to have multiple *potestates* for one *littera* is a deliberate adaption of the original theory, which Laing & Lass (2013) justify by claiming that it was common for medieval scribes to have multiple *potestates* for one *littera* and vice versa. Arguably, this decision somewhat weakens the justification for using a framework close to "what scribes would have experienced in their education" (Laing & Lass, 2013: 2.2.1.), because the rather practical requirement to have only one *potestas* for *each* littera seems to be a vital element of the theory. Still, as the authors themselves claim, the framework remains useful despite apparent differences between medieval and antique practices (Laing & Lass, 2013: 2.3.2).

"In order to create space for the treatment of variation in EME spelling, Laing & Lass extended the model with two new concepts. A *Litteral Substitution Set* (LSS) is a set of *litterae* which may be used to represent a given *potestas*. A *Potestatic Substitution Set* is a set of *potestates* which may be assigned to a given *littera* (Laing & Lass, 2013: 2.3.2)" (Vaňková, 2021: 5).[1]

The authors of The Middle English Grammar Project (to be discussed in section 2.3.3) reference the model of *litterae* and *potestates* in connection with their own model, which is similar in making a three-way distinction between its elements. Their model distinguishes between *letter*, *grapheme* and *realisation*. *Letter* is practically coreferential with *littera* and

[1] „A similar model is found in McLaughlin (as cited in Fisiak, 1986:13). The central term in this model is *fit*, which refers to the "relations between graphemes and phonemes" (Fisiak, 1986: 13). *Graphoneme* roughly corresponds in meaning to the *literal substitution set*. Thus, a *graphoneme* is a set of symbols each of which is called an *allographone*. In a *simple graphoneme*, one phoneme is represented by one grapheme, while in a *complex graphoneme*, there are more graphemes which may represent the same phoneme. This is a distinction analogical to the biunique/non-biunique representation discussed above." (Vaňková, 2016: 22)

realisation is a label for an individual instance of a *letter*. *Graphemes* are defined relative to one another based on contrasting sound values. For example, *w* and *p* can be two different letters with the same value, while <w> and <d>[2] are two different *graphemes. Grapheme* is in fact closest to the concept of *litteral substitution set*, because *letters* sharing the same value can be "assigned to a single *grapheme*" (Stenroos, 2004: 263). The definition of *graphemes* thus implies differing *potestates* but there is no direct equivalent of *potestas* and therefore no need to assign explicit (albeit approximate) sound values.

It is worth noticing that the association of the abstract *littera* with multiple *potestates* is reminiscent of the association of an abstract concept with multiple possible referents, as described in the structuralist model of the sign. Similarly, the mechanism behind the *logographic principle* can be also discerned in compounds and fixed phrases on higher levels of language, whereby the individual components lose their independent meaning and the compound is interpreted as a whole. In the case of collocations like the typical "utterly impossible" as opposed to anti-idiomatic "utterly beautiful" it is tempting to think about the phrase that the latter could be "equally well formed" (Laing & Lass, 2013: 2.2.1), mentioned in connection with *bright* and **brite*. These analogies support the understanding of written language as a sign system.

2.1.2.2.1. Classification of spelling systems

Within the model of *litterae* and *potestates*, "spelling systems may be characterised either as *economical* or *prodigal*. *Economical* systems are relatively close to the biunique representation (one *littera*, one *potestas*), while *prodigal* systems have a number of "unnecessary" correspondences (one *littera* for several *potentates* and vice versa (Laing & Lass, 2013: 2.3.2)" (Vaňková, 2016: 23). Despite the fact that such systems may appear chaotic due to the multiple non-biunique relations between *litterae* and *potestates*, it is important to bear in mind that the variation is not completely random (Laing & Lass, 2009: 30). In other words, the usage of a specific scribe in a specific copy usually has a somewhat internally consistent linguistic system, potentially different from systems of other texts. Such a text-specific system is called a *text language* and it has a similar role as a single live informant in synchronic dialectology (Laing & Lass, 2013: 1.1). This implies that the sounds repreented by a single littera may differ from text to text, therefore it is essential to consider each *text language*

[2] The brackets reflect the original conventions exaplained in Stenroos (2004: 263).

separately. An equally strong emphasis on the assumption of internal consistency is found in Black, Horobin and Smith (1999). The next part of the text deals with two concepts related to the development of correspondences between sounds and symbols – litteral substitution and speech segmentation.

2.1.2.2.2. Litteral substitution

"Prodigality" in the spelling systems can be "a product of intricate interactions between the scribe's interpretation of the symbols in his exemplar or other texts, which he has read, and his approach to copying. Assumed "meanings" of litterae can shift in similar ways as meaning of words do and multiple relations between sound and spelling develop. Such developments were explored by Laing & Lass (2009), who proposed several scenarios whereby multiple relations between sound and spelling originate" (Vaňková, 2021: 5). The general mechanism they describe is the so-called "extension" of literal substitution sets (Laing & Lass, 2009: 21), i.e. the addition of a new littera to a LSS. It is possible to distinguish between two kinds of extension, which differ in their motivation.

The first kind is based on similarity of letter shapes. Laing & Lass specifically mention the fact that *y/þ* and *þ/p* are indistinguishable in some manuscripts. Consequently, their functions can become "confused" (Laing & Lass, 2009: 3). The previously distinct litterae become members of the same LSS, i.e. both can be used to represent the same sound. The second kind is motivated phonetically. Laing & Lass (2009) give the example of change in spelling for OE intervocalic -*g*- from ȝ to *w*/*p*, reflecting the vocalisation of OE [ɣ]. "There are also 'mixed' cases in which combinations of phonological and orthographic change trigger alterations of sound/symbol mappings, creating what might be called "floating figurae" which are "'unanchored' from their original potestatic moorings and can therefore be redeployed.." (Laing & Lass, 2009: 16).

A specific motivation for spelling change is that the sound change produces an "intermediate" sound and if the scribe relies mainly on his ears, he finds none of the available symbols to be an adequate representation of the new sound, but he is nevertheless forced to choose between them.

Substitutions can be combined into sequences (Laing & Lass, 2009: 22), which can be invoked as explanations of a specific spelling variant. For example, the following explanation is given by Laing & Lass for the spelling *swo* (SHOE/N):

> ‘sh’ (beside usual ‘sc’) may represent [ʃ]; there is ‘þ/h’ substitution making ‘sþ’ theoretically possible for [ʃ]; via the postulated exemplar system, ‘þ’ and ‘ƿ’ are interchangeable (…therefore sƿ- is a possible spelling for [ʃ]; with substitution of <w> for <þ/ƿ>, sw- is a possible spelling for [ʃ] (Laing & Lass, 2009: 22).

The discussion of literal substitution underscores the instability of the mappings of symbols to sounds perceived by the scribes. On the one hand, this instability again calls for cautious interpretation of the symbols in phonologically oriented analyses. On the other hand, it invites research into changes in the writing systems in the period of little institutional regulation, which would otherwise act as a restrictive factor in their development.

2.1.2.2.3. The problem of segmentation

The previous section illustrated potential volatility of EME spelling systems, focusing on the links between sounds and symbols. This section briefly discusses the problem of speech segmentation, which can be another source of instability. The topic is of course particularly relevant for the present thesis because segmentation was the core procedure in the construction of the database.

Writing systems which do not represent at the level of higher units like words or morphemes by definition require segmentation of speech flow into separate units. The segmentation was regarded as objective until the 1930s, which of course partly shaped phonetic research of the time. The question whether speech segmentation is unequivocal or not remained a matter of debate until 1950. A major contribution to solving the dilemma, which may seem rather obvious today, was the stress on differences between so-called explicit (*lento*) and implicit (*allegro*) style of pronunciation proposed by Jakobson and Halle in 1956. Segments are clearly distinguishable only in the explicit style, i.e. slow and careful pronunciation (Vachek & Luelsdorff, 1989: 37-38).

In the light of these findings, it is reasonable to assume that medieval scribes faced with the task of segmentation did not always have the perfect explicit models, which means that spellings variants may differ even at the level of segmentation. For instance, a scribe might have used a single littera to represent what another scribe perceived as two segments.

Litteral substitution and speech segmentation both relate to the inner structure of a spelling system and their variation. Black, Horobin and Smith (1999) specify that “the variation which characterizes the set of Middle English spellings correlates with a range of definable factors” (Black, Horobin & Smith 1999: 14). Our present knowledge of such possible factors is going to be the dominant theme of the subchapter about scribal practice.

2.1.2.3. Scribal practice

Each text language is shaped by the linguistic resources available to the scribe, such as his native dialect, his perception of the sounds or knowledge of certain spelling conventions. To complicate matters further, Hudson (1966) points out that the oral dialect of the scribe does not necessarily correspond to his "written dialect", i.e. the scribes might have retained certain written forms which they would not have used in speech (Hudson 1966, 371-372). Another source, standing slightly apart from the others is the text language of the exemplar. The extent to which the scribe relies on his individual resources as opposed to the source text depends on his approach to copying or *scribal strategy*. While tracing the influence of the scribe's dialect as opposed to his reading is virtually impossible, inferences regarding scribal strategies can be made if there are multiple copies of texts in a single hand (see Laing, 2004).

2.1.2.3.1. Scribal strategies

It has been noted that copies of ME texts can display a mixture of the scribe's usage and variants from his exemplar(s). The ratio of forms from these two sources depends on the approach of the copyist. There basic types of scribal practice have been described, two of which were noted by Angus McIntosh: *translating*, *literatim copying* and *partial translating* (McIntosh as cited in Laing, 2004: 52). "A *translating* scribe converts the language of the exemplar into his own dialect. A *literatim* copyist transcribes the text word-for-word, preserving the dialectal features of the exemplar" (Laing, 2004 as paraphrased by Vaňková, 2016: 30). It is assumed that *literatim* copying originated with scribes "trained to copy Latin texts, such as Biblical texts, where the language was fixed and variation was not an option" (Horobin, 2010: 17). The result of *partial translating* is *Mischprache* – "linguistic output containing two or more elements that are mutually incompatible: that is, from non-contiguous areas within the established dialect continuum" (Laing & Lass, 2013: 1.4). Previous research suggests that EME scribes often copied texts literary rather than trying to "translate" them (Laing & Lass, 2013: 1.5.6). Although scribes can hardly be labelled as "pure translators" or "pure literatim copyists" the distinction between the approaches provides a useful conceptual framework for analyses.

It is important to note that variation does not automatically imply exemplar influence. A rare piece of evidence illustrating a certain randomness on Medieval writings was presented by Brook (1972), who analysed a short passage in MS Cotton Caligula A.ix (containing Laȝamon's

Brut) which the scribe accidentally copied twice and identified a number of differences between the two versions. The results made him express deep scepticism about the value of copies as evidence for the language of the exemplars and concluded that "a Middle English manuscript could contain a large number of spelling variations that were not due to the participation of a number of scribes writing in different dialects" (Brook, 1972: 28).

2.1.3. Changes, their progression and spread

The previous subchapters have been at least indirectly concerned with variation in written language in the context of individual spelling systems. The present section focuses on the nature of change in language, which is of course inherently connected with variation. In fact, an underestimation of the role of variation in language had been a major obstacle to our understanding of change for decades. It was associated with a misleading idea of language as a unified system shared by everyone in the speech community, which persisted until the 1970s (Aitchison, 2002: 42).

The concept of "one language" and the Saussurean *synchrony-diachrony* dichotomy were replaced with a more realistic account, i.e. that each of the individual speakers develops his own linguistic system. The "common language" is then nothing more than an overlap of the individual systems. In the words of Charles Lyell, "species are abstractions, not realities – are like genera. Individuals are the only realities" (Lyesll as cited in Lass, 2006: 30). The reality of individual systems has important implications for the diffusion of changes. All changes must necessarily spread from speaker to speaker, which means that there is no strict division between their progression in space as opposed to time.

It has been proposed to analyse this complex situation within the theoretical framework of *complex adaptive systems*, which was adopted by Ogura & Wang (2004) in their article about dialectology. *Complex adaptive systems* may be described as

> Systems made up of a large number of entities that by interacting locally with each other give rise to global properties that cannot be predicted or deduced from an even complete knowledge of the entities and of the rules governing their interaction (Ogura & Wang, 2004: 137).

The framework originated in physical and biological sciences and its use in historical linguistics was advocated by Kretzschmar (2015). His article, among other things, shows that several aspects of the framework in fact coincide with concepts already employed in linguistics, such as variation, Zipf's law, S-curve or Hopper's (1987) description of *grammaticalization* as an ongoing movement towards structure which is never complete. Another concept which is not explicitly mentioned by Kretzschmar but essentially responds to the same properties of language is the model of language centre a periphery (Daneš, 1966; Vachek, 1966). Kretzschmar himself states that

> The process at work in complex systems just explains better what we already knew: we tend to talk like the people nearby, either physically near or socially near, or both, and we tend to use the same linguistic tools that others do when we are writing or saying the same kind of thing (Kretzschmar, 2015: 281).

The contribution of complex system to historical linguistics thus seems to be mainly a matter of incorporating the previously developed models within a larger framework, possibly refining them and pointing out connections between them. Also, it can be a useful platform for interdisciplinary discussion of principles which language shares with other phenomena.

The special relevance of Kretzschmar's (2015) article for the present thesis is due to the fact that he openly challenges some of the methodological aspects of the construction of LALME (LAEME), which is going to be mentioned in the next subchapter. A number of observations made by the authors of LAEME in fact perfectly fit the theory of complex systems. For example, the following quote from the introduction to *Methods and Data in Historical Dialectology* essentially applies the concept of *scaling* to linguistic data: "an encoder whose collocation may seem peripheral in geographical terms may in fact prove quite central in social terms" (Dossena & Lass, 2004: 8). *Scaling* describes the distribution of variants in various subsystems, which would mostly correspond to dialects and registers in language.

2.1.3.1. The nature of sound change

Besides general theories of change, a number of theories focus on changes on a selected level of language. The present discussion is limited to a very short overview of two useful concepts related to sound change.

The study of sound change was a major concern for the Neogrammarian movement. The dominant feature of the Neogrammarian view of sound change was so-called *regularity hypothesis* (McMahon, 1994: 19), i.e. the assumption that precise rules can be formulated, which account for sound change, operating without exceptions. Apparent irregularities result from imperfections in the rule and disappear once the rule is formulated properly. The rules typically describe a change of a segment into another segment in certain phonological contexts and the change is supposed to act gradually and simultaneously in all the concerned words.

More recent theories introduced the term *lexical diffusion* to account for situations when a sound change (definable in Neogrammarian terms) seem to affect a limited number of words, which it is expected to affect according to the rules (McMahon, 1994: 47). If *lexical diffusion* is combined with the definition of "rules", changes can be described in terms of the affected segments and contexts in conjunction with list of specific words affected by the change.

Similar definitions of sound change are going to be used in the following section dealing with previously described developments in Middle English.

2.1.4. Phonological and orthographic developments in Old and Middle English

This subchapter discusses previously described phonological and orthographic developments which took place in the ME period as well as some of the OE changes which are considered relevant for the data in LAEME. The subchapter is based predominantly on the Corpus of Narrative Etymologies (CoNE). The special relevance of CoNE for the present project is going to be described in detail in the final subchapter of the theoretical part (see section 2.3.2.2).

Changes affecting the inventory of litterae are briefly summarized in the table below, which presents the inventory of graphemes available to the scribes at the beginning of the ME period and the inventory at the end of the 14th century.

OE	a	æ	b	c	d	ᵹ	h	i	k	l	m	n	o	p	r	s	t	þ	ð	u	ƿ	x	y	z	q				
ME	a		b	c	d	ȝ	h	i	k	l	m	n	o	p	r	s	t	þ		u		x	y	z	v	j	g	v	w

Table 1: Grapheme inventories in OE and ME (based on Fisiak, 1986:14)

As can be seen from the table, Middle English alphabet underwent several changes. While *æ*, *ð* and *ƿ* gradually went out of use, *j*, *g*, *v*, and *w* were added. Insular ᵹ developed into ȝ. Changes in the inventory of litterae were naturally accompanied with shifts of sound-spelling correspondences. The individual developments are going to be explained along with phonological changes.

The following overview focuses mainly on changes affecting consonants, because vocalic changes turned out to be less relevant for segmentation. The changes are grouped according to specific segments (phonemes or graphemes/litterae) as this arrangement is the most convenient one for a potential discussion of possible links (or confusion) between the changes. Obviously, there are multiple overlaps and connections between the groups.

2.1.4.1. Nasals

All CoNE changes concerning nasals are cases of segment deletion, weakening or addition. This section presents four such changes plus consonant insertion conditioned by a neighbouring nasal.

Final Nasal Neutralisation (FNN) accounts for the loss of the [m]/[n] contrast in the final position which is predominantly an OE development, although LAEME does contain some instances of preserved final [m] as well as a few instances of [n] becoming [m]. FNN could be further followed by Final Nasal Deletion (FND). Final nasals are known to be phonetically unstable and this factor operates together with morphological ones. As word-endings get affected, the change contributes to the disintegration of the paradigms, which in turn devoids final *-n/m* of morphological significance and makes it prone to loss. FND also has observable lexical conditioning (CoNE cites several lexels, including *weapon* and *burden* as examples of words which preserve final *-n*).

The loss of final [n] may also occur in the coda of weak syllables, which is discussed separately in CoNE under Weak Syllable Nasal Deletion (WSND), which sometimes operates along with the vocalic change Weak Vowel Neutralisation (WVN).

The fourth change labelled Nunnation (NN) consists in the addition of the so-called *parasitic -n* to the end of the word. Nunnation seems to be a rare feature found in the two samples of Laȝamon A (texts # 277 and #278 in LAEME). CoNE states that despite their similarity, NN can be distinguished from the analogical extension of the paradigms (cf. AE in CoNE).

The last two developments to be mentioned here are associated with nasals only indirectly. With Post-Nasal Stop Epenthesis (PNSE), nasals in consonant cluster can trigger the insertion of a consonant, e.g. LAEME *drempte* (DREAM). This reflects a cross-linguistic tendency and it is attested also in PDE. A similar phenomenon occurring word-finally is labelled Final Consonant Exerscence (FCE). This change differs from PNSE also in that it covers also instances of excrescent [t] after a velar, although these are very rare (only *inoht* – ENOUGH and *burgt* – BURG are cited in CoNE).

2.1.4.2. Liquids

The developments concerning [l] and [r] are of two kinds - dropping or metathesis. L-Loss (LL) typically occurs before [ʧ] in syllable codas and the change affected a group of very common adjectives and quantifiers (such as SUCH and EACH). It is unclear whether l-dropping in different environments represents the same development or not.

Early r-Deletion (ERD) differs from LL in that it can be found in syllable onsets (e.g. *specan* < *sprecan,* “speak”) as well as codas before coronals (e.g. *īsen* IRON < *īsern*). This change may trigger insertion of *r* in previously *r*-less words (e.g. *burðerne* BURDEN). The r-loss behind the rise of non-rhotic accents, which appeared on a much larger scale in the 15th century, shares a similar pattern as this early change so the two might in fact be parts of the same process. As for metathesis, there are two kinds proposed for *l* and one for *r*. The difference between the two *l*-metatheses is that Dental-l Metathesis (DLM) concerns only *tl* > *ld* and *dl* > *ld*. Although the change is categorised as OE, CoNE quotes also an example from LAEME (NEEDLE). The other kind, l-Metathesis (LM) is behind the change of clusters *VCl* to *VlC* in OE.

Unlike LM, R-Metathesis (RM) concerns the sequence of *r* + V (V + *r*). The tendency of *r* to change places with the neighbouring vowel is universal to Germanic languages and also appears in Slavic languages.

2.1.4.3. Dental fricatives and alveolar stops

Several changes in CoNE involve the change from [θ, ð] to [t, d] or vice versa. The individual changes differ in timing, contextual dependence and the sounds involved (voiceless [θ/t] or voiced [ð/d]). This creates an almost symmetrical configuration comprising two pairs of changes: Late dental hardening (LDH) of [ð] > [d], the analogical Theta hardening (TH) of [θ] > [t] plus a pair of two spirantization changes (Late dental spirantization - LDS and Late t-spirantization - LTS) going in the opposite direction. LTS is the only one of the four changes which is presented as primarily orthographical with possible phonological significance.

Obviously, the discrimination between the hardening and spirantisation relies on our knowledge of source forms, which is not always sufficient so the changes may be easily confused.

LDH and TH are not the only changes where a fricative becomes as stop. There are two more developments called Sonorant Cluster Hardening (SCH) and Fricative Cluster Dissimilation (FCLD). Both of them are contextually restricted to clusters but the mechanisms behind the change are different. SCH occurs in the vicinity of [n, r, l], e.g. ME *burden* from OE *byrðen*. The effect of FCLD is that “the second member [of a fricative cluster] dissimilates to a stop”, for instance LAEME forms of *sight* include *sihte* as well as *sichðe*.

Three more changes concern voicing of alveolar stops only, again going in opposite directions. Low Stress t-Voicing (LSTV) refers to optional voicing of t “in low stress environments”, i.e. [t] becomes [d]. Devoicing Weak Verb (DWV) is the label used for the change of [d] to [t] in verbal endings. Devoicing of stops in morpheme-final position is a more general change, which may also affect [b] and [g]. This process is labelled Final Devoicing 2 (FD2) in CoNE.

The last change in this group, Deaspiration (DA), was proposed to explain the peculiar spelling *-td* at the end of the word, which can occasionally be found in several texts in LAEME. The authors of CoNE suggest that the spelling might reflect unaspirated final stop, arguing that if aspiration was a universal feature of voiceless stops in ME, unaspirated stop would probably be heard as something between voiceless and voiced. This claim is supported by our knowledge of the perception of voicing and aspiration in contemporary speakers and possibly by the same final spellings in older German (CoNE, DA).

None of the changes described in this section seems to be regular or widespread, but the presence of variant spellings in some of the texts might suggest that the scribes were sensitive to the relatively small differences in pronunciation, especially in the case of DA.

2.1.4.4. Voicing of fricatives

The restructuring of the fricative system in ME is one of the major developments of the period. It consisted in the development of phonemic /v, ð, z/ from the OE voiced allophones of /f, θ, s/. CoNE dates the rise of the allophonic set to PrOE (cf. Medial Fricative Voicing - MFV). The phonemicization of /v, ð, z/ is discussed under Initial Fricative Voicing (IFV), whereby initial ./f, θ, s/ became /v, ð, z/. The entry focuses mainly on ME evidence for the change, pointing out that while there are multiple examples of *v*-, *-u* (*-w*, wynn) for expected *f*-. Attested *z*- for expected *s*- in Germanic words is restricted to three LAEME texts and it is consistently used only in one of them (#291, MS London, British Library, Arundel 57). As for initial /ð/, the expected change is not manifest in writing but this does not disprove the claim that there was a chance, as [ð] and [θ] are not distinguished in writing to the present day. Both *ð* and *þ* were used to spell the voiced as well as the voiceless variant but the use of *ð* was gradually abandoned in favour of *þ*. This was a later development completed in the second half of the 14th century (Fisiak, 1986: 14). An even later variant for thorn was the digraph *th*. The earliest attested instances come from the *Peterborough Chronicle* (LAEME #149, before 1200). However, both variants remained in use until ca. 1400, when *th* finally prevailed.

While MFV appears in all dialects, evidence for IFV is found mainly in texts from the South and South Midlands (CoNE, IFV).

The littera *v* was another new addition to the inventory, which came to be used interchangeably with *u*. According to the entry Emergence of 'v' (EOV), *v* was first introduced in the initial position, which corresponded to its natural placement in the square capital script.

2.1.4.5. Changes involving [k, sk]

This is the first of the two groups in this overview which deal with the output of a major OE change called Velar Palatalisation (VP). This section deals mainly with ME reflexes of OE /k/ and /sk/. The third possible input for VP *ɣ is covered in the next section. The group comprises ten changes - seven phonological and four orthographic ones. Given the complexity of the changes, this section is divided into three subsections. The first one focuses on changes behind the emergence of [ʃ] and [tʃ], the next one deals with subsequent development of [ʃ] and [tʃ] and the final one relates the phonological changes to innovations in orthography.

a) The origin of contemporary [ʃ] and [tʃ] can be traced to two major OE changes, namely Velar Palatalisation (VP) and sk-Palatalisation (SKP).

VP is categorized as a Proto-Old English development and it consists in the change *k > [ʧ], *ɣ > [j] "in palatal environments" (CoNE). The latter of the two changes is going to be discussed later on. The dating and progression of this prominent change is problematic. The authors of CoNE assume a gradual change of *k > [ʧ] with an unspecified number of intermediate stages which cannot be reconstructed with precision from written sources. The change apparently did not occur in the North, but "In ME it is not always possible to tell from manuscript spellings whether palatalisation has occurred or not." (CoNE). Although it is conventional to assume the values [ʧ] for the spelling *ch* and [k] for *k*, this cannot be generalized to all texts and an analysis of the given text language is needed to reconstruct the likely sound values.

Sk-Palatalisation (SKP) is close in dating and character to VP. It affected the cluster *sk, turning it into [ʃ]. Similarly to VP, the change was not abrupt. The likely progression can be reconstructed from reflexes of the original *sk in other germanic languages, which are [ʃ] in German and [sx] in Dutch. This distribution suggest a possible intermediate stage *[sç] followed by palatalisation and fusion into [ʃ]. While this change seems to be nearly exceptionless in the initial position medial *sk sometimes underwent metathesis to [ks] instead (e.g. *fixum* - *fish* cf. CoNE sk-Metathesis, SKM). The interpretation of ME spelling is again problematic. The authors of CoNE suggest treating ME *sc* as ambiguous. Whether palatalisation of [sk] was universal in OE, the contact with Old Norse seems to have been a source depalatalised forms.

Besides VP and SKP, there are two more sources of [tʃ] and [ʃ]. Dental Palatalisation (DP) consisted in the change of [tj] > [ʧ] (e.g. OE *fetian* > *fetch*). Palatal Fronting (PF) refers to the development of initial [ʃ] from [ç]. This change is relevant for the explanation of initial [ʃ] in the personal pronoun *she*. The ME forms [ço] are presumably represented by *ȝho*. The input for this change results from two preceding changes, namely Yod Epenthesis (YE) and Fusional Assimilation (FA). The transition of initial [h] of *héo* to [ʃ] had the following stages: [h] > YE > [hj] > FA > [ç] > PF > [ʃ].

b) The expected spelling *ch* for [tʃ] sometimes alternates with [g] in LAEME. The authors of CoNE interpret this as a reflection of the change of [ʃ] to [ʤ] (CoNE label is Affricate Voicing, AV). This development is rare and all but one of the attested examples appear in SWML (e.g. *gildre* for *children*).

Later developments of [ʃ] are also irregular and restricted. Palatal Hardening (PH) of [ʃ] > [ʧ] is evidenced by *ch*-spellings, e.g. *charpe* (SHARP). *chaw* (SHOW). Sibilant Depalatalisation (SD), i.e. the change of [ʃ] > [s], could account for *s*-spellings for expected *sh/sch* (e.g. final -*isc* spelled -*is*). Still, CoNE explicitly states that the *s* might as well in fact stand for [ʃ], which is going to be discussed in connection with orthographic developments.

c) The situation concerning [tʃ] and [ʃ] is further complicated by the introduction of new graphemes traditionally taken to represent the sounds. CoNE proposes two "Orthographic Remappings", one for palatal *c* (ORPC) and one for palatal *sc* (ORSC). First. *ch* was introduced to represent palatal [tʃ], following the practice of OF scribes, which presumably provided a model for the introduction of *sh* to spell [ʃ]. The data in LAEME contain a considerable number of variant spellings close to *ch* and *sh*. *Cch* is taken to reflect OE *cc*. *Sch, ssh, ssch*, *ss* and sometimes also *s* are regarded as alternatives of *sh*. Both *sh* and *ch* also have the reversed variants *hc* and *hs* (*hss*), which is considered a purely orthographic feature in CoNE.

The next change in this group concerns a novel use of *c* (Orthographic Remapping of c, ORC), which sometimes came to represent [s] based on Romance models.

The last orthographic strategy associated with the distinction between [k] and [tʃ] is labelled Diacritic final 'e' (DFE) and it consists in adding a morphologically unmotivated -*e* to the end of the word to mark that the preceding segment is “not a stop”, e.g. *chilce* should be read [ʧil(t)s] not [ʧilk] (CoNE).

2.1.4.6. Successors of OE g

This set of changes is perhaps the most complex one. Orthographic and phonological developments combine in intricate ways and there are clear connections with changes discussed in the previous (and the following) sections, the most prominent one being the importance of Velar Palatalisation.

Most of the developments started from the voiced velar fricative *ɣ. Early, Voiced Fricative Hardening (VFH) turned the sound into a stop after nasals and in gemination. Hardening in other positions occurred later.

The next change to be presented here, Velar Palatalisation (VP) has been already mentioned in connection with [k]. While *k palatalised to [ʧ] in the vicinity of front vowel, the velar fricative *ɣ became [j] in similar contexts. There has been some controversy over the fricativeness of the original *ɣ, which is rather asymmetrical to the stop *k, but the current view, held among others by the authors of CoNE, is that the phoneme was indeed a fricative one. The new /j/ merged with the older Germanic /j/.

Both VFH and VP are OE developments. Whatever the exact timing of the changes, a single grapheme, insular ᵹ is used to represent the original [ɣ], the hardened [g] as well as the palatalised [j] in OE. The uniform spelling *g* found in OE texts does not enable us to reconstruct the change in detail. Similarly to *c*, discussed in the previous section, the values of *g* are simply assumed to have been [j] close to front vowels. Also, according to one view described under Front Vowel as Palatal Diacritic (FVPD), *i* and *e* could in some cases serve as a diacritic marking palatal pronunciation in words like *geong* (YOUNG). Still, the digraph spellings could in fact represent actual diphthongs, resulting from the so-called Palatal Diphthongization (PD). Although Roger Lass (1994) previously argued against this explanation (Lass, 1994: §3.9.4), the current view presented in CoNE is that PD was “at least in part, a genuine phonetic change” (CoNE).

Another, minor OE development of [ɣ] was Final Devoicing 1 (FD1) evidenced as final *h* for the expected *g*, where [x] seems to have been its likely value. Examples include *fuhlas* (FOWL, OE *fuglas*) or *burh* (BURG, OE *burg*). Another change invoked to explain such spellings is Dorsal Continuant Deoralisation (DCDO). The difference between FD1 and DCDO is that

DCDO can have both [ɣ] or [j] as input and the output is [ɦ] (not [x]). This would be an alternative development to vocalisation.

In EME, we are to expect three different values for original OE *g*, i.e. the fricative [ɣ], the stop [g] or the approximant [j]. The spellings found in LAEME are so diverse that correspondences between spelling and sound need to be analysed separately for each text, but at least some rough generalisations can nevertheless be made. The first one is associated with a major orthographic innovation, namely the addition of *g* described under the Emergence of Caroline g (EOG) and Orthographic remapping of *g* (ORG). The new *g* gradually became a norm for [g] and [ʤ], while insular *ᵹ* and its later variant *ȝ* continued to be used for [j] and [ɣ]. The use of *ȝ* (*yogh*) for [j] in the initial position was abandoned after 1300, its successor being *y*. Yogh representing the velar or palatal fricative was replaced by the digraph *gh*. Nevertheless, *yogh* did not completely disappear until a much later period: it appeared in provincial texts and charters as late as the 15th century (Fisiak, 1986: 15).

Another orthographic development was the use of the digraph *ᵹh* (*ȝh, gh*) to distinguish the fricative from the approximant, in accordance with the general function of *h* as a marker of fricativeness (cf. CoNE, HDF).

“Later, in some systems, *ȝ* was also adopted (with or without the support of ‘h’) for a dorsal fricative, whether velar [x] or palatal [ç]” (CoNE, ORG). It should also be added, that *Orm* (LAEME text #301) and also the scribe of *The Bestiary* (#150) invented their own letter shapes based on *ᵹ* to represent the different sound values.

At this point, it is necessary to stress that besides being the output of VP, [j] could also be a product of a change called Yod Epenthesis (YE), which appears in OE as well as ME and there are also some PDE examples like *human*, *music*. YE consists in the insertion of [j] “word-initially or at the right edge of a consonantal word-onset” (CoNE). Perhaps the most prominent ME example of this change are the forms *yede*, *yode* of the OE verb *ēode*. YE can operate together with eo-Merger (EOM), whereby the original diphthong becomes [j] + monophthong. YE can provide input for the Fusional Assimilation (FA) mentioned above.

There is also a reverse change called Initial Yod Deletion (IYD). Examples of IYD in LAEME are scarce but it seems to be connected with the weakening of the OE prefix *ge-* to *i-* (cf. ge-Weakening, GEW).

As for further development of [g] and [ɣ] in ME, CoNE describes one change for each of the sounds. In Final g-Deletion (FGD) [g] drops when preceded by a nasal. This is a relatively rare phenomenon in ME but it is associated with the familiar development of the *-ing* suffix. FGD along with formally analogical Final Coronal Deletion (FCD) can explain a set of unusual forms in LAEME text #169 (Oxford, Merton College 248) , e.g. *thynd* for THING or *myge* for MIND. It seems that the scribe of this text confused the litterae *g* in *d* in some contexts. This suggests that the segments were no longer pronounced and he tried to "reconstruct" them without knowing their right values. CoNE discusses this phenomenon in a separate entry labelled Merton Merger (MM).

Gamma Weakening (GW) is the process which turns [ɣ] into [w] intervocalically (e.g. *dawes* from OE *dagas* (DAYS)) or after *r* (e.g. *sorwes* for OE *sorges* (SORROW)). The developments concerning [w] are the subject of the next section.

2.1.4.7. Changes involving [w]

This section partly overlaps with the previous one as the change Gamma Weakening might as well be included here. Moreover, [w] as an approximant is phonetically close to [j] in that there is a thin boundary between approximants and high vowels. This similarity is partly reflected in the nature of the changes.

The first CoNE entry to be presented here is, w-Absorption (WA) because it is firmly associated with GW mentioned above. WA is invoked to account for rather odd spellings with *w* in the final position, such as *burw* (from OE *burg*, [burɣ]). The authors of CoNE propose a sequence of two changes which are likely to have affected the output [w] of GW. First, a vowel is inserted in the final [rw] cluster via Sonorant Cluster Vowel Epenthesis (SCVE) and the result is subsequently vocalised (cf. Coda Vocalisation, CV). Alternatively, the final *-w* functions as a marker of secondary articulation of the preceding *r*.

The next change in this group is w-Deletion (WD), which occurs before rounded vowels and is perhaps best exemplified by the well-known variants *swo* and *so* (SO). The reverse, i.e. w-insertion is also possible. This process is labelled w-Epenthesis (WE) in CoNE and it is described as "formally parallel" to Yod Epenthesis (see above), which means that the insertion occurs "word-initially or at the right edge of a consonantal word-onset" (CoNE, WE). Examples of this change in LAEME are scarce (the type *hpu* for *how*), but they grow more numerous in later texts.

Another relevant entry in CoNE has the label Initial w-Insertion (IWI) but this change seems to be subsumed under WE as well. The last development in this group is the Emergence of 'w'

(EOW). The new littera originated as "two ligatured <v> figurae" and it gradually replaced the older variant *ƿ* (*wynn*). Unlike the introduction of caroline *g*, this change is unproblematic as far as the value of the new symbol is concerned.

The frequent clusters containing [w] appear under several entries in CoNE. This is why there is a separate section covering the development of these clusters.

2.1.4.8. Clusters with w

Most of the changes in this group concern the OE cluster *hw*, but *cw* is also briefly mentioned in relation with orthographic innovations. The first three changes affect the cluster [xw], which creates three separate strands of development. The most general (and least problematic) of the three is Cluster-x Lenition (CXL) responsible for the change of [x] to [h]. (This is a known lenition sequence also appearing in XW2.) CXL also affects [hn] and [hr] but only [hw] is discussed here. As the lenition is not reflected in spelling, dating remains problematic (CoNE gives the whole PrOE-ME period). As for regional restrictions, dialects in the North of the Midland area appear to have preserved [xw] into ME.

Continuing the lenition sequence [x] > [h] > [0], the [h] of [hw] gradually disappeared, which is called Cluster-h Deletion (CLHD) in CoNE. The clusters [hn] and [hr] again underwent analogical development. Unlike CXL, the loss of [h] is easily observable in spelling. *Hw-* is gradually replaced by *ƿ-* or *w-*, although *hw* survives alongside [w], especially in NE Midlands.

The reversal of *hw* (*hƿ*, *hl*) to *wh* (*ƿh*, *lh*) is treated as a separate orthographic change labelled Orthographic Metathesis Initial Cluster (OMIC). The authors of CoNE do not accept the interpretation of *wh* as a reflex of voiceless [ʍ] (CoNE, CLHD). Instead, they propose to view the reversal as a mere orthographic adaptation of the spelling by analogy with *ch, sh* and other new digraphs with *h*.

There are also a few instances of forms with initial *h* where *hw/w* would be expected, i.e. [w] seems to be the element that drops. This alternative development of the cluster is called Cluster w-Deletion (CWD) in CoNE and it is viewed as the trigger for initial *wh-* in words which originally had *h-* in OE. This completes the main strand of development for [xw].

It has been mentioned that lenition of [xw] into [hw] was regionally restricted and CoNE suggests two more possible courses of development. One of them is limited to two texts in LAEME (No. 66, 67), which have some instances of initial *fw-*, presumably reflecting [xw] > [fw]. The respective change in CoNE is labeled Initial Cluster Assimilation (ICA).

The other course is discussed extensively under xw-Fortition (XWF). This change of [xw] > [kw] was proposed in relatively recent study by Lass & Laing (2016) to account for the *qu*-spellings found in several texts in LAEME. Such development is plausible phonetically and the authors also present evidence from LAEME as well as later texts, showing that *q*- spellings are far from random and appear in later texts in a relatively well defined area in the North and NE Midlands. Lass & Laing's claim represents an alternative to the older interpretation of *q*-spellings reflecting the change of [hw] back to [xw].

An important argument supporting the claim is associated with another ME orthographic change - the Emergence of q (EOQ), which had been used only in Latin texts during the OE period to represent [kw] and it preserved the same quality as an alternative to OE *cp*.

The proposed fortition sequence of [xw] > [kw] is mirrored by CoNE kw-Lenition (KWL), whereby [kw] > [xw] > [hw] > [w], which shares the same pattern with the development of OE *hw*- words discussed at the beginning of this section. KWL is a later change and attestations in LAEME are scarce. One of the examples cited in CoNE is *hpakien* (OE *cwacian* QUAKE) (CoNE, KWL).

2.1.4.9. Changes involving h

This group comprises three changes, which led to weakening of a segment to [h] and three types of h-dropping. The weakening processes are reminiscent of the developments of [hw] discussed above. Final k-Weakening (FKW) involves the same lenition sequence [k] > [x] > [h]. The weakening occurs in morpheme-final or word-final position regardless of syllable type (strong or weak), examples include *kinhis* (KINGS) or *þinhes* (THINGS). The entry Final k-Palatalisation and Weakening (FKPW) describes an analogical weakening sequence for the cases where [k], following a front vowel, was first palatalised to [ʧ] (or [ç]). This development is behind the change of *ic* to *I*. Another sound which may leniate to [h] is the dental fricative [θ] and the process is called Theta Lenition (TL) in CoNE.

The four dropping changes, which may follow the lenitions, differ in the position of [h]. The loss of [h] may occur initially, finally or in syllable codas. Initial h-Dropping (IDH) is a very common phenomenon, which often triggers etymologically unmotivated h-insertion. This change seems conspicuously close to the disappearance of *h* from the initial clusters presented above, "on the other hand many authorities treat these as voiceless sonorants rather than clusters" (CoNE).

Final h-Deletion 2 (FHD2) is in fact the final step of a lenition sequence and it is (among other cases) exemplified by *þuru* (THROUGH). Coda h-Deletion (CHD) also affects outputs of the lenition of [x] and it is accompanied by lengthening or diphthongization of the preceding vowel, e.g. *knit* (KNIGHT), *ibrout* (BROUGHT).

2.1.5. Subchapter summary

The present subchapter discussed Early Middle English sources mostly in the framework of broad theoretical descriptions of the nature of written language, production of texts and the mechanisms of change. As such, it outlined the numerous variables, which should be ideally taken into account in research, and their interactions.

The final section about specific sound changes contrasts with the rest of the chapter in its concern with highly specific observations rather than general theories. Moreover, the definition of sound changes appears to be relatively regular in comparison with the complexity of Medieval spelling systems and their development. The intriguing question is how to best combine such orderly description of sound changes with the attempts at making sense of the apparently disorganised spelling systems.

2.2. Methodological problems of research into ME texts

The character of historical linguistics has been aptly captured by Labov (1994), who described the discipline as “the art of making the best use of bad data” (Labov, 1994: 11 as cited in Adams, 2015: 2). This subchapter aims to discuss selected issues from the field of historical dialectology and research into ME texts in general. It begins with a short introduction to the principles of historical dialectology and its main challenges. Then it moves on to specific principles and methods in the study of ME sources, commenting on how they try to overcome the problems.

2.2.1. Historical Dialectology

Williamson (2004) characterises historical dialectology as a “strongly empirical” discipline which is “both data-driven and theoretically oriented.” It draws on “on carefully observed and recorded linguistic forms. Its interest is in the variation in these forms and their relationships in Space and Time” and “its key character is that it neither evades nor idealizes away linguistic complexity, but seeks to engage with it” (Williamson, 2004: 98).

If we generalize from this definition a little, we may state that historical dialectology can be characterized by a considerable scope and an avoidance of simplification, which applies both to data and “linguistic complexity”. Both of these characteristics are inherently connected with its greatest challenge, which is lack of data. A large body of data which would be potentially useful is simply not available and some of it will never be. The absence of recorded sounds and the complicated relationship of writing and speech have been already explained in the previous subchapter. The lack of material does not concern only recorded speech. Written sources themselves are scarce, incomplete and their reliability often questionable (Smith, 2007: 30).

Given the scarcity of data, simplification is something that historical dialectologists cannot afford, they “have to work within the constraints of the data that survive” (Laing & Lass, 2013: 2.3.4). Laing & Lass (2013) further explain that the lack of data also necessitates reliance on a wide range of possibly relevant sources (ibid.). This strategy would of course be useless without a good understanding of the connections between the pieces of evidence. We might say that to some extent, historical dialectology compensates for the lack of data by focusing on meaningful relations between the scattered pieces which can be found. Examples of this approach are the assumption of “internal consistency” (e.g. Laing, 2004: 57) in the spelling systems mentioned in the previous subchapter, the concept of dialect continuum or the attention paid to the temporal and spatial planes and the spread of linguistic changes. All of such concepts impose desirable constraints on interpretation of the sparse data.

Dossena & Lass (2004) point out that close attention to genuine sources is something that differentiates historical dialectology from “the approach frequently adopted in typological studies, when (…) there is at attempt to identify general patterns by means of strategies that do not necessarily include the analysis of computerized corpora or authentic texts/utterances beyond ordinary inspection” (Dossena & Lass, 2004: 10).

2.2.2. Sources of evidence

The logical consequence of the lack of data is that historical dialectologists generally consult a number of sources, looking for all data possibly related to their questions. In the words of Laing & Lass (2013: 2.4.1), they “make claims on the basis of convergence or consilience of many different arguments from different temporal strata and theoretical positions” (Laing & Lass, 2013: 2.4.1). The various kinds of evidence can be roughly categorized as follows:

a) Evidence from related languages

The forms found in ME are compared with attestations of the same word in other Indo-European languages, such as German. Greek and Latin are moreover valuable in that there are some early phonetic descriptions of the languages.

b) English

OE forms as well as subsequent development of ME forms are important sources of evidence. This includes modern equivalents of the older variants, for which we often have recorded speech. Moreover, certain features as associated with a specific region and this dialectal evidence opens considerable possibilities, allowing us to reconstruct the development in the area.

c) Verse and alliterative evidence

Analyses working with rhythmical patterns in poetry are especially useful for reconstruction of suprasegmental phonetic features. Claims regarding sounds can be supported with evidence from rhymes and alliterations. For instance, Minkova (2003) proposed a re-evaluation of the process of palatalization based largely on an analysis of alliteration in late OE texts.

d) Contemporary comments

Secondary sources in the form of explicit comments on language use are almost non-existent for the EME period. Even so, reconstruction of sounds from much later periods based on contemporary comments may be invoked as evidence for an earlier development.

2.2.3. Methodological principles and concepts

This subchapter presents several methodological contributions to the field of historical dialectology. It begins with a short overview of research principles which are not part of a more specific framework or model. The following part deals with methods focused on studying and comparing text witnesses. It presents two frameworks for the construction of linguistic profiles and briefly explains the term *stratigraphy.* The next topic is the treatment of time and space. A separate section discusses the problem of sound reconstruction.

2.2.3.1. General principles

One of the vital tasks in historical dialectology is to identify pieces of data which can be compared in a meaningful way. In his article about principles in Middle English dialectology, McIntosh (1989) stressed the preference for accumulating as much comparable data as possible before interpreting it. For instance, he suggested treating spelling differences primarily as graphemic and plotting features on the map before interpreting their sound value (McIntosh et al., 1989: 24). He also advocated examination of individual words rather than isolated features (McIntosh et al., 1989: 23) and he was critical of studies which limited themselves to a single feature, e.g. initial *sh*- and disregarded the distribution of other features found in the analysed words (McIntosh et al., 1898: 24).

2.2.3.2. Linguistic Profile (based on McIntosh et al. 1989)

Angus McIntosh developed a principled approach to the study of ME dialects and his concept of Linguistic Profile (LP) was crucial for the construction of the Atlas of Late Medieval English. LP is essentially a set of selected items and their forms found in a specific manuscript. The set is based on a questionnaire. Ideally, LP should be accompanied by Graphetic Profile (GP), which would characterize the handwriting of the scribe but it is admitted that the construction of GP is technically much more complex than LP. Observations regarding GPs are not treated here because they are not relevant to the present thesis. The following discussion deals with LP only.

Each LP should characterise the unique language of an individual scribe with such precision and discriminatory power that it is possible to identify the work of the same person even if he uses different scripts. A well organised catalogue of LPs should allow systematic and relatively fast comparison of texts. Whenever a new text is to be included in the catalogue, it suffices to compare it with a limited number of texts which are closest to it (McIntosh et al., 1989: 38-39).

The quality of the LP largely depends on the selection of the right items. Each item is a unit appearing in a number of equivalent forms, which vary across space. An example of an item would be the noun FIRE with all of its variants, which include *fire, fyre, fier etc*. The usefulness of an item is determined by its *discriminatory yield* (Laing & Lass, 2013: 1.4). Items with high discriminatory yield are widely attested and their forms are highly diversified.

The exact nature of items depends on their type. According to the original design, the LP should comprise two sections: SLP covering spoken language features (S-features) and WLP covering written language features (W-features). This arrangement responds to the requirement of distinguishing between spoken and written language. The distinction between S-features and W-features must necessarily be unclear because "both are in a sense written language features" (McIntosh et al., 1989: 46) and deciding which of the orthographic differences mirror differences in speech is never straightforward.

Besides simply listing the items and forms, our observations regarding a spelling system can be often formulated as restrictions on the scribe's usage. The most general restriction would be the total absence of a certain grapheme from his inventory. *Positional* constraints can restrict the use of a certain grapheme to a specific position in the word, e.g. the initial position, while the same phoneme might be represented by a different grapheme in other positions. *Contextual* constraints are defined by neighbouring segments. For instance, a well-known practice of medieval scribes is to use *o* rather than *u* before minims. In other cases, the distribution of specific graphemes depends on the identity of the word, which usually but not necessarily reflects differences in sound. For example, "insular ȝ is used in words like *brought* or *might* but not in *you* or *yet*" (McIntosh et al., 1989: 52).

2.2.3.3. Scribal lexicon (Laing and Lass)

In the Introduction to LAEME, Laing & Lass (2013) present their model of so-called *scribal lexicon* based on the concept of *litterae* and *potestates*. The model comprises the list of *potestatic substitution sets* (PSS), *literary substitution sets* (LSS) and "a set of word and affix templates" (Laing & Lass, 2013: 2.5). Obviously, the model cannot be constructed without first assigning some *potestates* to *litterae*. The PSSs and LSSs represent the "material" available to the scribe, but the distribution of representations tends to be "lexically specific", i.e. some of the representations of a given *potestats* are associated with certain words. These associations also need to be included in the model. The example below includes the definition of one LSS, one PSS and the "lexical representation" of GOD/N:

LSS: [o:] ⇔ {'o', 'oi', 'ohi'}

PSS: 'oi' ⇔ {[o], [o:], [ɔ:], [u:], [u]}

good n.

#

[g] ⇔ {'g'}

[o:] ⇔ {'o', 'oi', 'ohi'}

[d] ⇔ {'d'}

# (Laing & Lass, 2013: 2.5)

Compared to LP, the description of scribal lexicons in the introduction of LAEME does not place an equally strong emphasis on mutual comparability of profiles. Still, given the clear and regular structure of the lexicon, this should be possible in theory. Thus, scribal lexicons might be a proper reaction to the requirement raised by Horobin and Smith (1999) that "a robust system of categorization is needed, which will allow not only for spelling to be treated independently of sound system, but also for spelling and sound systems to be correlated and compared in as transparent way as possible" (Horobin & Smith, 1999: 364).

As for the practical motivation of working with the model of litterae, Laing & Lass (2013) propose to group forms based on potestatic interpretation, which produces a smaller number of "abstract types", which may be subsequently plotted on the map (Laing & Lass, 2013: 2.3.3).

2.2.3.4. Stratigraphy

One strand of research into ME texts focuses on the identification of linguistic layers (strata) in medieval manuscripts or the establishment of the texts's *stratigraphy*. This task is especially worthwhile if we want to examine texts which clearly display linguistic mixture and as such cannot be regarded as representative of a specific language variety. Such texts used to be neglected by scholars as useless (Black, 1999: 155).

Successful identification of various layers of copying in the texts can not only provide data useable for sound reconstruction and mapping but it can also contribute to our knowledge of scribal practices and textual histories.

The main limitation of this method is that if it is to be used effectively, multiple copies of one texts or multiple texts written in the same hand are needed. The distinction between exemplar forms and forms introduced by the copyist is usually based on a comparison of two texts copied by the same scribe. Differences between the copies speak in favour of the conclusion that the scribe was a *literatim* copyist and the forms found only in one of the texts are the likely relicts taken from the exemplar. Hudson (1966) assumes that exemplar forms generally correspond to the less common variants (Hudson 1966, 361-362).

Laing & Lass further propose a somewhat finer classification of exemplar forms, describing two phenomena resulting in mixed language: *relict usage* and *constrained selection*. A *relict* is a piece of language which the scribe failed to translate, although he normally would.

Constrained selection occurs when the scribe does translate a form because it appears in his passive repertoire (Laing & Lass, 2013: 1.5.6).

It might be worthwhile to consider *constrained selection* with an emphasis on the distinction between written and spoken language. Although the original definitions of scribal strategies refer simply to "dialects", it is logical to assume that the scribes' approaches to the preservation of the spelling system of the original could be described in similar terms and the scribe could have a different approach to the replacement of symbols and translation of what he perceived as a different sound. For instance, a scribe who would normally use *w* to spell [w] could nevertheless preserve *wynns* found in his exemplar (which would be described as "literatim copying"). The same scribe may substitute *w* for *g* in words affected by the change [ɣ] > [w] (which would be described as "translating"). A study of several groups of related texts tagged for LAEME was published by Laing (2004).

2.2.3.5. Modelling time and space

The theoretical discussion briefly discussed the operation of language change in time and space, stressing the central role of interaction of individual speakers. The methodological challenge is to create a model which would reflect this reality as faithfully as possible. Moreover, the model has to allow us to focus on a meaningful subset of our data at a time, because the amount of data we are able to analyse at a given time is limited. The inadequacy of many previous accounts of changes as well as the difficulty or even impossibility to avoid simplification were discussed by Lass (2006).

Williamson (2004) addressed this problem, proposing the concept of *spacetime continuum* (Williamson, 2004: 110) and he also made a useful distinction between three kinds of "spaces" (Williamson, 2004: 119-120). The spacetime continuum replaces the traditional two-dimensional maps with a three-dimensional space, which enables to model temporal and spatial relations between witnesses in a single "picture".

The researcher can study "constellations" (Williamson, 2004: 110) or groupings of witnesses which appear to be close in time and space. He may also choose to focus on a specific *language extent* defined with reference to time as well as space.

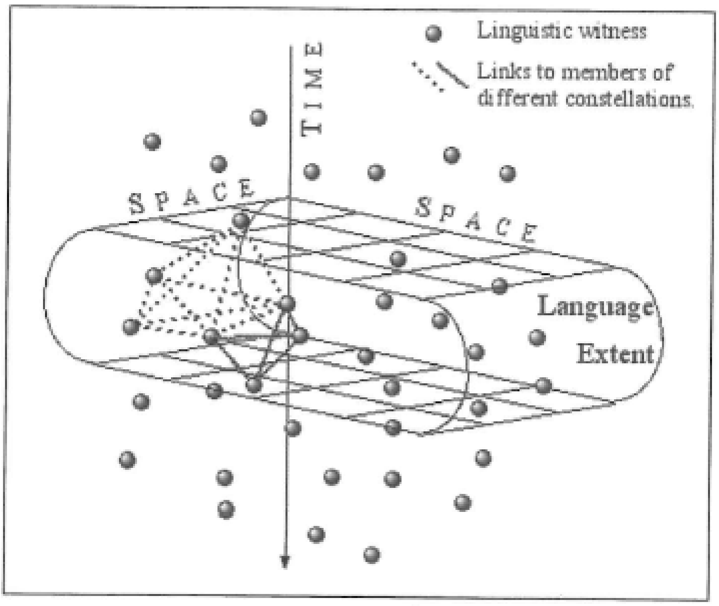


Figure 1: Characterisation of a "language extent" (Williamson, 2004: 110)

The above schema displays what Williamson calls *reticular space* (Williamson 2004: 120). Witnesses placed in reticular space are distributed based solely on their shared features suggesting their relative closeness. There is no definite reference point in real time or geographical location. The distribution in reticular space can be "projected to Geographical space" (Williamson, 2004: 122) , i.e. on the map. Williamson also explains that closeness of speakers in *geographic space* alone does not necessarily imply more intensive contact between speakers which is needed for changes to spread. Closeness in *real space*, on the other hand actually reflects contact. The shape of real space depends on factors like the location of settlements or mobility of speakers.

The task of a researcher working in the proposed framework is to "determine the types of relation between the witnesses within the language extent according to two principal types, extralinguistic and linguistic" (Williamson, 2004: 110-111). Linguistic relations seem to be the less problematic of the two, at least as far as Early Middle English is concerned, because their identification relies purely on a detailed analysis of the available witnesses (Williamson, 2004: 111). Extralinguistic features, such as closeness of the witnesses in time, should be established based on secondary sources, which can be rarely acquired.

2.2.3.6. *Visualisation of spacetime continuum*

Williamson (2004) accompanies his discussion of spacetime with two concrete methods of visualising the distribution of witnesses in this three-dimensional space. One of them is a rather complex graphical representation combining temporal and spatial axes, which will not be presented here in detail. The other option is to combine multiple kinds of markers on one map:

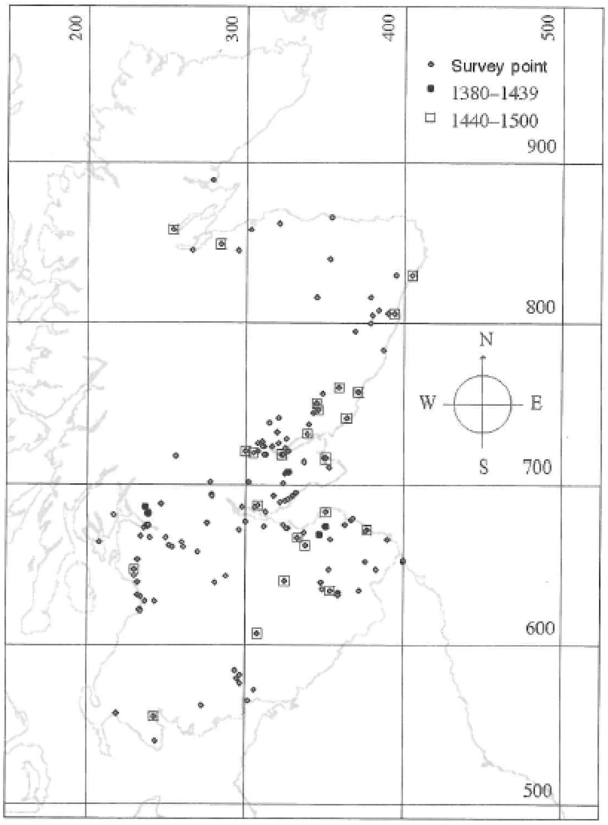


Figure 2: Spacetime map (Williamson, 2004: 126)

As the legend suggests, the witnesses for the feature from the period 1380-1439 are displayed as filled squares, while witnesses for 1440-1500 are represented by larger empty squares which can be placed over the black squares. The map is easy to read and its construction should not require complicated calculations.

2.2.3.7. Dialect continuum and the fit technique (based on Williamson, 2004: 129-131)

The term *dialect continuum* is used to describe the distribution of linguistic features in space. We can often roughly delimit the area in which a given feature appears at a given time. Areas of different features overlap one another and the whole is best regarded a *continuum* because the distribution does not allow to draw a clear boundaries between regions.

The concept of dialect continuum is vital for the placement of witnesses on the map using so-called *fit technique* developed by Angus McIntosh (McIntosh et al., 1989). This method consists in

> …comparing, map by map, spellings particular to an unlocalised text with those already placed in the localised matrix. For each map, areas where those or similar spellings are *not* found are then eliminated, until (in the ideal case) only a single, well-defined location is left where the whole assemblage of spellings could plausibly occur (Laing & Lass, 2013: 1.4).

The concept of *dialect continuum* is not without its opponents. Kretzschmar (2015) claims that it contradicts the nature of language as a complex system. According to his view, our knowledge of complex systems leads us to expect much less regularity that the concept of dialect continuum requires and he regards it as a mere "formal assumption", although not "bad in itself" and states that "the fit-technique (…) defines one dialect, one grammar, to fit all the texts of a place" (Kretzschmar 2015: 298), but this does not appears to be a proper description of what fit-technique is based on. The crucial principle of the fit-technique is the placement of texts relative to one another and the placements may shift as new witnesses are added (McIntosh et al., 1989: 27). In fact, text languages recognized as slightly different sometimes share the same location in LAEME.

2.2.3.8. Reconstruction of sound

Reconstruction of sounds is a major concern for a historical dialectologist and this challenging task should not be performed without prior consideration of the target level of precision. This issue was obviously duly addressed by McIntosh (McIntosh et al., 1989) as well as Laing & Lass (2013). McIntosh (McIntosh et al., 1989) stressed the impossibility to reconstruct the exact phonetic value from ME texts. He claims that "any such statement as '*swilk* and *tham* represent or stand for [swiłk] and [ðam] runs a grave risk of lacking any meaning whatsoever" (McIntosh et al., 1989: 2). His argumentation is based on a highly realistic account of the differences in pronunciation in the speech community. Regardless of how minute such differences might be, every speaker has his own pronunciation of, e.g. the word *swilk* and it is his individual pronunciation which governs what the word on the page "stands for", which entails that there is nothing like a truly "common" interpretation of the spelling, the individual pronunciations are simply close enough to enable communication. The same is true of word meanings, which opens up some space for analogy between our analyses of the two systems. Even if, for instance, the range of colours which could be described a *red* is slightly different for each member of the speech community, we usually work with a concept of "the meaning of red", which somehow represents the range of individual concepts solely for the purpose of our investigation. The place of "the meaning of red" in phonological analyses is sometimes taken by the concept of *phoneme*, but this is not the only possibility.

Laing & Lass (2013) characterise the target level of precision as “poorly resolved broad transcription”, which roughly means that “if a responsible phonetician equipped with a time machine were able to hear the items represented, the symbol in question would be a reasonable transcriptional response” (Laing & Lass, 2013: 2.4.1).

2.2.4. Phonemes and litterae – commentary

Laing & Lass (2013) repeatedly mention structuralism or structuralist concepts like *phoneme* or *contrast*, usually pointing out their inadequacy to the analysis of ME manuscripts. While most of their points are accepted without reservations in this thesis, the concept of contrast is considered highly relevant here. This short passage was included to explain what exactly is meant by the term in the context of the present project.

While *contrast* in the purely structuralist sense is associated with a position in the system, it can be also used to describe the perception of sounds on the part of the and there should be some degree of overlap between the two. Smith (2007) speaks about a logical connection between “minimal pairs” and “perceptual salience” (Smith, 2007: 35). Whenever a scribe uses the same *littera* in two different positions, it is reasonable to suppose that he perceived the two segments as identical (although it is by no means the only explanation). If, on the other hand, he uses two different *litterae*, there is a chance that he “heard” a difference between the two. Such behaviour is not “structuralist” but perfectly natural. This does not imply that two different representations of what we suspect to be the “same” sound must not be taken as “identical” for the purpose of mapping. In fact, the point is much more relevant for our understanding of the writing process than it is for a reconstruction of developments of sounds.

It is interesting to notice that speakers are not forced to deal with “contrast” in this sense, until they need to represent their speech in writing. One can express oneself fluently and effectively with no definite idea of which sounds in his speech are somewhat similar to other sounds. Contrarily, a person trained to employ the “accepted” spellings for certain words can easily miss differences in pronunciation of what s/he believes to be “the same letters”. These points are perhaps best illustrated by spellings invented by pre-school children learning to write. Their creations were the subject of studies performed by Chomsky (1971) and Wood (1982). The children were not taught any codified spellings, they merely learned the “values” of the individual letters in English. The collected data display some spelling choices, which are perfectly understandable but nevertheless striking to the eye. For instance, the word *train* was spelled with initial *ch* and without *i* i.e. *chran*, and the form *jran* represented *drain* (Wood, 1982: 711). Chomsky (1971) analysed the children's strategies of recording vowel sounds and described them as "very systematic", for instance, *e* was regularly used for [i] and *i* for [ʌ] because of their names pronounced [i:] and [ai] (Chomsky, 1971: 513-514).

Obviously, medieval scribes did receive proper training, still, if this training only covered French and/or Latin and little or no English (Horobin, 2010: 20), their situation might not have been as radically different from the children's, as it could appear at first sight. Although we do not have any prior knowledge of a specific scribe's training, his reading experience etc. which could all account for apparent "contrast" in his writing, the possibility that some of his choices faithfully reflect his perception of the sound should at least be taken into account.

2.2.5. Subchapter summary

The apparent lack of regularity in Early Middle English, which was a dominant theme of the first subchapter, stands in contrasts to the high level of systematicity found in the models presented above. The complex relations between speech a writing can be described in an orderly manner in the form of a scribal lexcion. Linguistic profiles were designed to enable effective comparison and mapping of many witnesses. The concepts of spacetime continuum and language extents respond to the complex nature of language change.

The need for regularity is inherently connected with lack of data. Although the expectations of "regularity" were sometimes described as far-fetched and unrealistic, it should be pointed out that there is clearly a strong effort to avoid simplification and artificial distinctions, such as the "traditional" boundaries between dialects. The said "regularity" is in fact expected at a very general level, which allows for much surface variation. The rather modest target level of precision in sound reconstruction follows this principle. Although the consulted books and articles do not explicitly discuss specific ways of integration of the models, e.g. scribal lexicons in spacetime, this seems to be perfectly possible.

2.3. Electronic sources in historical dialectology

The theoretical part of this thesis concludes with a presentation of electronic resources in historical dialectology which exemplify novel uses of computers in analyses of medieval texts. A substantial part of this subchapter deals with the projects of Angush McIntosh centre in

Edinburgh, because they are closely related to LAEME and therefore the present thesis. The characterisation of LAEME is not included in this chapter because it is going to be discussed in the methodological chapter. The chapter first presents several specific projects and resources and then it comments on their shared features and connections with the methodological issues from the previous subchapter.

2.3.1. Sound Comparisons

Sound comparisons is a project responding to the challenges of the concept of so-called "Big history" (Christian, 1991), which calls for largescale comparisons of data. In the words of the authors of the project,

> Our method starts from the (rather challenging) assumptions that we should ideally be able to compare many varieties all at once; to compare both social and geographical variation at the same time; and to include similarities regardless of whether they reflect common ancestry, or parallel development, or contact (McMahon & Maguire, 2012: 145).

The project included a quantitative comparison of pronunciation variants for 110 English words across different dialects. The primary material was provided by live informants. Their pronunciations of the words were transcribed and compared by a sophisticated programme analysing a number of phonetic features. The crucial methodological aspect which this project shares with the proposed spelling database is the preparation of the input data for automatic analysis, which consisted in splitting the sound stream into segments (slots) and relating them to corresponding segments in their supposed ancestral form.

This procedure is very similar to the method called *grapho-phonological parsing*, which was employed in the production of FITS (Kopaczyk et al., 2018) introduced below (see subchapter 2.3.2.4).

One of the outputs of the project is a website[3] presenting pronunciation variants of selected words spanning across many regions and periods. Recordings are provided for the modern variants. The website comprises nine sections covering different groups of languages. Early Middle English data is included in the section "Englishes". The variants can be searched by word and plotted on the map or displayed in a tabular format.

[3] https://soundcomparisons.com/

2.3.2. Projects of Angush McIntosh Centre for Historical Linguistics

The Angush McIntosh Centre for Historical Linguistics (AMC)[4] is the successor to the Institute for Historical Dialectology at the University of Edinburgh founded by Angus McIntosh. The projects of the AMC include a few resources related to LAEME.

2.3.2.1. Linguistic Atlas of Late Middle English

The first version of *A Linguistic Atlas of Medieval English* was published in 1986. The electronic version is called *eLALME: A Linguistic Atlas of Late Medieval English* and it was published online in 2013.[5] The printed version of the atlas comprised Linguistic Profiles (LPs) of more than 1000 texts from the period ca 1325-1450. Each LP lists the forms of over 300 pre-selected items. The space for maps was limited in the printed version, some 1200 maps based on the LPs were included but others had to be left out.

All the LPs found in the printed version as well as the questionnaire are available as static web pages in eLALME. The extra possibilities of electronic media were realized in the presentation of maps. eLALME offers about 1700 maps corresponding to the dot maps from the printed version including those which could not be published in print. The data from multiple maps can be combined and displayed as one map. For instance, the map showing the distribution of THE with initial *y* can be combined with the map for thorn in the same position. The result is displayed as a picture with coloured dots showing the localisation of manuscripts in which the feature in question is present.

Construction of fully customized maps is made possible by a special tool. This tool displays a map along with selection of items and features to be plotted. The user can select multiple items and multiple forms one-by-one. The dots on the map are interactive and work as a quick link to the associated LP. A click on the dot displays a box with the number of the LP and the complete list of forms for the selected item, including visualisation of their frequency.

Another interactive mapping tool was designed specifically for the purpose of "fitting"[6] a new manuscript on the map. The researcher can simply check all the items and forms that s/he is able to find in the text and a computer programme automatically changes the colour of the dots on the map based on the size of the overlap with the text being localised. The most similar

[4] http://www.amc.lel.ed.ac.uk/

[5] http://www.lel.ed.ac.uk/ihd/elalme/elalme.html

[6] http://archive.ling.ed.ac.uk/ihd/elalme_scripts/mapping/fitting.html

LPs are displayed as darker dots and the least similar ones as lighter dots. As more data is added, the calculation becomes more precise until (ideally) there appears a discernible concentration of dark dots in one region.

2.3.2.2. The Corpus of Narrative Etymologies (CoNE)

The Corpus of Narrative Etymologies (CoNE) is a sister project of LAEME. It was developed by Roger Lass, Margaret Laing, Rhona Alcorn and Keith Williamson in the years 2010-2013 and it is to a considerable extent based on LAEME data. CoNE is conceived as a comprehensive database of phonological, morphological and orthographic changes manifest in the spelling variants in LAEME. The individual changes are labelled, categorised and described with references to primary data (specific lexical items in LAEME) as well as secondary sources. The scope is not limited to Middle English and some of the developments are traced to earlier stages of development.

CoNE database consists of two interwoven sets of data. The data in the first set is structured according to changes, i.e. the basic unit is a single linguistic change. The basic unit in the second set is a specific tag, usually a lexical unit and grammatical tag. The data about changes should explain the variation in EME forms of specific items. As this variation is not always a result of genuine linguistic change, CoNE also uses a set of special codes for processes and strategies which can be invoked separately or in connection with changes to account for certain variants. For example the code ([MAF]) stands for "Modelling after French" and the code suggests that "The form of a word to a greater or lesser degree resembles its French cognate" (CoNE, [MAF]). The structure of the entries for changes and tags is given below.

2.3.2.2.1. Change

Each change was assigned a 2-4 letter abbreviation (code) usually composed of the initial letters in the name of the change. Thus "Emergence of v" has the code "EOV", "Analogical Extension" has "AE" etc. Each code is unique and it provides a quick way of referencing a change from anywhere on the website. The description of the change is given as a single stretch of text ranging in length from a couple of lines to several paragraphs. This description includes references to literature, related changes and entries for items affected by the change. Each change also has several "descriptors", i.e. categories. "General descriptors" characterize the change in terms of dating and regularity (e.g. "OE, ME, dialectally restricted, variable or irregular, "). "Domain descriptors" correspond to the affected linguistic level and type of segment (consonant/vowel). Most of the changes are categorized as phonological, but orthographic and morphological developments are also included. The distinction between phonological and orthographic change reflects the general concern about the distinction between written and spoken language.

2.3.2.2.2. Tag

Etymologies describe the development of individual words, which implies that each entry can be linked to a LAEME tag (i.e. a combination of a lexel (lemma) and grammel (grammatical tag), i.e. `$child/n`, "*child* as noun"). The tag is given as the top field in the entry. The next field, *dictionary box*, holds a definition of the word and its forms found in dictionaries, which work as hyperlinks to the original source. Typically, there is the form found in OED, The Middle English Compendium and The Dictionary of Old English. *Classification* gives OE morphological information.

Etymological information is divided into several sections. *Old English Etymology* describes pre-ME development of the word, which is followed by *Introductory notes* to *Middle English etymology* and a complete lists of the forms of the base found in LAEME. *Middle English Etymology* has separate sections for phonology, morphology and *Probable Old English Input Paradigm to Morphology*. The last field treats etymologies of derivations and compounds associated with the base. The "core" section *Middle English Etymology* proposes sequences of changes or special codes leading from OE input to each variant form found in LAEME.

Besides being based on LAEME, the methodology of CoNE shares an important aspect with the present project, namely its focus on the segmental level. It traces changes of segments (in specific contexts), which occur in different lexical items. Specific suggestions regarding the integration of data from the two sources is going to be discussed in the methodological chapter of the resent thesis.

2.3.2.3. Linguistic Atlas of Older Scots (LAOS)

Linguistic Atlas of Older Scots[7] developed by Keith Williamson is methodologically and structurally almost identical to LAEME. The corpus comprises legal documents in Older Scots from the period 1380-1500 and its compilation was largely motivated by the need to make up for the limited coverage of the area in LALME. The corpus per se is not particularly important for the present project but it was used to produce a grapho-phonologically parsed corpus, which is very similar to the spelling database based on LAEME (see below).

[7] http://www.lel.ed.ac.uk/ihd/laos1/laos1.html

2.3.2.4. From Inglis to Scots: Mapping Sounds to Spelling (FITS)

FITS is another project of the AMC in Edinburgh. The purpose of FITS is to "elucidate the language's underlying sound system, via the orthographic alternations within the Germanic morphemes of the corpus, as well as suggesting how their sound and spelling features developed from proposed sources" (AMC website). FITS was developed from LAOS data but it shares an important methodological strand with CoNE, as it focuses on etymological developments of individual segments and it includes a separate database of changes.

The chief task in the production of FITS was to reconstruct the sound, link individual segments in the words from the corpus to the corresponding sound in their source forms (from earlier stages of the language) and propose a sequence of phonological developments accounting for the change.

The database constructed using FITS methodology enables searches for what would be termed *litteral substitution sets* and *potestatic substitution sets*, i.e. lists of variant spellings linked to one sound or possible "values" of a chosen symbol. Moreover, users can search for alternatives of a certain segment in an ancestral form or attestations of a certain change. The data can be easily quantified and visualised as networks.

The core concept of the methodology, so-called *grapho-phonological parsing* refers to the segmentation of forms into units, which can be linked to corresponding units in the ancestral form. This was performed manually using a tool designed specifically for the purpose. The analysis focused on root morphemes only, but the preceding or following segments had to be taken into account when assigning sound values.

2.3.2.5. Towards an Inventory of Middle English Spelling Systems (TIMESS, project not realized)

TIMESS is a project proposed by Rhona Alcorn (2016) for LAEME specifically. It shares the core methodology (i.e. *grapho-phonological parsing*) with FITS, but, as the title suggests, it focuses primarily on the analysis of the individual spelling systems. As such, TIMESS is indeed very close to this thesis, both in its methodology and objectives. The long-term goals of TIMESS project were to construct "a set of grapho-phonological profiles, each an inventory of the reconstructed and contextualised correspondences between the units of spelling and units of sound for a particular early ME specimen", which could subsequently be used to identify regional patterns and the ultimate product would have been "a taxonomy of ME spelling-system types" (Alcorn, 2016: 4). A more specific description of the profile as such is not included in the grant application quoted here. The text merely references a work in progress, namely *grapho-phonological profiling* developed within FITS (Kopaczyk et al., 2018). The present project is less ambitions in its goals compared to TIMESS in that its proposed final product is a tool usable for analysing spelling systems rather than complete spelling profiles.

2.3.3. Middle English Grammar Project

The purpose of the Middle English Grammar Project is to publish updated resources describing all linguistic levels of Middle English (Black, Horobin & Smith, 1999: 9-10). The project is a joint effort of researchers at the universities in Stavanger and Glasgow and the work on it started already in 1997, i.e. before the publication of LAEME. Its goals comprise the publication of new electronic resources, of which two corpora have been published (The Middle English Grammar Corpus and A Corpus of Middle English Local Documents).

Soon after the work on the project began, Horobin and Smith (1997) presented a description of a large spelling database to be constructed from ME texts and covering both the Early and Late Middle English periods. Although the database was not (yet) published online, its design is certainly of interest for the present thesis.

Black Horobin & Smith (1999) propose to structure the database around so-called *Standandard Orthographic Sets* (SOS) (Black, Horobin & Smith, 1999: 14; Horobin & Smith 1999: 361). This way of structuring the data was originally devised by Venezky (1970) and it consists in grouping forms according to sets of spelling variants in PDE. For instance, the words spelled with *gg* would constitute one group and words with *gh* would be another group (Venezky, 1970: 72). The reason why the authors selected PDE rather than OE as their reference point is that extant OE forms are not very numerous and a large proportion of the preserved manuscripts is written in the West Saxon dialect, which may lead us to relating ME forms to OE forms from which they were not originally descended (Horobin & Smith, 1999: 366).

Each entry in the database should be related to a particular SOS and it comprises a number of fields, e.g. "GROUP, PDE SPELLING, ME SPELLING, FREQUENCY, MS REF., DATE, SCRIPT" (Horobin & Smith, 1999: 368).

The examples of possible searches in the database include mainly listing of reflexes of a given PDE word or ME spellings belonging to a particular orthographic set, e.g. all forms containing a reflex of PDE *th* and the search can be further restricted to a particular manuscript, county etc. (Horobin & Smith, 1999: 370). A pilot study testing the potential of the database was carried out by Stenroos (2004). The database was used to describe patterns of usage for *th* and related letters (*þ*, *y*), their change in time and differences between documentary and literary texts. The results revealed clear differences in the development between the South and the North as well as a markedly higher incidence of *th* in documentary texts.

2.3.4. The Wycliffe corpus with Orthographic Annotation

This project seeks to respond to the need for detailed orthographic information, such as capitalisation, insertions, corrections, line spacing etc. The necessity of including this kind of data for the purpose of spelling analyses was advocated by Diemer (2012a, 2012b). A special tagset was developed for this purpose (Diemer, 2012a: 28-30). The corpus links a simple transcription of the manuscript texts with a tagged version and manuscript images. The three "layers" of data can be displayed in a simple interface, which allows to switch between the layers (Diemer, 2012a: 31).

2.3.5. Commentary

The projects share a number of common approaches and concepts. The most prominent one is probably the focus on segmental level, rather than the level of the word, which led to the introduction of *grapho-phonological parsing*. The importance of segments is also evident in the design of CoNE. The extra potential of electronic processing is exploited to provide quantification and visualisation of data, which is most developed in the FITS network visualisations and the interactive maps in Sound Comparisons and eLALME. The fitting feature is especially interesting in that it goes beyond data retrieval in computerizing the logic of the fit-technique.

The fitting tool is definitely not the only example of a specific method smoothly transferred to the electronic medium. The FITS project essentially materializes the concept of *Litteral Substitution Sets* and *Potestatic Substitution Sets*, developed for scribal profiles although analyses of individual text languages are not elaborated on in the article about the corpus. Linguistic profiles are of course still used in eLALME.

Although considerable deal of work has been done, not all the concepts and models described in the previous chapter have been incorporated in the electronic resources. For instance, the spacetime maps which should be relatively easy to generate are not yet a standard part of the available tools. The regular structure of scribal lexicon could be turned into a machine-readable dataset, but grapho-phonologically parsed data would be needed to achieve this.

A separate topic would be the integration of data from the related sources, which is obviously highly desirable. A notable example of integration is the FITS database which links phonological changes to the actual forms found in the corpus. CoNE provides useful links to dictionaries. Another contribution to integration is Laing's (2015) article encouraging the use of LAEME maps in combination with LALME maps. Also, CoNE was designed specifically to complement the data from LAEME. Still, the integration could be carried further, e.g. by adding interactive links to LAEME to CoNE or the construction of a mapping tool capable of combining data from multiple databases in one map.

2.4. Chapter summary

The highly realistic accounts of the complex nature of written linguistic systems and the development of language in space and time raise methodological requirements which are sometimes difficult to meet, but a number of obstacles have been already overcome. The theoretical chapter of the present thesis outlined the complexity of research into (Early) Middle English texts and explained how various aspects of this subject of study shape the methods and approaches adopted to cope with its challenges. The subsequent brief survey of the available electronic resources showed which of the theoretical and methodological concepts were involved in their production and commented on possibly useful ideas which may yet await incorporation into an electronic tool. If the insightful theoretical analyses of gifted linguists like Angus McIntosh raised methodological requirements which inevitably hit purely technical limitations in their time, electronic processing opened new avenues towards models and methods realistic enough to realize some of the visions presented more than 50 years ago.

This final assessment should prepare the ground for the upcoming discussion of LAEME and methodological principles observed in designing the tool proposed in this thesis.

3. Material and method

The methodological chapter provides a detailed description of the transformation of LAEME data into the new database, which consisted mainly in a segmentation of the spelling variants and alignment of the segments. The segmented forms function as an additional layer of data linked to the original data from LAEME.

The chapter opens with a presentation of LAEME, focussing on its characteristics which were particularly relevant for the construction of the tool. The next part explains general methodological principles behind the tool. The definition of the principles was informed by the theoretical and methodological considerations discussed in the theoretical chapter.

The third part of the chapter briefly summarizes the first attempt at processing the data from LAEME, which was restricted to a limited number of corpus files. The results of this pilot project were used to design the methodology which was ultimately applied to the whole corpus. The fourth subchapter describes the structure of the new spelling database and the process of its construction. The final subchapter deals with the various queries and the structure of data retrieved from the database, including a few experimental features.

3.1. Linguistic Atlas of Early Middle English

LAEME was designed as an electronic research tool for analyses of Early Middle English texts and dialects. It was envisaged from the beginning that LAEME would be useable in combination with LALME (introduced in the theoretical chapter, section 2.3.2.1) but the construction of the tool required a different methodology. The extant texts from the EME period provide a much less voluminous data sample as opposed to LME texts and if the atlas was limited to a set of linguistic profiles based on a questionnaire, the already poor amount of data would be reduced even further. This is why the basis of LAEME is an corpus of Early Middle English texts.

3.1.1. Corpus sources and structure

The size of the corpus is approximately 650,000 tokens and it has detailed lexico-grammatical tagging. A searchable index of sources and information about the included manuscripts are also available (Vaňková, 2016: 34). Each corpus file represents a maximally homogenous *text language* (see subchapter 2.1.2.2.1). This means that the work of each scribe contributing to a manuscript is stored in a separate file. If stretches of texts in a single hand can be recognized as distinct types of language, the work of the scribe is split into multiple files and the placement of the texts on the map may differ. For instance, the text of the *Trinity Homilies* (MS Cambridge, Trinity College, B 14.52) copied by scribes A, B and C is split into three files. Another file is reserved for version T of *The Poema Morale,* found in the same manuscript and copied by scribe A, because its text language differs from the parts of *Trinity Homilies* copied by the same scribe.

"The LAEME corpus contains almost all of the available texts from the EME period (some of the longer texts, however, are not transcribed in their entirety) plus several slightly later northern texts, which also appear in LALME. These texts were included in order to make up for the absence of earlier texts and provide a better coverage of the whole territory (Laing & Lass, 2013: 1.3), which is nevertheless very patchy. The only area with a number of texts sufficient to create a real continuum is the West Midlands (WM)" (Vaňková, 2016: 35).

"The placement of texts on the map proceeded from the identification of so-called anchor texts, i.e. texts with an explicitly indicated place of origin. Extralinguistic data enabling localisation are scarce and often unreliable, the notable exception being MS Arundel 57 (containing the *Ayenbite of Inwyt*). LAEME distinguishes between "literary anchor texts" and "documentary anchor texts". A table listing all texts serving as anchors is available in Appendix 7.1. The remaining texts were localised using the so-called fit-technique discussed in the theoretical part of this thesis (subchapter 2.2.3.7)")" (Vaňková, 2016: 35).

"Due to the lack of anchor texts in EME, fitting sometimes relied also on texts already localised in LALME" (Vaňková, 2016: 35). Text languages which were considered too heterogeneous to be placed anywhere on the map remain available for analysis, they are simply not assigned any map coordinates. The same concerns very short texts which do not provide enough linguistic data. The locations of all texts included in the corpus are shown in the picture below:

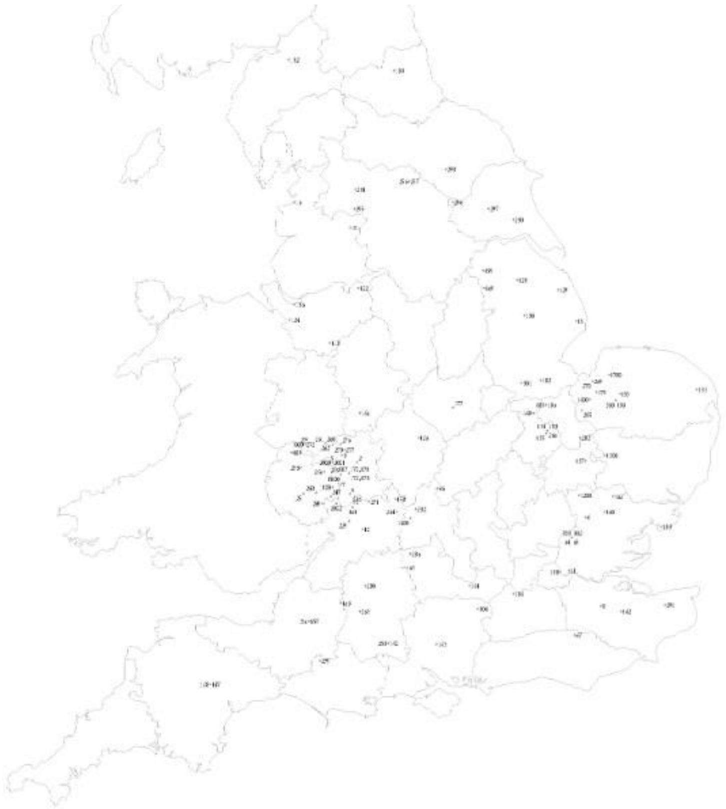


Figure 3: LAEME key map

It is clear from the picture that the distribution of texts is highly uneven. There is a conspicuous concentration of texts localised in the West Midlands, which provides better coverage in comparison with the Eastern part of England. The Southern and Northern and especially the central Midlands regions have a rather poor coverage. Moreover, some regions are represented only by texts covering only a short period of time and some texts are very short so they seldom provide useful data.

3.1.2. Tags

The basic unit of the LAEME corpus is the *tag*. Each word in an actual MS is represented by one tag in the corpus. Each tag consists of the actual *form* found in the MS, the so-called *lexel*, which serves to identify the lexeme and a *grammel*, which is the grammatical tag. For example, the lexel AFTER appears with several different *grammels* including *aj* (adjective), *av* (adverb) and *pr* (preposition) and has a wide range of forms, for instance *aftir*, *hafter*, *eftre affeter*, *hefteir* etc.

Lexels are taken primarily from Present Day English (PDE) reflexes of the lexeme in question. OE forms serve as *lexels* if PDE forms are not available or ambiguous. These two sources were sufficient to cover the vast majority of *lexels*. The remaining *lexels* are either Old Scandinavian words or ME words and there are also some composite *lexels* (Laing & Lass, 2013: 4.3.). Some *lexels* have so-called *lexel specifiers*, which explicitly mark a functional or semantic aspect of the *lexel*, such as temporal/spatial use of a preposition. *Lexel specifiers* have the form of a code in curly braces attached to the lexel, e.g. BEFORE{T}, BEFORE{P}. Some grammatical words, such as articles or personal pronouns, are clearly identifiable by their grammatical categories and have no *lexel* in the corpus.

Grammatical tagging is based on the "traditional" categories (nouns, adjectives, number, gender) (Laing, 2013: Grammel Commentary). The system of tags is quite complex and very detailed and some tags include also syntactic information, for example the symbol "<" "indicates postposition of an expected preposed form and points backward to a syntactically connected word" (Laing, 2013: Grammel Commentary).

Words consisting of more than one morpheme have one tag for the whole word plus a separate tag for each of their constituent morphemes except the root. Morpheme boundaries are marked with "+" or "-" signs. "+" indicates that there is no space between the morphemes and "-" indicates that there is a space. For instance, the main tag for *gladly* below is followed by a separate tag for *-ly*:

```
$gladly/av_GLAD+LICHE $-ly/xs-av_+LICHE
```

Not all elements found in the manuscripts receive a separate tag in the corpus. Proper names and place names are preceded by special characters which set them apart from the tagged words. French and Latin words or glosses and other additions to the manuscript in different hands are included as comments in curly braces.

3.1.3. Transcription

The transcription of manuscript forms follows a rich set of conventions designed to enable maximally faithful representation of the original manuscript. Non-roman characters are of course never replaced, as is sometimes the case with editions. *Insular* ᵹ is distinguished from *yogh* (ȝ). Features like capitalisation, abbreviation, superscripts or special letter shapes are preserved in the transcribed texts.

Since LAEME was compiled at a time when it was technically problematic to use other than ASCII characters, it employs a special system of uppercase and lowercase letters to cover the required range of characters. Latin characters are normally transcribed as uppercase letters and lowercase letters always have a special function. The most important use of lowercase is the transcription of *ash*, *yogh*, *insular ʒ, wynn*, *edh* and *thorn* (represented by *ae*, *z*, *g*, *w*, *d* and *y* respectively).

3.1.4. Querying

The electronic version of LAEME allows searching the corpus by lexels (lexical items), grammels (morphological tags), forms (actual words in the text) or a combination of the three. It can also generate a complete lists of forms for a particular item (e.g. lexel) or group forms by text or county.

The online interface of LAEME also includes a set of pre-defined feature maps as well as a tool for the construction of custom maps. The researcher defines a feature or multiple features to be plotted on the map and selects the shape and colour of markers used to represent each feature. The result is a dot map with a legend, which is identical in design to LALME dot maps.

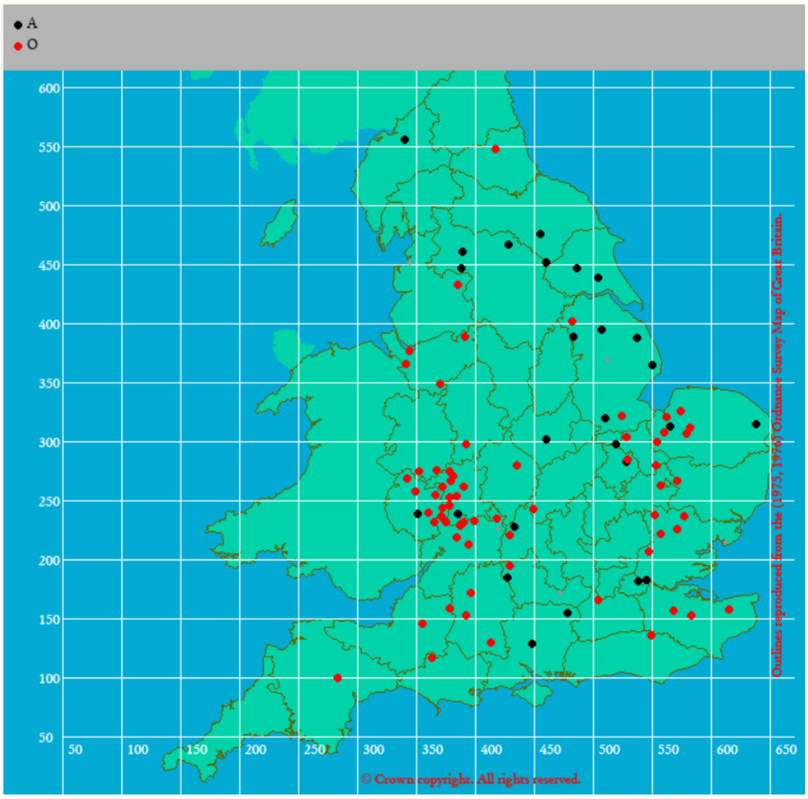

● A
● O
600
550
500
450
400
350
300
250
200
150
100
50
50
100
150
200
250
300
350
400
450
500
550
600
650
Outlines reproduced from the (1975, 1976) Ordnance Survey Map of Great Britain.
© Crown copyright. All rights reserved.

Figure 4: LAEME custom map for *a/o* in LAND, MAN, STRONG

LAEME data can be generally characterised by its exceptional level of detail and highly systematic and consistent tagging and transcription. The structure of the data offers rich querying possibilities, which go beyond the searches currently available in the online interface. The next section explains the principles and basic concepts of the methodology proposed for the present project, the objective of which is to open new querying possibilities and construct an interface tailored to the enriched data structure.

3.2. Chief points and principles behind the methodology

The purpose of this project from the very beginning has been to contribute an additional layer of data to the LAEME corpus, which would facilitate research into Early Middle English texts and dialects. The original plan to achieve this was to create a database of sound-spelling correspondences, which would have been very similar to the FITS project introduced in the theoretical chapter (see section 2.3.2.4). This approach would respond primarily to the requirement of "subdividing the attested material into more abstract types" (Laing & Lass, 2013: 2.3.3).

This intention was reconsidered in the light of the complexity of LAEME data and the diversity of spelling systems employed by the scribes. One of the most problematic aspects of potestatic interpretation is that it is often unclear what the assigned potestates actually represent. For instance, LAEME text #280 (London, British Library, Cotton Otho C xiii containing *Laȝamon B*) has both *w* and *ȝ* as reflexes of OE *g* in LAW/N (*laȝe, lawe*) and it is not unreasonable to suppose that the alternation might be due to exemplar influence and the scribe of the exemplar might have had a different sound in mind. Even if this assumption was confirmed (which can in fact prove impossible), should the two litterae be assigned different potestates or a single potestas, say [w], because it is the likely sound which the scribe of #280 pronounced in LAW/N? A possible solution would be to do the former and explicitly mark exemplar provenance. Another solution could be to do the latter and state that potestates in fact represent a reconstruction of the scribe's sound system, but this would mean that the connections between potestates and litterae, which would all look identical in the database, would in fact describe qualitatively different phenomena (intentional representation as opposed to accidental representation). In any case, different items in a single text could require detailed and lengthy analyses.

It is beyond the scope of this project to analyse *all* the text languages in LAEME and assign sound values to the litterae without failing to check all the relevant data. Still, the assignment of sound values is only one of the potentially useful upgrades of LAEME and a decision was taken to approach the problem from a different direction.

The basic concept of segmentation into smaller units was not abandoned but instead of taking up the interpretative challenge, the main question was how to best address the notorious problems encountered in historical dialectology, mainly the need to compensate for the lack of data by combining many different sources and perspectives. Rather than “adding” pieces of data to the imaginary network of relations between sound, spelling, scribal strategies, time and space, the task became to better navigate the network and in fact, devise ways to postpone interpretation so that each decision may be as well informed as possible. After all, one of the main virtues of electronic processing exploited by all language corpora is not that it produces additional data as such, but that it provides faster access to data. Moreover, it can reveal connections and patterns, which are otherwise difficult or even impossible to notice. The paragraphs below briefly explain the general principles and requirements which guided the design of the tool:

a) Enable identification of unusual features

The tool should provide means of highlighting (potentially meaningful) unusual features in the text, which are not easy to notice, such as conspicuously low frequency of a littera in a text. Moreover, if an unusual form is noticed it should be possible to check for similar features elsewhere in the data without too lengthy searching and reading through the texts.

b) Postpone interpretation

Fast access to possibly relevant data should allow the researcher to consult a relatively greater volume of data before taking interpretative decisions. Furthermore, the method used to construct the tool should involve as few interpretative choices as possible.

c) Re-usability, future compatibility with more data

The methodology proposed for the present project should be re-applicable to other sources of data, such as OE or LME texts, which could be stored in the same database as LAEME data. The design of the online interface should be data-neutral, i.e. useable to display data from another database with identical structure and/or output data.

d) Zooming (data scaling)

Relatively large-scale data should serve as a possible point of departure for explorative analyses and it should be as simple as possible to access specific pieces of evidence (specific forms), which constitute the larger picture.

The principles described above do not imply that less time should be necessarily needed to perform analyses using the tool. The goal is rather to spare the researcher's time so that s/he is able to consult more data and does not spend too much time on mechanical tasks.

The next subchapter explains the concept of segmentation without explicit assignment of sound values, which is at the heart of the methodology and serves as the foundation of all subsequent calculations and queries.

3.2.1. The concept of slots

The theoretical part mentioned a few projects and studies which work with segmentation of words into smaller units and this method is vital also for the construction of the spelling database. The crucial difference between the present project and the previously mentioned ones is that no explicit connection is established between specific *litterae* and *potestates*. Moreover, neither OE forms, nor PDE forms are chosen as reference points, although the tool does include experimental parsed OE forms for a limited number of items (see section 3.7.1).

All the slots are defined solely by their position in a specific word, e.g. the initial segment of FELLOW/N, which generally corresponds to *f*, *v* or *u* is one slot and the third position in LOVE/N (*f*, *u*, *v*, *w*) is also a slot. This structure of data allows us to observe specific litterae found in a given slot in a specific text or region etc. but we cannot perform queries like "what are all the representations of [v]?", which is possible with FITS (see section 2.3.2.4), because the explicit interpretation of sound values is not available.

In order to compensate for this, slots can be grouped dynamically based on the assumption that all the slots in which a specific littera appears can be regarded as potentially related. For instance, both of the above-mentioned slots FELLOW/N (1) and LOVE/N (3) are related because *f* and *u* can be found in both of them. Slots can be grouped within a single text as well as across texts and possible queries are formulated in the following manner:

- List all the litterae used interchangeably with *e* (in text #173).
- List all slots (and associated items (words)) in which the scribe of text #8 uses *h*.
- List all the slots (and associated items) in which *j* appears anywhere in LAEME.
- List all the slots (and associated items) in which the scribe of text #304 uses *g* interchangeably with *ȝ*.

The decision whether, e.g. the *g* and *ȝ* from the last example in fact represent the same sound or different sounds in the text language of #304 can be then based on the consideration of the list of items and other data available in the database. The data only provides a systematic framework for such decisions, which is in fact structurally close to the *scribal lexicon* and the idea of LSS described in the theoretical chapter (see subchapter 2.2.3.3), except the *potestates* are replaced by the abstract groups of slots. For instance, where the complete scribal profile would have a *literal substitution set* {v, f} for [v] in the initial position, the tool would have an abstract slot alternately filled with {*v*, *f*} and defined by a list of items and positions in which the litterae appear, e.g. FIRE/N (1), FROM (1), FOR (1). A list of such abstract slots can be generated for a single text, but also for a given region, period or the whole corpus. Another possibility would be to look at slots in a group of specific texts but this has not been implemented in the tool.

3.2.1.1. Slot alignment

In order to generate data in the format just outlined, it was necessary to align the diverse spelling variants (forms) of individual items, such as all forms of FIRE/N, separating the strings of characters into segments (slots) and identifying the segments which roughly correspond to each other. The target result can be displayed like this (selected forms only):

f | ie | r | e

f | uy | r | e

v | e | r | _

u | u | r | _

The underscores in this notation represent an empty slot. The alignment is based predominantly on the comparison of the individual forms in the group. OE or other source forms are normally not taken into account nor aligned with the rest of the forms, although this would be possible. The example of fire is one of the more straightforward ones in terms of segmentation and there seem to be only one preferred solution, but this is definitely not true of all the groups of spelling variants in LAEME. The following paragraphs explain the approach to the segmentation and alignment adopted here and some of the arbitrary decisions which needed to be taken to solve the difficult cases.

It is assumed that the individual spelling variants represent different sound streams (i.e. strings of non-discrete sounds), which can be roughly aligned with one another, even though they are not identical. The alignment of spelling variants then follows the alignment of the sound streams.

The differences between the individual variants may reflect either alternative representations of approximately the same sound or different sounds and therefore imply sound change. Any considerations regarding sound values were limited to the decision whether the assumed sound change can be reasonably analysed as affecting one segment (e.g. voicing, lenition, diphthongisation etc.) or whether it is better understood as involving two segments (e.g. insertions, deletions).

Another obvious but important assumption is that despite great variability of the spelling systems, there are limits to which litterae can be reasonably expected to represent similar sounds. Sequences of vowels are taken as one segment, with the exception of spelling variants, where the sequence of vowels seems to correspond to a sequence in another form involving a consonant. Sequences of consonants traditionally recognized as digraphs are usually considered single segments. Although it is preferable to create segments close to the familiar dihraphs e.g, align *sh* and *sch* rather than *s/s* and *h/ch*, the alignment should above all reflect correspondence between the segments.

These principles can be illustrated on the example of selected forms of THOUGHT/N:

þ | o | h | t

þ | ou | _ | t

þ | ou | s | t

dh | o | g | t

The first form, *þoht*, is by far the most frequent one and its rough structure seems to be CVCC (probably dental fricative – vowel – fricative - stop). It is not necessary (nor it is immediately possible) to decide whether the initial *dh* in the last form reflects a sound change (most likely voicing). Aligning it with *þ* makes more sense that (potentially) creating a separate slot for the *h*. The second segment is a vocalic one and *o* alternates with *ou*, which is likely to correspond to a diphthong or potentially lengthening, but there is no need to distinguish between the two possibilities at this point because both would be parsed the same, i.e. *o/ou* would share the same slot. The third segment seems to be missing from the second form and there are three different representations of it in the other forms (*h*, *s*, *g*). A loss of the segment is assumed for the second variant and, accordingly, the slot is left empty and *s*, *g* and *h* are aligned with it, while the interpretation of their sound values remains an open question.

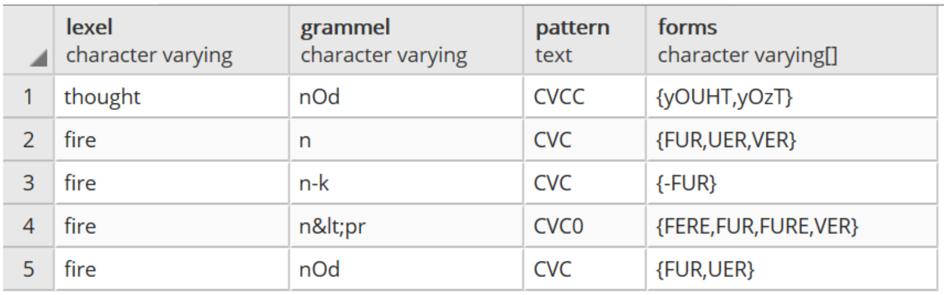
As the number of spelling variants in LAEME is high enough to make partly automated processing worthwhile, the main task at the initial stage of the project was to explore possibilities of such automatization, which is the subject of the next subchapter.

3.3. Pilot version – Poema Morale

The first attempt to process LAEME data, thereby transforming them to the desired database structure was limited to the seven texts of the *Poema Morale*.[8] The task was performed using a provisional methodology which was assessed and updated before being applied to the whole corpus. This section briefly describes the original methodology and presents a summary of its shortcomings which was considered before designing the final version of the methodology.

3.3.1. The steps of the analysis

The procedure was semi-automatic. At the outset, the unique combinations of lexel + grammel were stored in a separate table along with the list of forms found in the seven texts of the *Poema Morale* and complemented with slot "patterns" (to be explained shortly). The picture below shows several lines from the table for illustration:



	lexel character varying	grammel character varying	pattern text	forms character varying[]
1	thought	nOd	CVCC	{yOUHT,yOzT}
2	fire	n	CVC	{FUR,UER,VER}
3	fire	n-k	CVC	{-FUR}
4	fire	n<pr	CVC0	{FERE,FUR,FURE,VER}
5	fire	nOd	CVC	{FUR,UER}

Figure 5: Items of FIRE and THOUGHT in the pilot project

[8] The choice of the *Poema Morale* was motivated by the fact that some of its versions had been previously analysed using LAEME (Vaňková 2016).

Beside the patterns, the processing script also required a list of litterae and a list of sets, i.e. litterae which are likely to appear in the same slot. The next section discusses these three kinds of input in more detail.

a) Patterns

The first step of the analysis was performed manually. Each of the lexel + grammel combinations was assigned a pattern representing the assumed structure of the sound stream as a sequence of letters “C” (standing for *consonant*), “V” (standing for *vowel*). However, this notation soon proved insufficient for the subsequent automatic processing, mainly due to the frequent occurrence of “empty” slots. For instance, if some of the forms had initial *h* while others did not, the error rate of the processing script rose dramatically.

The solution to this problem was to make use of the fact that the alternation of an empty slot in one form with a littera in another is far from random and it is possible to identify specific sets of litterae, which are likely to be missing in one or more forms in a group, e.g. *e*, *h*, *s/n* etc. The initial list of such sets was based on secondary literature, but it was necessary to add more sets and litterae as the analysis proceeded. Each of such sets was assigned a label (a single capital letter), which could then be used in the pattern. For instance, OUT-/XP-V, which has three forms: *hut-*, *ut-* and *vt-* had the pattern “HVC” (instead of “CVC”). The patterns were stored in the table with lexel + grammel combinations and used as input for the processing script.

b) List of litterae

The list of litterae was taken from literature about ME spelling (Fisiak, 1986; Upward & Davidson 2011). The only step required here was to transform the information found in handbooks to a machine-readable format. JSON[9] was used for this purpose. The data corresponded to a table with two columns: “littera” (the actual littera, e.g. *f*, *ch* etc.) and “category”, which had three possible values: “C” (consonant), “V” (vowel) or “A” (ambiguous).[10] The third category was introduced to deal with litterae which are known to represent consonants in certain positions and vowels in others, mainly *i*, *y*, *u*, *w*.

[9] JavaScript Object Notation – a lightweight data format commonly used in web applications.

[10] The reason why there are only three categories as opposed to the wider range of labels used in the patterns (e.g. ”H” mentioned above) is that the more specific labels only make sense in the context of a specific word. For instance, *l* is categorized as “C” on the list of litterae and only some of the slots in which it appears are labelled “L” in the pattern because the segment drops only in a quite restricted number of words (e.g. *each*, *much*).

Such categorisation is obviously based on a supposed sound value and as such already involves interpretation, which is, however, practically undisputed in previous research and the data obtained in this way is very useful for the script.

c) Sets of litterae

Similarly to the list of litterae, the initial version of the list of sets was copied from previously published resources on ME spelling and the data was simply typed into the computer. An example of such set are the various spellings for [ʃ], i.e. *sh*, *sch*, *ss* etc. Litterae possibly representing two different sounds involved in a sound change, e.g. *k* and *ch* were kept apart at this stage.

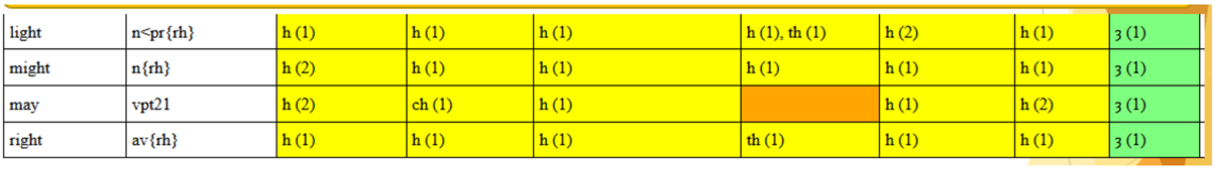
The next step was to feed the input described above to a computer script. The script processed the forms one-by-one and tried to fit the form into the prescribed slot pattern. Besides the number of slots, it also checked whether the category of each littera corresponds to the label of the slot ("C" or "V") or if the littera is a member of the set "prescribed" for the given slot, e.g. [h, 0]. There were three possible outcomes of the analysis:

1. The script found only one possible solution which satisfied the criteria (the number of slots as well as categories matched). This was the prevalent outcome (over 90 %). The results could be saved straight into the database.
2. The script identified multiple (typically two) "solutions" which satisfied the criteria. In such cases, the correct solution had to be selected manually before saving, which was relatively quick and easy.
3. The script was unable to produce an alignment which would satisfy the criteria. This happened if there was a problem with the input data, i.e. a littera (usually a digraph) or a set was not present on the list or the manually provided pattern was incorrect (usually due to oversight). Therefore, the input data had to be amended accordingly before proceeding with the processing.

3.3.2. Data testing

The experimental processing produced a small database covering the seven texts of the Poema Morale, which was structurally very close to the final version based on the whole corpus. In order to check the data, an SQL query was used to retrieve a complete list of "sets" (groups of litterae sharing the same slot) and the list was checked manually. A number of unlikely sets, suggesting an error in the data, were identified, e.g. {_, m, s, t}, {r, u, v}.

A very simple webpage was created for the purpose of testing the actual use of the data. The webpage displayed the inventory of litterae for each of the seven texts and it was able to generate a table comparing the usage of a selected littera across texts. The script first generated a list of slots in which the selected littera appeared in *any* of the compared texts and displayed the littera (or multiple) litterae appearing in the given slot in a specific text. The results were displayed as follows:



light	n<pr{rh}	h (1)	h (1)	h (1)	h (1), th (1)	h (2)	h (1)	ȝ (1)
might	n{rh}	h (2)	h (1)	h (1)	h (1)	h (1)	h (1)	ȝ (1)
may	vpt21	h (2)	ch (1)	h (1)		h (1)	h (2)	ȝ (1)
right	av{rh}	h (1)	h (1)	h (1)	th (1)	h (1)	h (1)	ȝ (1)

Figure 6: Slot comparison in the seven texts of the Poema Morale (pilot project)

The first two columns of the table specify the item and the remaining columns list the litterae in the compared texts (one column per text). The cells which contain the littera which was used to trigger the comparison have a green background.

This simple tool was used to perform a comparison of the usage of litteare associated with palatalization (*c*, *ch*, *k*, *g*, *j*). The results of the analysis are not presented here because its primary objective was to test the usefulness of the data and check for errors. The evaluation of the methodology is presented in the next section.

3.3.3. Assessment of the original methodology

a) Patterns

The speed of processing was satisfactory but the this might have been partly due to the limited scope of the pilot project. The average number of different forms of each item was obviously lower when restricted to the seven texts, therefore, less time and effort were needed to write the slot patterns. As the analysis proceeded, the list of "optional" sets of litterae became disorganised and difficult to manage, which was considered to be another drawback of manually provided patterns. Moreover, there was a large number of items with a relatively straightforward structure without empty positions etc., which nevertheless required manually written pattern. All of these observations led to the consideration of the possibility to generate patterns automatically.

b) Items

The grouping of forms under different items turned out to be a major flaw of the original methodology, which grossly underestimated the variability of grammels in LAEME. The forms of a single lexel which should be comparable phonologically were often split into multiple groups based on slight differences in grammels marking phonologically irrelevant features. As a consequence, statistics became distorted. For example, if a littera was found in "5 different items" according to the data, there could be only one or two distinct lexels involved.

Another problem was that items were tag-based, i.e. the forms of the whole tags, including endings were grouped together. As a result, morphological differences in endings, such as *-es* vs. *-en* became mixed with purely graphological or phonological differences. For instance, *n* was counted among the alternatives of *s* along with *z*, *ss* etc. and the distinction between the two types of alternation (morphological and phonological) was not formally marked in the database.

c) Sets

The incidence of errors revealed by the analysis of sets of alternating litterae retrieved from the database was higher than expected (ca 20 %).[11] On the other hand, the list of possible sets retrievable from the database was clearly more comprehensive than the list initially taken from literature and as such provided a useful input for further processing (its use is to be described below).

3.3.4. Requirements for the updated methodology

Considering the problems outlined above, a modification of the original methodology was deemed worthwhile. The main requirements for the final version can be summarized as follows:

a) Seriously reconsider the definition of items. Morpheme should be the basic unit instead of word and phonologically irrelevant differences in grammels should be disregarded.

b) Generate patterns programmatically, at least for the less problematic sets of forms. Keep patterns maximally simple. On a more general level, manual intervention

[11] This is the ratio of clearly "wrong" sets like {*f*, *n*, *u*}, {*r*, *t*, *n*} calculated from the total number of all possible unique sets regardless of their frequency. Only the "largest possible" sets were counted. For instance, if {*f*, *u*} and {*f*, *v*} were in fact subsets of {*f*, *u*, *v*}, only the last set appeared on the list. This implies that the error rate calculated in this way is higher than the ratio of wrongly aligned *forms*.

should be required to resolve problematic cases rather than be a standard part of the process for each item.

c) Make a good use of the list of possible sets of litterae.

3.4. Postgres (tabular) version of LAEME and database structure

First of all, the structure of the spelling database will be described. This will hopefully make it easier to understand the individual steps of the updated parsing process to be discussed later on. This subchapter essentially lists all the database tables and briefly characterises their content. It proceeds from the tables constructed directly from LAEME data to the tables populated with the segmented data. The final section of the subchapter lists all the derived tables, which were constructed programmatically from the primary tables, such as tables with statistical data.

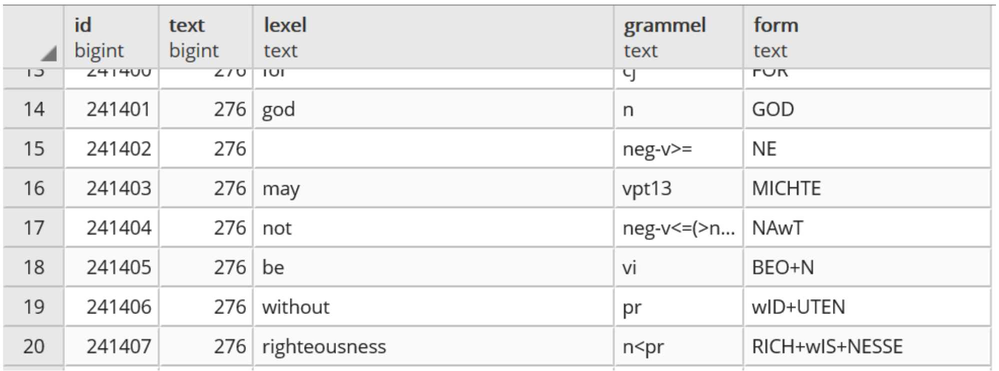
3.4.1. Tables with original LAEME data

The original data from LAEME is contained in the database alongside the added (segmental) data. Database tables cover the tagged texts constituting the LAEME corpus, including metadata and also selected pieces of information available in the LAEME Index of sources, namely the titles of texts contained in the individual files (e.g. *Poema Morale, Ancrene Riwle* etc.) and a tabular representation of links between corpus files like (supposed) common exemplars or copyists.

Raw files from LAEME were downloaded, parsed and stored on a local machine in the form of an PostgreSQL relational database (i.e. tables). Four tables were created in this way:

a) Tags

The table has five columns: *lexel*, *grammel*, *form*, *text* and *id*. The data in the first three columns correspond to the beginning of a single row in a LAEME file, i.e. the lexel, grammel and form of one word, including affixes and endings. *Text* gives the id of the manuscript sample, which remains the same as in LAEME. *Id* is a unique identifier of the row assigned automatically. The table preserves the order of words in the LAEME text. The picture below shows several rows taken from text #276 for illustration. The corresponding passage in the text reads *god ne michte napt beon piduten richpisnesse*:

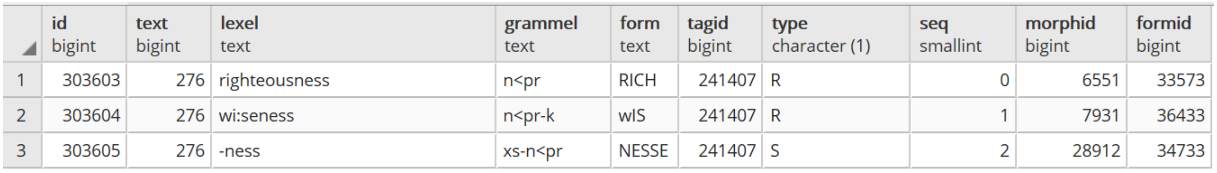


	id bigint	text bigint	lexel text	grammel text	form text
13	241400	276	for	cj	FOR
14	241401	276	god	n	GOD
15	241402	276		neg-v>=	NE
16	241403	276	may	vpt13	MICHTE
17	241404	276	not	neg-v<=(>n...	NAwT
18	241405	276	be	vi	BEO+N
19	241406	276	without	pr	wID+UTEN
20	241407	276	righteousness	n<pr	RICH+wIS+NESSE

Figure 7: LAEME data as a table - tags

b) Morphemes

This table has a very similar structure as the first one. In addition to the previously mentioned columns, it has the column *tag_id* , which references the *id* of the previous table. The table stores lexels, grammels and forms of the individual morphemes constituting a single tag in the previous table. Additionally, a column labelled *type* indicates whether the morpheme is the root (R), affix (A) or ending (E) and the column *seq(uence)* indicates the order of the morpheme within the word. Two more columns were left blank at the beginning – *morphid* and *formid*. These columns are needed to link LAEME data to the new database (see below). The picture below shows the rows linked to the last row in the previous picture (RIGHTEOUSNESS):



	id bigint	text bigint	lexel text	grammel text	form text	tagid bigint	type character (1)	seq smallint	morphid bigint	formid bigint
1	303603	276	righteousness	n<pr	RICH	241407	R	0	6551	33573
2	303604	276	wi:seness	n<pr-k	wIS	241407	R	1	7931	36433
3	303605	276	-ness	xs-n<pr	NESSE	241407	S	2	28912	34733

Figure 8: LAEME data as table - morphemes

c) Comments

This table has only two columns – *tag_id* and *comment*. It contains comments, line breaks or other information given in curly braces in the original LAEME file, referencing *id* of the tag in table *tags*.

d) Text index

This table was constructed from the data found in the header of each file (i.e. information about the text) combined with statistical data retrieved from the spelling database. The columns with LAEME data are: *manuscript, fols*, *hand, localisation*, *script*, *date* and *anchor*. As the name of the columns imply, the original LAEME data from the field "manuscript" were split into three separate fields. The field *manuscript* identifies the manuscript (e.g. "Oxford, Bodleian

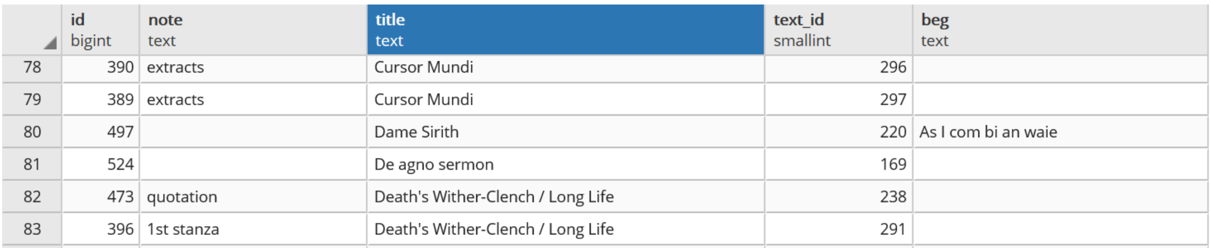
Library, Digby 86”), while the information about the actually transcribed folios was moved to the column *fols* and the siglum identifying the scribe is in the column *hand*.

As the purpose of this table is to enable filtering and grouping of the data from the spelling database, which means that it should be machine-readable, rather than human-readable, the original data from LAEME was slightly modified manually (see section 3.6.7 below). The titles of the texts found in the manuscripts were stored in a separate table (see below). The column *anchor* was added manually to enable easy identification of anchor texts. Possible values are “A” for *anchor*, “D” for *documentary anchor* and “L” for *literary anchor*.

Besides the tables created programmatically, there are also two tables created mostly from the LAEME Index of sources. Similarly to the table *text index*, these tables do not add anything to the original data, they merely enhance search possibilities, especially quick reference to possibly related texts. The first table holds titles of texts linked to the corresponding corpus files. The second table lists all the connections between the files (shared exemplars etc.). The structure of these two tables is described below.

e) Text titles

The table has four columns: *text_id*, *title*, *note* and *beg* and it holds data about all the texts in LAEME. *Text_id* references *text_index*, *title* gives the title of the text or its general label used in the index, such as “*a verse on the vanity of the world, a song of Passion*” etc. The column *beg* specifies the beginning of the text, which is also frequently found in the LAEME Index of sources. Wherever the index also specifies whether the manuscript contains only fragments / quotation from the text etc., this information is stored in the column *note*.). As the table was designed to quickly group manuscripts containing the same text, it was sometimes necessary to adjust the titles so that one text does not appear under variant titles on the list. If there were slightly differing “general labels” referring to the same text, only one of the labels was used for all the texts. If a text appeared under different titles e.g. *Trinity Homilies* vs. *Lambeth Homilies*, which in fact share some of their content, both titles separated by a slash were used. See the picture below for illustration:



	id	note	title	text_id	beg
	bigint	text	text	smallint	text
78	390	extracts	Cursor Mundi	296	
79	389	extracts	Cursor Mundi	297	
80	497		Dame Sirith	220	As I com bi an waie
81	524		De agno sermon	169	
82	473	quotation	Death's Wither-Clench / Long Life	238	
83	396	1st stanza	Death's Wither-Clench / Long Life	291	

Figure 9: LAEME data as tables - text titles

f) Manuscript links

The Index of sources mentions a considerable number of connections between the manuscripts (e.g. "text A was copied by the same scribe as text B"), however, this data cannot be easily and systematically targeted in searches. The table *manuscript links* was constructed as a formalized list of connections sorted into a few categories. It has three columns: *a_id*, *b_id* and *type*. The first two columns hold LAEME ids of the related texts and the last column specifies the type of connection. The possible types include:

- *ms* – The texts appear in the same manuscript. This information is retrievable also from *text_index* (and it was in fact copied from there), but it was included for the sake of structural integrity of the database.
- *scribe* – The texts were copied by the same scribe.
- *exemplar* – The texts (probably) shared an exemplar.
- *similar L(anguage)* – Previous research suggests a connection between two texts based on similar text language or common unusual forms.

3.4.2. Spelling database structure

The construction of the spelling database proper consisted in creating an index of items and an index of forms from LAEME data and subsequent segmentation of the forms. The spelling database comprises the two indices plus a table containing the parsed data. These new tables are linked to the tables generated straight from LAEME and described above. This section briefly outlines the structure of the tables. The relations between the three "core" tables are shown in figure 10 below:

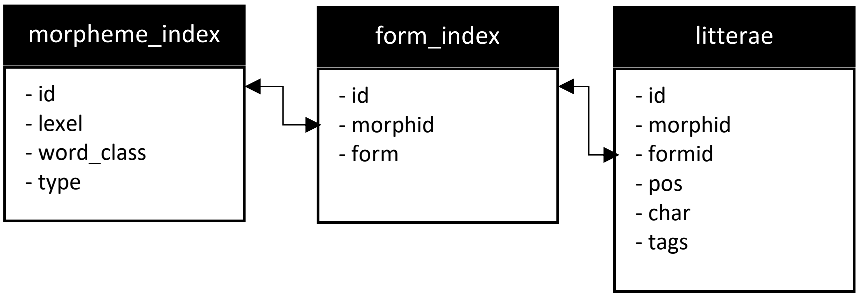


Figure 10: Spelling DB core tables

Morpheme index lists all the items, *form index* lists all the forms and the table *litterae* holds information about all the litterae found in a specific slot. Each row in the table *morpheme_index* corresponds to a single item, each row in the table *form_index* corresponds to a single form related to a particular item and each row in the table *litterae* corresponds to a single segment in one of the forms. This structure is almost identical to the FITS database (presented in subchapter 2.3.2.4). The individual tables are described in the following section.

3.4.2.1. Core tables

a) Morpheme index

Each row in this table corresponds to a unique item, which is defined by a combination of a lexel and simplified grammel. There are four columns: *lexel*, *word_class*, *type* and *id*. The first three columns hold data taken from the table *morphemes*. The label *word_class* (rather than *grammel*) is used to indicate that the values is not an exact copy of *grammel*. The value of *id* is the automatically assigned unique identifier, which is necessary for linking each item with its various forms stored in the table *form_index*.

b) Form index

This table references *morpheme index* and stores a set of forms for each combination of *lexel* and *word_class*. It has four columna: *morphid*, *form*, *corpus_form* and *id*. The column *morphid* holds the reference to the *id* column in *morpheme index*. The individual forms are stored in the original LAEME ASCII format as well as the updated format in the columns *corpus form* and *form*. *Id* is again the unique identifier of the form in question, which can be referenced from other tables.

The structure of the table implies that all occurrences of a single form always share only one slot structure. Although it appears theoretically possible that two versions of alignment should be used for a single form to reflect its different use in different text languages, this was never actually done, because the effort needed to reveal and verify such cases would not be adequate to the value of the result.

c) Litterae

This is the table which holds the parsed data. While in the previously presented tables, one row corresponded to one morpheme, ech rows in this table corresponds to a *slos* or "position" in the morpheme. A slot is always defined by combination of the unique *id* from the table *form_index* and a number (1-n) specifying the position. It follows from this that three columns are needed: *formid*, *position* and *littera. Formid* references the *id* from *form_index* and *littera* gives the littera found at the given *position* (*slot*). (There is also the unique *id* for each row but unlike with the indices, this column is included to satisfy the formal requirements of the database and plays no role in the queries.)

Joining the three tables together produces the following result:

morpheme_index			***form_index***		***litterae***	
lexel	**word_class**	**Id/morphid**	**Form**	**Id/formid**	**pos**	**char**
righteousness	n	6551	Reyt	33566	1	r
righteousness	n	6551	reyt	33566	2	ey
righteousness	n	6551	reyt	33566	3	_
righteousness	n	6551	reyt	33566	4	t
righteousness	n	6551	rihht	33567	1	r
righteousness	n	6551	rihht	33567	2	i
righteousness	n	6551	rihht	33567	3	hh
righteousness	n	6551	rihht	33567	4	t

Table 2: Table join of the three core tables in the spelling DB

The table shows data for two different forms of RIGHT- in RIGHTEOUSNESS/N – *reyt* and *rihht*. The scope of the individual tables is indicated in the topmost row, which shows that the columns "id/morphid" and "id/formid" are shared between two tables.

The order in which the tables were presented reflects the order of their construction which proceeded in a cascade-like manner, i.e. an item had to be present first in *morpheme index* before its forms could be listed and parsed.

The database also includes three more "experimental" tables, which were included to test a way of linking database data to sources other than LAEME. These tables are fully integrated in the database and can be queried but only a slight amount of data is currently available. Their structure is briefly outlined below.

d) Source forms

This table is structurally very similar to *form_index* and it holds mainly OE but also PDE forms of LAEME items. Besides the columns *formid, morphid* (referencing *morpheme index*) and *form*, it includes the columns *language* (specifying the period of form attestation) and *dilalect*, which is available to specify regional provenance of the form if it is known. In theory, any number of forms from various sources can be inserted into this table. The forms currently present in the table were extracted from CoNE or looked up in the OED and inserted manually. About 20 forms are available (see Appendix 7.10.7).

e) Source litterae

This table has a structure identical to *litterae* and it holds parsed source forms. It references the columns *formid* and *morphid* from the previous table.

f) CoNE sets

This table contains selected codes of phonological and orthographic changes described in CoNE (column *label*) and relates them to a set of litterae, which may potentially indicate the presence of the change (column *set*). For instance, l-dropping is related to the set {*l*, _}. The columns *position* and *context* hold positional tags or contextual specification if they are defined in CoNE. The mechanism is going to be explained in more detail in subchapter 3.7.2.

3.4.2.2. Derived tables

The last group of tables to be presented in this section are tables generated by transformations of the data from the "core" tables. This means that the data in these tables could be calculated straight from the "base" tables, but the calculations would be so complex, that they would seriously affect performance of the tool. Therefore, all the necessary calculations were done in advance and the results were stored for future use.

a) Litterae statistics

This table lists all the litterae (segments) found in the corpus and gives their average normalized frequency, i.e. the average of the normalized frequencies for the individual text languages. The table also holds phonetic information about the "prototypical" sounds represented by the litterae, which can be used in filters (see subchapter 3.6.7.4 for details). Table sample is available in Appendix 7.10.1.

b) Text-litterae statistics

This table stores raw frequency plus normalized frequency of specific litterae in specific texts. Table sample is available in Appendix 7.10.2.

c) Rare uses

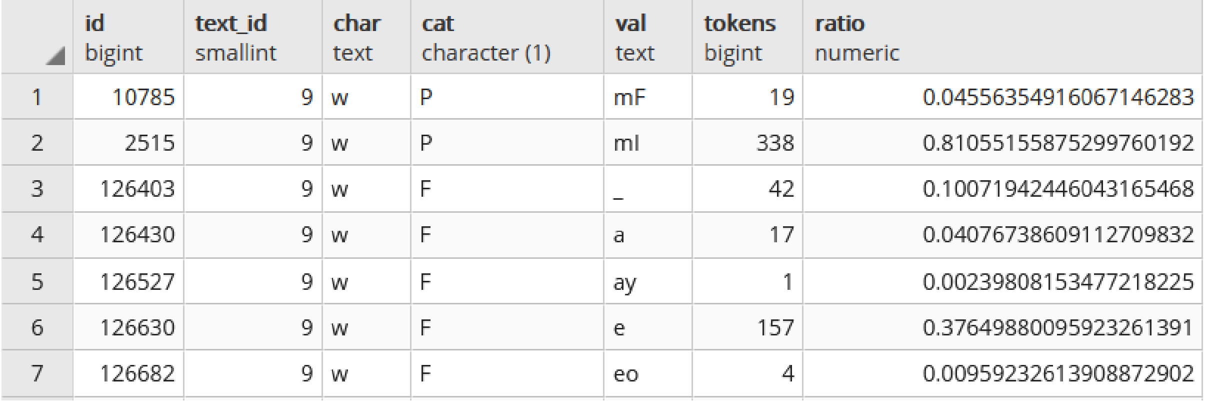
This table lists rare combinations of slots and litteare in a given text. Table sample is available in Appendix 7.10.3.

d) N-grams

The table lines up every three subsequent rows from the table *litterae* as one row, which makes it possible to filter the results of the search by the previous or the following littera, e.g. search for all cases where *c* is followed by *e*.

e) Constraints

Following the proposed model of scribal lexicon (discussed in subchapter 2.2.3.3), positional and contextual constraints for the individual litterae were calculated and stored in a separate table. The table gives the ratios between the use of a littera at a specific position or in a specific context and the total number of its occurrences in a given text. The statistics for positional constraints are calculated from positional tags in the table *litterae*, which means that they include morpheme-initial and morpheme-final positions. Contextual constraints are calculated from *N-grams* and they are defined by the litterae following or preceding the littera in questions. The screenshot below shows several positional constraints for the littera *w* in text #9 (a version of the *Poema Morale*):



	id	**text_id**	**char**	**cat**	**val**	**tokens**	**ratio**
	bigint	smallint	text	character (1)	text	bigint	numeric
1	10785	9	w	P	mF	19	0.04556354916067146283
2	2515	9	w	P	ml	338	0.81055155875299760192
3	126403	9	w	F	_	42	0.10071942446043165468
4	126430	9	w	F	a	17	0.04076738609112709832
5	126527	9	w	F	ay	1	0.00239808153477218225
6	126630	9	w	F	e	157	0.37649880095923261391
7	126682	9	w	F	eo	4	0.00959232613908872902

Figure 11: Positional constraints on <w> in text #9 (sample)

The first column gives the littera to which the constraints apply, the second column *cat* indicates constraint category – "M" for instance, stands for "position in the morpheme", *val* gives the positional tag for positional constraints of littera for contextual constraints. The column *tokens* gives the number of instances of *w* in the respective context or position and the column *ratio* gives the ratio between *tokens* and total number of instances of the littera in the given text. The id of which is found in the last column *text_id*.

e) Chunks

This table was generated from the table *litterae* and it is identical to it in structure. The value of the field *char* merges the litterae at two subsequent positions. The reasons for including this table are going to be discussed in subchapter 3.6.6 below.

f) Special Features

"Special features" is a cover term for graphic features of the text which are potentially relevant for analysis but lack direct phonetic significance, e.g. insertions, deletions, unexpanded abbreviations. It has the columns *feature* specifying the type of feature (e.g. superscript, deletion etc.), *char*, which gives the littera to which the feature is related, *text*, which relates the feature to a specific text language, and *mid* referencing the *id* column in table *morphemes*.

The next section explains the construction of the tables presented in this section.

3.4.3. Morpheme index

The table *morpheme_index* was constructed primarily with SQL queries, which selected a part of the data from the table *laeme_morphemes* (typically unique combinations of *lexel* + the initial part of the *grammel*) and inserted them into a new table.

The fundamental principle governing the construction of the index was that in the ideal case, *all* comparable forms should be linked to a single item. It has been explained above that since the grammatical information from LAEME is very detailed, it would have been inappropriate to simply group unique combinations of lexels and grammels because this would result in treating comparable forms as separate items. For instance, no connection would be made between forms of a single noun that only differ e.g. in position (rhyming/non-rhyming) etc. In order to avoid this problem, the lexels and grammels from LAEME were slightly modified. Lexels were stripped of lexel specifiers and predefined sets of grammels were subsumed under a simplified label, which means that the respective forms fell in a single group instead of two or more groups. The analysis focused primarily on lexical words, i.e. words having a specific lexel in LAEME and did not cover all grammatical items, which was partly due to lack of time and partly due to the fact that differences between the forms often have grammatical rather than phonological significance.

In the vast majority of cases, the labels corresponded to the initial section of the grammel, which typically indicates word class. The following section gives an overview of the labels generated in this simple way. The categories are ordered by the number of items they comprise.

a) Nouns (“n”, 5324 items)

b) Adjectives (“aj”, 1564 items)

c) Adverbs (“av”, 1016 items)

d) Numerals (“q”, 142 tems)

e) Suffixes (“xs”, 46 items)

f) Prefixes (“xp”, 41 items)

g) Interjections (“i”, 34 items)

h) Lexical verbs

The items for lexical verbs distinguish between tenses (i.e. there are separate items for present, past and the subjunctive form), because their roots may differ.

k) Modal verbs and *to be*

The reflexes of PDE modal verbs *may*, *shall* and *will* and the verb *be* have different forms depending on person and number and they are transcribed as single morphemes in LAEME. Therefore, separate groups of forms for different numbers and persons were needed. As for modal verbs, all the plural forms were considered comparable and received *word_class* “vps2” and “vpt2”. Singular forms have distinct items for every person, i.e. “vps11”, “vps12” and “vps13” for present tense forms and “vpt11”, “vpt12” and “vpt13” for past tense forms.

Present forms of the verb *be* have separate items for every combination of time and person and sometimes also a given type of form. For instance, second person plural forms can be of the *beo-* type or the *are* type, spelling variants of which are of course not comparable, so distinct items were created for each of the two types.

With certain uninflected items (e.g. TO, FOR), it was preferable to group them by lexel only rather than the combination of lexel and word class. For instance *after* or *out* can appear as adjectives, adverbs, conjunctions or prepositions but there is no obvious reason to suppose that the scribe systematically used different spellings for the different word classes. Moreover, the original word class remains accessible in the database all the same.

Adjectives and adverbs are slightly problematic in that their comparative and superlative forms are usually transcribed as separate morphemes, e.g. *fair+est* (FAIREST), *briht+ere* (BRIGHTER) but there are exceptions to this, e.g. *heihste* alongside *heg+este* (HIGHEST). A few of such words which required it were split into two or three separate items (see subchapter 3.4.3.1), but a vast majority of adjectives and adverbs are covered by a single item, because the ending has a separate tag.

Numerals were found to have too diverse forms for the items to be reasonably aligned and they were excluded from analysis for the time being.

Grammatical words, which have no lexels in LAEME required more manual work, especially because shortened grammels were not always suitable as *word class* labels so the value for the column had to be defined manually. The complete list of labels and corresponding grammels can be found in appendix 7.5. The forms of grammatical items in a single group often are highly variable and it does not always make sense processing them. This is why only some of the items were segmented. A complete overview of grammatical items along with the number of their forms and the number of processed forms is available in appendix 7.6.

It should be noted that the reason for preferring larger groups of items, as opposed to smaller groups defined by different grammels, is that the data is more flexible. The original grammels remain accessible in the database, which means that the items can be regrouped if there was a reason to do that. For instance, each adjectival item could be split into three separate items – the positive, the comparative and the superlative.

As new items were added to the table *morpheme index*, their *ids* were inserted into the column *morphid* of the table *morphemes*, which holds the data taken from LAEME. This enabled to join the two tables and retrieve all the forms linked to a given item. For example morpheme id (morphid) 6551 was generated for the root morpheme of the previously mentioned RIGHTEOUSNESS/N. The column *morphid* in table *laeme morphemes* for all the rows having lexel RIGHTEOUSNESS and grammel beginning with "n" was updated using an SQL query. All eleven distinct forms subsumed under this item could be subsequently retrieved.

Another query was used to count the number of distinct forms related to each of the items. Single forms (one form per item) were not subject to processing, as they are markedly less useful because we can never compare, e.g. different representations of a specific sound in the given item. There is a small number of exceptions to this rule and these are forms containing extremely rare sequences of graphemes. Some of these forms were parsed to provide more examples of such unusual spellings. The forms were identified at the final stage of the analysis dealing with exceptional spellings.

3.4.3.1. Split items

Some of the items had forms which were not suitable for comparison as one group because the alignment would create correspondences whose phonetical significance was doubtful at best and/or too many correspondences between multiple litterae and empty slots. Typical examples of such items are AS or THENCE. Separate items for this kind of lexels (e.g. the forms of THENCE of the type *thene* as opposed to *thede*) were created manually because the different groups do not have distinct grammels.

The next step consisted in retrieving and parsing the forms, thereby populating the tables *form index* and *litterae*.

3.4.4. Form index and litteare

The tables *form_index* and *litterae* were populated in the course of the parsing process. One group of forms associated with a single item was loaded and parsed at a time and the result was immediately saved to the two tables. The parsing method can be characterised as semi-automatic. The majority of items were processed on the command line using a Python script. The script was a complete rewrite of the one used in the pilot project based on the *Poema Morale*. The main difference between the two is that the new script only takes as input the list of litterae, a list of forms for a specific item and a list of "valid" sets of litterae. "Patterns" in the form of sequences of capital Cs and Vs remained in use, but they were not prescribed manually but generated by the script. The input is described in more detail below:

a) The list of litterae

Two more columns were added to the list used in the pilot roject:

- *Length*, which indicates the number of characters, e.g. "1" for *o,* "2" for *sh*, and "3" for *sch*.
- *Default digraph*, whose value is either *true* or *false*. This value tells the script, whether the sequence of graphs should be automatically considered a digraph when parsing. Out of the 108 digraphs on the list 63 were marked as *true* ("digraphs proper", e.g. *ch*, *sh*, *ee*, *ea*) and 45 as *false* ("contextual" digraphs, e.g. *ov*, *ng*, *gn*, *iw*).

b) The list of forms

The list of unique forms was retrieved from the table *morphemes*. The forms were stripped of some special characters used in LAEME to signal insertions, deletions etc. (see appendix 7.8 for details). The use of uppercase and lowercase letters employed in LAEME to avoid special characters (see subchapter 3.1) is not preserved in the new database. The combinations of uppercase vowel + *x/v* used in LAEME to transcribe diacritics were replaced with the corresponding symbols, e.g. "Ax > á, Ex > é" etc. The lowercase letters used to transcribe special characters, such as *thorn* and *wynn*, were replaced with the actual characters and capital characters were converted to lowercase.

c) Valid sets of litterae

The output of the pilot project was used here. First, all the sets found in the processed data based on the *Poema Morale* were retrieved. Highly unlikely combinations of litterae which were probably due to errors in parsing were deleted manually. New sets were added in the course of the analysis.

3.4.5. Processing with script

The general principle behind the script is to generate several possible alignments of litterae in the slots, evaluate them and return the "best" alignment. One set of forms is processed at a time and there are four stages to the task:

a) The list of forms is retrieved. A set of all possible "patterns" (CV sequences) is generated based on the list of forms and the list of litterae. The number of possible patterns depends on the number of ambiguous litterae and potential digraphs. If a form contains an ambiguous littera, two different patterns are generated. For instance, the status of *y* in the form "yfel" of *evil* is undetermined in the input data. Therefore, the script generates two patterns: "**C**CVC" and "**V**CVC". Similarly, multiple alternatives are generated if a sequence of litterae might constitute a digraph. Each form is processed individually at this stage, which means that at the end of this stage, each form has a list of potential patterns attached to it and only some of the patterns appear with all of the input forms. This procedure cannot ensure that the correct pattern will always be present, i.e. it has a high recall but possibly low precision.

The patterns are stored along with the sequence of litterae which fits the pattern, e.g. `{pattern: VCVC, split: [y,f,e,l]}`.

b) All proposed patterns are filtered and ranked and the "valid" ones are chosen as the basis for analysis. There are two filtering criteria:

- *Minimal length*: It is common for the patterns to vary in length because there can be empty slots in some of the forms. Minimal length corresponds to the shortest pattern generated for the longest form, i.e. the lowest number of slots needed to parse the longest form. Shorter patterns are excluded from analysis. For example, the form *alf* of HALF would be associated with the pattern "VCC" but this form is found in one group with *half*, which is clearly too long to be parsed as "VCC", therefore "VCC" is discarded.

- *Phonotactic criteria*: In some cases, automatic generation produces patters which are not in accordance with the possible configurations of vowels and consonants in the sound stream, e.g. sequences of multiple vowels, four consonants in syllable onset etc. Such patterns are also excluded.

The remaining patterns are considered potentially correct and an alignment is produced for each of them at the following stage.

c) The script works with one of the patterns at a time, filling the individual slots in the pattern one by one. First it fills the first slot for all forms and then it moves to another slot. The forms which do have the selected pattern on their list of potential patterns generated in step are processed first. The order in which the slots are filled is indicated in the following table:

C	V	C	C
h^{1}	a^{3}	l^{5}	f^{7}
_^{2}	a^{4}	l^{6}	f^{8}

Table 3: Filling slots - the order

The first form on the list serves as a model for the others. The segmentation of the first form is silently assumed to be correct. Once there is a littera present in the column it becomes possible to check whether the next littera to be placed in the column is a likely alternative of the already present litterae. This can be achieved by checking the potential new set of litterae in the column against the provided list of valid sets. In the example above, the script first placed the "h" from "half" in the first slot. Then the script took "a" from the beginning of "alf" (which has only one possible segmentation: `[a, l , f])` and checked whether $\{a, h\}$ appears anywhere on the list of valid sets. As this was not the case, the first slot for "alf" was left empty.

The script is also capable of managing multiple working options of parsing for the individual forms, some of which may prove wrong in the course of the analysis. Moreover, it keeps track of "rejected" litterae which were considered as "candidates" for a given slot but the projected new set was not found on the list of valid sets. Thus, *a* from the example above was listed as a rejected littera at position 1. At the end of the analysis, the script either succeeds or fails in imposing one of the patterns on all of the individual forms. The version with the lowest number of incorrectly parsed forms is returned as the preferred solution along with the chosen pattern and the list of rejected litterae.

Considering the high variability of Middle English spelling, it was expected that only items with smaller groups of relatively unproblematic forms will be processed correctly without manual intervention, which will be necessary in the case of more complex group of forms. In order to prepare for this situation, the script was designed so that possible causes of its failure would be maximally predictable. This enabled to propose "standard" ways of dealing with the individual causes of failure, which are going to be discussed shortly. The analysis could be performed in three different "modes" with varying level of automatization.

3.4.5.1. Semi-automatic

This mode was used at the beginning of the analysis so that the script could be tested and debugged, and also at a later stage to deal with the forms which were too complex for the script to handle without manual intervention.

The script was run manually from the command line. If the result was satisfactory, it was immediately saved to the database. If not, it was necessary to identify the cause of the failure and perform manual adjustments. The following problems were expected to occur:

a) The script picks the wrong pattern as the best basis for parsing, because it achieves better results with it. This error should only occur in combination with another one, otherwise the right pattern should produce better results. If it does, it is possible to set the pattern manually. The researcher simply writes the correct pattern and re-runs the script.

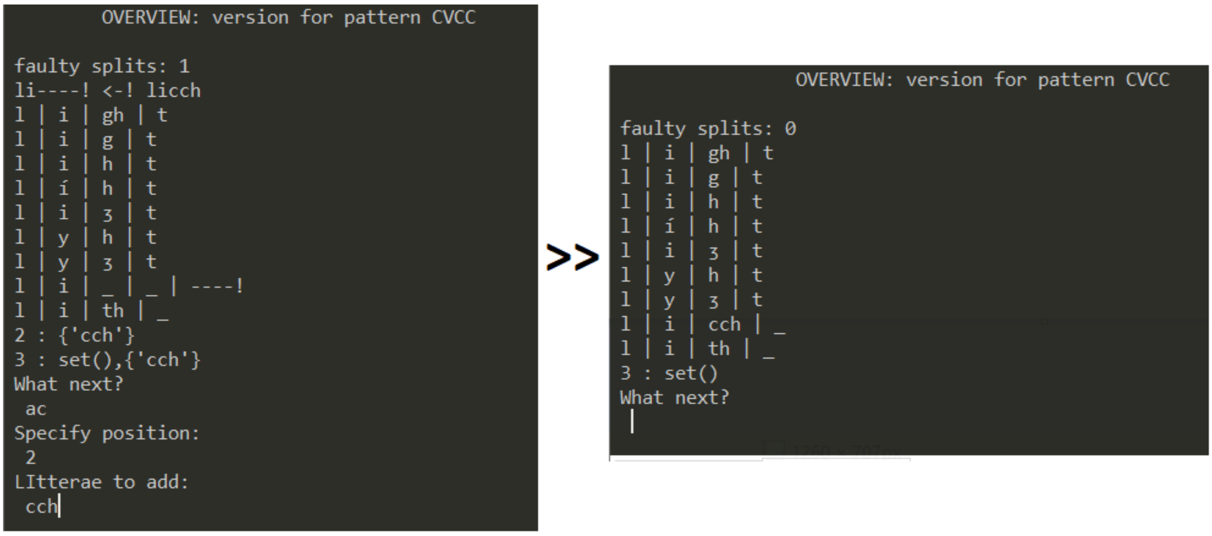
b) A set is missing from the list of valid sets which leads to the rejection of a littera at a certain position. As the rejected litterae are stored and displayed at the end of the analysis, the researcher can use a simple command to "allow" a littera at a certain position. This produces a new valid set which is immediately added to the list of valid sets. For example, when parsing the form *lowe* of LOVE/N (the model being *lu**u**e*), the combination {*u*, *w*} was not found on the list of valid sets and the processing failed and the set {*u*, *w*} had to be added to the list.

c) There is a mistake in the list of litterae. Either a littera is missing from the list or the value of "default digraph" is incorrect. In such a case, it is necessary to fix the error and rerun the script.

d) The correct pattern is missing from the list of automatically generated patterns. In theory, this can happen if the group of forms has two or more slots which can remain empty (such as initial *h* and medial *l*) and none of the forms has both slots filled. This was the most complex issue. The pattern had to be written manually and it was also necessary to provide a form which should serve as the initial model.

The analysis may require more than one of the adjustments described above. If the predictable adjustments were not enough to parse *all* the forms in a group, but the alignment of some of the forms as well as the pattern were correct, the parsed forms were saved to the database and the remaining forms had to be processed manually.

The picture below shows two subsequent outputs of analysis for LIGHT/N:



```
        OVERVIEW: version for pattern CVCC

faulty splits: 1
li----! <-! licch
l | i | gh | t
l | i | g | t
l | i | h | t
l | í | h | t
l | i | ȝ | t
l | y | h | t
l | y | ȝ | t
l | i | _ | _ | ----!
l | i | th | _
2 : {'cch'}
3 : set(),{'cch'}
What next?
 ac
Specify position:
 2
LItterae to add:
 cch
```

>>

```
                   OVERVIEW: version for pattern CVCC

faulty splits: 0
l | i | gh | t
l | i | g | t
l | i | h | t
l | í | h | t
l | i | ȝ | t
l | y | h | t
l | y | ȝ | t
l | i | cch | _
l | i | th | _
3 : set()
What next?
 |
```

Figure 12: Sample semi-automatic analysis output

The first screenshot shows the initial output. The script processed 8/9 forms correctly[12] but failed to process *licch* because {*cch*, *gh*} was not on the list of valid sets and was rejected as an option during the analysis. *Cch* is listed as a rejected littera at positions 2 and 3[13]. The solution was to allow *cch* at position 2. The second screenshot shows the result after the scrip was re-run.

3.4.5.2. Automatic

The purpose of automatic processing was to identify and parse forms which do not require manual intervention. Before running automatic processing, a few hundred items were processed semi-automatically to check for bugs in the script and fix them.

The script was run automatically and only successfully parsed forms were saved in the database. The items were marked as “A” (automatically processed). In order to be considered satisfactory, 85 % of the forms had to be parsed correctly (100 % for sets with a few members

[12] The parsing of *lith* his questionable but impossible to control through the script, as {th, h} as well as {th, t} are attested in the data and therefore considered valid.

[13] Positions are numbered from 0

only). If the automatic analysis failed, the item was marked as problematic and requiring manual processing or a modification of the script. Such items were subsequently processed semi-automatically.

A random sample of the automatically processed items was checked manually at a later stage.

3.4.5.3. *Fully manual*

If the analysis proved to be too complex for the script to perform even with manual intervention, the items were processed manually using a very simple web interface instead of the command line. Typically, only isolated forms or a few forms from a large group needed to be parsed in this way.

Items were loaded into the interface one by one. Forms already processed by the script were displayed in slots, while problematic forms were displayed only as strings in an input field. These strings were manually segmented by spaces and underscores were added to represent empty slots. The result was then saved to the database.

This procedure proved effective and sufficient for a large number of forms unanalysable by the script, still, some of the forms seemed to offer multiple possibilities of segmentation and others were too idiosyncratic to fit in the same pattern as the rest of the forms in the same group.

It was assumed that some of the most unusual forms could be errors but others could be in fact merely exceptional spellings restricted to a single text or a small number of possibly related texts. In order to decide between the two, it was necessary to examine the forms together and look for similarities between them. This final step of the processing is described in the next subchapter.

The table below shows the approximate number of items for each of the processing modes:

mode	**number of forms**	**percent**
automatic - success	3700	33
semi-automatic	6000	54
manual	1500	13

Table 4: Number of items by processing mode

3.4.6. Processing by text

The main task was to decide whether the idiosyncratic forms are likely to be accidental or whether they are instances of a more regular and systematic practice. Also, potential ambiguities in segmentation needed to be resolved. A list of unprocessed forms was generated for each text and inserted into a spreadsheet.

The individual lists were grouped according to connections between the texts, e.g. the same hand, manuscript etc. and each group had a separate sheet in the file, so that potential unusual forms from one text could be easily associated with similar forms in a related texts. For instance, all the seven versions of the *Poema Morale* were analysed together because shared unusual forms are good candidates for exemplar forms. Similarly, unusual forms found in one of the texts taken from MS Digby 86 could be related to similar forms from the other texts found in the same MS. A full list of the groups of texts can be found in Appendix 7.7.

Forms apparently sharing the same feature, e.g. an unusual digraph or extra letter were considered instances of a non-accidental spelling practice and marked with the same colour. For example, text #1600 (Oxford, Bodleian Library Laud Misc 108, part 1 containing the *South English Legendary*) contains multiple instances of *-thþ-* and all the forms sharing this feature were treated together.

Problematic sequences of litterae identified in the course of the analysis are going to be presented in chapter 4 (Results). Forms which could not be easily grouped in this way were checked against the already parsed forms and sometimes marked as likely errors and excluded from processing. For example, the form *mulchel* of *much* found in #9 appears only once in the text and no similar cases of l-insertion were found. Therefore, this form is considered a scribal error.

The majority of the analysed forms were eventually parsed manually and saved to the database. Likely errors were simply stored without parsing.

3.5. Character encoding and treatment of special features

As the description of LAEME at the beginning of this chapter already mentioned, manuscript transcriptions in LAEME were made using a special system of characters of the ASCII format, which combines uppercase and lowercase letters (see subchapter 3.1.3). While uppercase letters retain their normal value, lowercase letters are reserved for special functions, mainly the representation of non-latin characters. Although it is not very difficult to learn the system, it was preferred to "transliterate" it into the UNICODE character set. Moreover, LAEME is rich in information about genuine manuscript forms and the tagged texts preserve many features which would be omitted or normalized in editions. Naturally, it was considered important to transfer this kind of information into the spelling database. Two possible solutions to this problem were considered and each of them was applied to a particular group of features:

a) Storing at the level of forms and litterae

This strategy consists in storing the forms with special features as separate entries in the table *form index* and subsequently *litterae*. For instance, the abbreviated form for LORD LAUerD or Orm's *eo*-words with deleted *o* (transcribed as "E<O<") in LAEME would be entered into the table *form index* along with *lauerd* and *heo*. This arrangement is not without its advantages. It allows to keep the structure of the DB maximally simple and easily retrieve data for analysis of an individual spelling system. The undesirable effect of this approach is that the difference between e.g. "regular" *o* and dele*ted o* becomes equivalent to the difference between *o* and another littera. Consequently, the codes for superscript litterae, inserted litterae etc. would appear as separate litterae on the lists of litterae, maps etc.

This solution seemed the preferable one for expanded abbreviations. The abbreviated forms were entered into the index along with full forms. As the spelling generally uses lowercase letters, the expanded parts were written in uppercase letters. Thus, it is possible to distinguish between full and abbreviated forms in the data, while having something representative of "phonetic substance" (Laing & Lass, 2013: 2.3.6.).

b) Storing the information in a separate table

A new table would provide an additional level of information, which would allow to treat e.g. the superscript forms and normal forms as equivalent in some contexts without completely losing access to the information about special features. It was assumed that detailed information about phenomena like capitalisation, insertions, use of superscripts etc. is mainly relevant for analyses of individual text languages, while it can be rather superfluous if a more global perspective is taken. For instance, it should not be included (by default) in the construction of maps. This solution was applied to the remaining features.

The entries in the table *special features* (introduced in subchapter 3.4.2.1) can be linked to particular litterae and particular tags in the corpus. Nine types of features are distinguished:

a) “Capital” marks capital letters coded as “*+letter” in LAEME. It is by far the most common feature.

b) “r+superscript” / “u+superscript” are linked to instances of superscript letters being used for “r+letter” / “u+letter” and coded with ^ in LAEME. For example, LAEME “Gr^ACE” (*grace*) stands for “g^{a}ce” in the manuscript. The feature “r+superscript” is linked to the littera *a*.

c) The practice of writing the doubled consonants vertically is almost exclusive to the *Ormulum*. This usage is marked “stacked”.

d) The label “insertion” replaces LAEME “>letter>” convention for marking inserted letters.

e) “Reconstruction” is linked to illegible or poorly legible litterae written in square brackets in LAEME.

d) “Deletion”, marked with “<letter<” in LAEME is again found almost exclusively in the *Ormulum*.

e) The label “de nexus” allows to identify instances of the special figura.

f) “Flourished S” marks the “raised version of ‘s’” (Laing & Lass, 2013: 3.4.7).

3.6. Queries and calculations

The first part of the chapter described the database and the process of data parsing. This part moves on to the structure of the data accessed by the researcher who uses the tool. First of all, it introduces the basic units (*sets*, *slots* and *lists of litterae*) which are not used with the LAEME corpus and explains the relation of those units to the familiar concepts of *item* and *form*. It also describes how the units are manipulated to produce more complex output such as maps or network visualisations and what kinds of quantitative data and filtering options are available. The following subchapter introduces partially implemented features to be tested and assessed. The chapter concludes with a short discussion of the recommended approach to searching, pointing out the connections between different types of data and outlining possible paths of navigating it.

Generally, all the output data is structured around three basic units – *litterae*, *sets* and *slots* - and their relations, which are illustrated by the figure below. The three units interact with the already familiar concepts of *item* and *form*.

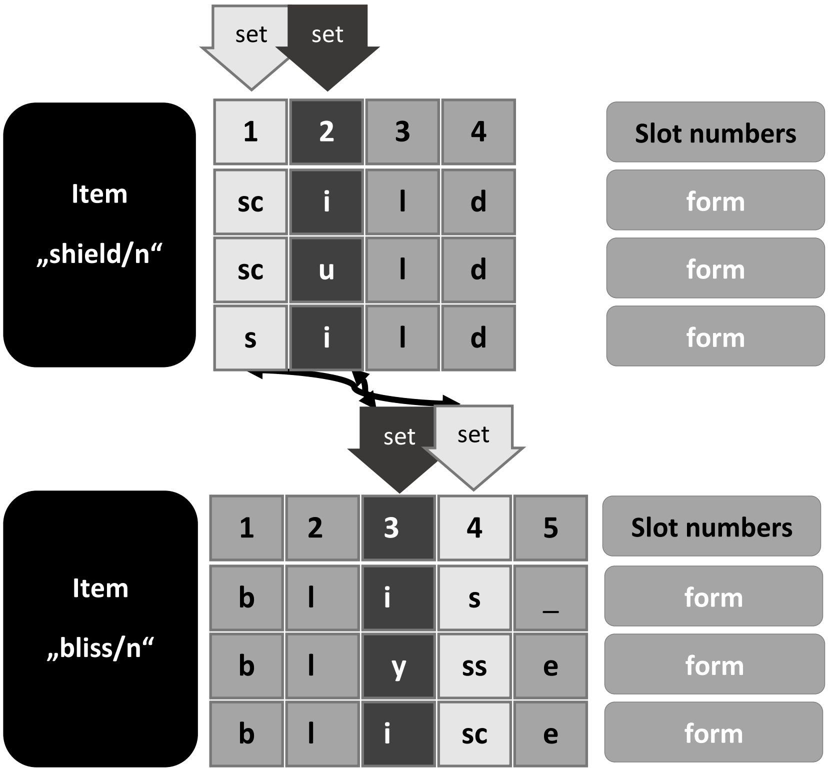


Figure 13: The relations between litterae, slots and sets

The diagram shows selected forms of the items SHIELD/N and BLISS/N. Each form is displayed on a separate row and each column represents a separate slot. A slot is defined as a position in an item, e.g. slot SHIELD/N (1) corresponds to the first position in SHIELD/N (light grey column). Each form under the item may have a different littera in the slot. In the case of SHIELD/N (1), there are two forms with *sc* and one form with *s*. Whatever is found in a single slot may be called *littera*, which implies that digraphs are structurally equivalent to *literae*.

Sets can be defined relative to a single slot, e.g. the set {*sc*, *s*} for SHILED/N, position 1 or a group of slots (e.g. SHIELD/N (1) + BLISS/N (4)). Conversely, it is possible to retrieve a list of slots based on littera or a set of litterae. For example, a search for alternation of *s* and *sc* in a single slot will return BLISS/N (4) as well as SHIELD/N (1).

3.6.1. A note on frequency data

Despite the differences in data structure, the familiar terms type frequency and token frequency remain useable. *Type* corresponds to a single slot. Token frequency corresponds to the actual number of occurrences. For instance if *ch* appears once in CHILD(1) and twice in MUCH(4) (in a specific text), its *type frequency* is 2 and its token frequency is 3. Type/token frequency can be calculated for a single littera or a set of litterae.

3.6.2. Basic units

The following paragraphs describe the three basic units (*litterae*, *slots*, *sets*) in more detail.

a) (Lists of) litterae

The term littera has been already defined in LAEME documentation and discussed in subchapter 2.1.2.2. Litterae as the output data type of the database also include polygraphs. They are usually retrieved as lists and each row on the list typically comprise three fields: the actual symbol, *type frequency* and *token frequency*. *Type frequency* corresponds to the number of distinct slots in which the littera appears and *token frequency* corresponds to the total number of its instances. The frequencies can be relative to the whole database or to a specific subset of data such as a single text, a certain region etc. For instance, the littera *þ* has a total frequency of 823 types / 51895 tokens in in the database, 79 types / 291 tokens in text #7 and 93 types / 407 tokens in text #10.

More detailed statistics are available within the context of a specific text language. These include positional and contextual constraints and *rare uses*.

b) Items and slots

Items correspond to the rows in the table *morpheme index*, i.e. combinations of lexel and simplified grammel. Items are usually retrieved as lists defined by the presence of one or more litterae in their forms, e.g. all the items in which *eo* and *ea* are used interchangeably. Any position in a specific item is called *slot*.

c) Sets

The term set was chosen to reflect the connection with *litteral substitution set* found in LAEME documentation but it has a broader meaning. While in LAEME, *literal substitution set* refers to the set of litterae associated with a certain *potestas* (sound), *sets* in the DB are defined as groups of litterae which at least once appear in the same slot.

Queries for sets significantly vary in scope. The researcher may be interested in all the litterae appearing in a specific slot, litterae used interchangeably in a specific text or a complete list of sets in which a certain littera or alternation of litterae occurs.

Similarly to litterae, type and token frequency can be calculated for each *set*.

When displayed in the interface, the three basic units – *litterae*, *slots/items* and *sets* – appear in different variations and contexts. For instance, a list of *litterae* may be a list of polygraphs containing a given littera or complete inventory of *litterae* in a specific text. A list of *items* may be defined by its association with a certain *littera* or an alternation of litterae etc. Moreover, *sets* are important for the generation of maps and network visualisation. The next section explains these more complex uses of the data.

3.6.3. Maps

Every mapping query in fact returns a *set* of litterae for every text in the database. The simplest kind of map shows the distribution of litterae found in a specific slot (position in an item), e.g. all the possible representations of the initial vowel in UN-. More complex queries can combine sets from multiple slots and the list of items may be defined merely by the presence of a certain littera or alternation of litterae. For instance, it is possible to search for all the items where *l* sometimes drops. The script identifies all such items in each text and combines the respective sets into one.

The frequency data needed to draw the pie chart on the map is currently based on token frequency.

3.6.4. Networks

There are three kinds of networks. The first kind shows the sets found in a single text. The second kind combines litterae from two texts. The third kind visualises global data, i.e. alternation of litterae across texts.

Two sets of data are needed to draw a network: nodes and links between them. Nodes represent the litterae from the given text(s) or all litterae in the DB. A link is established if two litterae appear together in one *slot*. The strength of the link depends on the type frequency of the alternation.

3.6.5. Inventory of litterae

Besides type-token frequency of the individual litterae, an inventory of litterae from a single text offers two more pieces of statistical data, namely (*littera*) *frequency comparison* and the incidence of *rare uses*. These numbers are intended to facilitate quick orientation in the data, highlighting litterae that potentially deserve more attention than others.

Littera frequency comparison compares the relative frequency of a littera in the examined text with its relative frequency in other texts. The two relative frequencies are calculated and the difference between them indicates whether the littera is relatively more/less common in the text. By default, the comparison is based on “global” frequency, i.e. the average relative frequency in the whole corpus, but in theory, it is possible to define a custom group of texts to perform the comparison. For example, if we look for litterae in text #4 (version T of the *Poema Morale*) which stand out compared to the other texts localised in Essex, the “reference frequency” could be calculated from these texts instead of the whole corpus. When comparing two texts, the relative frequencies in the individual texts serve as reference for one another.

For example, the relative frequency of *w* in text #10 (5 tokens only) is markedly lower when compared to the rest of the corpus, but if text #4 (4 tokens, comparable length) is chosen as the reference point, the two relative frequencies are very similar.

It is important to look at frequency comparison in connection with type/token frequency because highly unusual litterae, such as the idiosyncratic polygraphs *dþ* or *hv* will be marked as “exceptionally frequent” despite their low number of occurrences.

Comparing frequencies of litterae may be useful, still, the number is not sufficient to reveal cases of abnormal *usage* of a littera, i.e. cases where the littera is found in items where it does not commonly appear in other texts. For example, *ch* is quite a common littera, but its appearance in *spichen* (SWICA/N (4)) is quite rare, the “usual” littera at this position being *k*. The number of items in which the littera rarely appears is called the incidence of *rare uses*. Slots are marked “rare” if a littera appears in the slot in no more than 20 % of texts[14].

3.6.6. Chunk search

The theoretical part of this thesis devotes some space to the problem of representing continuous sound stream as a sequence of units, which inevitably must be simplificatory to a certain extent (see subchapter 2.1.2.2.3). The alignment of the corresponding segments in different forms inevitably runs into this problem. Deciding which segments in fact correspond to each other is not always straightforward. This is of course largely due to the fact that the compared forms represent dialectal and diachronic varieties, so the differences between the represented sound streams can span across multiple segments and even affect the “CV structure”. While some changes, such as voicing, can be relatively safely considered to affect a single segment, other changes occur at the boundary between phonemes. If the change

[14] This treshold can be easily adjusted.

supposedly involves the emergence of a new phoneme and a subsequent loss of another (both of which are not necessarily visible in the data), there are two basic parsing possibilities:

a) Place the original and the new phoneme in the same position, e.g. the forms *ehte* and *eite* of EIGHT would be aligned as follows:

   e | **h** | t | e

   e | **i** | t | e

   This analysis obscures the fact that [h] might not be the direct source of the new sound. Moreover, it is unclear, since when should the new sound be read vocalically, forming a diphthong with the preceding [e]. Lastly, this sort of parsing is virtually impossible to perform automatically using the script, because once {*i*, *h*} is considered a valid set, the script will not be able to analyse forms where *i* and *h* are found next to one another. Still, in terms of searching for changes and mapping them, this analysis conveniently reflects the change "of [x] into [j]".

b) Keep the two phonemes in separate slots and use empty slots to indicate their emergence/disappearance, i.e.

   e | _ | **h** | t | e

   e | **i** | _ | t | e

   This analysis appears more realistic and also makes it possible to align these forms with forms like *eihte*, which possibly capture the intermediate stage at which both of the sounds might have been heard by the scribe. However, searches and mapping become less neat, because we would have to search for the opposition of *i/_* or *h/_* and the litterae at the neighbouring positions would not be visible on the map.

The concept of *chunk search* responds to this problem, trying to preserve the virtues of both approaches. Option (b) (separate slots) was selected for parsing. The parsed data was then used to construct the table *chunks* (see subchapter 3.4.2.2) analogical to the table *litterae*. This table merges the neighbouring slots together, displaying what happens at the boundary between two slots. From this alternate perspective, the old and the new segment share the same position:

1	2-3	4	5
e	h	t	e

e	i	t	e

Table 5: Illustration of chunk alignment

3.6.7. Filters

The previous section presented the different kinds of units that can be retrieved from the database, namely (lists of) litterae, sets and items. It has been mentioned that it is possible to restrict the search to a certain subset of data, e.g. search for litterae in a specific text or a set in a specific region. This section describes various filters which can be applied to the queries.

The filters can be divided into two basic categories: a) filters based on manuscripts and b) filters based on adjacent litterae.

3.6.7.1. Filters based on manuscripts

It has been explained that the parsed data is linked to the original LAEME data and every item and form can be traced to specific manuscripts in which they appear. Every manuscript is in turn linked to its metadata such as date and localisation. Accordingly, filters can be applied at two levels: manuscript *id* or manuscript metadata. If a text *id* is specified, it becomes pointless to add filters based on metadata.

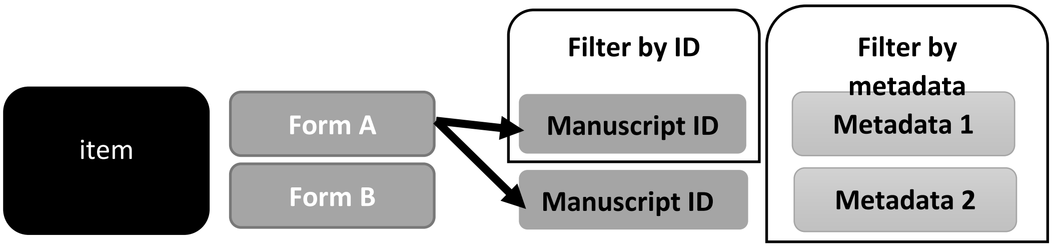


Figure 14: Filtering by LAEME file metadata

It is possible to combine multiple fields on the level of metadata. The currently available fields are *date* and *localisation*. This sort of filtering can be used to generate maps which display only texts from a specific period, alternatives of a littera found in a specific region etc.

LAEME codes for *date* were replaced with numbers, but this notation remains invisible in the interface, which continues to use LAEME codes. Thus, the earliest texts dated to the last quarter of the 12th century ("12b2" in LAEME) are marked "1", first quarter of the 13th century corresponds to "2" etc. If a text is not dated to a single quarter-century, it is tagged with multiple numbers, e.g. "2, 3" for LAEME "13a" (first half of the 13th century). Such a text will then fall into one group with texts dated to "13a1" as well as "13a2".

Localisation was reduced to a code for a single county with no further regional specification (e.g. “Gloucs” is used instead of “N Gloucs” etc.). Similarly, the field *script* which sometimes includes comments on the nature of the hand in question was simplified to the name of the script (e.g. “textura semiquadrata”). The purpose of the modification of the original tags was to group the manuscripts into larger groups rather than a lot of groups with few members.

3.6.7.2. Filters based on adjacent litterae

Filters based on adjecent litteare can be freely combined with filters from the previous group. They operate at the level of litterae. Analogically to the filters based on manuscript, it is possible to filter by adjacent litterae or their metadata and there is also a positional filter. The different fields available for filtering are shown by the schema below:

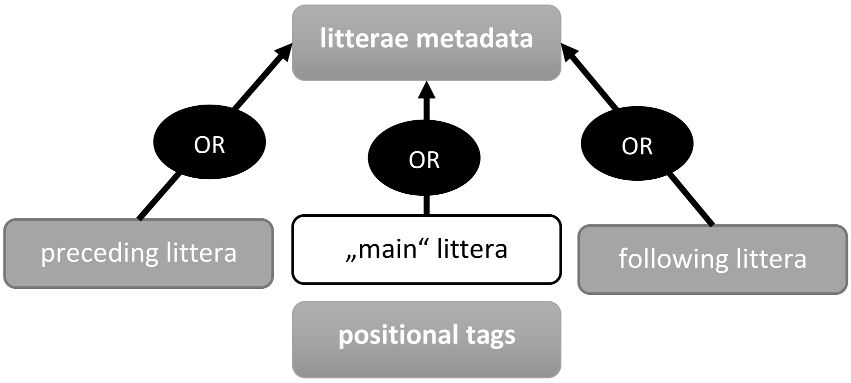


Figure 15: Filtering by adjacent litterae

In simple searches, only the white field “main littera” would be available, which could be used to run queries like “get all the items where *ch* appears, get all the sets where *c* alternates with *k*” etc. The field “positional tags” adds a positional constraint, e.g. “get all items with *ch* in the initial position”. The fields for adjacent litterae can be used to specify which littera should follow/precede the main littera, e.g. “all items with *ch* followed by *o*”. Litterae metadata offer a set of tags which can be used instead of a specific littera, e.g. “get all items with *ch* followed by a high vowel”.

The filters also allow to search for sets of litterae while leaving the field “main littera” blank. This functionality can be used to list adjacent litterae, e.g. “all sets of litterae following *ch*”. The system of concrete tags available for filtering is described below.

3.6.7.3. Positional tags

Positional tags are found in the table *litterae* and they were assigned programmatically. They include “morpheme-initial position” and “morpheme-final position”.

3.6.7.4. Litterae metadata

Litterae metadata is stored in a separate table. It is based predominantly on assumed sound values found in literature on historical phonology, i.e. a standard classification of phonemes. It is vital to stress that this table does not present *interpretation* of sound value, which can vary across text for the individual litterae. It is a mere filtering tool. Wherever sources suggest multiple sound values for a single littera, *all* the corresponding tags are linked to it. For example, *u* has the tag for “consonant” as well as “vowel”, *g* has tags for “plosive” as well as “approximant” etc.

3.7. Experimental features

This section describes features which were proposed as potentially useful components of the tool and partially implemented but their completion would exceed the scope of the present project. The first feature is the integration of forms from external sources with the rest of the spelling database and the second feature are links to CoNE data.

3.7.1. External forms

Subchapter 3.4.2.1 mentions two special tables available to store spelling variants from different sources. This could include source forms from OE texts, PDE forms as well as LME forms provided that they can be linked to one of the items defined in the table *morpheme index*. Such forms obviously need to be parsed to be accessible. So far, the tool can display such forms as aligned with LAEME forms. A code marking the source of the form (e.g. “OE”) is displayed instead of form frequency. The data of this sort could further be used as a query filter, which would allow the researcher to submit queries like “list all sets of litterae found in place of OE *c*”. So far, only 21 external forms have been included in the DB (see Appendix 7.10.7).

If the data from LALME was added to the database, it would in theory be possible to combine LAEME data and LALME data in a single map.

3.7.2. CoNE

It has been stated in the theoretical part that the relevance of CoNE for the present project consists not only in its direct connection with LAEME but also in its focus on the segmental level (see subchapter 2.3.2.2). A number of sound changes and orthographical changes described in CoNE naturally imply a set of litterae. For instance, "Orthographical Remapping of c" (ORC) consisting in the novel use of *c* as a representation of [s] is likely to be found in items in which *c* alternates with *s*. Therefore, the label ORC used in CoNE can be linked to the set {*c*, *s*}, which is a concept "understood" by the tool described here. Connections of this sort can be specified in the spelling database, which results in:

a) The possibility to use the CoNE code as input for a query returning a list of "candidate" lexels potentially involved in the given change.

b) The possibility to display links to CoNE along with sets or items potentially involved in the changes.

While context-free changes can be linked to sets of litterae, contextual changes need to be associated with specific filters, but this is technically possible. For instance. If searching for *f/v* alternation, the code for Initial Voicing of Fricatives (IVC) can be displayed only for items which have the concerned slot marked as "initial position". Filter codes defining the context of specific changes can be stored in a separate column of the table *cone sets* (see subchapter 3.4.2.1).

3.8. Zooming

As the description of the data shows, the structure differs from the more familiar model of item-forms. This naturally implies a slightly different approach to querying and work on the data in general. This subchapter discusses the suggested broad approach, which could be labelled "zooming" and subsequently moves on to outlining specific paths of navigating the data.

Generally speaking, the key features of the database are high levels of detail and interconnectedness. Taken together, these features significantly improve the access to quantitative data, which can nevertheless be misleading if not properly refined. In other words, the researcher has to look into many details at the lowest level of data i.e. actual texts (which should nevertheless be easily accessible). This method of rather fast and repetitive "travelling" from the higher level of statistics to the level of individual items or even stretches of text in a manuscript is labelled "zooming" here because it is similar in principle to a situation when we look at a picture, trying to notice a pattern and when we do, we have to zoom in to check whether the apparent pattern makes sense.

The possibility to "travel" in this way is largely due to the introduction of *sets*, which adds a new dimension to the data, creating links or "paths" between items. However, not all of these links are meaningful and useful. It may be also said that the database materializes data about some of the sound changes described in literature, namely connections between items involved in a change. This can be demonstrated on an example of a well-known sound change such as the voicing of fricatives. A good corpus of Middle English such as LAEME provides very useful material and querying possibilities. It is easy to get a complete list of *forms* of a particular item in which we expect this change to occur, because the crucial link in the database leads from the item (lexel) to its forms. What the spelling database adds, is a link between items based on a shared set of litterae. This situation can be illustrated graphically as follows:

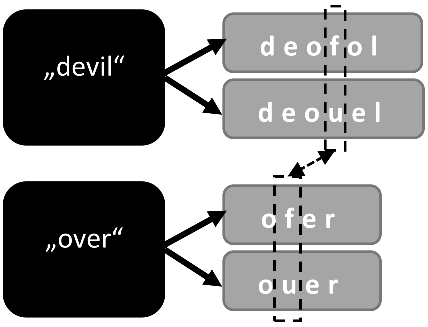


Figure 16: An illustration of the linking function of sets

One of the possible uses of the newly added link (dashed arrow) would be to get the list of items (potentially) involved in the change. If the link between items were not available, we could either take a list from literature or compile our own, which might take a long time. One way to achieve this would be to list all items with medial *f* (*ff*) and check for those in which *f* actually alternates with *u*, *v* or other likely representation of the voiced variants. This task would be even more demanding if we dealt with a change which involves multiple spelling variants for both of the phonemes involved in the change. The connection between items in the spelling database makes it possible to search directly for a list of items in which the alternation of certain litterae ever occurs.

3.8.1. Suggested approaches to searching

Broadly speaking, "zooming" consists in navigating the network of data. The network can be entered from different sides and some of the paths are relatively fixed. This subsection briefly outlines the typical points of departure and directions in which analyses may proceed. The following chapter (Results, see subchapter 4.3) demonstrates the use of the queries and functions mentioned here on specific examples and actual pieces of data retrieved from the database.

a) Start from a littera

One of the possible kinds of studies of ME spelling consists in analysing the use of a specific littera. Such studies may ask questions like "How *x* was used in EME? What are the most common alternatives of *x* in EME? Which sounds were represented by *x*?". If we take a specific littera as our point of departure, we may immediately retrieve frequency data for the littera and a list of alternatives of the littera found across texts. The list of alternatives can in turn be used to generate a list of items, in which *x* alternates with a given alternative. The forms associated with the items can in turn be traced to specific texts.

b) Start from a combination of litterae (possibly indicating sound change)

If the researcher is interested in a specific sound change, s/he may choose the alternation of specific litterae as his/her starting point. This alternation of litterae is sufficient to run a query for a complete list of items in which the alternation ever occurs. The researcher may then choose to plot the variants found in specific items on the map, display the variants in context (Key Word in Context), or examine the forms of the items in a specific text.

c) Start from a text

Analyses focusing on a specific text may begin with consulting the list of available texts. The researcher may immediately access the original description in LAEME and quick links to related texts, if there are any. Then s/he may display the text profile of the chosen text, quickly identify conspicuously rare/frequent litterae or unusual digraphs, highlight their occurrences in the text and examine them. S/he may also analyse the list of alternating litterae, display the alternations graphically as a network and access lists of items in which the alternation occurs. Any of the items may of course also be highlighted in the text.

Text profiles can be also displayed side by side, in which case the tool automatically calculates and visualises differences in the relative frequencies of individual litterae. This is useful for text comparison.

d) Use item lists

The concept of item list is a familiar one. The usual method is to define a list of units based on shared historical sound value, such as [f] in the initial position and analyse the occurrences of the items in a selected text or multiple texts. The tool does not (yet) include data about OE source forms or presumed sound values, which would enable construction of item lists of this sort. It does, however, enable compilation of item lists based on shared littera or a combination of litterae, which can be further filtered by context of the littera, its position or occurrence in a manuscript or manuscript metadata. For instance, it is possible to define a list of "all items containing initial *sch* in text #242" or "all items with *h* followed by *t*". Item lists can be stored and used to search for instances of the items in the manuscripts or to construct maps.

As this description implies, there are certain fixed paths available to navigate the data. For instance, a slot in an item displayed in whatever context (text profile, simple DB search etc.) can always be immediately plotted on the map, a form can be always displayed as a Key Word in Context etc. The following schema shows the relations between the pieces of data.

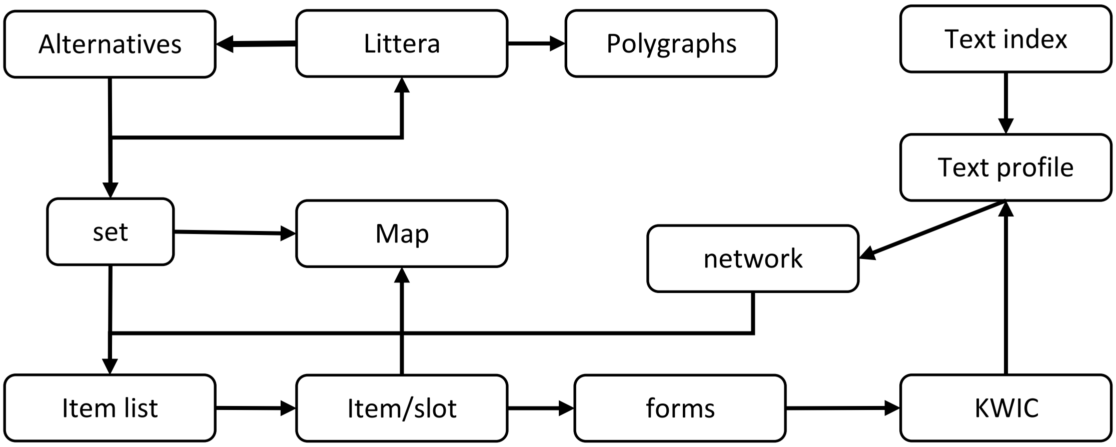


Figure 17: Relations between pieces of data in the database

3.9. Chapter summary

The methodological chapter provided a detailed description of the spelling database and the process of its construction. The final part of the chapter outlined querying possibilities and links between different kinds of searches to be demonstrated on practical examples in the next chapter.

4. Results

The first part of this chapter discusses spellings which could not be parsed in a straightforward manner. The rest of the chapter is conceived predominantly as a demonstration of application of the presented tool and its assessment. It opens with a description of the interface designed to access data from the database, presenting its individual screens and features one by one. This introductory part is followed by a series of practical examples or "micro analyses", demonstrating how specific tasks can be approached using the tool, what kinds of data can be retrieved and how they relate to some of the methodological concepts discussed in the theoretical chapter. Besides explaining the possibilities of the tool, the section also comments on its limitations. The final part of the chapter has the form of a general commentary on the process of construction of the tool, including some theoretical and methodological observations inspired by the project and possible directions, which further development of the tool might take.

4.1. Problematic segmentation

The variants which seemed difficult to parse were left out from the processing at first in order to identify recurrent patterns (problematic sequences of litterae) and devise a consistent way of parsing for each sequence. Such sequences fall into two basic categories – highly repetitive patterns appearing across a large number of texts (e.g. swapped letters) and idiosyncratic forms restricted to a small number of texts. The first part of this subchapter focuses on the former group.

4.1.1. Repetitive patterns

4.1.1.1. Swapped letters

This section discusses items associated with change of position or apparent change of position of a littera in the word. Three specific patterns can be identified.

The most common one by far was the occurrence of forms with *-re-* alongside *-er-* and *-ere-* under the same item, e.g. *neuer/nevre/neuere* (NEVER/AV). Groups of forms of this kind were considered cases of syncope rather than metathesis and parsed as follows:

n | e | u | e | r | e

n | e | v | _ | r | e

n | e | u | e | r | _

The same strategy was applied also to some items for which the “full form” with both *e*s was not attested but the cluster *er/re* occurred in the final position, e.g. *number/numbre* (NUMBER/N).

If similar cluster appeared in the medial position, the segments were aligned to create apparent correspondence between e.g. *e/r*, which enabled to collapse the whole group in a single slot in chunk searches, e.g. GOLD/N:

g | o | l | d > g | o l | d

g | l | o | d > g | l o | d

The obvious drawback of this solution is that it creates sets like {*r*, *e*} in the database, nevertheless, such sets are rare and unlikely to be confused with “genuine” sets like {*r*, *rr*} or {*e*, *ea*}.

4.1.1.2. qu or q+u

The cluster *qu* is often described as a digraph in literature. The occurrences in LAEME can be sorted into two major categories based on typical corresponding litterae, i.e. sets in which *qu* appears. The first set comprises reflexes of OE *hƿ*, {*ƿ*, *hw, wh, w, etc.*} and also *q* alone. The other group of sequences alternating with *qu* are multiple reflexes of OE c*ƿ* {*kƿ*, *ku*, *kw*, *cu,* etc.}. These two basic uses were treated differently when parsing. The instances corresponding to *hw* etc. were parsed as digraphs, e.g. WHOM/P:

hƿ | a | m

qu | ai | m

w | a | m

q | a | m

The remaining instances were treated as two separate litterae, because the assumed represented sounds seem reasonably distinct, while the *hw* type is sometimes interpreted as voiceless [ʍ]. For example, *qu-* in CWEALM/N was aligned as follows:

q | u | a | l | m

c | ƿ | a | l | m

c | u | a | l | m

4.1.1.3. sc or s+c

A large group of lexels have forms with alternating *sc* and typical digraphs (trigraphs) for [ʃ] like *sh, ssh, sch* etc. and *sk*. Some of the concerned words underwent palatalisation and others did not. *Sc* was aligned with the corresponding digraphs and *sk* whether or not it appeared in typical palatalisation contexts, so that the instances in different contexts could be easily compared. The noun SKILL/N, for instance, has forms with *sc*-, *sch*- and also *sk*-:

sc | i | l

sk | i | ll

sch | i | l

There were only several items in which, *sc* was split into two slots (e.g. BASKET/N and SCRIPTURE/N). These items are etymologically distinct from the rest and most of them are of Romance origin.

4.1.1.4. cu, gu or c+u, g+u

Cu and *gu* are sometimes considered to be digraphs indicating "hard" pronunciation in literature, but this approach was not adopted here because it is disadvantageous for automatic processing.[15] Wherever *c* or *g* are followed by *u* + vowel, the *u* is placed in one slot with the vowel. The *u* can still be understood as a diacritic for "hardness". See the example of LONG/AJ for illustration:

l | o | n | g | ue

l | o | n | k | e

l | o | n | g | e

4.1.1.5. The littera x

The obvious problem with *x* is that can correspond to sounds perceived (and sometimes written) as two segments, e.g. *cs*. For the sake of simplicity, the two slots corresponding to *x* were merged into one, as this does not have any obvious disadvantages. For example, NEXT:

n | e | x | t | e

n | i | s | t | _

n | e | cs | t | _

[15] The sequences *gu*, *cu* are often instances of C+V , which would be difficult to distinguish from the digraphs using the script.

4.1.1.6. Reflexes of OE g

The diverse spellings of words with attested *g* (and sometimes also *h*) in OE are without doubt the most complex ones and the most difficult ones to process. The reflexes of *g* include all the g-shapes (*g*, *ȝ*, *ᵹ*), sometimes in combination with *h*, *ch* and also *i*, *y, u* and related litterae, for instance, EYE/N (selected forms only): *eiᵹene*, *eaᵹen, eᵹan*, *eghe*, *egen*, *ehe*, *eyen*, *eihen*, *éᵹen*. Phonological developments associated with these segments have been discussed in subchapter 2.1.4.6.

The general pattern usually is that most of the forms have either only *g/h* , e.g. *é**ᵹ**en* or *i/y* (*u*) , e.g. ***ey**en* and some forms have both, e.g. ***eih**en*. The standard solution of this situation was to use the special method of parsing (cf. *chunk search*, 3.6.6) and create special slots for *i* and *g*, which could be merged into one in boundary search mode. For example, DAY/N:

d | e | i | ᵹ

d | a | _ | ᵹh | e

d | æ | i | _ | e

This alignment may appear somewhat overcomplicated, still, considering the variability of the forms it was decided to keep the segmentation as flexible as possible.

As for examples of interpretation of sound from literature, Laing & Lass (2009) interpret the sequence in *geihet* (GÉGAN) as *e* plus a combination of *i*+*h* Laing & Lass (2009: 26). In the same article, they suggest that the *h* in *ehnen* (EYE/N) might in fact stand for [ɦ] (Laing & Lass, 2009: 28). If this is so, the analysis of the similar form *eihnen* as *ei*+*h*+*nen* should at least be considered. At the same time, there is no obvious reason to suppose that the *eih*- in *eihnen* must be different from *geihet*. The current segmentation allows to look at the chunk *ei*- as well as -*ih*- and the segments aligned with them.

In a number of cases the sequence -*iᵹe*- (especially word-finally) corresponds to -*ie*- in other forms. The standard way of parsing such forms is to align the segments as follows:

i | ᵹ | e

i | _ | e

The probable sound represented by such sequences is [ie] or almost identical [ije].

4.1.2. Rare features of the texts

While the parsing procedure turned out to be relatively straightforward for the majority of items and forms, there were also spelling variants which could not be easily analysed without reference to related forms from the same text language and similar ones in other texts. These problematic spellings usually represented low-frequency variants appearing alongside more common ones and they occurred in a small number of texts, sometimes even only one text.

The following part of this chapter discusses spelling variants which were subject to manual processing at the final stage of the analysis. As specified in the methodological chapter, the concerned items and forms were observed in the context of specific text languages in which they appeared and partly also other manuscripts possibly related to them. The forms were sorted into multiple small categories, some of which are possibly related to one another. The individual categories are going to be characterised below.

4.1.2.1. hV spellings, double ii and iy

Several texts, notably #246, sometimes insert an extra vowel in between final *h* and *t*. For instance, NOT, commonly spelled *naht*, *noht* is spelled *noh**u**t*. The scribe of #246 furthermore sometimes employs the sequence -*hit (-hid)* at the end of words, which end in simple *t* or *d* in most cases, e.g. FEET, which has the high frequency variant *fet* is spelled *fehit* or *fehid*. The forms with *h* sometimes alternate with *h*-less forms, e.g. *fehid/feit* (FEET, #246), *þohut/þout* (THOUGHT, #218). A possible explanation could be that the whole sequence -*ehi*- in fact corresponds to the medial vowel (diphthong) and the *h* is a marker of breaking. A possible reading would be e.g. [fe:-it] rather than [feit]. The whole sequence corresponding to the diphthong was parsed as a single segment. For example, the slots in *wight* were aligned in the following manner:

v | ichi | _ | t

v | ii | _ | t

w | i | h | t

This solution was eventually chosen over the possibility to align the medial *h/ch* with the *h* in the "common" forms, e.g.

v | i | ch | i | t

w | i | h | _ | t

The decision is justified by the following arguments: a) the sequence -*ehi*- is not restricted to items with etymological *h*, b) there is no account of a change consisting in the development of a sound within the historical -*ht* cluster and c) the rejected solution would introduce a slot in between *-h* and *-t*, which would be empty in the vast majority of cases and the slot structure of the concerned items would diverge from most of the items with historical *-ht*.

While the spellings with inserted *h* are very rare, the sequence *ii/yy/iy* or *ij* appears over 100 times in the corpus. Besides historical *h* or *t/d*, it sometimes precedes *f* (*wijf* - WIFE), *s* (*wiis* - WISE), *k* (*liik* - LÍCGAN) or *l* (*viil* - VILE), but it is unclear whether all the forms are related. There seems to be at least a weak connection between the two spellings (*h* and *h*-less) as the texts with multiple occurrences of *h*-spellings (namely #246 and #2002) also have some cases of *ii*.

4.1.2.2. y/i + h in the initial position

A few texts, notably #1400 use *y* (or *i*) + *h* at the beginning of words with presumed initial palatalization, such as GEORNAN/VI, YEAR/N etc. In text #295, these forms sometimes alternate with simple *y*, e.g. *iher* (YEAR/N) appears alongside *yeir*. The variants with *h* are relatively rare and alternate mostly with initial *g*. Interestingly, in two lexels in #1400, namely *ihwhat* (WHAT) and *ihu* (HOW), the initial *ih-* does not correspond to historical *g-*.

The correspondence between *h* and *g* is relatively common in LAEME, the correspondence between i | h and _ | g seems doubtful at best and the position of *y* is almost impossible to decide in this configuration. This is why initial *ih/yh* was parsed as a single segment:

ih | e | l | d

g | e | l | d

4.1.2.3. The tht cluster and related variants

The lexels RIGHT, MAY, LIGHT, NAUGHT and several others containing the historical cluster -*ht* have a number of variants, which were eventually parsed manually. They sometimes contain the sequence -*(ȝ)tht-* corresponding to the much more frequent -*ht-*. The forms with -*(ȝ)tht-* can alternate with plain -*th-* or -*ȝth-* in the same text or the same manuscript. The first pattern is found in texts #136, #137 and #285. For instance #136 has *noth* for NAUGHT/AJ and *fiytht* for FIGHT/VI. The second pattern appears in MS Laud Misc 108 (#282, #285, #1600). Text #1600 has mostly -*ȝtht-*, while #282 has -*ȝth-*. For example, #1600 has *riȝtht* for RIGHT/V-IMP and #282 has *niȝth* for NIGHT/N. There are also forms where either *h* or *t* or both are missing and there is a certain overlap between the texts containing such cases of missing litterae and those having the cluster -*tht-*, specifically texts #1800, #246, #285 and #137 sometimes drop the *h/ch* and #129 drops *t*. The forms with missing final *t* are markedly less frequent.

The analysis of this group of spelling variants had to answer several questions. The first question was whether to read *-th-* in *-tht-* as a digraph or whether the whole sequence is in fact a confused variant of *-ht-*.

The examination of alternatives in different texts revealed that if we interpret *th* within the cluster as a unit, it can sometimes alternate with *ch,* e.g. *drichtin/drithtin* (DRYHTEN/N, #1400). This is especially the case of text #1400. Moreover, the description of text #285 available in LAEME explicitly states that the shapes of *t* and *c* tend to be very similar, which might account for the apparent *th* in #285.

The sparse occurrences of *tht* in texts #246 (1 instance) and #249 (2 instances) both alternate with *st*. Interestingly, neither text consistently uses *th* (unlike the previously mentioned texts), which speaks in favour of exemplar provenance. Given these observations, the sequence *tht* was parsed as *th-* | *t*, corresponding to the canonical *h* | *t*.

The problem with *th* consisted in deciding whether to align it with *h* (which would put it in the same slot with the *th* in *tht*) or with *t*. The profiles of texts #137 and #285, both of which have the variants in *-th-* showed that *th* in fact alternates with *t* even in contexts without historical *h*, for instance NEAT/AJ in #285 is sometimes spelled *neth*. Moreover, the concerned texts often also contain instances of missing *h*, like *naut* (NAUGHT/AJ) in #137.

The variants *-ȝth-* (#282) and *-ȝtht-* (#1600) were parsed as *ȝ* | *th* and *ȝ* | *tht* respectively, because as such they best fit the pattern shared with the other variants and no arguments against this choice were found during the analysis. The common variant alternating with the highly idiosyncratic (and very rare) *-ȝtht-* in #1600 is *-ȝht-*.

4.1.2.4. Ow versus oh and hg

The sequence *-ouh-* appears (not exclusively) in lexels with historical [*ɣ] which changed into [w], such as SLAY/VSPT (*slouh*), BOUGH/N (*þouh*). The sequence *ouh* (or its variants *owh, oug* etc.) alternate with simpler *ow/ou* or *oh/og*. While the alignment of *o* | *o* and *h* | *h/g* seems straightforward, a decision needed to be taken regarding the medial *u*. The most natural reading for *ou* followed by *h* would probably be a diphthong (or a long vowel), e.g. *pouh* (WÁG), but the *u/w* in forms like *wawe* (WÁG/N) can also have consonantal reading, which speaks against simply aligning the diphtongal *ou* with the V+C sequence *aw*. On the other hand, aligning the consonantal *w* with the consonantal *h* obscures the potential connection between the second element of the diphthong and the new vowel, which might have evolved from it.

ƿ | ou | h

w | a | w | e

Therefore, wherever the forms suggested consonantal reading for *w/u*, an extra slot was added for it. For instance, WÁG/N was parsed in the following way:

ƿ | o | u | h

w | a | w | _ | e

A small group of seven texts (plus 6 texts with single occurrences) sometimes spell words with historical *g* with *hg* (*chȝ*), e.g. *inohg* (ENOUGH) appears alongside *inoh*, *inoge* and *inoph*. It is questionable whether *hg* should be interpreted as a reverse variant of *gh* or whether the *h* rather modifies the pronunciation of the vowel, the whole form representing something like [ino:x] or [inoux] rather than [inox]. As this is difficult to ascertain within limited time and the number of texts having this feature is relatively restricted the solution was chosen based on practical grounds, i.e. the digraph parsing being simpler and allowing to keep the slot structure of the concerned lexels less complicated. Thus, *hg* was aligned with *g*, *gh* etc., for instance (WÓH/N):

w | o | _ | hg | e

ƿ | o | u | h | e

ƿ | o | _ | ȝh | e

The possibly related variant -*chȝ*- is unique to text #273.

4.1.2.5. Vocalic w and vu-w confusion

The littera *w* occurs a few times in vocalic positions, e.g. *swn* for SUN/N in text #297. The concerned instances of *w* were obviously aligned with the other vowels at the same position. Another unusual feature associated with *w* is its use in the initial position without a following vowel (a consonant appears instead), for example WOUND/VPSP is spelled *wnde*. A likely explanation of this usage is that the doubled *v* should in fact be read as [vu], similarly to double *uu* in forms like *luue* (LOVE). This spelling as a variant corresponding to PDE *wV* is restricted to texts #170 and #246, where *vu* in fact alternates with *w*, e.g. *vundeN* and *wndes* (WOUND/NPL).

The initial *w* was aligned with the other litterae in the initial position, e.g.

ƿ | u | n | d | e

w | _ | n | d | e

This alignment is of course inaccurate but it was considered preferable to collapsing the first two slots into one in all the concerned words.

4.1.2.6. Initial suw-

Text #282 has rather atypical forms of SUCH, SWEET/AV and SWELL/VI, all beginning in *suw-* (*suwech*, *suwilk*, *suwete*, *suwell*). A possible explanation of such spellings is that they reflect a change [swete] > [suete] and the medial *w* corresponds to an almost silent element between [u] and [e]. However, *uw* and *w* could also be mere orthographic variants. An argument in favour of the latter explanation is the occurrence of the same sequence in *mouwen* (MAY/VPS2). The *uw* was eventually aligned with the more usual *w* (*p*).

4.1.2.7. Initial sw- for sh-

Texts #278 and #276 contain forms spelled with initial *sw-* (*p*) in place of the expected *sh-* (*swahte* (SEHTAN), *swome-* (SHAMEFAST), *spap* (SHOW) and *swoððen*, *swuððen* (SINCE) and text #67 spells SHOE/N as *swo*. These forms were discussed by Laing & Lass (2009), who explain them by a sequence of litteral substitutions of *w* for *h*, pointing out that the instances are found in too many manuscripts to be dismissed as scribal errors. Their claim is further supported by the presence of "reverse" forms (*sh-* for *sw-*) in the concerned manuscripts (Laing & Lass, 2009: 22). The initial *sw-* is aligned with *sh-*.

4.1.2.8. Vhl, Vhn, Vhr

Three texts (#182, #285 and #278) sometimes use *h* after vowels where no consonantal element is expected, e.g. *wahr* (WHERE, #278). As the most likely interpretation of the V+*h* combination is a long vowel, the *h* was placed in the same slot with the preceding vowel, e.g.

wh | a | r

Qu | a | r

w | ah | r

4.1.2.9. Insertion of p

Some forms of NAME appear with *p* inserted after *m* (*nempn*) and the same occurs once with HÉRSUMNESS (*hersump+nes*), which is a patterns corresponding to "Post-Nasal Stop Epenthesis" (CoNE, PNSE). According to the rules proposed for parsing, the pattern for these forms should have an extra slot for the inserted *p*. However, considering the extremely low frequency of the forms, it was considered preferable to parse *mp* as a single segment, e.g.:

h | e | r | s | u | m | nesse

h | e | r | s | u | mp | nes

4.1.2.10. Idiosyncratic polygraphs

This category comprises forms, where a single littera apparently corresponds to two litterae, which cannot be easily recognized as a common digraph described in literature, e.g. the form *kingke* (KING, #246), *ðhanc* (THANK, #155). Many of the concerned litterae can be interpreted as variants of the "canonical" digraphs but at least some of them may result from the effort of the scribe to capture sounds which could not be easily represented by single litterae and which he possibly perceived as bisegmental.

When processing, the clusters of litterae were aligned with the corresponding single litterae same as common digraphs, because they are generally rare and it is usually impossible to decide whether the single littera should share the same slot with a specific member of the unit, e.g. whether the *g* in SING/VPS should correspond to *g* or *k* in *singk+et*:

s | i | n | gk

s | i | n | k

s | i | n | g

The polygraphs can be sorted into several categories:

4.1.2.10.1. Clusters corresponding to PDE dental fricatives

The clusters in this category generally correspond to *þ*, *ð* or *th*. They include *tþ*, *þh, ðh, dh, dð, ðþ, td, tð, tȝ, thþ* and *thz*. The majority of the instances are occasional occurrences scattered across several texts. Litterae with the highest frequency in a single text are *ðh* in text #155 and *thþ* in text #1600. Both more often than not alternate with other digraphs or single litterae in the same slots. The digraph *tȝ* (6 occurrences) appears almost exclusively in the 3rd person singular and plural verbal endings in #161 and it alternates with *t*, *þ* and the reverse variant *ȝt*.

Some of the digraphs from this group sometimes appear in the same slots, but this concerns a relatively small number of lexels, mainly SOOTH, SINCE and DEATH.

4.1.2.10.2. gk, kh

Gk is found in four texts only and three of the four have only a single instance of it. It always appears after *n* and before final *e* in endings and alternates with simple litterae, such a *c*, *g*. The cluster *kh* is also very rare (5 occurrences in 3 texts).

4.1.2.10.3. (s)ꝫc

The cluster *sꝫc* is found only at the beginning of *shall* in text #146 (5 instances). It was parsed as a trigraph in the initial slot of *shall* corresponding to *sh*, *ss* etc. Unfortunately, #146 is too short to offer more useful data, but the use of insular *ꝫ* in positions typical for *h* (*sh*, *sch*) is reminiscent of the use of *ȝ* in such positions in other texts. A similar cluster *ꝫc* appears as the second element of *ac* in text in #2000, which in fact has some instances of *h*/*ꝫ* alternation.

4.1.2.10.4. td, dt

These digraphs appear in the final position, alternating with *t*. They are very rare and mostly appear as single occurrences in different lexels. Texts #263 and #160 have 3 instances each, text #160 has two and the remaining 13 texts have only one. The only connection between any of these texts found in LAEME data is that the language of text #227, which has one instance of *nastd* (NOT) is similar to the languages of texts #248 and #249. According to CoNE, these spellings may indicate a change called Deaspiration (CoNE, DA).

The instances of *-td* #263 appear alongside *-þt* (e.g. *brytd/bryþt*, BRIGHT/AJ)

4.1.3. Summary

Although the original intention was to split the forms into single litterae, “canonical” polygraphs or more or less obvious variations thereupon, the solution adopted for the idiosyncratic forms like *vichit* (WIGHT/N) was to create a larger chunk in a single slot. The recurrent general problem is that creating more extra slots can make the alignment more precise but empty slots overly increase the complexity of the slot structure, making it less predictable and therefore less user friendly.

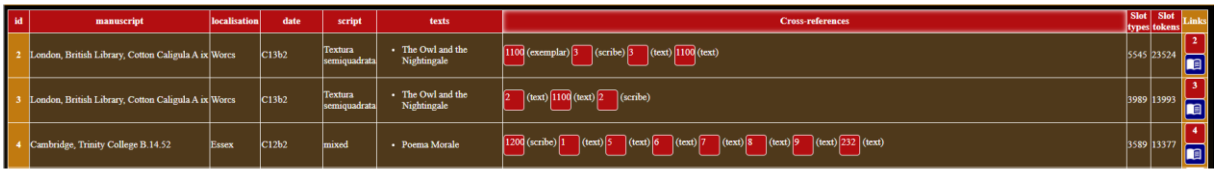
4.2. The interface

The description of the interface briefly explains its functions, which are going to be referred to in the following section (Micro analyses). Although it mentions some of the concepts already introduced in the methodological chapter, it differs from the methodological description in that its perspective is primarily practical and user-focused. Also, a major part of the description deals with "physical" realization of functionalities, which have been described in rather theoretical and general terms so far.

The interface provides several features (screens) which can display data from LAEME and the new spelling database. The individual features are going to be presented here one by one.

4.2.1. Browse files

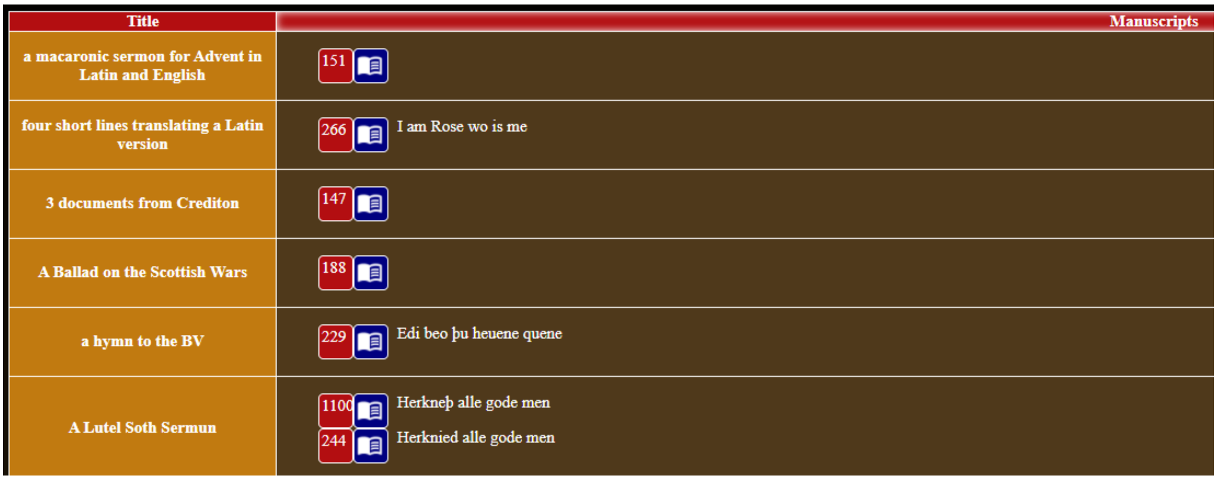
The list of texts found in LAEME can be viewed as a sortable table. The table lists all the files included in the LAEME corpus and gives basic information about them. The columns *manuscript*, *localisation*, *date*, *script* and *texts* are based directly on original LAEME data. The column *cross references* displays links to related texts extracted from LAEME Index of sources and stored in a special table introduced in subchapter 3.4.2.1. Cross references are displayed as clickable links to the full description of the respective files in LAEME.



id	manuscript	localisation	date	script	texts	Cross-references	Slot types	Slot tokens	Links
2	London, British Library, Cotton Caligula A ix	Worcs	C13b2	Textura semiquadrata	- The Owl and the Nightingale	1100 (exemplar) 3 (scribe) 3 (text) 1100 (text)	5545	23524	2
3	London, British Library, Cotton Caligula A ix	Worcs	C13b2	Textura semiquadrata	- The Owl and the Nightingale	2 (text) 1100 (text) 2 (scribe)	3989	13993	3
4	Cambridge, Trinity College B.14.52	Essex	C12b2	mixed	- Poema Morale	1200 (scribe) 1 (text) 5 (text) 6 (text) 7 (text) 8 (text) 9 (text) 232 (text)	3589	13377	4

Figure 18: Screenshot - browse manuscripts

A similar table is available for browsing by text rather than manuscript/file. Here, the first column gives text title and the second column lists corpus files in which the given text appears. The list of corpus files again includes links to LAEME full description and also links to text profiles (see section 4.2.3 below).



Title	Manuscripts
a macaronic sermon for Advent in Latin and English	151 📖
four short lines translating a Latin version	266 📖 I am Rose wo is me
3 documents from Crediton	147 📖
A Ballad on the Scottish Wars	188 📖
a hymn to the BV	229 📖 Edi beo þu heuene quene
A Lutel Soth Sermun	1100 📖 Herkneþ alle gode men 244 📖 Herknied alle gode men

Figure 19: Screenshot - browse texts

4.2.2. Custom database searches

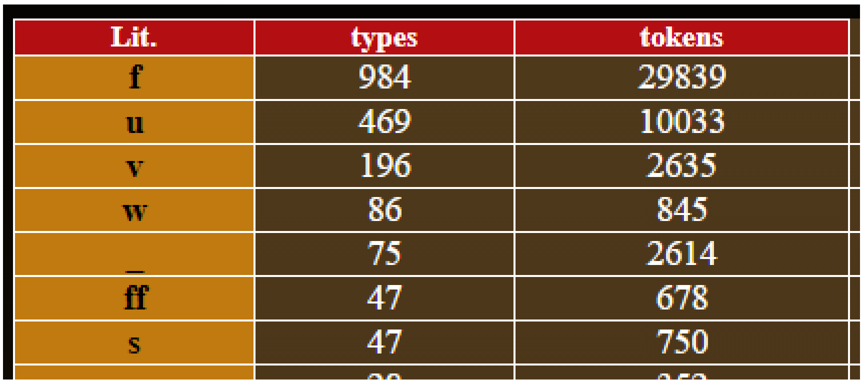
This screen allows the user to search directly for literae, *sets* of litterae or *items*. All searches performed at this page require a single littera or a comma-separated list of litterae as input and the user can switch between simple and advanced queries, which include custom filters. The individual types of searches are described below.

4.2.2.1. Searches for litterae (simple search only)

Searches for litterae are intended to provide basic statistical data about the litterae in the DB, which could serve as a starting point of more complex and in-depth analyses. It is possible to search for alternatives of a given littera or polygraphs containing a given character.

A search for alternatives returns a complete list of litterae which alternate with the input littera at least once in the same slot. For instance, the search for alternatives of *f* returns *v*, *u*, *w* etc.

A search for polygraphs returns a list of polygraphs in which the input littera is present. For instance, the search for *w* returns *hw*, *wh*, *aw* etc. Search results are in both cases displayed as tables and include frequency data for the individual litterae, as illustrated by the picture below.

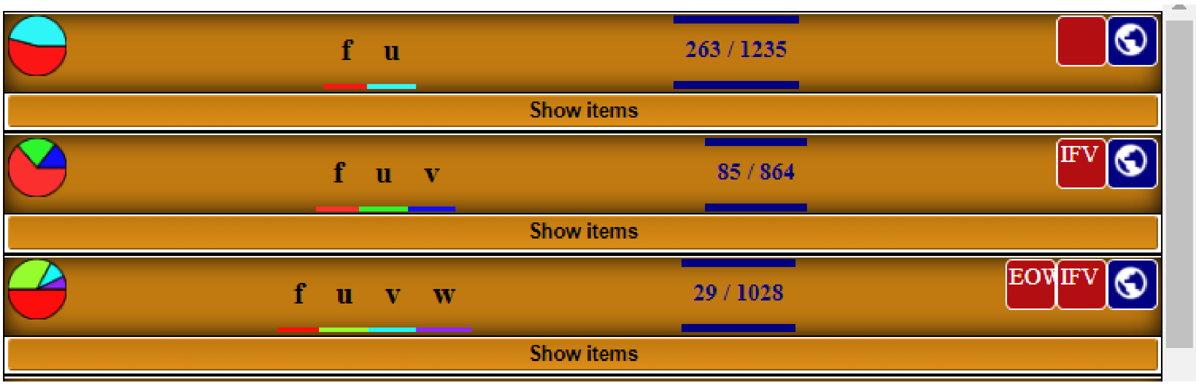


Lit.	types	tokens
f	984	29839
u	469	10033
v	196	2635
w	86	845
_	75	2614
ff	47	678
s	47	750

Figure 20: Screenshot - alternatives of <f>

4.2.2.2. Searches for sets

A search for sets returns all the *sets* (groups of litterae) which at least once occur at the same position as the input littera(e). The input for the search can be a single littera or a comma-separated list of litterae. For example, if the user is interested in the sets in which the littera *f* alternates with *u*, s/he can run a search for "f ,u".



f u 263 / 1235

Show items

f u v 85 / 864

Show items

f u v w 29 / 1028

Show items

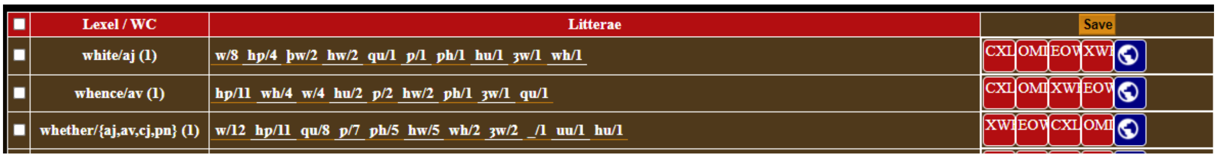
Figure 21: Screenshot - sets containing f/u

Sets are displayed as boxes. The ratio of the litterae in the set is visualised as a simple pie chart and total number of items / total number of tokens in the set are also displayed. The list of slots (items) in which the given range of litterae appear can be displayed immediately for each set.

Whenever the alternation of litterae in the set seems to correspond to a specific sound change described in CoNE, quick links to CoNE are also displayed. Links to CoNE are displayed as red icons with the code of the change, as illustrated by "IFV" (Initial Fricatives Voicing) and "EOV" (Emergence of *v*) in the picture above. This does not indicate that the set is necessarily representative of the change in question. The sets are linked automatically whenever the alternation of litteare in the set corresponds to the pattern expected for the change. The link functionality is an experimental feature, which was not meant to by fully developed within the scope of the present project. So far, only a couple of sets potentially implied by the sound changes have been inserted to the database and the application is not yet able to work with positional and contextual constraints, although this feature could be added.

4.2.2.3. Searches for items

Every set is by definition connected with a list of items (slots) in which the alternation of litterae appears. For example, the set {*w*, *wh*} will be connected with WHENCE/AV (1), WHO/P (1) etc. Therefore, search parametres for item lists are analogical to sets. Search for items is preferrable to search for sets if the user is not interested in displaying separate item lists for each possible set but rather a single list of items. For example, all items exemplifying the *w/wh* alternation are displayed together in one table instead of separate tables for each possible combination of litterae, e.g. {*p*, *w*, *uu*, *wh*, *v*} as opposed to {*qu*, *hw*, *w*, *wh*} etc.



■	Lexel / WC	Litterae	Save
■	white/aj (1)	w/8 hp/4 þw/2 hw/2 qu/1 p/1 ph/1 hu/1 ʒw/1 wh/1	CXL OMI EOV XWI
■	whence/av (1)	hp/11 wh/4 w/4 hu/2 p/2 hw/2 ph/1 ʒw/1 qu/1	CXL OMI XWI EOV
■	whether/{aj,av,cj,pn} (1)	w/12 hp/11 qu/8 p/7 ph/5 hw/5 wh/2 ʒw/2 _/1 uu/1 hu/1	XWI EOV CXL OMI

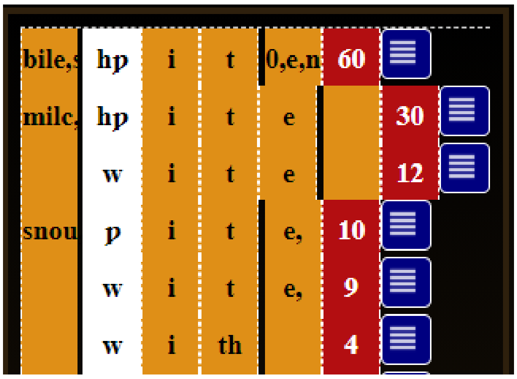
Figure 22: Screenshot - items with alternating w/wh

Items are displayed in a table in the usual format (i.e. lexel/word class plus a bracketed number which stands for the position of the littera in the word, e.g. "WHITE/AJ (1)"). The next column lists the litterae found at the given position along with their token frequency. The last column is reserve for links to CoNE and the link to map.

It is possible to select specific items from any list and save the list locally (local storage in the browser). The use of item lists stored in this manner is going to be demonstrated further on.

4.2.2.3.1. Displaying forms

Actual forms linked to any item on the list and their token frequencies can be loaded directly from any item list. Forms are normally displayed in a table so that the litterae appearing at the same position are aligned in one column. If relevant, one of the columns corresponding to the selected slot may be highlighted (such as the first column for WHITE/AJ (1)) in the picture below.

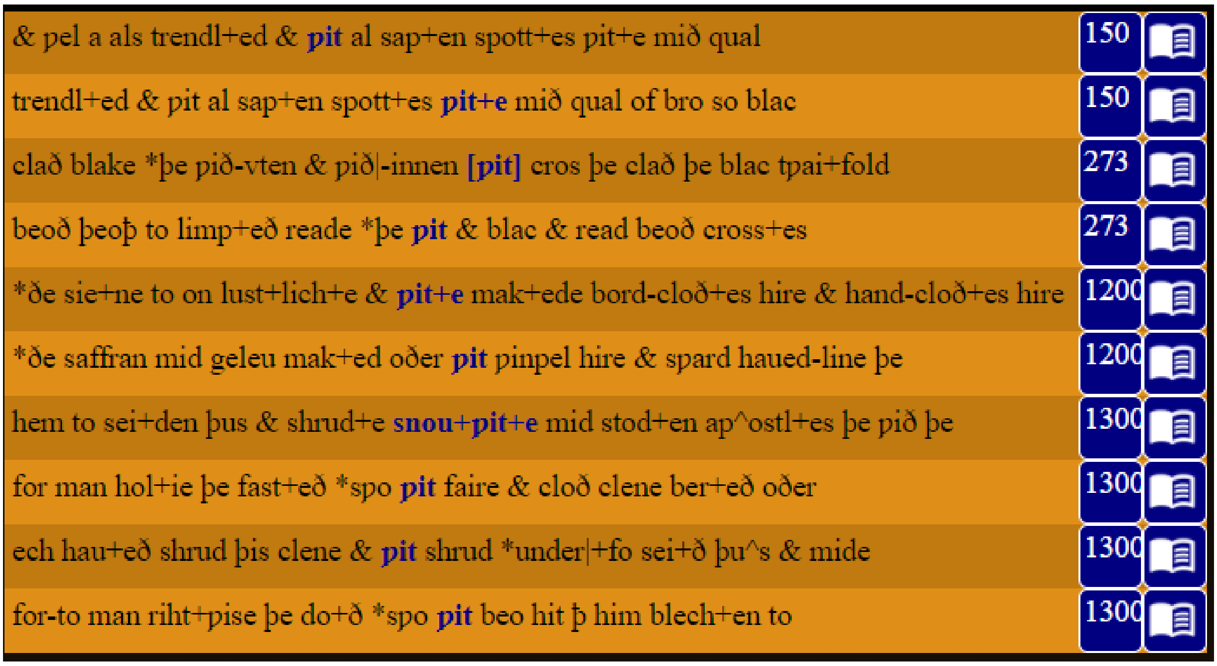


bile,s	hp	i	t	0,e,n	60	
milc,	hp	i	t	e		30
	w	i	t	e		12
snou	p	i	t	e,	10	
	w	i	t	e,	9	
	w	i	th		4	

Figure 23: Screenshot - the forms of WHITE/AJ (1)

As items are morpheme-based, the preceding/following morphemes attached to each form anywhere in the corpus are displayed along with it. For instance the fourth form *pit* in the picture above appears as a part of *snou+pit+e* (SNOW-WHITE).

If any forms of the selected item taken from external sources are available in the database, they can be displayed along with LAEME forms. This is an experimental feature and only 21 OE forms have been included in the database so far. Forms are always displayed along with a blue icon serving as a link to KWIC, which in turn includes links to Text profile and LAEME description. The picture below shows all the occurrences of *pit*, including *snou+pit+e*. The links to text profiles are displayed as blue book icons.



& pel a als trendl+ed & **pit** al sap+en spott+es pit+e mið qual	150
trendl+ed & pit al sap+en spott+es **pit+e** mið qual of bro so blac	150
clað blake *þe pið-vten & pið|-innen **[pit]** cros þe clað þe blac tpai+fold	273
beoð þeoþ to limp+eð reade *þe **pit** & blac & read beoð cross+es	273
*ðe sie+ne to on lust+lich+e & **pit+e** mak+ede bord-cloð+es hire & hand-cloð+es hire	1200
*ðe saffran mid geleu mak+ed oðer **pit** pinpel hire & spard haued-line þe	1200
hem to sei+den þus & shrud+e **snou+pit+e** mid stod+en ap^ostl+es þe pið þe	1300
for man hol+ie þe fast+eð *spo **pit** faire & cloð clene ber+eð oðer	1300
ech hau+eð shrud þis clene & **pit** shrud *under|+fo sei+ð þu^s & mide	1300
for-to man riht+pise þe do+ð *spo **pit** beo hit þ him blech+en to	1300

Figure 24: Screenshot - KWIC

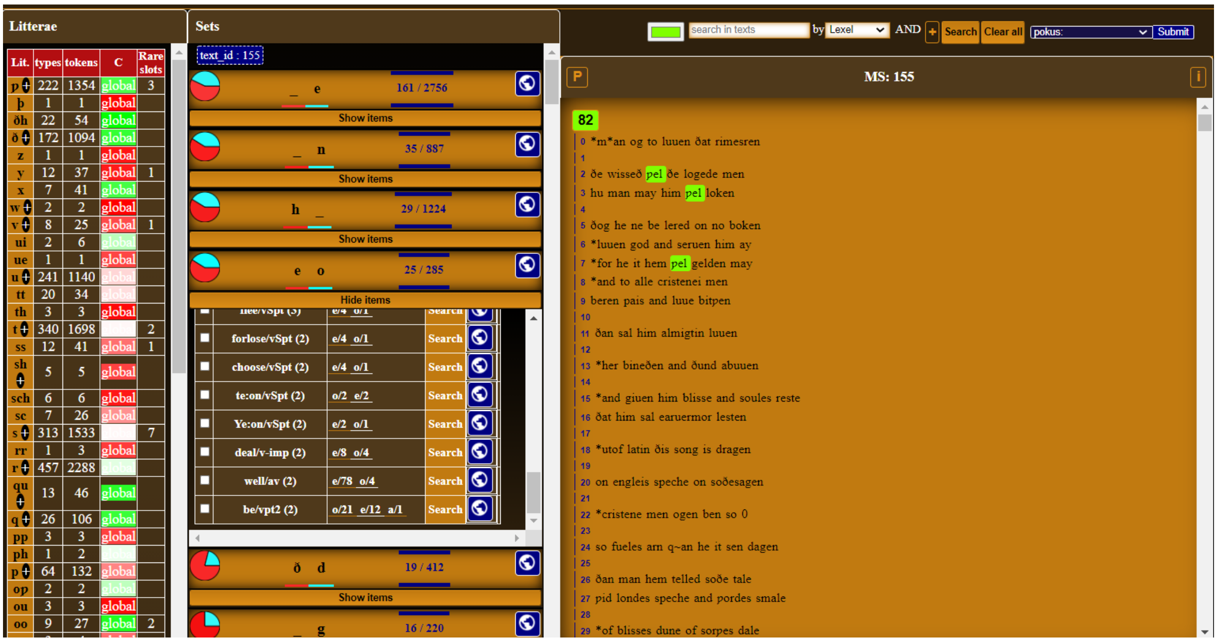
4.2.2.4. Advanced search

Advanced search allows to filter the sets or items by different criteria, described in subchapter 3.6.7.

4.2.3. Text profile

“The purpose of *text profile* is to offer a good starting point as well as tools for a comprehensive analysis of a text language, but it can also be used as a brief overview of the spelling features of the manuscript” (Vaňková, 2021: 13).

The screen *text profile* combines some of the components introduced above (lists of litterae, sets, item list) and also displays the actual text. The basic version of text profile displays three components: the inventory of litterae, sets and the complete text of the MS, each of which is going to be discussed separately. The components are linked together so that the user can use litterae inventory and items (displayed with sets) to navigate the text of the manuscript (see below). The following picture shows *text profile* screen of text #155 (Cambridge, Corpus Christi College 444 containing *Exodus* and *Genesis*).



Lit.	types	tokens	C	Rare slots
p ⊕	222	1354	global	3
þ	1	1	global	
ðh	22	54	global	
ð ⊕	172	1094	global	
z	1	1	global	
y	12	37	global	1
x	7	41	global	
w ⊕	2	2	global	
v ⊕	8	25	global	1
ui	2	6	global	
ue	1	1	global	
u ⊕	241	1140	global	
tt	20	34	global	
th	3	3	global	
t ⊕	340	1698	global	2
ss	12	41	global	1
sh ⊕	5	5	global	
sch	6	6	global	
sc	7	26	global	
s ⊕	313	1533	global	7
rr	1	3	global	
r ⊕	457	2288	global	
qu ⊕	13	46	global	
q ⊕	26	106	global	
pp	3	3	global	
ph	1	2	global	
p ⊕	64	132	global	
op	2	2	global	
ou	3	3	global	
oo	9	27	global	2

_ e 161 / 2756
Show items

_ n 35 / 887
Show items

h _ 29 / 1224
Show items

e o 25 / 285
Hide items

□	forlose/vSpt (2)	e/4 o/1	Search
□	choose/vSpt (2)	e/4 o/1	Search
□	te:on/vSpt (2)	o/2 e/2	Search
□	Ye:on/vSpt (2)	e/2 o/1	Search
□	deal/v-imp (2)	e/8 o/4	Search
□	well/av (2)	e/78 o/4	Search
□	be/vpt2 (2)	o/21 e/12 a/1	Search

ð d 19 / 412
Show items

_ g 16 / 220

MS: 155

82

0 *m*an og to luuen ðat rimesren
1
2 ðe wisseð pel ðe logede men
3 hu man may him pel loken
4
5 ðog he ne be lered on no boken
6 *luuen god and seruen him ay
7 *for he it hem pel gelden may
8 *and to alle cristenei men
9 beren pais and luue bitpen
10
11 ðan sal him almigtin luuen
12
13 *her bineðen and ðund abuuen
14
15 *and giuen him blisse and soules reste
16 ðat him sal earuermor lesten
17
18 *utof latin ðis song is dragen
19
20 on engleis speche on soðesagen
21
22 *cristene men ogen ben so 0
23
24 so fueles arn q~an he it sen dagen
25
26 ðan man hem telled soðe tale
27 pid londes speche and pordes smale
28
29 *of blisses dune of sorpes dale

Figure 25: Screenshot - text profile #155

4.2.3.1. Litteare inventory

The inventory of litterae for the given manuscript is displayed in tabular form and it is very similar to the lists of alternatives and polygraphs (see above), except there are two more columns in the table (see subchapter 3.6.5 for the description of the data).

A coloured rectangle in the third column (“C”) “reflects the relative frequency of the littera compared to average relative frequency in the remaining texts in LAEME. As such, it points to litterae which are either conspicuously rare in the text (marked with red colour) or, contrarily, comparatively more frequent (marked with green colour) and therefore likely to deserve the researcher’s attention” (Vaňková, 2021: 14-15). For instance, *qu* is relatively frequent in text #155 and therefore displayed with a green rectangle, while the relative frequency of *sch* (5 instances only) is clearly below average and therefore displayed with a red rectangle. The word “global” indicates that the frequencies are compared against the average frequency in the whole database.

The last column gives the number of slots in which the littera in question appears in less than 20 % of texts (incidence of *rare uses*). The reason for including this information is that the relatively higher or lower frequency of a littera does not necessarily help to discover cases when a given littera is used in an unusual way. For example, relative frequency of the littera *h* in text #155 is average, still the list of “rare slots” for this littera shows that it sometimes appears in the initial position in EARTH/N (*herðe*), which is relatively uncommon. This measure does not work very well for low-frequency items, e.g. if the total frequency of an item is 3, one occurrence with the given littera is enough for the usage to be marked as rare. A better result could be achieved with a more sophisticated formula for the calculation of “rarity score”.

Litterae which sometimes appear as capital letters, superscripts etc. in the examined text are displayed with a small “+” icons which can be used to show additional information available for the littera, such as the frequency of capitalised occurrences.

Whenever a littera is selected in the inventory, all sets which do not contained the littera are automatically hidden. At the same time, all relevant items (having the selected littera or one of its alternatives) get highlighted in the text on the right. For instance, *th* in text #155 is found in two sets – {*t*, *th*} and {*ð*, *th*}. The column *Rare slots* points straight to the list of items in which the littera rarely appears.

4.2.3.2. Sets

Sets displayed within *text profile* show which litterae sometimes alternate with one another in the same slot. Text #155 has a number of such alternations, e.g {*e*, *o*} (25 slots, including OLD/AJ (2) and WELL/AV (2)), {*c*, *k*} (15 slots, including come/vSpp (1) and BOOK/N (3)). Analyses of the scribal system under examination should, among other things, give explanations for the different cases of alternations found in the text. The complete list of items relevant for each set can of course be loaded straight into the *text profile* screen and instances of the individual items can be highlighted in the text of the manuscript on the right. The picture above shows a part of the list of items with alternating {*e, o*}. The occurrences of WELL/AV have been highlighted in the text.

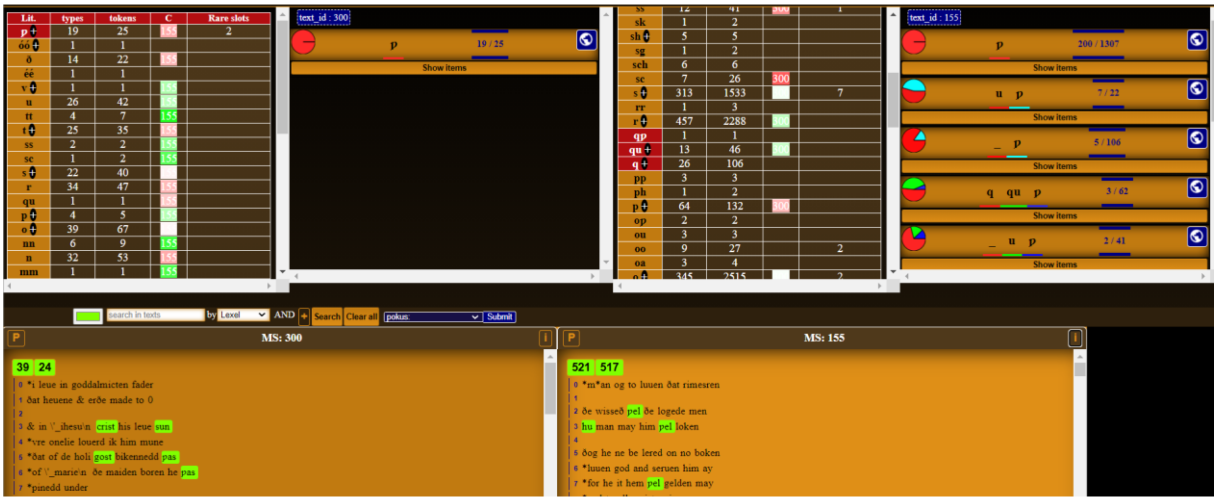
4.2.3.3. Manuscript text

The third component of the *text profile* screen is the actual text of the manuscript. A simple box with basic information about the file corresponding to the data included in the overview table (see subchapter 4.2.1) can be opened if needed. Words in the text can be highlighted in different colours either by selecting items as described previously or by searching the text by lexel, grammel, form (or a combination of the three). Regular expressions can be used in these searches" (Vaňková, 2021: 15). Hovering over a word displays a tooltip showing the lexel and grammel associated with the word.

Furthermore, it is possible to highlight all items from a locally stored item list (see subchapter 4.2.2.3). For instance, after storing a list of all items associated with the set {*wh*, *w*} in the initial position, the user may highlight all the items present in text #155 and examine their realisation straight in the running text.

4.2.3.4. Text comparison

Multiple *text profiles* can be displayed side by side and "the functionalities are very similar to *text profile* of a single text, except any actions (such as searches in the text or filtering of sets) affect all the displayed profiles. The visualisation (red-green rectangle) of littera relative frequency is based on relative frequencies in the compared texts instead of the average values for LAEME as a whole" (Vaňková, 2021: 16). The picture below shows comparison of profiles #155 and #300 after clicking *p* in the inventory of #300. *Qp*, *qu* and *q* are highlighted in the inventory of #155 because they can appear in the same slots as *p* in #300.



Lit.	types	tokens	C	Rare slots
ƿ +	19	25	155	2
óó ⇕	1	1		
ð	14	22	155	
éé	1	1		
v ⇕	1	1	155	
u	26	42	155	
tt	4	7	155	
t ⇕	25	35	155	
ss	2	2	155	
sc	1	2	155	
s ⇕	22	40	155	
r	34	47	155	
qu	1	1	155	
p ⇕	4	5	155	
o ⇕	39	67	155	
nn	6	9	155	
n	32	53	155	
mm	1	1	155	

text_id : 300

ƿ 19 / 25

Show items

ss	12	41	300	1
sk	1	2		
sh ⇕	5	5		
sg	1	2		
sch	6	6		
sc	7	26	300	
s ⇕	313	1533	300	7
rr	1	3		
r ⇕	457	2288	300	
qƿ	1	1		
qu +	13	46	300	
q +	26	106		
pp	3	3		
ph	1	2		
p ⇕	64	132	300	
op	2	2		
ou	3	3		
oo	9	27		2
oa	3	4		
o ⇕	345	2515		2

text_id : 155

ƿ 200 / 1307

Show items

u ƿ 7 / 22

Show items

_ ƿ 5 / 106

Show items

q qu ƿ 3 / 62

Show items

_ u ƿ 2 / 41

Show items

MS: 300

39 24

0 *i leue in goddalmicten fader
1 ðat heuene & erðe made to 0
2
3 & in \'_ihesu\n crist his leue sun
4 *vre onelie louerd ik him mune
5 *ðat of de holi gost bikennedd ƿas
6 *of \'_marie\n ðe maiden boren he ƿas
7 *pinedd under

MS: 155

521 517

0 *m*an og to luuen ðat rimesren
1
2 ðe wisseð ƿel ðe logede men
3 hu man may him ƿel loken
4
5 ðog he ne be lered on no boken
6 *luuen god and seruen him ay
7 *for he it hem ƿel gelden may

Figure 26: Screenshot - text profile comparison, #155 and #300

4.2.4. Maps[16]

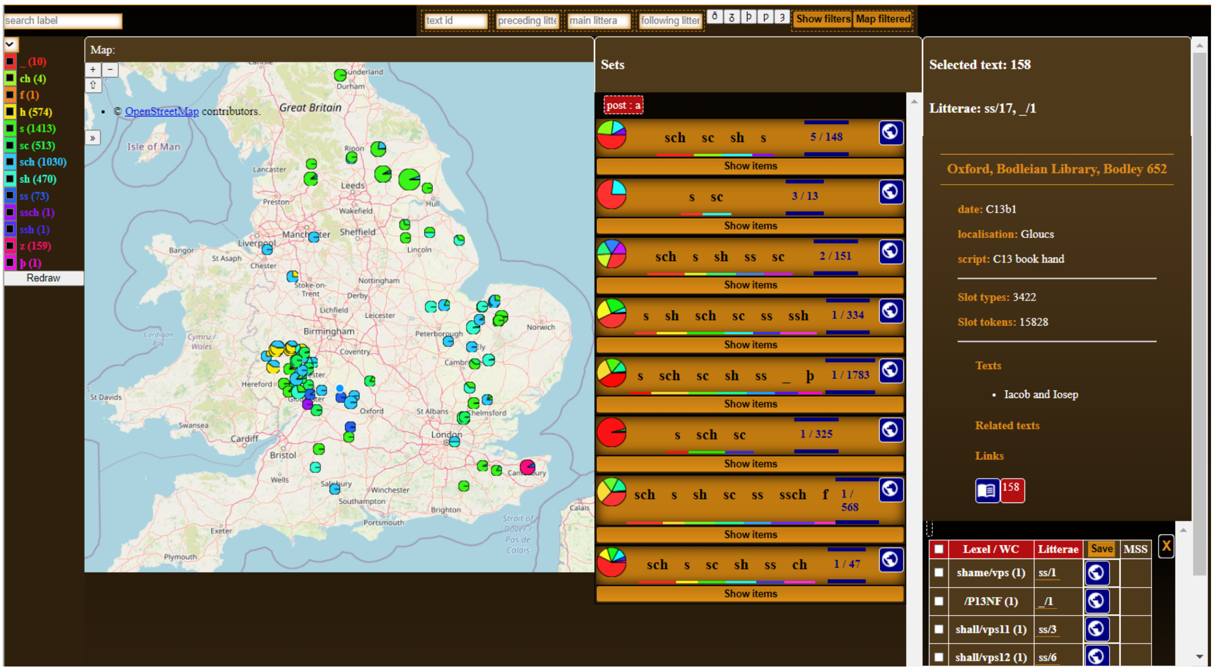
Maps generated from the spelling DB data work slightly differently in comparison with the custom maps in LAEME. In order to generate a custom map in LAEME, the researcher has to define a search for a specific form of a specific lexel (group of lexels) and the result of the search is subsequently plotted on the map. Naturally, multiple realisations of the same feature can be used to make a single map. For instance, when analysing the reflexes of OE *sc*, the researcher may plot various realisations of the initial element of SHALL, i.e. *sc*, *sh*, *s*, *ss* etc. one by one.

A similar analysis using the parsed data can be performed simply by selecting a slot or specifying a list of slots (e.g. position 1 in all items of SHALL) and the tool automatically plots all the possible realisations on the map. The list of slots can be defined by search input analogical to normal searches for sets in the database or by making a custom list from a list of items.

4.2.4.1. Maps generated from sets (defining lists of slots indirectly)

Maps generated from sets take input analogical to search for sets, including filtering options, e.g. "*s*, *sc* before *a*" can be used to search for all slots in which this set occurs. The application looks up all slots (items) having this alternation of litterae and counts all the litterae in such slots for each text. The calculation is currently available only in token frequency.

[16] Maps are generated using the open source library OpenLayers (https://openlayers.org)



Set	Count
sch sc sh s	5 / 148
s sc	3 / 13
sch s sh ss sc	2 / 151
s sh sch sc ss ssh	1 / 334
s sch sc sh ss _ þ	1 / 1783
s sch sc	1 / 325
sch s sh sc ss ssch f	1 / 568
sch s sc sh ss ch	1 / 47

Selected text: 158

Litterae: ss/17, _/1

Oxford, Bodleian Library, Bodley 652

date: C13b1

localisation: Gloucs

script: C13 book hand

Slot types: 3422

Slot tokens: 15828

Texts

- Iacob and Iosep

Related texts

Links

158

□	Lexel / WC	Litterae	Save	MSS
□	shame/vps (1)	ss/1		
□	/P13NF (1)	_/1		
□	shall/vps11 (1)	ss/3		
□	shall/vps12 (1)	ss/6		

Figure 27: Screenshot - map for "s,sc before a"

The data is displayed on the map in the form of pie charts with a different colour for each littera. The size of the chart reflects the total number of tokens included in the calculation. Colour legend is displayed along with the map on the left. If required, it is possible to change the colours and redraw the map. This can make the map easier to read if the user is interested in sounds and assumes identical sound value for multiple litterae, which can be then displayed in the same colour. For instance, *sc, sh* and *sch* may be displayed in red and *s* in blue rather than having a separate colour for each of the variants.

The pie charts themselves function as links to the manuscripts. A click on the chart displays basic information about the manuscript along with the list of items included in the calculation for the pie chart. In the picture above, text #158 has been selected.

If the map is based on sets, all of its subsets are also displayed (on the right).

Simple searches (no filter) can be run in two modes: “Map set” and “Map strict set”. The difference between the two is that the first mode adds all possible corresponding litterae alternating with user input, e.g. if the user runs a search for “*s*, *sc*”, the result will also include *sh*, *sch*, *ss*, *ssc* etc. The second mode will only include items which have the alternation of *s*, *c* but no other litterae.

4.2.4.2. Maps generated from slots

Besides mapping whole sets, the user can display data for a specific slot or a custom list of slots. This can be done straight from any item list. The function “Map set” (see above) displays sets and lists of items directly within the *maps* screen and these lists include links to maps.

One of the expected uses of the map is combining “set maps” and “slot maps”. For instance, in the example above, the map for set “s, sc” is based on many slots, including SCEAFT/N (1), SHADOW/N (1), SHAME/N (1) etc. The researcher can use the item list to examine maps for the individual slots and get separate maps for SCEAFT/N (1), SHADOW/N (1), SHAME/N (1) etc. or select a smaller group of slots and combine them into one map.

Another way is to generate a map from a stored item list.

4.2.4.3. Map sequence

Maps based on simple searches (without filters) can be easily transformed into a sequence of maps for different time periods, which may provide insight into the progress on changes in time as well as space. The picture below shows such separate maps for the initial segment in SHALL. The first map displays only texts from the periods C12b2-C13a2 and the second map texts from the periods C13b1-C14a1.

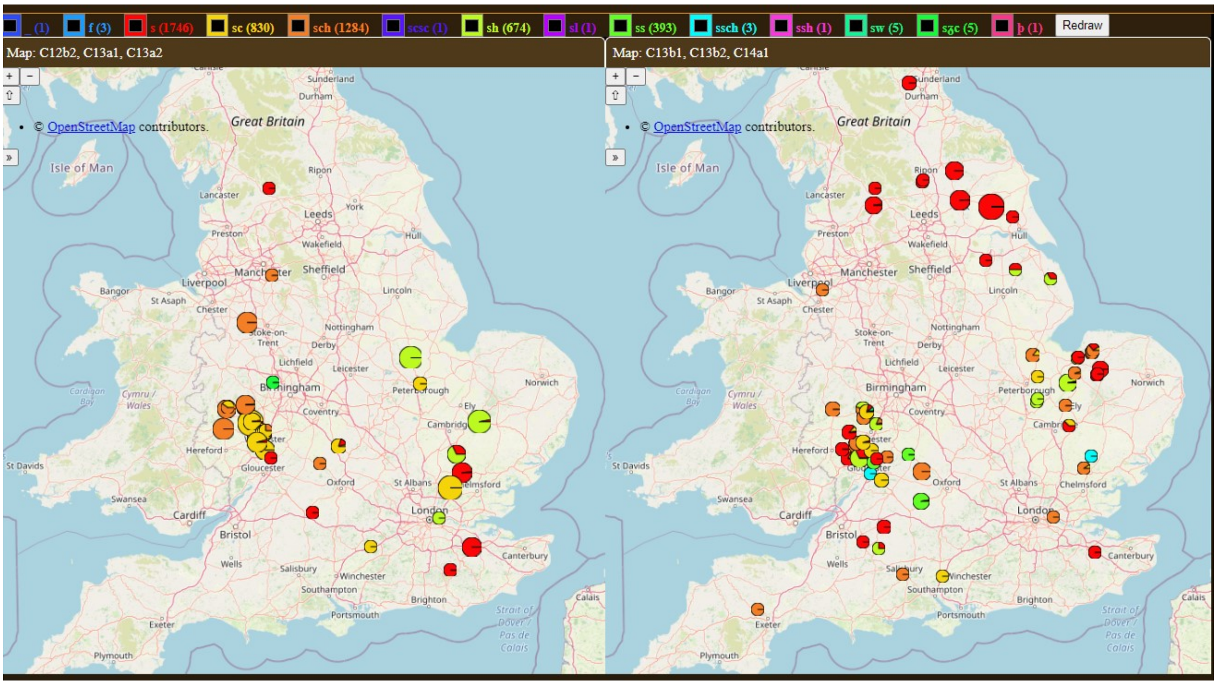


Figure 28: Screenshot - map sequence

Maps can be easily stored for future reference. The simplest way is to copy and save the URL. E.g. the map for *sc*, *s* before *a* can be accessed directly through the following link: http://laeme-spelling.silent3.ff.cuni.cz/#/map/mapfilteredSet/sc,s/%5B%7B%22level%22:%221%22,%22field%22:%22post%22,%22operator%22:%22equals%22,%22values%22:%5B%22a%22%5D%7D%5D .

4.2.5. Network visualisation[17]

In some cases it may be more convenient to look at the sets of alternating litterae in the form of network rather than a table. Network diagrams can be generated for a single text, for all sets in the database or two different texts. The last type displays litterae from the individual texts in different colours. For instance, the above mentioned correspondences between *p* in text #300 and *qp*, *qu*, *q* in #155 are visualised as folows:

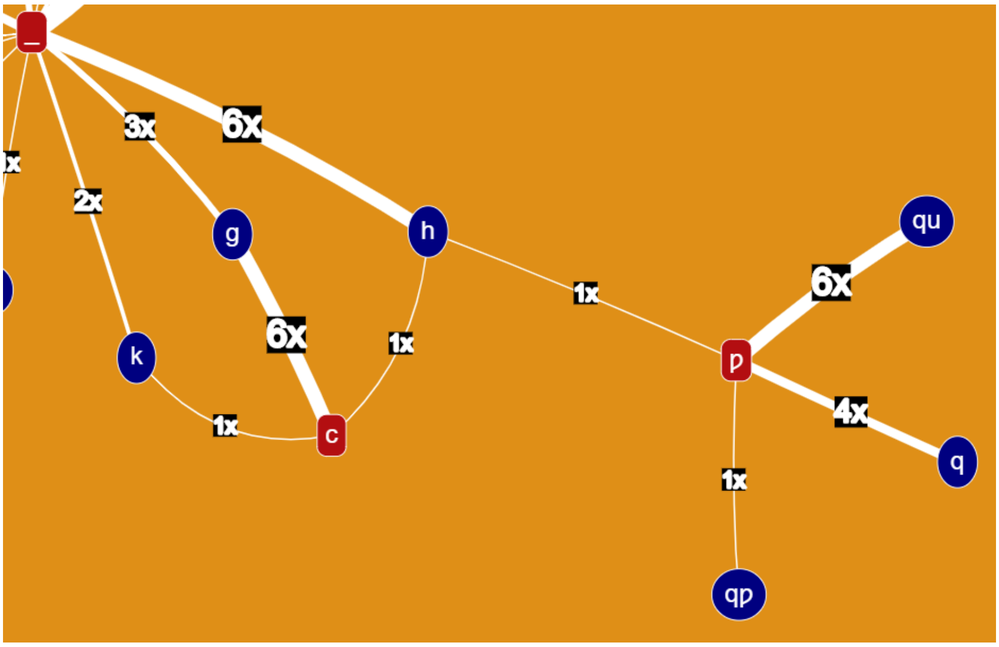


Figure 29: Screenshot - network visualisation for #155 (blue) and #300 (red)

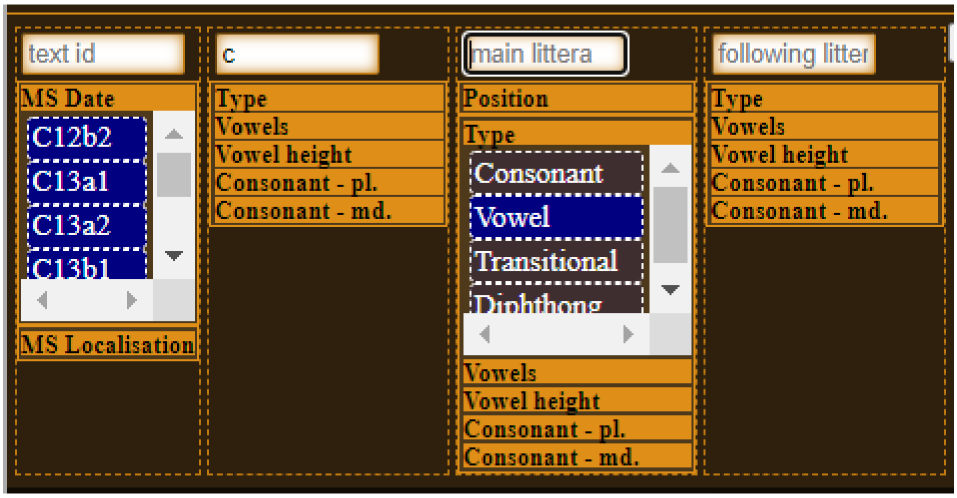
Each node in the network represents a littera and edges connecting the nodes show which litterae sometimes correspond to one another. The width of the edges reflects token frequency of the correspondences of two connected litterae and the actual frequency is displayed as a number in a black box.

Clicking a node or an edgde triggers a search for relevant items and the result is displayed as an item list. Item lists derived from the large network covering all texts are not available (yet).

[17] The networks are generated by the openSource library vis.js (https://visjs.org)

4.2.6. Filters

Filters can be applied in searches for sets, items or maps. The possible filters have been already described in subchapter 3.6.7. The field “main littera” does not necessarily have to be filled in advanced search. For example, the user may search for all sets of vowels following *c* in the texts from the periods C12b2-C13b1 (see the picture below).



c

C12b2
C13a1
C13a2
C13b1

Vowel

Figure 30: Screenshot . filter setup

4.2.7. Quick links

The description mentioned several types of quick links, which can be found across the application. Links always look like blue or red icons and always open a new tab. The possible links include text profile, map, network, KWIC, LAEME file description and CoNE change description.

4.3. Micro analyses

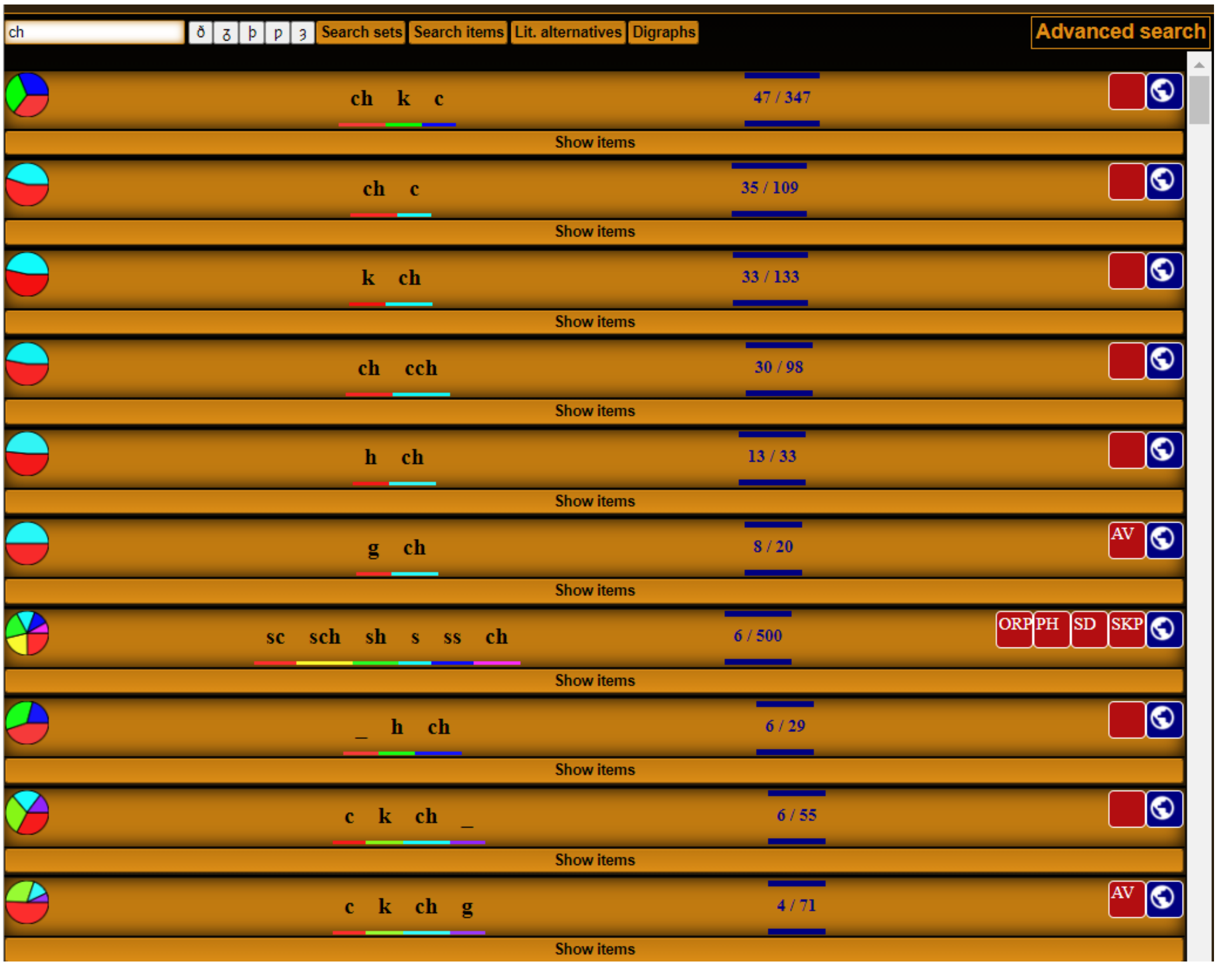
The purpose of this subchapter is to demonstrate the use of the tool in practice. This is done in a series of sample micro analyses. Each of the micro analyses focuses primarily on a specific component of the tool (e.g. DB searches, networks, maps). The last micro-analysis combines multiple functions, illustrating the possible directions of navigating the available data as outlined at the end of the methodological chapter (subchapter 3.8.1). Some of the micro analyses are followed by a brief note on connections between the feature of the tool and a specific methodological concept introduced in the theoretical chapter. As for the choice of specific problems or manuscripts examined within the micro analyses, the more familiar and well researched topics are preferred so that the output from the database can be compared with results obtained with more traditional methods.

Additional examples of the application of the tool can be found in an article about a separate study dealing with selected litterae in a group of related texts (Vaňková, 2021).

4.3.1. Sets and custom filters

The first micro study focuses on working with *sets*, because the concept of *set* is essential to the whole structure of the database. Sets may reflect sound changes as well as varying spelling practices, including their development. Various querying possibilities will be demonstrated on the example of *sets* with *ch*.

All the examples of queries discussed in this section were submitted through the feature *search DB*. The simplest possible query is to list all the sets in which *ch* appears (type *ch* in the search box and click "Search sets"). The picture below shows the output of the query (only the topmost part of the result is visible):



ch

ch k c 47 / 347

ch c 35 / 109

k ch 33 / 133

ch cch 30 / 98

h ch 13 / 33

g ch 8 / 20 AV

sc sch sh s ss ch 6 / 500 ORP PH SD SKP

_ h ch 6 / 29

c k ch _ 6 / 55

c k ch g 4 / 71 AV

Figure 31: Screenshot - sets containing <ch>

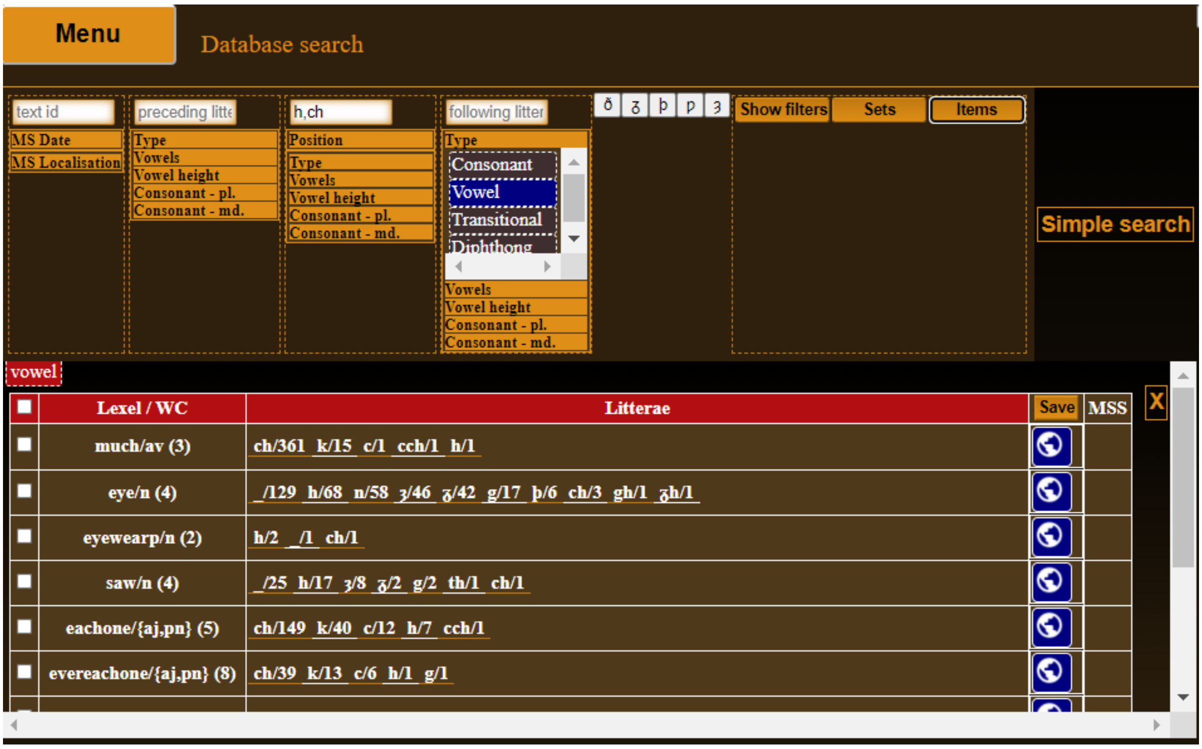
The result is essentially a list of possible combinations of litterae which sometimes appear at the same position as *ch* sorted by frequency. Note that the numbers on the right give type/token frequency of slots with the exact combination of litterae but there may be more items having the combination plus more litterae. For instance, there are 35 types of slots which have either *ch* or *c* and nothing else, but the item list associated with this set will also include slots with sets like {*ch*, *c*, *k*} etc.

Different sets potentially reflect different uses (and therefore *potestates*) of the litterae in question or merely variant representations of a single sound. As for the sets in the picture, there are several sets in which *ch* alternates with {*c*, *k*}, two sets where it alternates with {*h*, _}, one set with {*sc*, *sch*} and two sets containing *g*. Larger but less common sets further down show that the sets with *h* may also include *ȝ*. Litterae found in place of *ch* rather rarely include also *ȝh*, *hh*, *s*, *th* and several others.

There seem to be two basic kinds of sets, namely {*ch*, *c, k*} and {*ch*, *h*}. There are two basic explanations for this. Either *ch* represents roughly the same sound value in both cases and each of the set reflects a different sound change, e.g. [k] > [t͡ʃ] and [t͡ʃ] > [x] respectively or *ch* represents two different sounds, e.g. [t͡ʃ], [x] in the two sets and the sets may or may not reflect sound changes. Obviously, prior knowledge of OE and ME speaks strongly in favour of the latter explanation. The sample analysis will further focus on the sets with *h*.

4.3.1.1. Examining item lists

A new search was performed. The input was "*ch*, *h*". This search returned only those sets directly relevant for the examination of the given set type. Items for the set were displayed to identify the contexts in which the set appears (only the first half of the list was examined for the purpose of this analysis). This step confirmed the expected occurrence of {*ch*, *h*} before *t* in items like BRIGHT/AJ, AE:HT/N or RIGHTLY/AV, but the same set also appeared in other positions, especially in place of OE *h* in other positions and OE *g*, usually in morpheme-final position. The frequencies of these uses were low, which suggests that this usage could be limited to a small number of texts. In order to examine the associated items separately from those with {*h*, *ch*} before *t*, two separate advanced queries for items were used: {*ch*, *h*} in morpheme-final position and {*ch*, *h*} before vowels. Negative filter (i.e. {*ch*, *h*} "not before" *t*) is not (yet) available. The picture below shows filter setup and query results of the latter query:



h,ch

vowel

□	Lexel / WC	Litterae	Save	MSS
□	much/av (3)	ch/361 k/15 c/1 cch/1 h/1	🌐	
□	eye/n (4)	_/129 h/68 n/58 ȝ/46 ʒ/42 g/17 þ/6 ch/3 gh/1 ʒh/1	🌐	
□	eyewearp/n (2)	h/2 _/1 ch/1	🌐	
□	saw/n (4)	_/25 h/17 ȝ/8 ʒ/2 g/2 th/1 ch/1	🌐	
□	eachone/{aj,pn} (5)	ch/149 k/40 c/12 h/7 cch/1	🌐	
□	evereachone/{aj,pn} (8)	ch/39 k/13 c/6 h/1 g/1	🌐	

Figure 32: Screenshot - items with {h,ch} before vowels

Further analysis of the items could involve mapping or examination of specific text languages and will not be pursued further at this point.

4.3.1.2. CoNE references

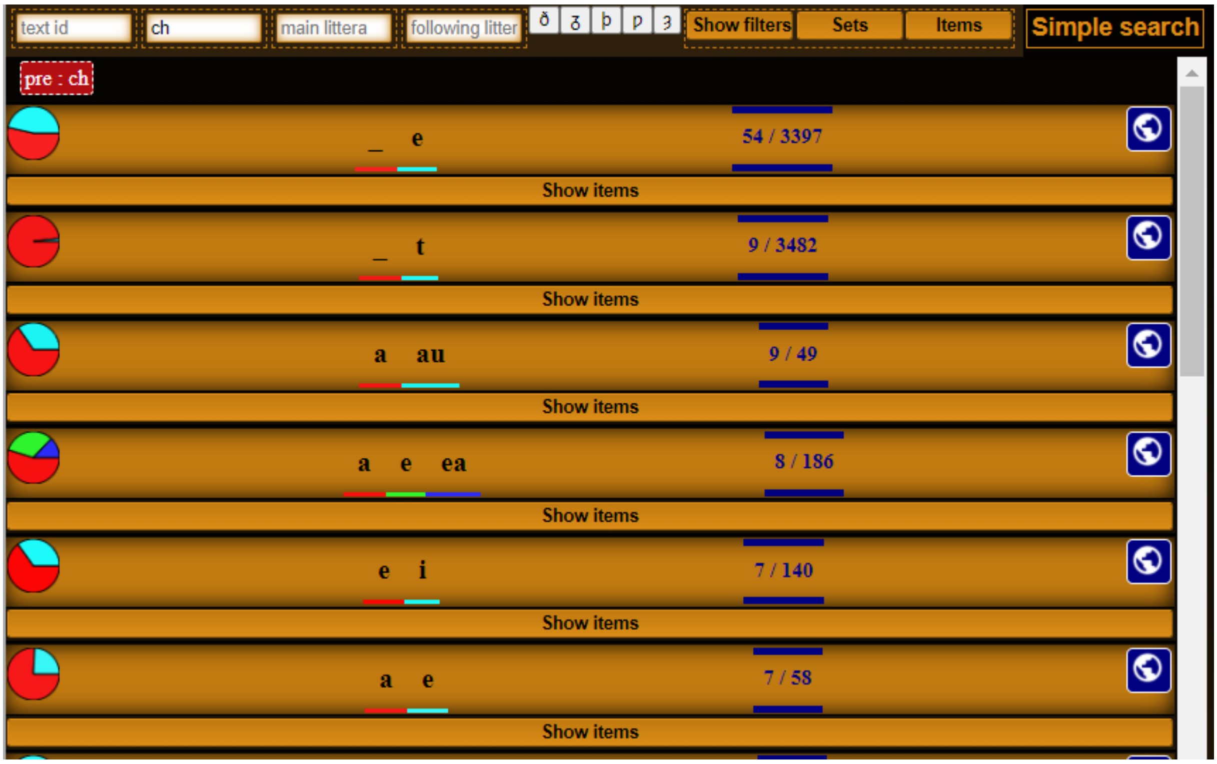
Another possibility of working with sets, which is going to be demonstrated here, involves references to CoNE. Some of the sets in the original picture are displayed along with red icons with two or three-letter codes. The sets {*g*, *ch*} and {*c*, *k*, *ch*, *g*}, for instance have the icon "AV", which is CoNE code for "Affricate Voicing" and serves as a direct link to CoNE. The tool can be used to access the items potentially affected by the change and plot the variants on the map.

As for "Affricate voicing" specifically, the proposed change consists in [ʧ] becoming [ʤ] and the entry quotes several examples of the change, namely *gildre* (CHILDREN), *cherge* (CHURCH), *heouerige* (KINGDOM OF HEAVEN), some forms of EACH, *ig* (first person pronoun) and DITCH, SUCH plus "a number of spellings in text #263" (CoNE, AV). The items listed under the set {*ch*, *g*} include the examples from CoNE plus a number of other items sharing the pattern of alternating *ch*, *g*. If we exclude those with the group *-cht-/-gt-* , only a few candidates for the proposed change remain: *-lige* (-LY/XS) in #280 (Wiltshire), *dringen, dringes* (DRINK/VI) in #2000, #280 and *swinge* (SWINC/N) in #2001.

Note that multiple links to CoNE may be displayed with a single set, which is the case of {*sc*, *sch*, *sh*, *s*, *ss*, *ch*} in the initial picture. All the entries in CoNE were checked to see which of the changes is in fact the most relevant one for the given set. The result was that the items potentially exemplify "Palatal Hardening" (PH), whereby [ʃ] > [ʧ] (CoNE). The list of associated items could again be used as a basis of further analyses.

4.3.1.3. Filtering

The last example of working with sets to be given here concerns filtering options. Advanced filters can be used to retrieve sets of litterae following or preceding a given littera. The list of such sets for *ch* was obtained with the following query:



ch

pre : ch

_ e — 54 / 3397 — Show items

_ t — 9 / 3482 — Show items

a au — 9 / 49 — Show items

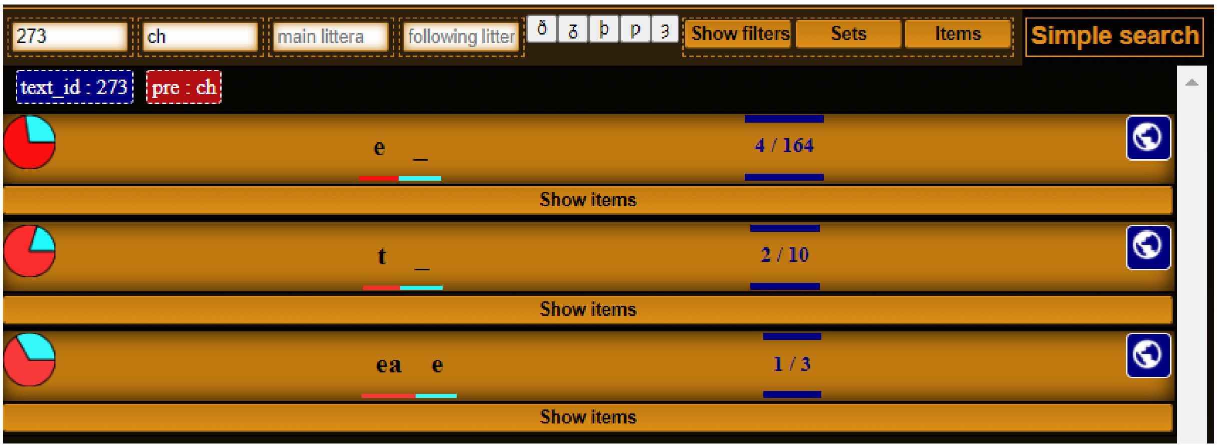
a e ea — 8 / 186 — Show items

e i — 7 / 140 — Show items

a e — 7 / 58 — Show items

Figure 33: Screenshot - sets following <ch>

Ch is given as the "preceding littera" and the field "main littera" is left blank. The sets show that *ch* is almost always followed by vowels, the only exception being *t*, which sometimes drops (see the set {*t*, _}). A brief inspection of the associated lists of items reveals that some of them form reasonably homogenous groups, notably items under the set {*a*, *au*}, all of which are Romance lexemes. A similar query targeting a particular text language can be used to observe contextual constraints governing the use of a littera in the given text. For instance, the results of the query below show that in text #273, *ch* appears before *e*, *ea* as well as *t* (which sometimes drops):



273 | ch

text_id : 273 | pre : ch

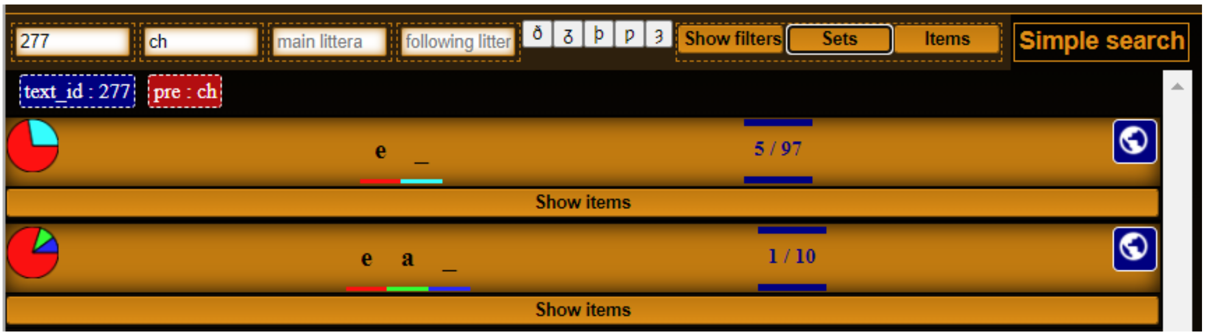
e _ 4 / 164

t _ 2 / 10

ea e 1 / 3

Figure 34: Screenshot - sets following <ch> in #273

The same query for text #277 returns only vocalic sets, which entails that *ch* never appears before *t* in this text:



277 | ch

text_id : 277 | pre : ch

e _ 5 / 97

e a _ 1 / 10

Figure 35: Screenshot - sets following <ch> in #277

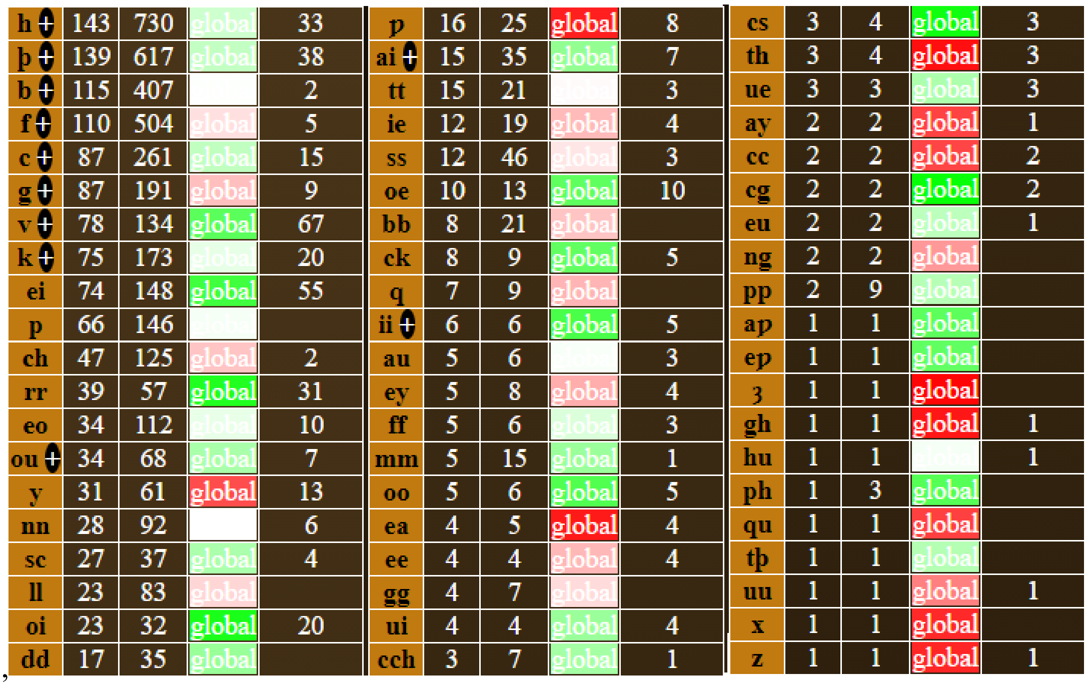
Sets of litterae which do appear before *t* in #277 can in turn be listed with a similar query ("text=277 & following littera=t").

4.3.1.4. A note on litteral substitution

Searching for sets appears to be compatible with Laing & Lass's (2013) concept of *litteral substitution* (discussed in subchapter 2.1.2.2.2) and can be used to perform analyses within this framework. The *sets* essentially suggest potential sequences of litteral substitution. For instance, if a scribe found *g* when he would normally expect *ch* (regardless of the intended *potestas*), a possible consequence would be the use of *g* instead of *ch* on the part of this scribe and this practice did not have to be limited to the item in which *g* was first seen.

4.3.2. Text profile

The next micro analysis tests the *text profile* screen. The text chosen for analysis is Cambridge, Trinity College B14.39, scribe A (#246), which is notorious for the prodigality of its spelling system (Laing 2003). The goals of the analysis were (a) to identify rare litterae (suggestive of exemplar influence), (b) examine their distribution in the text, (c) identify the litterae with which the rare litterae alternate and (d) examine the list of their rare uses.



h ⊕	143	730	global	33	ƿ	16	25	global	8	cs	3	4	global	3
þ ⊕	139	617	global	38	ai ⊕	15	35	global	7	th	3	4	global	3
b ⊕	115	407		2	tt	15	21	global	3	ue	3	3	global	3
f ⊕	110	504	global	5	ie	12	19	global	4	ay	2	2	global	1
c ⊕	87	261	global	15	ss	12	46	global	3	cc	2	2	global	2
g ⊕	87	191	global	9	oe	10	13	global	10	cg	2	2	global	2
v ⊕	78	134	global	67	bb	8	21	global		eu	2	2	global	1
k ⊕	75	173	global	20	ck	8	9	global	5	ng	2	2	global	
ei	74	148	global	55	q	7	9	global		pp	2	9	global	
p	66	146	global		ii ⊕	6	6	global	5	aƿ	1	1	global	
ch	47	125	global	2	au	5	6	global	3	eƿ	1	1	global	
rr	39	57	global	31	ey	5	8	global	4	ȝ	1	1	global	
eo	34	112	global	10	ff	5	6	global	3	gh	1	1	global	1
ou ⊕	34	68	global	7	mm	5	15	global	1	hu	1	1	global	1
y	31	61	global	13	oo	5	6	global	5	ph	1	3	global	
nn	28	92		6	ea	4	5	global	4	qu	1	1	global	
sc	27	37	global	4	ee	4	4	global	4	tþ	1	1	global	
ll	23	83	global		gg	4	7	global		uu	1	1	global	1
oi	23	32	global	20	ui	4	4	global	4	x	1	1	global	
dd	17	35	global		cch	3	7	global	1	z	1	1	global	1

,

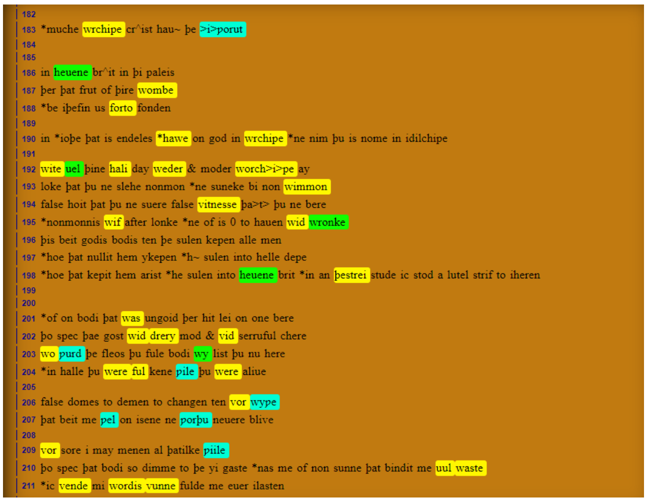
Figure 36: Screenshot - inventory of litterae in #246

The first task was to identify the rare litterae in the text. These litterae should be displayed in the inventory with a red rectangle, indicating relatively lower normalized frequency. According to the data in the tool, such rare litterae are: *y* (31/61), *ƿ* (16/25), *ea* (4/5), *th* (3/4), *cc* (2/2) and *ay* (2/2), *ȝ* (1/1), *x* (1/1), *z* (1/1), *qu* (1/1), *gh* (1/1). Litterae with markedly higher normalized frequency are *rr* (39/57)., *oi* (23/32), *cs* (3/4) and *cg* (2/2).

4.3.2.1. *The distribution of litterae*

One of the questions to be asked about rare litterae is whether the instances are scattered across the whole text or whether they are concentrated at one place. The quickest way to answer this question with the tool is to highlight all the items in which a given littera appears at least once and subsequently highlight all the actual occurrences of the littera with a different colour (using regular expression search)[18]. In this particular case, items sometimes spelled with *w* were first highlighted in yellow, the items which sometimes have *ƿ* were highlighted in green and a regular expressopm search was used to highlight all words with actual *ƿ* in turquoise. The picture below shows a part of the manuscript with the highlighted words:

[18] The tool currently does not offer distribution visualisation, which would provide a faster way of dealing with the task at hand.



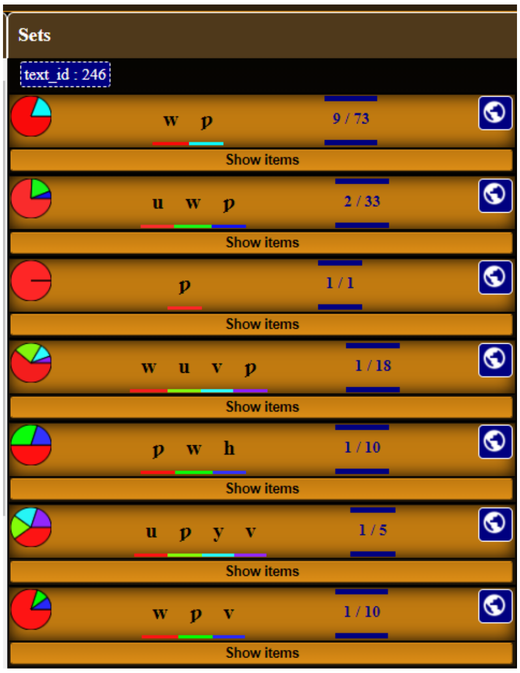
```
182
183 *muche wrchipe cr^ist hau~ þe >i>ƿorut
184
185
186 in heuene br^it in þi paleis
187 þer þat frut of þire wombe
188 *be iþefin us forto fonden
189
190 in *ioþe þat is endeles *hawe on god in wrchipe *ne nim þu is nome in idilchipe
191
192 wite uel þine hali day weder & moder worch>i>pe ay
193 loke þat þu ne slehe nonmon *ne suneke bi non wimmon
194 false hoit þat þu ne suere false vitnesse þa>t> þu ne bere
195 *nonmonnis wif after lonke *ne of is 0 to hauen wid wronke
196 þis beit godis bodis ten þe sulen kepen alle men
197 *hoe þat nullit hem ykepen *h~ sulen into helle depe
198 *hoe þat kepit hem arist *he sulen into heuene brit *in an þestrei stude ic stod a lutel strif to iheren
199
200
201 *of on bodi þat was ungoid þer hit lei on one bere
202 þo spec þae gost wid drery mod & vid serruful chere
203 wo ƿurd þe fleos þu fule bodi wy list þu nu here
204 *in halle þu were ful kene ƿile þu were aliue
205
206 false domes to demen to changen ten vor ƿype
207 þat beit me ƿel on isene ne ƿorþu neuere blive
208
209 vor sore i may menen al þatilke ƿiile
210 þo spec þat bodi so dimme to þe yi gaste *nas me of non sunne þat bindit me uul waste
211 *ic vende mi wordis vunne fulde me euer ilasten
```

Figure 37: Screenshot - highlights in the manuscript

The results show that most of the *wynns* (*ƿ*) are found between the lines 183 and 204, which can partly be seen from the picture above (words with *ƿ* are in turquoise). The yellow and green highlights are helpful in determining whether the absence of *ƿ* from some passages is due to the absence of items in which *ƿ* can be expected to appear or not.

4.3.2.2. Interchangeable litterae

Another possible task is to identify the litterae which alternate with the littera in question, in this case *ƿ*. The easiest way is to display all the sets in which *ƿ* occurs in #246 (click *ƿ* in the inventory of litterae). The picture below shows all of such sets:



Sets

text_id : 246

w ƿ 9 / 73
Show items

u w ƿ 2 / 33
Show items

ƿ 1 / 1
Show items

w u v ƿ 1 / 18
Show items

ƿ w h 1 / 10
Show items

u ƿ y v 1 / 5
Show items

w ƿ v 1 / 10
Show items

Figure 38: Screenshot - sets with *ƿ* in #246

The results show that the most common alternative is (not surprisingly) *w*. Apart from *w*, *p* sometimes alternates with *u*, *v*, *h* and *y*. The slots and items under the individual sets can be loaded by clicking "Show items", as shown above. For instance, the item list under {*u*, *w*, *p*} comprises HEAVEN/N, OVER-/XP and WELL/AV.

4.3.2.3. Rare uses

The last functionality to be tested here is the identification of rare uses. This is best demonstrated on litterae whose relative frequency is average. One of such litterae in #246 is *s*. According to the inventory displayed in the tool, *s* has 52 "rare uses". The corresponding item list comprises a number of items in which *s* appears in place of historical *h* before *t* (SIGHT/N, THINK/VPT, NIGHT/N etc.)[19]. There are also a few instances of *s* for expected *f* (*aster* – AFTER, *ges* - GIVE/V-IMP). LOFSONG/N and OFFSPRING/N where the usual *f* s followed by *s* are written *lo**sf**ong* and *o**sf**pring*. Lastly, *s* is sometimes found initially for expected [ʃ] (SHALL, SCEAFT/N, SCENE/AJ) but *s* in this position is relatively more common. More rare uses include *v* for expected initial *w* (e.g. *vende* (wendan/vi)), or *t* in place of the more common *d* in *heuet* (HEAD/N), *srout* SHROUND/N.

The list of *rare uses* of *s* is a good example of how the measure of "rarity" is supposed to work. However, more tests of the feature suggest that the results are not always equally satisfactory. The measure is less reliable when applied to extremely rare or extremely common items. For instance, the use of *y* in BY or WHY were marked as rare. Obviously, if the littera itself is rare, the list of "rare slots" can comprise all the slots in which it appears. For instance, *oi* from #246, which is quite uncommon in itself, has all of its occurrences marked as rare.

4.3.2.4. A note on scribal lexicon

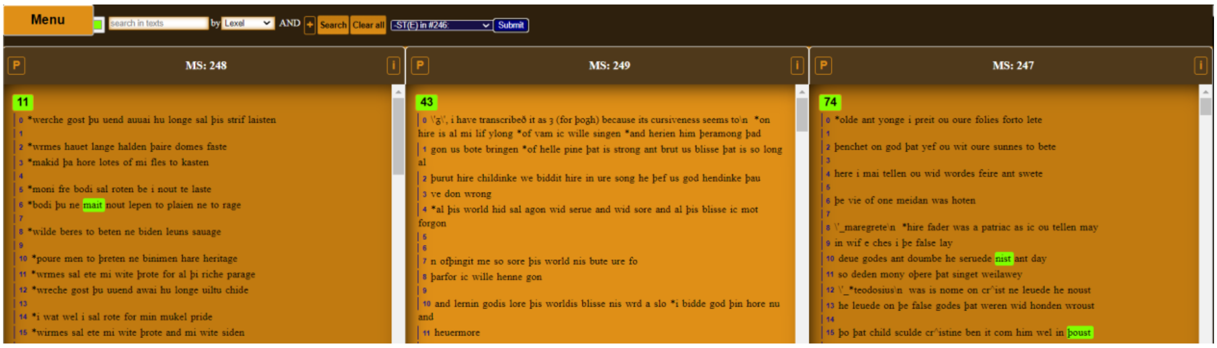
The inventory of literae along with the list of alternations can be used as a point of departure for the construction of a *scribal lexicon* in the form of LSSs and PSSs discussed in subchapter 2.2.3.3. For each littera in the inventory, the tool offers a list of alternating litterae, a list of items in which the littera is used and a quick way of finding the items in the text.

[19] This use of -*st* is explained in Laing & Lass (2003) as a case of backspelling „based on Old French sound change [sr] > [xt~c¸t~ht] (Laing & Lass, 2003: 262).

4.3.3. Item lists

Any item list loaded into the interface, such as the list of *rare uses* above or any item list associated with a set in the first micro analysis can be stored and re-used. This micro analysis demonstrates how this method can be used to check for equivalents of the final *-st(e)* spelling employed by scribe A of #246 in other texts found in the same MS but copied by different scribes (#247-249).

First, the list of *rare uses* was loaded into the interface. The items with final *-st* for *-ht* were selected and saved under the label "-ST(E) in #246". This item list was subsequently used to instantly highlight all the relevant items in *text profiles* of #247-249 (see the picture below).



MS: 248

11

0 *werche gost þu uend auuai hu longe sal þis strif laisten
1
2 *wrmes hauet lange halden þaire domes faste
3 *makid þa hore lotes of mi fles to kasten
4
5 *moni fre bodi sal roten be i nout te laste
6 *bodi þu ne mait nout lepen to plaien ne to rage
7
8 *wilde beres to beten ne biden leuns sauage
9
10 *poure men to þreten ne binimen hare heritage
11 *wrmes sal ete mi wite þrote for al þi riche parage
12 *wreche gost þu uuend awai hu longe uiltu chide
13
14 *i wat wel i sal rote for min mukel pride
15 *wirmes sal ete mi wite þrote and mi wite siden

MS: 249

43

0 \'ʒ\', i have transcribeð it as ȝ (for þoȝh) because its cursiveness seems to\n *on hire is al mi lif ylong *of vam ic wille singen *and herien him þeramong þad
1 gon us bote bringen *of helle pine þat is strong ant brut us blisse þat is so long al
2 þurut hire childinke we biddit hire in ure song he þef us god hendinke þau
3 ve don wrong
4 *al þis world hid sal agon wid serue and wid sore and al þis blisse ic mot forgon
5
6
7 n ofþingit me so sore þis world nis bute ure fo
8 þarfor ic wille henne gon
9
10 and lernin godis lore þis worldis blisse nis wrd a slo *i bidde god þin hore nu and
11 heuermore

MS: 247

74

0 *olde ant yonge i preit ou oure folies forto lete
1
2 þenchet on god þat yef ou wit oure sunnes to bete
3
4 here i mai tellen ou wid wordes feire ant swete
5
6 þe vie of one meidan was hoten
7
8 \'_maregrete\n *hire fader was a patriac as ic ou tellen may
9 in wif e ches i þe false lay
10 deue godes ant doumbe he seruede nist ant day
11 so deden mony oþere þat singet weilawey
12 \'_*teodosius\n was is nome on cr^ist ne leuede he noust
13 he leuede on þe false godes þat weren wid honden wroust
14
15 þo þat child sculde cr^istine ben it com him wel in þoust

Figure 39: Screenshot - equivalents of -ST(E) in #246

Scribe B (#247) has 74 instances of items from the list and it is highly consistent in using *-st* in the same positions as scribe A. There are only a few exceptions, namely LIGHT/VI (*litte*) and THOUGHT (*þout*).

Scribe C (#248) does not use *-st* in the examined positions at all. There are only 11 instances of the items and all with the exceptions of *acite* (AÉHT/N) are spelled with final *–(t)t*. The nuclear vowel in NIGHT/N, AÉHT/N and MAY/VPS12 is sometimes changed to *-ai-*.

Scribe D uses a range of alternatives, including *-st*, in the positions defined by the item list. There are 43 instances the items from the list. Examples of alternatives include: *licte* (LIGHT/N), *nitf* (NIGHT/N), *rid (*RIGHT/N), *cnith* (KNIGHT/N), *achte* (AÉHT).

The same data could be gathered either by reading through the texts or by searching for the forms of the items in LAEME one by one. While reading may be preferrable for analyses focussing on a single (short) text, the advantage of item lists is that they enable to gather the forms from multiple texts in a relatively short time.

4.3.4. Network visualisation

Network visualisation of correspondences between litterae essentially displays the same data as the list of alternating litterae available within *text profile* but it can be more convenient in that it shows all the links in a single picture. Networks are a good starting point of analyses aimed at understanding of a specific spelling system. The screenshot of network visualisation for *The Ormulum* illustrates that alternation of litterae at the same position are common even in the most “regular” writing systems.

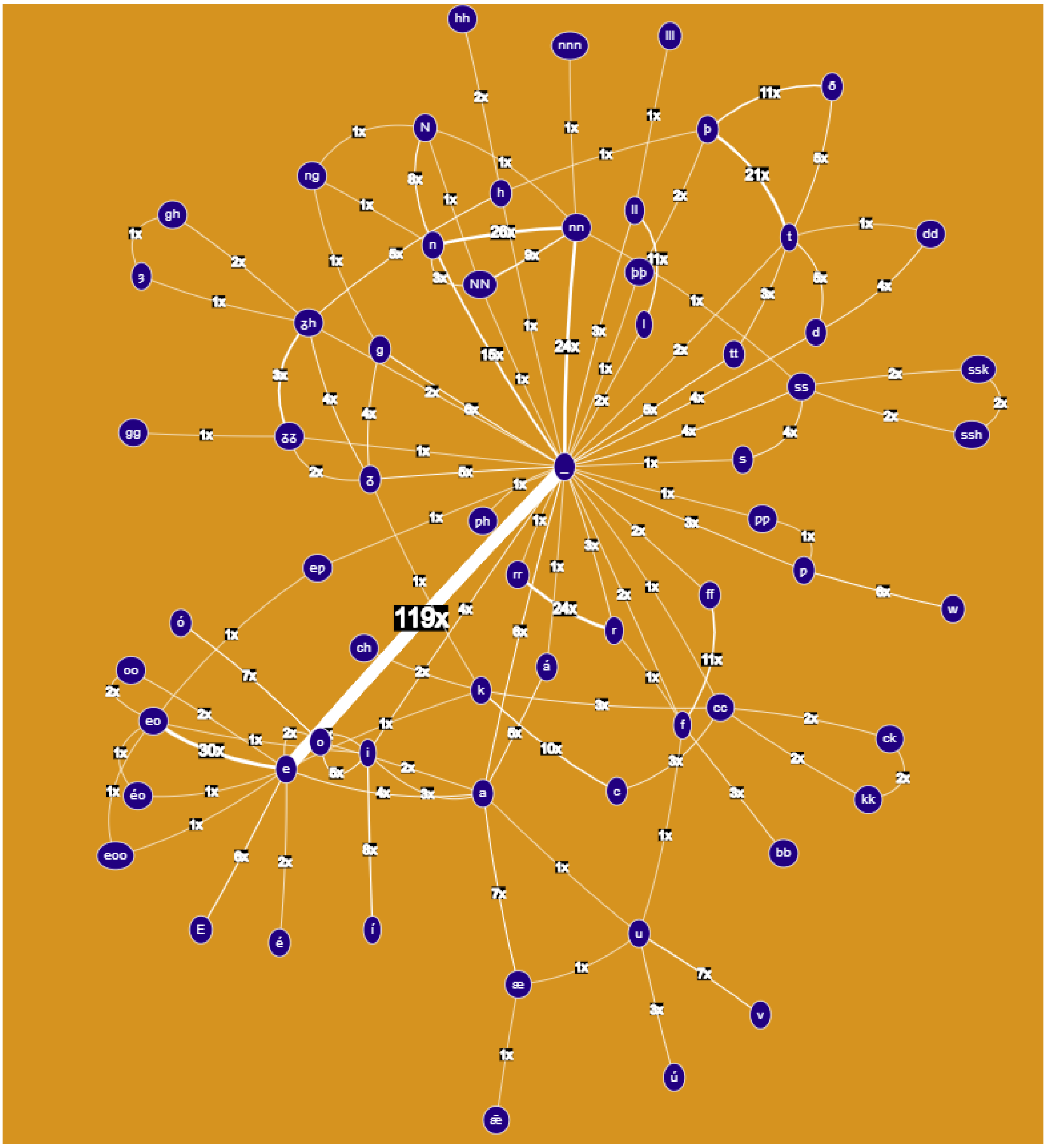


Figure 40: Screenshot - network visualisation of the spelling system of *The Ormulum*

It is important to realize that some of the alternations may be in fact governed by positional or contextual constraints, which needs to be verified. A closer look at the alternations in *The Ormulum*, shows that *v/u* and *c/k* are in fact regular (*v* is used as a capital letter only, *c* finally and *k* before *e*). Orm's doubled letters sometimes alternate with single letters and diacritics is not used consistently (*ó* alternates with *o* etc.) *G* alternates with insular *ꝫ* in (FOR)GIVE, even though the two theoretically should have distinct functions in Orm's system (Laing & Lass, 2013: 2.2.1.) and the alternation probably suggests that they represented very similar sounds. The alternation of *a* and *i* appears in NIGHT, MIGHT and MAY, *a* being very unusual at this position.

The relative economy of *The Ormulum* can be contrasted with the extreme level of prodigality found in the previously analysed MS Cambridge, Trinity College B14.39 (scribe A, #246). The complexity of the network clearly reflects the difference between the two spelling systems:

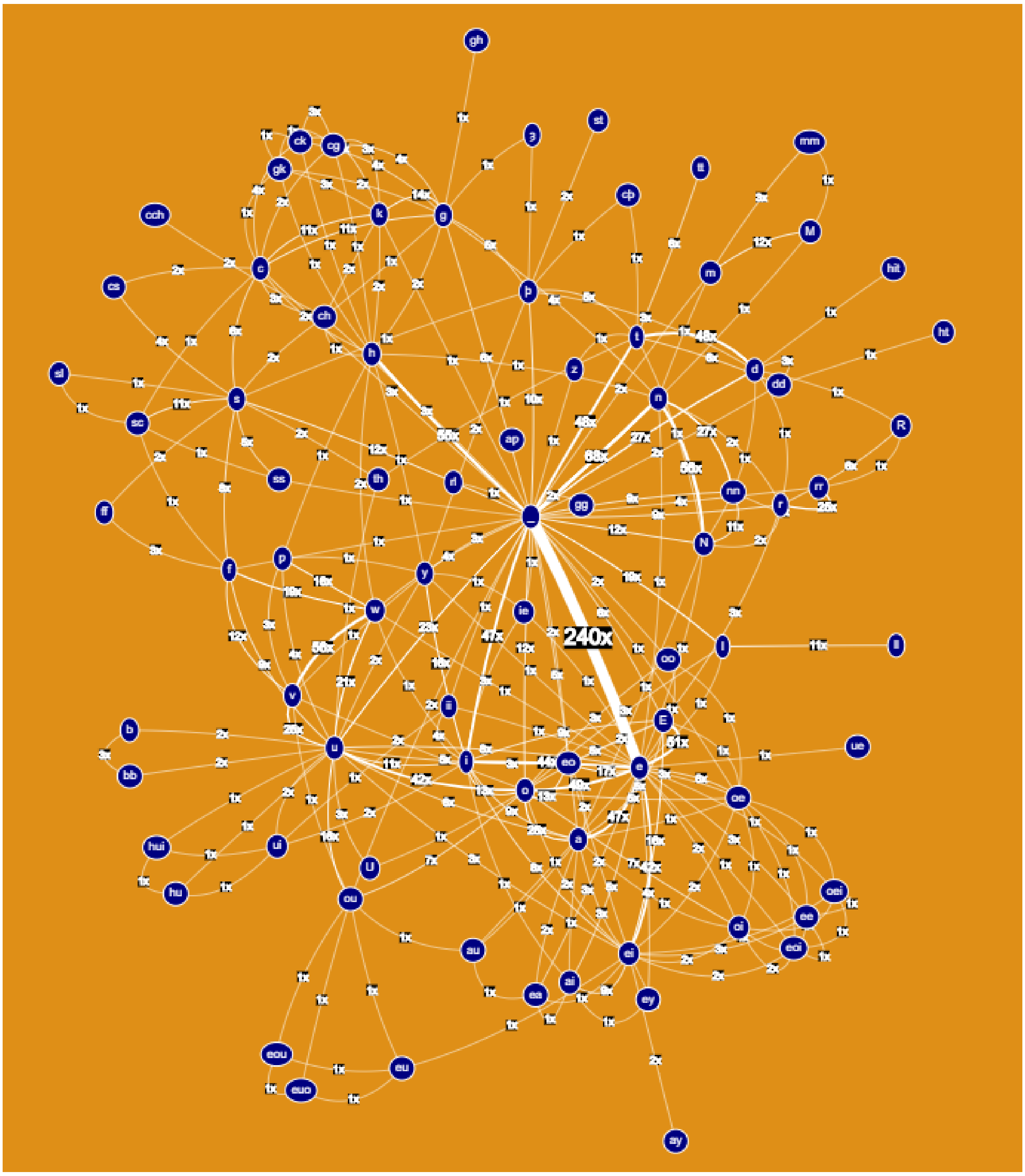


Figure 41: Screenshot - network visualisation, text #246

The network reveals that besides sets similar to those in *The Ormulum* (doubled/single letter, {*þ*, *t*} etc.), the system of scribe A has a number of much less predictable and often rare alternations like {*st*, *þ*}, {*h*, *þ*} or {*g*, *ck*}. Moreover, the sets form quite complex clusters and "chains". For instance, *gh* alternates with *g* (1x), which in turn alternates with *h* (2x), which alternates with 10 different litterae or zero, among others, with *w* and *p* which also alternate with each other etc. Most of the alternations occur only once or twice. The overall impression is in agreement with results of previous research (Laing 2003), which suggest that the scribe worked by his ears and his perception of what counts as the same sound is rather approximate and loose and/or let through a lot of varying spellings from his exemplars without changing them.

4.3.4.1. Litterae alternating with s

This micro analysis covers sets including *s* plus some other related sets, thereby revisiting the example of LSS in #246 employed by Laing & Lass in the Introduction in LAEME (Laing & Lass, 2013: 2.3.2.). First, item lists for the selected sets were loaded and saved. The items were highlighted in the text displayed within *text profile* so that not only the forms but also their distribution in the manuscript could be observed. Some of the items were also plotted on the map to compare the variants with texts localised nearby.

The alternation of litterae was sometimes accompanied by other changes in the form as a whole, for instance, *scrut*, *srout* (SHROUD/N), *suiniz*, *scinet* (SHINE/VPS), *scauit*, *sauit*, *scuiþe* (SHOW). This might suggest that the scribe copied the form as a whole

S is used extensively in typical *h*-contexts (RIGHT, MIGHT, NIGHT etc.) but, perhaps curiously, it never alternates with *h*. As *s* alternates with *f*, *ff*, *th* and also *þ*, a possible explanation seems to be that the final *-st(e)* for historical *-hte* in #246 in fact reflects a transitional stage towards the loss of [x], perhaps something closer to [ni:(θ)te] rather than [ni:xte]. Data from maps shows that texts #248 and #249 (the same manuscript, localised nearby) have forms like *rid* (RIGHT), *nitf* (NIGHT), *litht* (LIGHT) (#249) *nit/naite* (NIGHT), *mitte* (MIGHT).

It is also interesting to look at the multiple alternations between *s*, *c*, *sc*, *cs*, *ch* and *k* observable in the picture below (the relevant litterae are displayed in green for convenience):

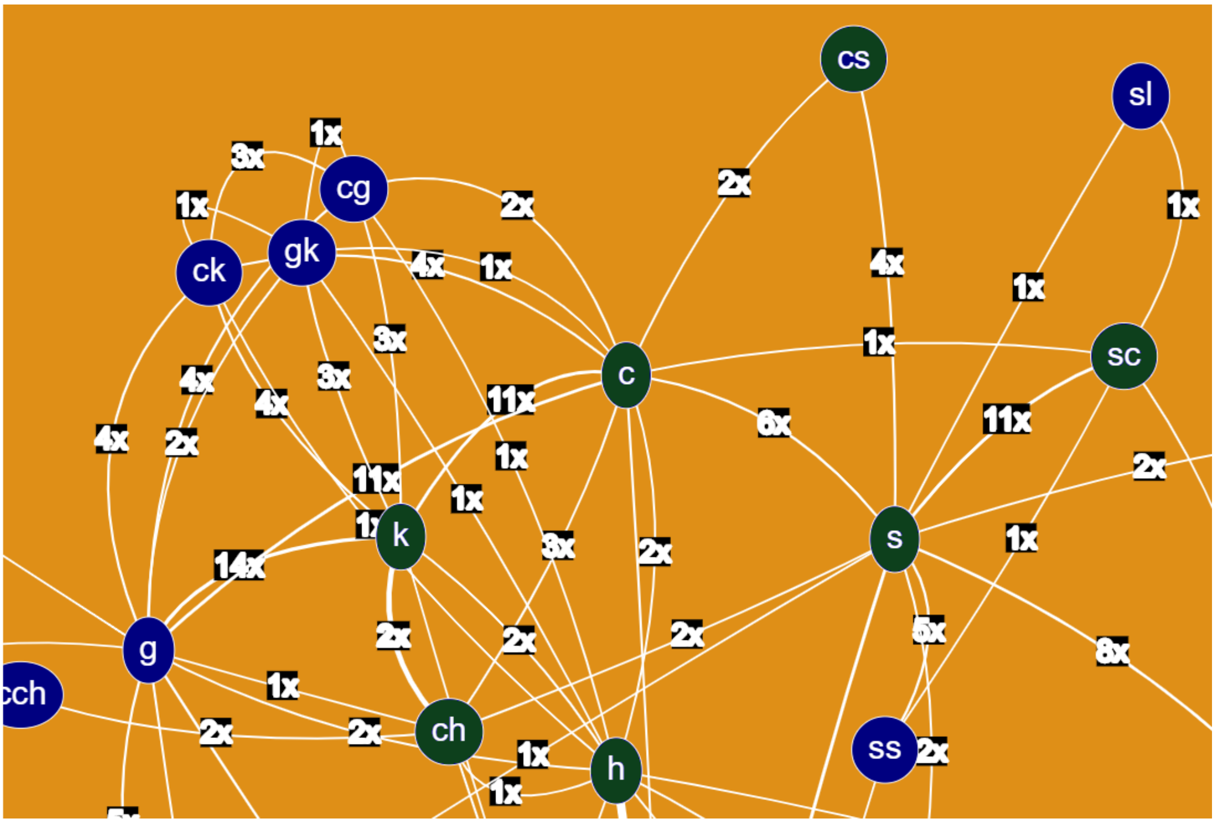


Figure 42: Screenshot: network relations between selected litterar in #246

Besides the set {*s, c*}, there are also {*s, cs, c*}, {*s, sc, c*} (but not {*cs, sc*}) and {*s, c, ch}*. *C* also alternates with *k* and *ch* which *s* does not and {*k, ch*} are used interchangeably.

Observations regarding the distribution of the forms in the text are rather less reliable because of low frequency of the items. SHALL, which is by far the most frequent item is spelled predominantly with initial *s*, the forms in *sc*- are only occasional and the greatest concentration of this variant was found in the text of *Doomsday* (6 forms out of 13 between the lines 1070-1150 have *sc*-). WIGHT appears 5 times and the forms *viit* and *vichit* are found in the first quarter of the text, while *wist, viste* occur later on. The same applies to the forms of RIGHT and NIGHT ending in -*cst*. The last of such forms is found close to *vichit* (line 403). FLESH has a number of variants and similar variants always appear together. The first occurrence reads *fleisc*, the next two *fleos*, the next two *flece* and the last *fles*.

4.3.5. Network comparison

In addition to networks based on a single text, the tool comprises network visualisations of correspondences between litterae in two selected texts. The nodes and edges of the network serve as clickable links to the already familiar item lists. The use of the network will be demonstrated on the example of texts #1100 (Oxford, Jesus College, 29, *The Owl and the Nightingale* plus several shorter pieces) and #3 (London, British Library, Cotton Caligula A ix, language 2 of *The Owl and the Nightingale*). A common exemplar has been proposed for the two texts and previous analyses suggest that the Cotton scribe was a literatim copyist while the scribe of Jesus College was a translator (Laing, 2004). Therefore, reasonably regular correspondences should be visible in the network. As the network is too large to fit in a single image, only a part of it is included for illustration.

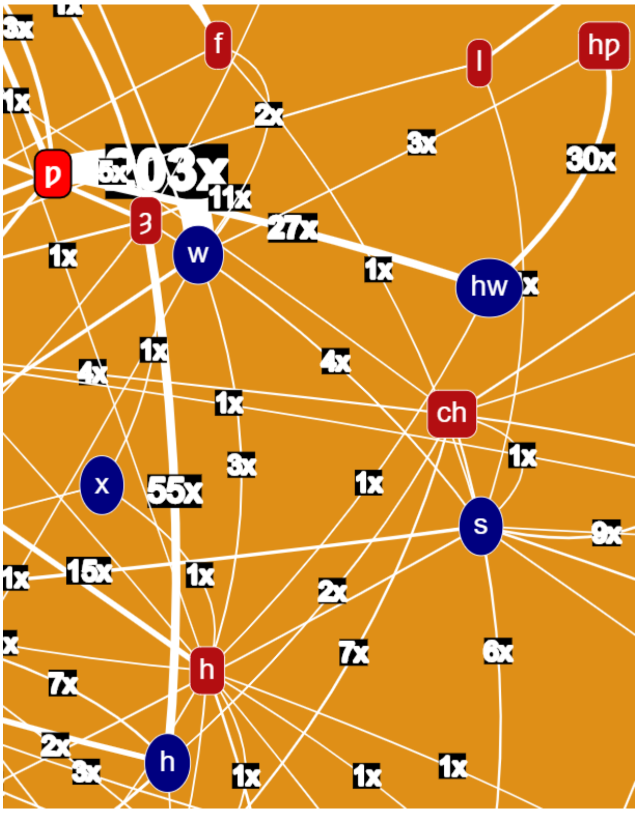


Figure 43: Screenshot - network comparison (texts #1100 (blue), #3 (red))

Blue nodes represent litterae in text #1100 (Jesus College 29) and red nodes represent litteare in #3 (Cotton MS). The picture shows some of the most prominent correspondences between litterae in the two texts, which are summarized in the table below:

The variant in #1100	**Corresponding variant in #3**	**Token frequency**
w	*p*	206
hw	*hp*	30
h	*ʒ*	56

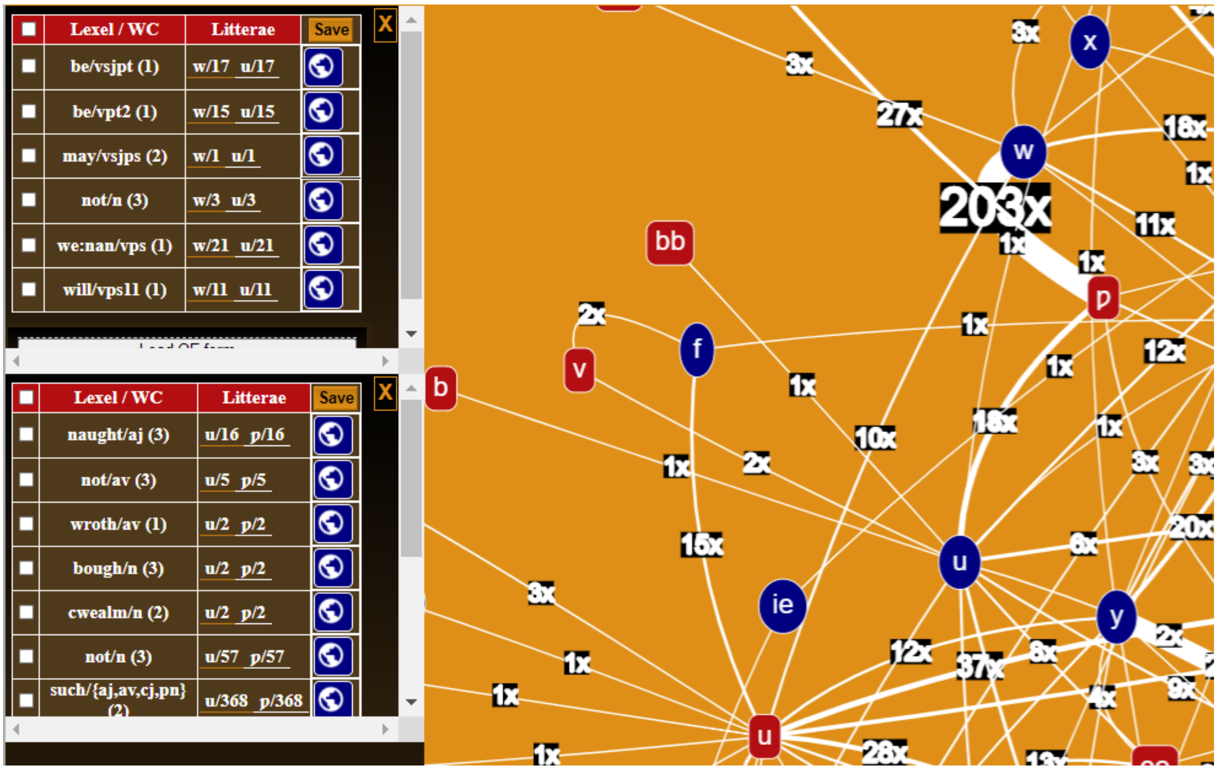
Table 6: Correspondences between litterae in texts #1100 and #3

One of the possible systematic ways of working with the network is to look at the individual correspondences between litterae, trying to decide whether the differences in spelling are best explained as pure orthographic variations or whether they have some phonological implications. The correspondences in the table above seem to be cases of orthographic variation suggesting that the scribe of #1100 systematically replaced the older *p* and *ʒ* with *w* and *h*, which is in accordance with the assumption that he was a translating scribe.

A number of other correspondences could be identified. The present sample analysis only focused on more correspondences involving *w*, namely the correspondence between *w – u* and *u – p* and correspondences between *f – v – u* in both texts.

4.3.5.1. *Correspondences between w – u – p*

The network shows 10 instances of *w* in #1100 for *u* in #3. At the same time, *wynn* in #3 does not always correspond to the usual *w* but *u* in #1100 (18 instances). In order to discover a possible pattern, item lists for the two types of correspondence were loaded into the interface.



□	Lexel / WC	Litterae	Save
□	be/vsjpt (1)	w/17 u/17	🌐
□	be/vpt2 (1)	w/15 u/15	🌐
□	may/vsjps (2)	w/1 u/1	🌐
□	not/n (3)	w/3 u/3	🌐
□	we:nan/vps (1)	w/21 u/21	🌐
□	will/vps11 (1)	w/11 u/11	🌐

□	Lexel / WC	Litterae	Save
□	naught/aj (3)	u/16 p/16	🌐
□	not/av (3)	u/5 p/5	🌐
□	wroth/av (1)	u/2 p/2	🌐
□	bough/n (3)	u/2 p/2	🌐
□	cwealm/n (2)	u/2 p/2	🌐
□	not/n (3)	u/57 p/57	🌐
□	such/{aj,av,cj,pn} (2)	u/368 p/368	🌐

Figure 44: Screenshot - correspondences between litterae w and u in #3 (red) and #1100 (blue)

NOT/N is the only item present on both lists but otherwise the correspondences are found in different items. The clearest tendency identifiable based on the lists is the (possible) replacement of initial *cp-* with *qu-* in #1100. The preference for *u* over *w* when replacing *p* in the initial position of WROTH/AV seems to be rather an exception to the rule. The likely explanation of the other forms is that *w* is used for consonantal quality and *u* for vocalic, e.g. *nawiht* as opposed to *nouht* (NOT/N).

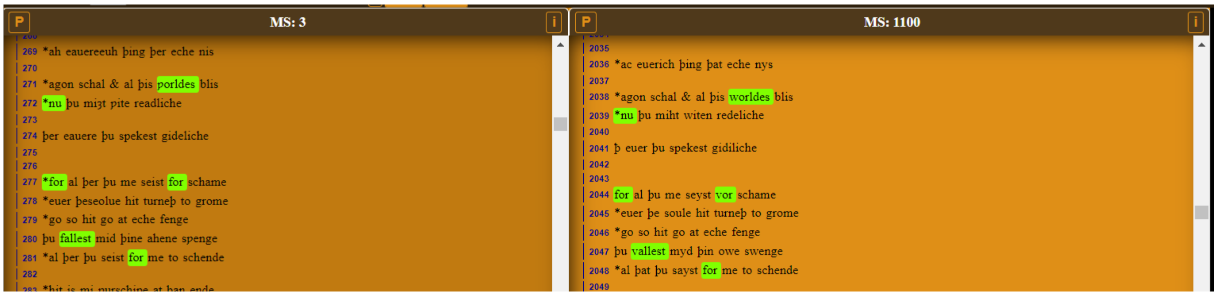
This part of the analysis showed how item lists can be used to search for patterns behind the correspondences visible in the network. The next part illustrates the limitations of data obtained through the network and explains how to complement the data from the network with manual analyses of texts.

4.3.5.2. *Correspondences between f – u – v*

It can further be seen from the network, that *w* in #1100 corresponds to *f* in #3 twice and at the same time, *f* in #3 frequently corresponds to *u* or *v* in #1100 (20 and 43 occurrences respectively). At the same time, *f* in #1100 appears in place of *u* or *v* in #3 (15 and 2 occurrences respectively), so the correspondence seems to work in both directions. Unlike the case of *w – p* the choice between *f – u* or *f – v* might reflect a sound change, specifically voicing (see CoNE IFV and MFV).

In order to check whether there is any pattern governing such replacements, lists of items for the individual cases of correspondence (*v – f* and *f – v* etc.) were again loaded into the interface but this time they were not immediately useful. A number of items (e.g. FOR, FARAN, FAIR) actually appeared on both lists, which means that #1100 sometimes has *f* in slots where #3 has *v* and vice versa. The lists are limited in that they do not tell anything about the distribution of the forms in the text. This means that if, for instance, the *v*-forms of FOR in #3 always corresponded to *v*-forms in #1100 and the same would be true of *f*-forms of FOR, the network and the item lists would be the same as if *f*-forms in #3 corresponded to *v*-forms in #1100 and vice versa.

A comparison of the actual instances of words with alternating {*f*, *v*, *u*} has to be performed manually with the tool. In the case of the present analysis, the texts were displayed side-by side in *text comparison* and all the forms of items ever spelled with *v* in #1100 were highlighted and compared with the forms in #3 directly corresponding to them (the items get highlighted when *v* in the inventory of #1100 is clicked).



MS: 3

269 *ah eauereeuh þing þer eche nis
270
271 *agon schal & al þis porldes blis
272 *nu þu miȝt pite readliche
273
274 þer eauere þu spekest gideliche
275
276
277 *for al þer þu me seist for schame
278 *euer þeseolue hit turneþ to grome
279 *go so hit go at eche fenge
280 þu fallest mid þine ahene spenge
281 *al þer þu seist for me to schende
282
283 *hit is mi purschipe at þan ende

MS: 1100

2035
2036 *ac euerich þing þat eche nys
2037
2038 *agon schal & al þis worldes blis
2039 *nu þu miht witen redeliche
2040
2041 þ euer þu spekest gidiliche
2042
2043
2044 for al þu me seyst vor schame
2045 *euer þe soule hit turneþ to grome
2046 *go so hit go at eche fenge
2047 þu vallest myd þin owe swenge
2048 *al þat þu sayst for me to schende
2049

Figure 45: Screenshot - examining instances of *f* - *v* in texts #1100 and #3

The comparison was limited to lines 1-30 of text #3 and the corresponding lines in #1100. The most common pattern was *f* in both texts (19 instances) followed by *f* in #3 and *v* in #1100 (15 instances). The latter pattern grew more common towards the end of the analysed sample. These results suggest possible replacement of *f* with *v* by the Jesus scribe (#1100), which is however far from universal.

This micro analysis illustrated how network visualisation can be used in combination with *text comparison*. Note that for more complex analyses, it might be worthwhile to highlight multiple groups of items in different colours and compare the distribution of all of them at the same time. Alternatively, the items can be highlighted using item lists rather than search in the text for a selected littera.

4.3.5.3. A note on text comparison

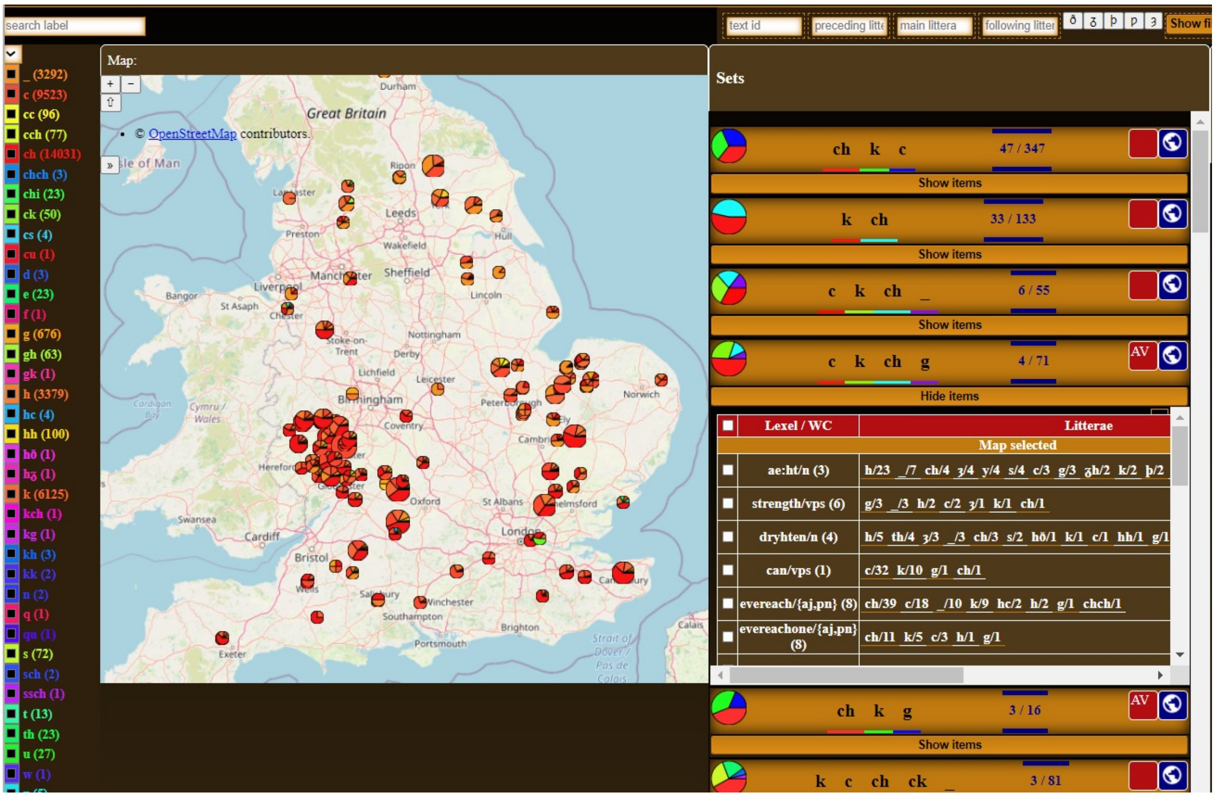
The data used to construct the network is in fact structurally similar to a scribal lexicon. LSSs in scribal lexicons are often based on historical sound values. For instance, we could postulate a LSS [f] <=> {‘f’, ‘v’, ‘u’} in the initial position for #1100 after examining all items with historical *f* in the initial position. Such a LSS would be very close to the correspondences between *f* in #3 and {*f*, *v*, *u*} in #1100, which could be described using a similar notation. The only difference is that the *f* (and the underlying item list) serving as reference for the correspondence was taken from text #3 and historical sound is not taken into account. As for the distinction between initial position and other positions, the network feature could in theory filter the correspondences by position but such a function is not (yet) available in the tool.

If a text with relatively regular spelling, such as *The Ormulum* was selected as the reference text, the LSSs should significantly overlap with LSSs based on historical sounds The obvious disadvantage is that the data would be restricted to items present in both of the compared texts. The data can be of course retrieved also in tabular format, although this has not yet been implemented in the tool. See Appendix 7.13 for a table showing correspondences between *The Ormulum* and Cambridge, Trinity College B.14.39, scribe A (#246).

4.3.6. Mapping

The next micro analysis focuses primarily on the mapping tool. The point of departure for the analysis is the set {*ch*, *k*, *(c)*}, which is associated with the well-known velar palatalization, (VP) described in CoNE. The general assumption is that OE [k] became [t͡ʃ] in palatal environments, i.e. in the vicinity of front vowels (CoNE, VP). When interpreting spellings in the manuscripts, *k* spellings are usually believed to represent [k], *ch* spellings [t͡ʃ] and *c* remains ambiguous.

The simplest way to generate a map is to start from a set of litterae. In the case of palatalization, *k* and *ch* are the obvious choice. This approach usually produces a map which is not directly usable and needs to be refined. The map for {*k*, *ch*} is shown in the picture below:



☐	Lexel / WC	Litterae
		Map selected
☐	ae:ht/n (3)	h/23 _/7 ch/4 ȝ/4 y/4 s/4 c/3 g/3 ʒh/2 k/2 þ/2
☐	strength/vps (6)	g/3 _/3 h/2 c/2 ȝ/1 k/1 ch/1
☐	dryhten/n (4)	h/5 th/4 ȝ/3 _/3 ch/3 s/2 hð/1 k/1 c/1 hh/1 g/1
☐	can/vps (1)	c/32 k/10 g/1 ch/1
☐	evereach/{aj,pn} (8)	ch/39 c/18 _/10 k/9 hc/2 h/2 g/1 chch/1
☐	evereachone/{aj,pn} (8)	ch/11 k/5 c/3 h/1 g/1

Figure 46: Screenshot - map for {k,ch}

When mapping a simple set, the tool automatically includes spelling variants alternating with the input litterae, in this case "*k*, *ch*". As a result, a large number of (minor) spelling variants can be added, which is obvious from map legend on the left side of the screen. As the colours are assigned randomly, the most frequent litteare can be displayed in similar colours and the map becomes difficult to read, which is exactly what happened with *c*, *k* and *ch* on the map above. The next section explains several possibilities of refining the data in the map and generating more readable maps.

4.3.6.1. Changing map legend colours

Map readability can be increased by picking contrasting colours for the frequent litterae. The picture below shows the map with *ch* displayed in bright green, *k* in turquoise and *c* in white.

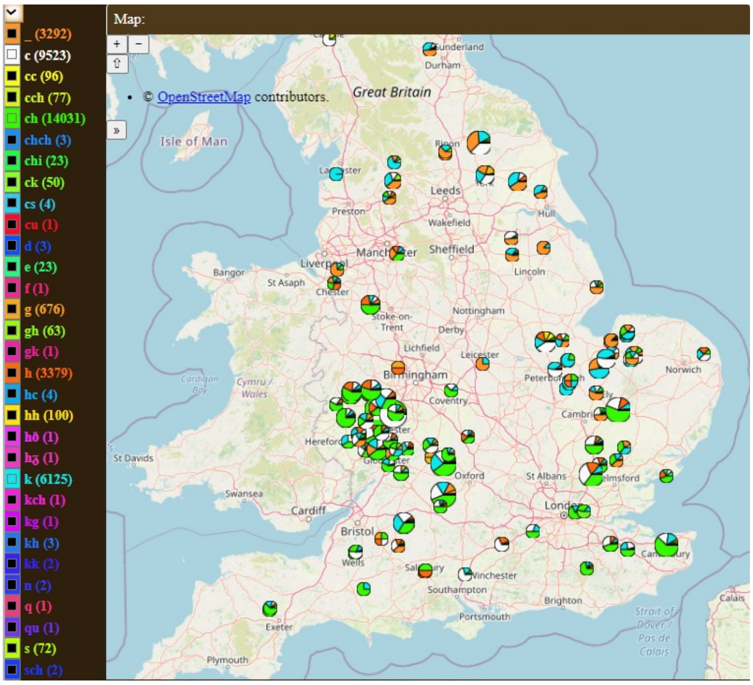


Figure 47: Screenshot - map for {*ch*, *c*} wth modified colours

This simple manipulation was enough to show the expected tendency of *ch* appearing in the south and (to a lesser extent) *k* in the north. Variants with *c* are spread across a large territory and display no clear regional tendency.

4.3.6.2. Map strict set

The original map was created with the basic mapping function "Map set", which automatically identifies and adds alternative spellings. There is also another function called "Map strict set", which maps only slots with the input litterae. For instance, MUCH/N (3) or CHAIN/N (1) would be selected for mapping with this functioin, but THINK/V-IMP (4) would be excluded because *k* and *ch* in this slot alternate also with *h*, *g* and *c*. The picture below shows the map for the strict set {*c*, *k*, *ch*}:

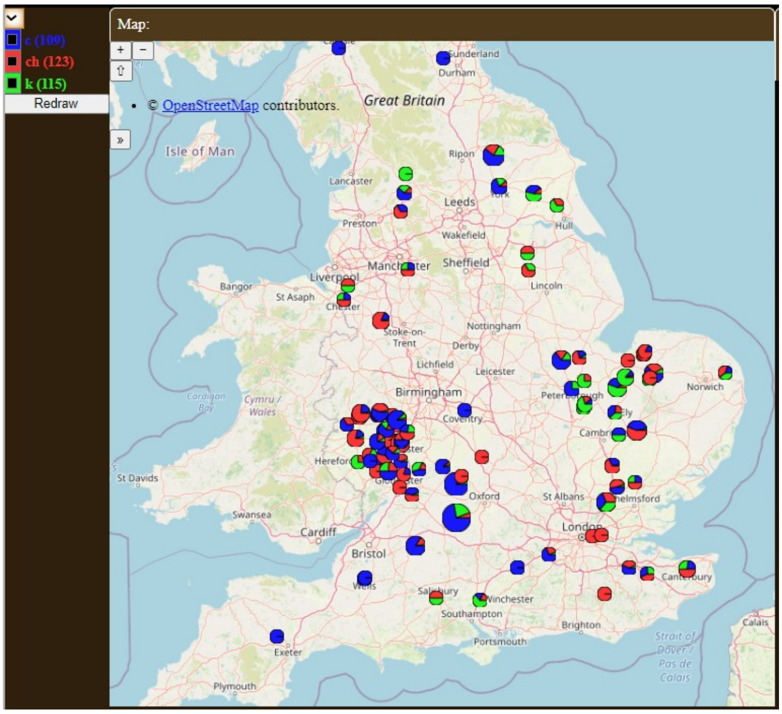


Figure 48: Screenshot - map for {*ch*, *c*, *k*} as a strict set

Compared with the original map, the map for the strict set looks more tidy at first sight. There appears to be a conspicuous concentration of *k* spellings in the North East Midlands and there are rather more *c* spellings in the SWM and the Southwest compared with the rest of the territory, which is probably due to the relatively more conservative character of spelling in the SWM. Still, the tendencies are not markedly clearer. In particular, the group of "green" texts in the NEM area is surrounded by several "red" texts, which does *not* confirm the expected tendency for *k* to appear in the North. However. what the map does not show is to what extent the distribution reflects regional as opposed to temporal tendencies, i.e. whether the apparent divergence from the expected regional pattern might be explained by different dating of the texts. This issue is addressed in the next section.

4.3.6.3. Map sequence

All maps except filtered maps can be transformed into a sequence of maps in the tool. The sequence generates separate maps for different time periods as indicated in LAEME metadata. By default, a separate map is generated for each half-century, which means that there are usually three maps, but it is possible to generate as many as seven (one for each quarter century). The default sequence generated from the map above is shown below:

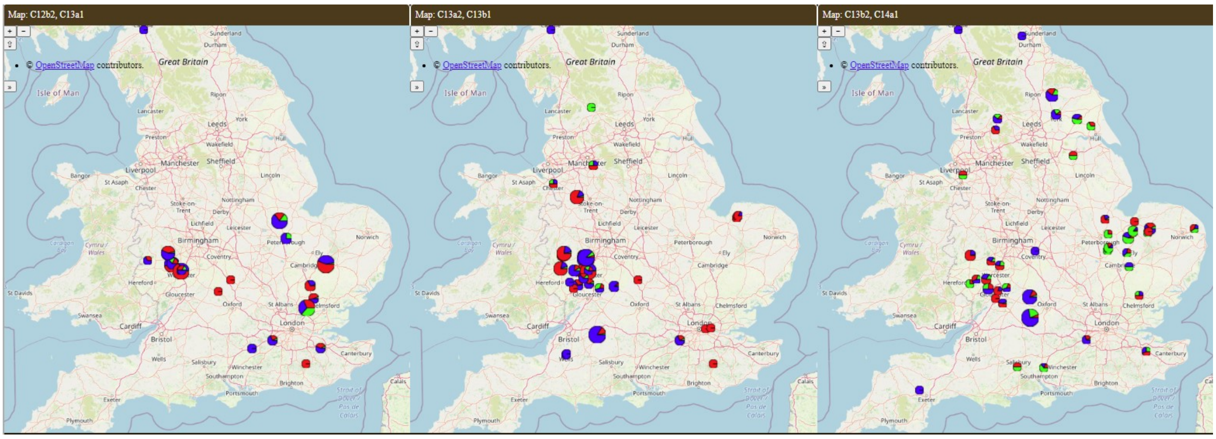
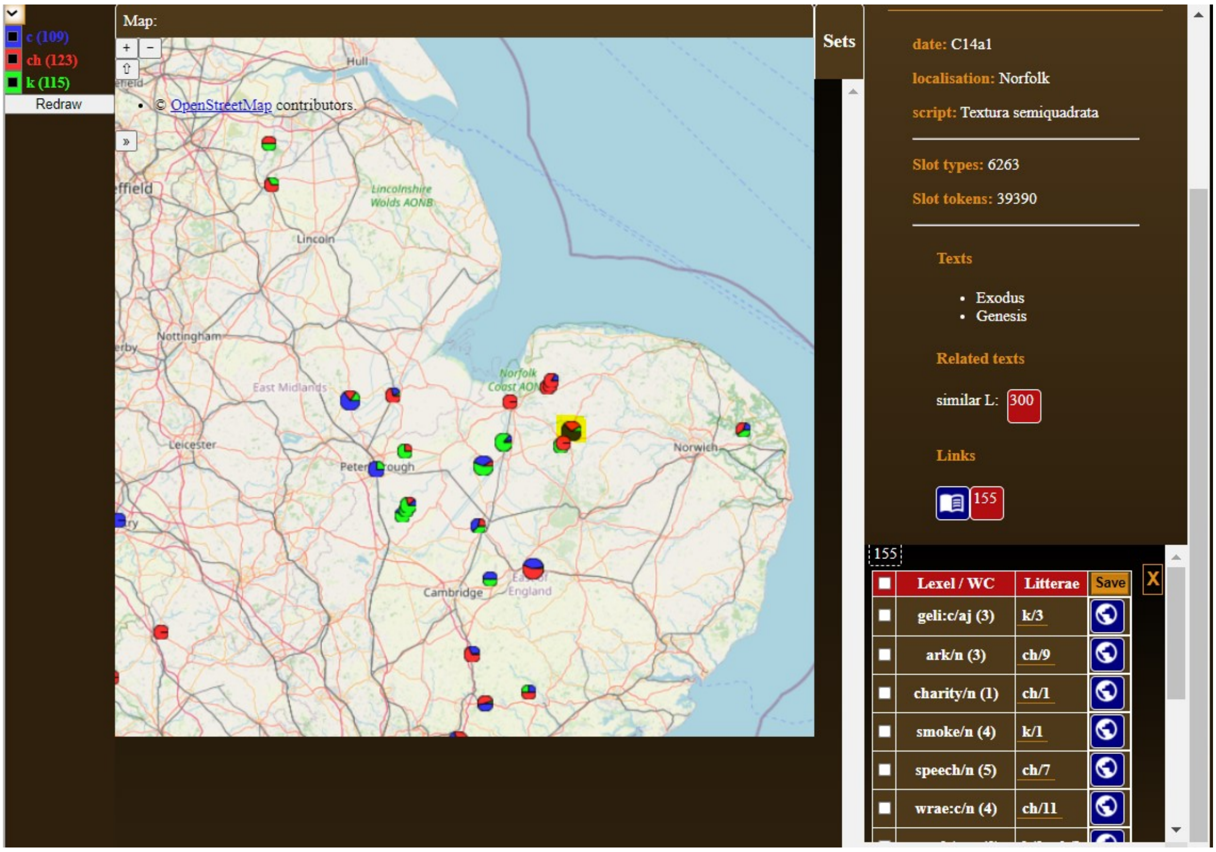


Figure 49: Screenshot - map sequence for the set {*ch*, *k* ,*c*}

Among other things, the sequence clearly shows the lack of northern texts from the early periods, mentioned in the description of LAEME, which complicates interpretation. As far as the aforementioned group of the North East Midland texts is concerned, they belong to the same time period, which means that the differences between the texts could not be accounted for in terms of different dating and another explanation had to be found. The next step in the analysis was to look at the actual items used to produce the maps.

4.3.6.4. Isolating items

The main limitation of set-based maps is that the pie charts for the individual texts can be in fact based on completely different items. The items from a particular text can be easily listed by clicking at one of the charts. The picture below shows the map along with the list of items for text #155 (Cambridge, Corpus Christi College 444 containing the copy of *Exodus* and *Genesis*). The corresponding chart on the map is marked with a yellow square and the list of items is found in the bottom right corner.



date: C14a1

localisation: Norfolk

script: Textura semiquadrata

Slot types: 6263

Slot tokens: 39390

Texts

- Exodus
- Genesis

Related texts

similar L: 300

Links

155

155

☐	Lexel / WC	Litterae	Save
☐	geli:c/aj (3)	k/3	
☐	ark/n (3)	ch/9	
☐	charity/n (1)	ch/1	
☐	smoke/n (4)	k/1	
☐	speech/n (5)	ch/7	
☐	wrae:c/n (4)	ch/11	

Figure 50: Screenshot - map with text data

The blue icons displayed with each item can be used to generate maps for the individual slots and check whether the distribution of spellings for the given item preserves or breaks the expected north-south divide. Two of such maps are given for illustration here – SPEECH/N (4) and MUCH/VPS (3):

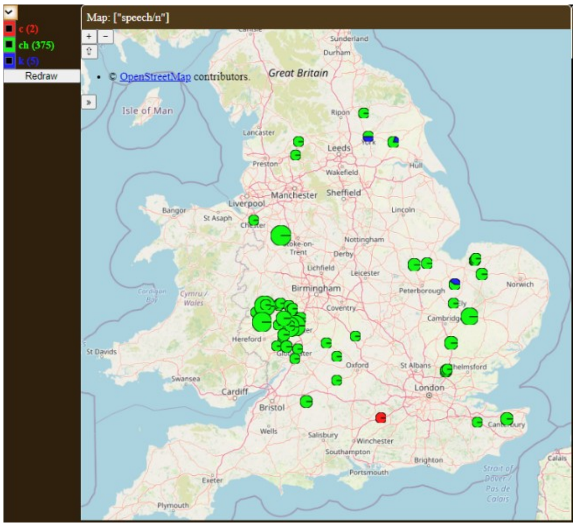


Figure 51: Screenshot - item map for SPEECH/N (4)

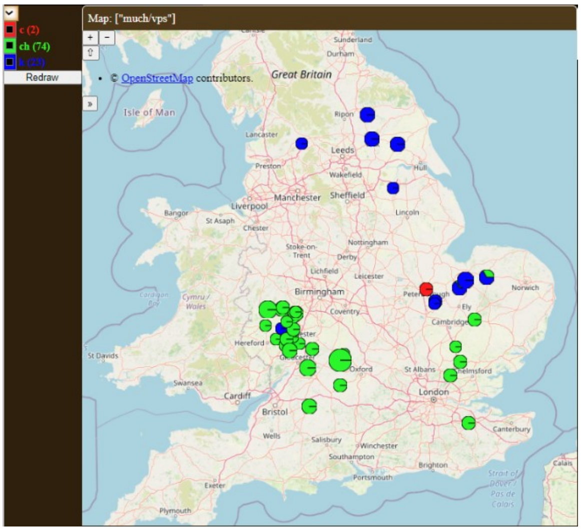


Figure 52: Screenshot - slot map for MUCH (3).

The maps show marked differences between the two items. While *k* spellings in SPEECH/N are highly exceptional and restricted to a few predominantly northern texts, the divide between *k* and *ch*-spellings in MUCH is almost flawless. A far as text #155 is concerned, the variants of speech with *ch* are in accordance with the distribution of forms. The *ch* in MUCH appearing alongside *k* seems to be slightly out of the *ch* area, which strongly suggests constrained selection (see subchapter 2.2.3.4) on the part of the scribe. A much more marked disruption of the regional pattern is found in text #246 (note the blue chart in the SWM area), but this will not be inquired into here.

4.3.6.5. Grouping items

While the maps based on sets can only provide a very rough picture which needs to be refined, the drawback of the maps based on individual slots (items) is that the coverage can be very limited. Consider for instance the following map for STARK/AJ (5):

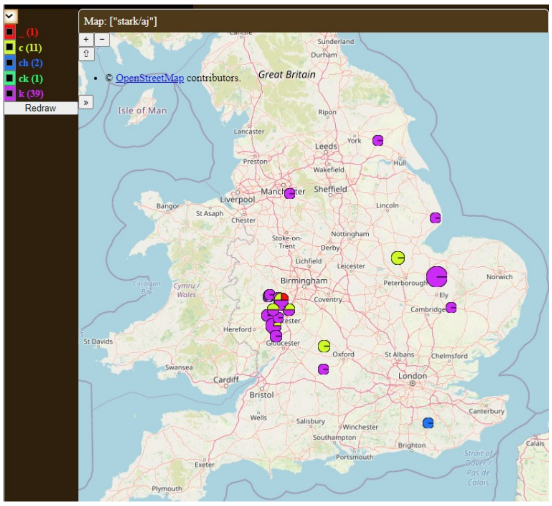
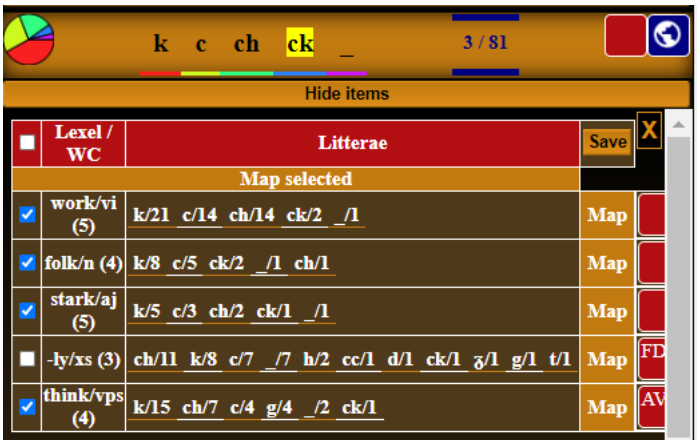


Figure 53: Screenshot - item map for STARK/AJ (5)

The “middle way” proposed to combine the benefits of both approaches is to group items manually. Any locally stored item list can be used for this purpose and item lists displayed within the mapping tool can be filtered and used directly. For instance, the set comprising STARK/AJ (5) was actually displayed with the first map for {*k*, *ch*} and the litterae in the set were {*k*, *c*, *ck*, *ch,* _}. This set as associated with five items shown in the picture below:



k c ch ck _ 3 / 81

Hide items

☐	Lexel / WC	Litterae	Save	X
		Map selected		
☑	work/vi (5)	k/21 c/14 ch/14 ck/2 _/1	Map	
☑	folk/n (4)	k/8 c/5 ck/2 _/1 ch/1	Map	
☑	stark/aj (5)	k/5 c/3 ch/2 ck/1 _/1	Map	
☐	-ly/xs (3)	ch/11 k/8 c/7 _/7 h/2 cc/1 d/1 ck/1 ʒ/1 g/1 t/1	Map	FD
☑	think/vps (4)	k/15 ch/7 c/4 g/4 _/2 ck/1	Map	AV

Figure 54: Screenshot - mapping items from a list

Four of the five items were selected, *-ly* was excluded because it does not represent a root morpheme and has a markedly higher frequency than the other items, which might distort the picture. The button "Map selected" at the top of the lists generates the following map:

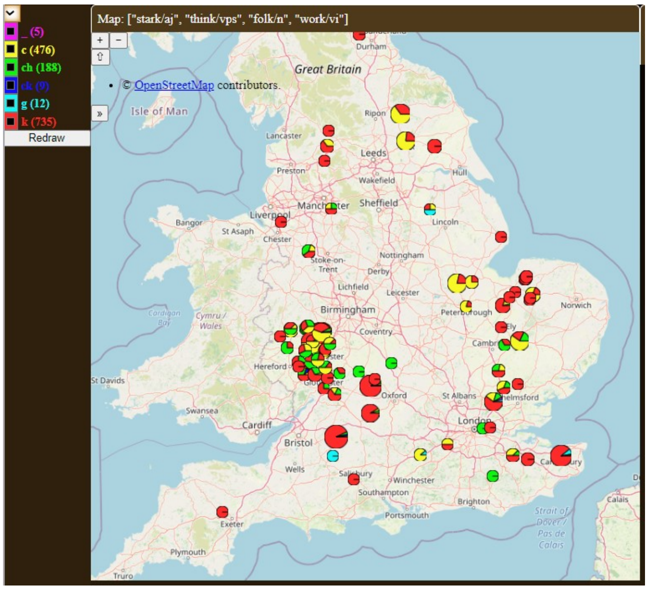


Figure 55: Screenshot - item list map: stark, think, folk, work

The map shows the prevalence of *k* in the examined slots and also a certain weak tendency of *ch* to appear in the northern part of the SWM. The coverage is clearly better than in the case of STARK/AJ (5) above.

4.3.6.6. *Applying filters*

The last strategy of refining the data is to filter the items. As palatalisation is a change conditioned by context, it makes sense, for instance, to filter the items by the segment following *ch*. The map below was generated from items where *ch, k* is followed by *e*, which should in theory trigger palatalisation:

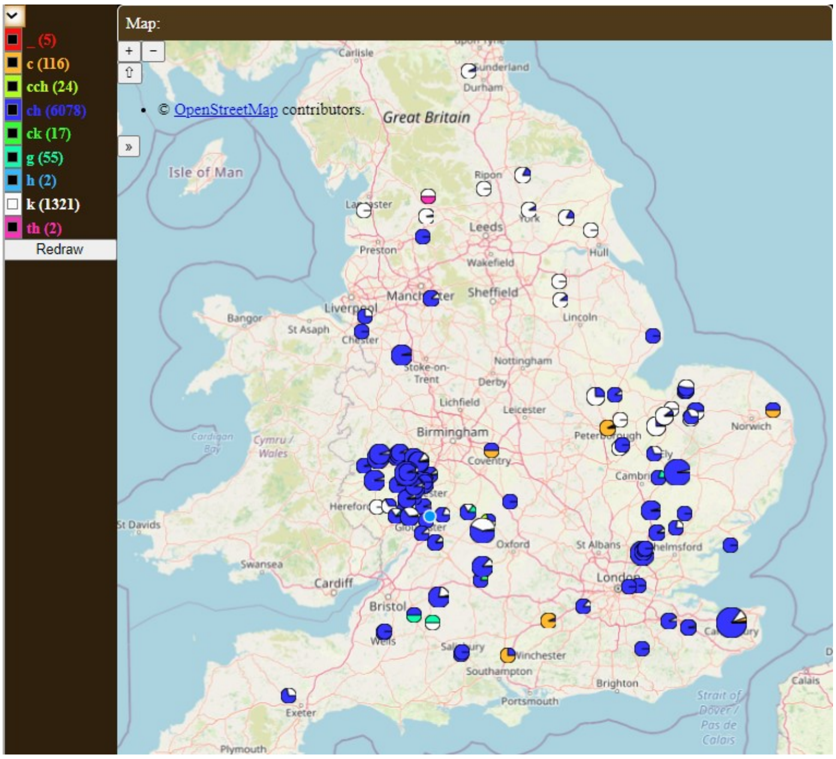


Figure 56: Screenshot - {ch, k} before <e>

The North-South divide is again discernible on this map. A few texts in the SWM area seem to stand out a little, because of relatively higher incidence of *k* (white). The map also shows which texts have mostly *c* (yellow), but the amount of data is very small judging by the size of the pie charts.

4.3.6.7. A note on mapping

The mapping tool is user friendly, however, the data have to be interpreted with caution. Any attempts to improve coverage by displaying data for multiple items should be ideally based on a manually checked item list because the items behind maps from sets may not always be comparable. Still, *set* maps can be useful in that they can suggest what the focus on in the analysis.

The main advantage of pie charts is that all the variants are displayed in one map, which makes it possible to consider their potential sound values in the context of the neighbouring variants. Variants assumed to represent the same sound can be assigned the same colour. The fact that the charts serve as links to texts (similarly to LALME tool) enhances the tool's capability to refer the user to potentially useful pieces of data.

The sequencing function (separate maps for different periods) exploits the concept of *spacetime* as described by Williamson (2004) and discussed in section 2.2.3.5. It works well in principle but its usefulness largely depends on the amount of available data which can be plotted on the maps.

4.3.7. The use of *x*

This micro analysis partly follows the path proposed for the study of a specific littera at the end of the methodological chapter (3.8.1), working with the example of *x*. This particular littera was chosen because its use is relatively restricted, so the analysis can cover a reasonable portion of the available data. At the same time, the use of *x* seems to be variable enough to deserve some comment. The analysis involved a number of possible searches. The first part of the analysis was performed using basic database searches available under *search DB* and the second part dealt with the use of *x* in selected text using *text profile* and *text comparison*, including the possibility to store custom lists of items and use them in searches.

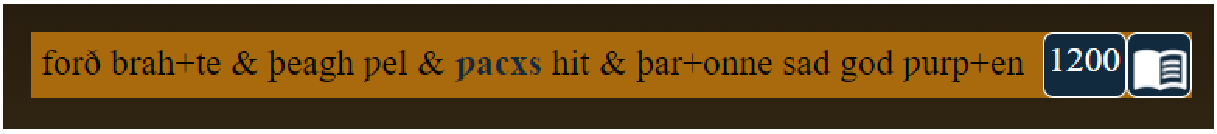
First of all, all the polygraphs containing *x* were listed using the function "Polygraphs". The search returned the following table:

Littera	**types**	**tokens**
cx	1	1
cxs	1	1
xi	2	2
xs	5	5
x	55	133

Table 7: Polygraphs containing *x*

Predictably, there are few polygraphs with *x* and three of the polygraphs occur only once in the whole coprus. The next step was to identify the items and texts in which the polygraphs appear. The item lists showed that the single occurrences of *cx* and *cxs* appear in ASH/N and WAX/N, respectively and *xi* appears in FIGHT/VSPP (*fexit*) and ASK/VPSP (*axi+ende*). The latter seems to be a case of inconsistent morphological analysis of the form, which is elsewhere found a*s ax+idende*. Therefore, there is no reason to expect a connection between the two forms and they will not be discussed further here.

The forms of ASH/N and WAX/N were displayed as KWIC, which enabled the identification of texts in which they appear.



forð brah+te & þeagh ƿel & **ƿacxs** hit & þar+onne sad god ƿurp+en 1200

Figure 57: Screenshot - KWIC for the form *pacxs* (WAX/N)

Cx and *cxs* were found in texts #1300 and #1200, respectively, both of which represent MS Cambridge, Trinity College B.14.52 containing *Trinity Homilies* copied by two different scribes. Unfortunately, WAX and ASH were not found in the related SWM copy of the *Homilies* (*Lambeth Homilies*), nor the *Poema Morale* copied by the same scribe as text #1200, which would be good candidates for comparison.

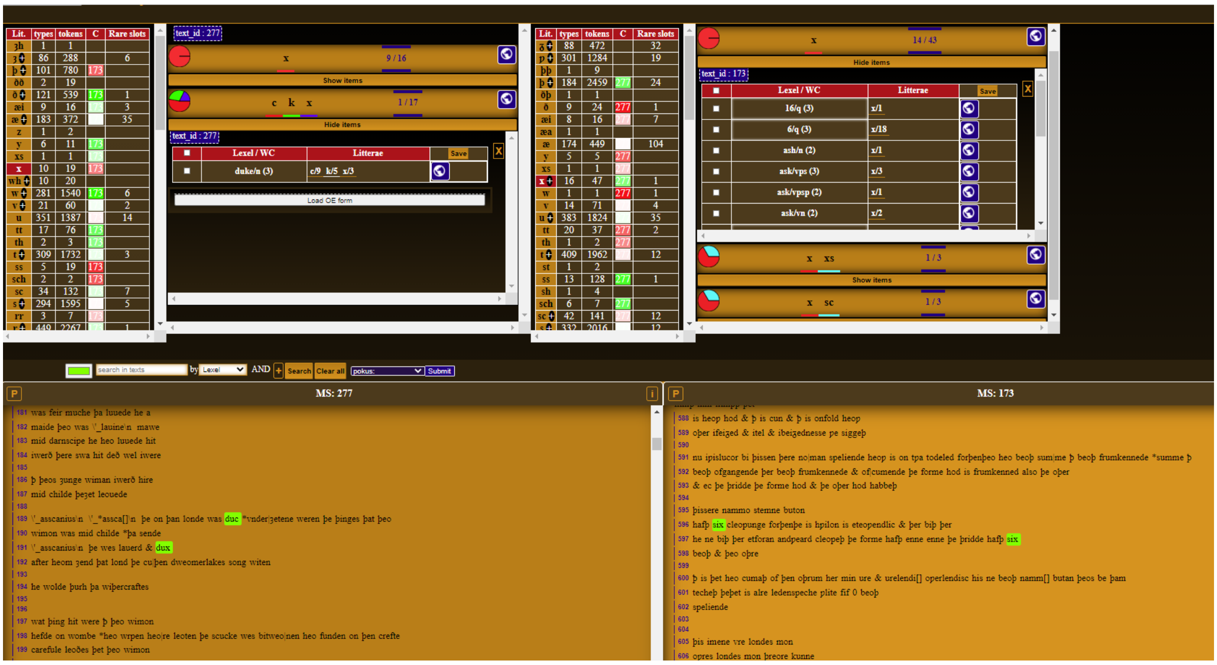
The items containing the five instances of *xs* were listed using the query for "Items" with *xs* as input:

Lexel/WC	Litterae	texts
next/{aj,av,pr} (3)	x/10, s/3, **hs/3**, xs/1, **cs/1**	#295
flesh/n (4)	ss/9, s/7, sch/7, sc/7, _/6, sh/4, ch/4, **chs/3**, c/2, schs/2, **hs/2**, ssc/1, cs/1, **hc/1**, shs/1, xs/1, ssch/1	#295
hnesce/aj (3)	sch/2, x/1, ssh/1, sh/1, s/1, ch/1, xs/1, ss/1	#64
ask/vn (2)	sk/2, sc/2, xs/1, x/1, cs/1	#173
high/ajs (4)	x/4, s/3, h/2, g/2, hs/2, cs/1, hȝ/1, ȝ/1, ks/1, xs/1, _/1	#277, #1100

Table 8: Items with <xs>

A closer look at the litterae corresponding to *xs* suggests that *xs* may be an alternative of *chs*, *hs* and *x* could therefore be used to represent [x] in certain words in the concerned text languages. The IDs of texts containing the forms were added manually to the table. As the digraph is very rare, the list of texts ("Browse MSS") was consulted to identify possible links between the texts. The only link between the concerned texts mentioned in LAEME catalogue is "similar language" for texts #173 (Worcester Cathedral, Chapter Library F 174, *Ælfric's Grammar and Glossary*) and #277 (London, British Library, Cotton Caligula A.ix, part 1, *Laȝamon A*).

A thorough analysis of the use of *x* should ideally discuss the use of the littera in all the text languages in which it appears. As the primary purpose of the present subchapter is to demonstrate the application of the tool, only a sample analysis of the related texts #173 and #277 is included. The feature *text comparison* was chosen as the main tool for this task.



Lit.	types	tokens	C	Rare slots
ʒh	1	1		
ʒ +	86	288		6
þ +	101	780	173	
ðð	2	19		
ð +	121	539	173	1
æi	9	16	173	3
æ +	183	372		35
z	1	2		
y	6	11	173	
xs	1	1	173	
x	10	19	173	
wh +	10	20		
w +	281	1540	173	6
v +	21	60		2
u	351	1387	173	14
tt	17	76	173	
th	2	3	173	
t +	309	1732	173	3
ss	5	19	173	
sch	2	2	173	
sc	34	132	173	7
s +	294	1595		5
rr	3	7	173	
r +	449	2267	173	1

text_id : 277

x — 9 / 16

Show items

c k x — 1 / 17

Hide items

text_id : 277

□	Lexel / WC	Litterae	Save
□	duke/n (3)	c/9 k/5 x/3	

Load OE form

Lit.	types	tokens	C	Rare slots
ʒ +	88	472		32
p +	301	1284		19
þþ	1	9		
þ +	184	2459	277	24
ðþ	1	1		
ð	9	24	277	1
æi	8	16	277	7
æa	1	1		
æ	174	449		104
y	5	5	277	
xs	1	1	277	
x +	16	47	277	1
w	1	1	277	1
v	14	71		4
u +	383	1824		35
tt	20	37	277	2
th	1	2	277	
t +	409	1962	277	12
st	1	2		
ss	13	128	277	1
sh	1	4		
sch	6	7	277	
sc +	42	141	277	12
s +	332	2016		12

x — 14 / 43

Hide items

text_id : 173

□	Lexel / WC	Litterae	Save
□	16/q (3)	x/1	
□	6/q (3)	x/18	
□	ash/n (2)	x/1	
□	ask/vps (3)	x/3	
□	ask/vpsp (2)	x/1	
□	ask/vn (2)	x/2	

x xs — 1 / 3

Show items

x sc — 1 / 3

search in texts by Lexel AND + Search Clear all pokus: Submit

MS: 277

181 was feir muche þa luuede he a
182 maide þeo was \'_lauine\n mawe
183 mid darnscipe he heo luuede hit
184 iwerð þere swa hit deð wel iwere
185
186 þ þeos ʒunge wiman iwerð hire
187 mid childe þeʒet leouede
188
189 \'_asscanius\n \'_*assca[]\n þe on þan londe was duc *vnder|ʒetene weren þe þinges þat þeo
190 wimon was mid childe *þa sende
191 \'_asscanius\n þe wes lauerd & dux
192 after heom ʒend þat lond þe cu|þen dweomerlakes song witen
193
194 he wolde þurh þa wiþercraftes
195
196
197 wat þing hit were þ þeo wimon
198 hefde on wombe *heo wrpen heo|re leoten þe scucke wes bitweo|nen heo funden on þen crefte
199 carefule leoðes þet þeo wimon

MS: 173

588 is heop hod & þ is cun & þ is onfold heop
589 oþer ifeiȝed & itel & ibeiȝednesse pe siggeþ
590
591 nu ipislucor bi þissen þere no|man speliende heop is on tpa todeled forþenþeo heo beoþ sum|me þ beoþ frumkennede *summe þ
592 beoþ ofgangende þer beoþ frumkennede & of|cumende þe forme hod is frumkenned also þe oþer
593 & ec þe þridde þe forme hod & þe oþer hod habbeþ
594
595 þissere nammo stemne buton
596 hafþ six cleopunge forþenþe is hpilon is eteopendlic & þer biþ þer
597 he ne biþ þer etforan andpeard cleopeþ þe forme hafþ enne enne þe þridde hafþ six
598 beoþ & þeo oþre
599
600 þ is þet heo cumaþ of þen oþrum her min ure & urelendi[] operlendisc his ne beoþ namm[] butan þeos be þam
601 techeþ þeþet is alre ledenspeche plite fif 0 beoþ
602 speliende
603
604
605 þis imene vre londes mon
606 opres londes mon þreore kunne

Table 9: Screenshot - comparison of #173 and #277

The picture shows the comparison screen after *x* was clicked to display its possible alternatives and highlight relevant items in the texts. Also, item lists for the set {*x*, *k*, *c*} in #277 and {*x*} in #173 were loaded into the interface. *Text comparison* should mainly provide basic frequency data for the individual litterae, an overview of sets to be accounted for and quick access to the associated lists of items.

The comparison of relative frequencies suggests that *x* is slightly more frequent in #173. The list of sets further shows that there are 9 items with *x* in #277 and 14 in #173 and both texts have one slot in which *x* alternates with another littera (DUKE/N in #277 (3) and FISH/VPS (3) in #173). Actual forms of each item were quickly checked to confirm the assumption that *x* in DUKE/N in #277 (3) and FISH/VPS (3) are likely to be a less common spelling compared to *x* in the other items, which were mostly words spelled with *x* in PDE (FOX/N, WAX/N etc.) and in which *x* is clearly the dominant, if not the only spelling found in LAEME.

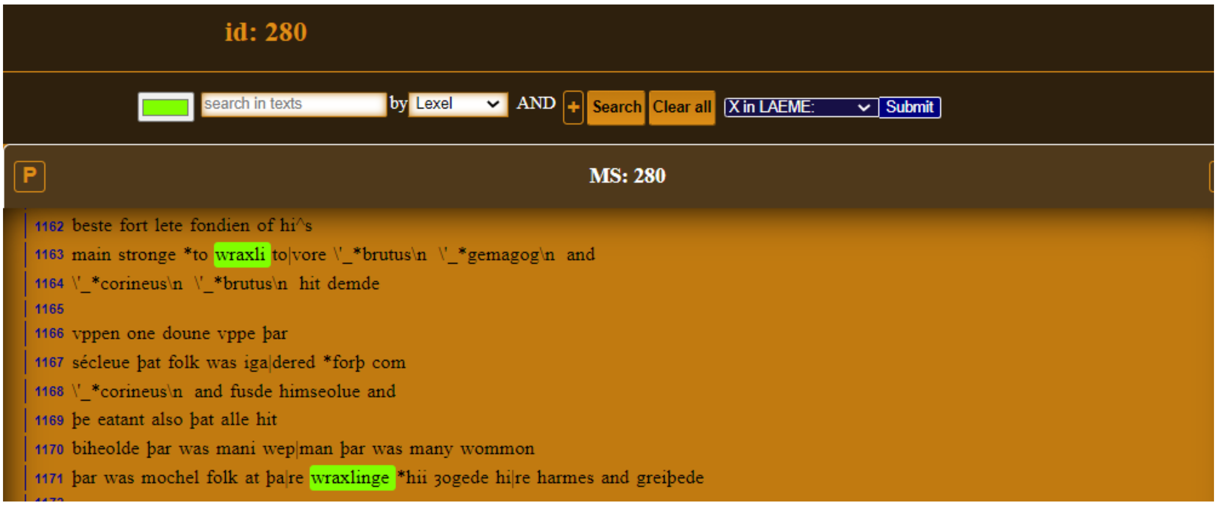
This part of the analysis revealed that the form *fixie* (FISH/VPS) is confined to text #173 alone and *dux* appears in two texts only, #277 and #280 (London, British Library, Cotton Otho C xiii), which is the text of *Laȝamon B*.

The use of the digraph *xs* in the two texts was compared in a similar way. The single occurrence *hexte* (HIGH/AJS) in #277 alternates with *hs*. The single occurrence of *axsunge* (ASK/VN) alternates with simple *axunge* and other forms of ASK spelled with *x*.

Such co-occurrence of an occasional variant with more common ones is a typical consequence of exemplar influence. In the case of #277 this explanation could be further supported by the fact that *dux* (DUKE/N) is also found in another copy of the same text (#280). On the other hand, *dux* is the standard Latin spelling (not an idiosyncratic form) so its co-occurrence in the two texts may as well be a mere coincidence. As there are related texts available for both #277 and #173 they were searched for further evidence supporting or contradicting the possibility of exemplar influence. Also, it was considered worthwhile to check, whether the two texts under examination and potentially also the related texts use *x* in all words where it can be expected. The fastest way to answer such questions using the tool is to store a list of items and subsequently use the list to search for the items in the texts. Considering the needs of the present analysis, three separate lists labelled “X in 277”, “X in 173” and “X in LAEME” were created. The first two lists were created straight from the comparison screen. All the items under the set {*x*} were selected, saved and assigned the desired label.

Any list based on the whole LAEME corpus can be obtained with the function “Search Items” (the input is *x* in this case) available on the screen *database search*. As there was no need to include the items already found on the lists for #173 and #277, only the remaining items were selected and stored.

Stored item lists can be accessed directly from *text profile*. The following picture shows a part of #280 within the *text profile* screen after the search for relevant items was performed. The list “X in LAEME” selected from the dropdown menu can be seen in the top right corner.



```
1162 beste fort lete fondien of hi^s
1163 main stronge *to wraxli to|vore \'_*brutus\n  \'_*gemagog\n  and
1164 \'_*corineus\n  \'_*brutus\n  hit demde
1165
1166 vppen one doune vppe þar
1167 sécleue þat folk was iga|dered *forþ com
1168 \'_*corineus\n  and fusde himseolue and
1169 þe eatant also þat alle hit
1170 biheolde þar was mani wep|man þar was many wommon
1171 þar was mochel folk at þa|re wraxlinge *hii ȝogede hi|re harmes and greiþede
```

Table 10: Screenshot - item list search in #280

The words *wraxli* and *wraxlinge* (WRESTLE) are highlighted because WRESTLE/VI and WRESTLE/VN appeared on the stored list. This procedure was applied to all of the examined texts and stored item lists. Also, the basic data for *x* in the related texts (frequency, sets, items) was checked analogically to the analysis of #277 and #173 described above.

The results for #171 and #172 (copied by the same scribe as #173) are not very useful because none of the items from the list are present in #171 and there are was only one in #172 namely WAX/N, spelled with *x*. A for #280, the searches were a little more fruitful. Besides the occasional forms *dux* of DUKE/N, the texts share the form *hexte* of HIGH/AJ. As the form appears at the exact same place in both texts, it is conceivable to assume exemplar influence for the two forms.

The final micro-study demonstrated how global statistical data can be combined with examination of individual instances of words in the manuscripts. The analysis could also proceed in the opposite direction. For example, any idiosyncratic use of a *littera* found in the manuscripts could prompt new searches of a more global character.

4.4. Discussion

This subchapter deals with three topics. The first part summarizes the main problems and limitations encountered during the construction of the tool. The second part presents a broad theoretical discussion of the *scribal lexicon* (see subchapter 2.2.3.3) in the context of the functions available in the tool. The discussion is an attempt to outline a path towards a model of *text language* which would more fully exploit the possibilities of the new database. The final part briefly comments on possible future upgrades of the tool.

4.4.1. Limitations, weak points and problems with tool construction

This subchapter comments on several problems which emerged during the processing of LAEME data. Most of the problems led to the exclusion of some forms from the database. Some of the issues were of technical nature, but others are indicative of the limits of the methodology.

The first problem concerned the conversion of LAEME files to tables in the relational database. While the identification of whole tags was straightforward, correct parsing of individual morphemes proved to be more difficult. The failure of the parsing script was either due to high complexity of the input (e.g. `$manifoldly/av_MON+I+FOLD+LICHE $-`

`ig/xs-aj-k_+I+ $-fold/xs-aj-k_+FOLD+ $-ly/xs-av_+LICHE\n`)[20] or slight inconsistencies in the LAEME data. For instance, `$fae:tels/nOd_`**`FETEL`**` $-els/xs-nOd_`**`+EL`**`\n` has a separate tag for the final *-el* but the morpheme boundary is not indicated in the form *fetel*. The parsing problem occurred with 534 tags out of the total ca. 650 000. Whole tags were stored in the table *tags* as usual but they have no corresponding rows in the table *morphemes* and as such could not be processed (see Appendix 7.4 for the complete list).

Minor inconsistencies of similar nature also produced extra empty slots at the end of some forms, which in fact correspond to the endings of other forms. For instance, the form *hyalde* of HOLD/VI stands alongside *heald+e*. Consequently the final *-e* in *hyalde* apparently corresponds to and empty slot in *heald+e*.

As not all of the data was checked manually, not all the errors introduced by the parsing script were fixed. The error rate calculated from a random sample of automatically processed forms is 3.5 %. The sample comprised 254 items with a total number of 2136 forms.[21]

4.4.1.1. Parsing

The discussion of problematic variants has already shown that the segmentation and alignment were not always straightforward. The original intention to make parsing maximally realistic and as interpretation neutral as possible sometimes clashed with the need not to overcomplicate the structure of the forms. Moreover, the correspondences between segments could not always be identified with a reasonable level of certainty, so some of the parsing choices were purely arbitrary.

A crucial point to bear in mind is that even the most "successful" and unproblematic alignments should be treated carefully because the likely value represented by each littera should be interpreted in the context of the whole form. For instance, it is conceivable that the sounds behind *h* in *ma**h**t* and *mai**h**t* (MAY) might have been different. An even more problematic case is *nauicht*/*nouht* (NAUGHT), where the sounds represented by *u's* appear to correspond to each other from the perspective of development but one of them is consonantal and the other

[20] The problem was that the script had to be able to handle forms wherein the morphemes with separate tags are nested, e.g. `$ateli:c/ajpl_ATE+LICH+E $-ly/xs-ajpl_+LICH`**`+E`**` $/plaj_`**`+E`**`\n`. As an unwanted consequence, it sometimes wrongly identified short elements like *i* as parts of another morpheme.

[21] The errors in parsing typically occur in a few forms in the whole group subsumed under one item. Taken together with the high diversity of forms, this means that 3.5 % of faulty *forms* translates to roughly 15 % of *items* the forms of which are not all parsed correctly.

one is vocalic. The difference between the two types becomes observable and can be mapped if we look at the chunk *-ui/u_-* instead of looking at a single slot.

Despite all of these imperfections, the extra dimension available thanks to segmentation has considerably improved the possibilities of data navigation and quantification. The parsing does not always have to be flawless to be able to show connections between possibly related forms in different texts or generate a useful map.

4.4.2. Theoretical and methodological observations

It has been pointed out that segmentation of the forms is inevitably arbitrary to a certain extent. Despite prior awareness of this fact, the parsing proved to be even less interpretation-neutral than expected. On the other hand, the difficulties connected with the attempt at segmentation of the forms can inspire deeper thinking about the limits of analysing spelling systems primarily in terms of correspondences between litterae and potestates.

The model of PSS ("littera *x* represents sounds [a}, [b}…[n]") is reminiscent of dictionary entries which usually simply give one or more definitions of a word. The meanings are not discrete of course and their range is determined by the range of contexts in which the word may appear. The list of definitions is an adequate model in that it essentially captures this range of meanings, still, speakers usually do not think about definitions when they use the language.

If we assume that written language as a system behaves analogically to spoken language at least in the case of scribes who rely primarily on their ears, it is reasonable to expect that litterae will be treated similarly as words in the spoken language. The "represented reality", i.e. sounds in the spoken language, is no doubt less complex, still the abstract phonological inventory definitely does not do justice to the acoustic differences between different realisation of the "same" sound.

While the notion of LSS, i.e. multiple representations of a single sound (e.g. *-st*, *-ct*, *-cst* for historical [xt]) seems relatively unproblematic and "plausible". Some of the different representations are likely to be inspired by the exemplar and/or the representations may in fact reflect very slight differences in pronunciation. The idea of different sounds represented by a single littera appears somewhat less natural and plausible.

Judging by the multiple alternations between different litterae, it would seem that the "image" of the potestas in the mind of the scribe is rather fuzzy but it does have a limited scope.

After all, the realisations of what would be considered a single phoneme are definitely not the same because articulation is influenced by context. Moreover, if the scribe really works by ear, he may need to perform the segmentation of the sound stream on his own and the result might be different from what we would expect. The different *potestates* of a given littera may in fact reflect differences in segmentation rather than different sounds associated with the littera.

For instance, if *ȝ* in the system of scribe D of MS Cambridge, Trinity College B.14.39 apparently stands for [h] in *ȝu* (HOW), [x] in *driȝten* (DRIHTEN, “lord”) and [w] in *roȝen* (TO ROW) (Laing & Lass, 2013: 2.3.2), it does not necessarily mean that the scribe associated the littera with sounds as different as the proposed PSS would suggest. It might be that he simply perceived little or no difference between the sounds descended from OE [h] and [x]. As for *roȝen*, if the represented sound was something like [roɦen] (Laing & Lass, 2009: 28), it would be perceptually close enough to [ro^{w}en] written as *rowen* in other texts.

Apparently, two sounds which are close enough perceptually to be represented by the same littera are described as two distinct potestates in a PSS. The case of *ȝ* suggests that if we renounce the idea that the spelling system needs to be interpreted in terms of pairing litterae and potestates, one of the effects is that the spelling system begins to look somewhat more consistent. Going back to the analogy with spoken language, we grasp the “meaning” as a unified whole rather than as a list of definitions.

4.4.2.1. The problem of modelling written language

A particularly acute problem with any models of language (spoken or written) is how to treat the “meaning” represented in language. Any attempt at capturing something like “units of meaning” or to distinguish between different “senses” can hardly avoid simplification. “Meaning” in the case of written language with predominantly phonemic level of representation mostly equals sounds. Therefore, the key question is how to include the sounds in the model, specifically in a model applicable to language with no attestations of the sounds.

One of the possible answers is to work with broad historical values and/or present day values, i.e. the “reality of sounds” is represented as a set of sounds characterised by their phonetic properties. The structuralists moreover tried to ascertain the phonemic status of the units represented in writing, but phonemic inventories constructed in this manner are abstractions somewhat removed from the surface variation of real life speech. As Laing & Lass (2013: 2.3.1.) pointed out, it seems that strict structuralist identification of phonemes was not involved in the construction of medieval spelling systems.

The idea of pairing the litterae with approximate values is no doubt justified by the impossibility to reconstruct the sounds with a higher level of precision and it probably is the best approach if the reconstruction itself is the ultimate goal. On the other hand, such a simplified template might not be the best option for analyses focused at better understanding of written language because it can partly predetermine our interpretation of the written sources and also make us disregard cases of variation, which may be insignificant in isolation but important in a wider context.

One of the useful strategies in avoiding such bias, proposed already by McIntosh (McIntosh et al., 1989) is to focus primarily on the full range of spelling variants and *relations* between them before interpreting the sounds (McIntosh et al., 1989: 24, see subchapter 2.2.3.1). The present thesis does not define a fully developed model of the written language but some of the functions of the tool support dynamic comparisons of variants which are independent of the predefined historical values. After all, historical values sometimes serve as reference points rather than an actual interpretation of sound.

Instead of starting from lists of items defined historically, it is possible to group slots by the actual litterae employed by a specific scribe. The list of slots (item lists) can be viewed as "territories" or ranges of sounds defined by their mutual relations rather than a fixed reference point. This does not mean that historical values should be disregarded, the difference between the approaches is that the initial grouping reflects the similarities between the attested variants in a given text rather than an "external" point of reference. The diagram below shows the "territories" of *ch*, *k, c* and *cch* in text #1300 (Cambridge, Trinity College B.14.52, *Trinity Homilies*, scribe B), only selected items are included:

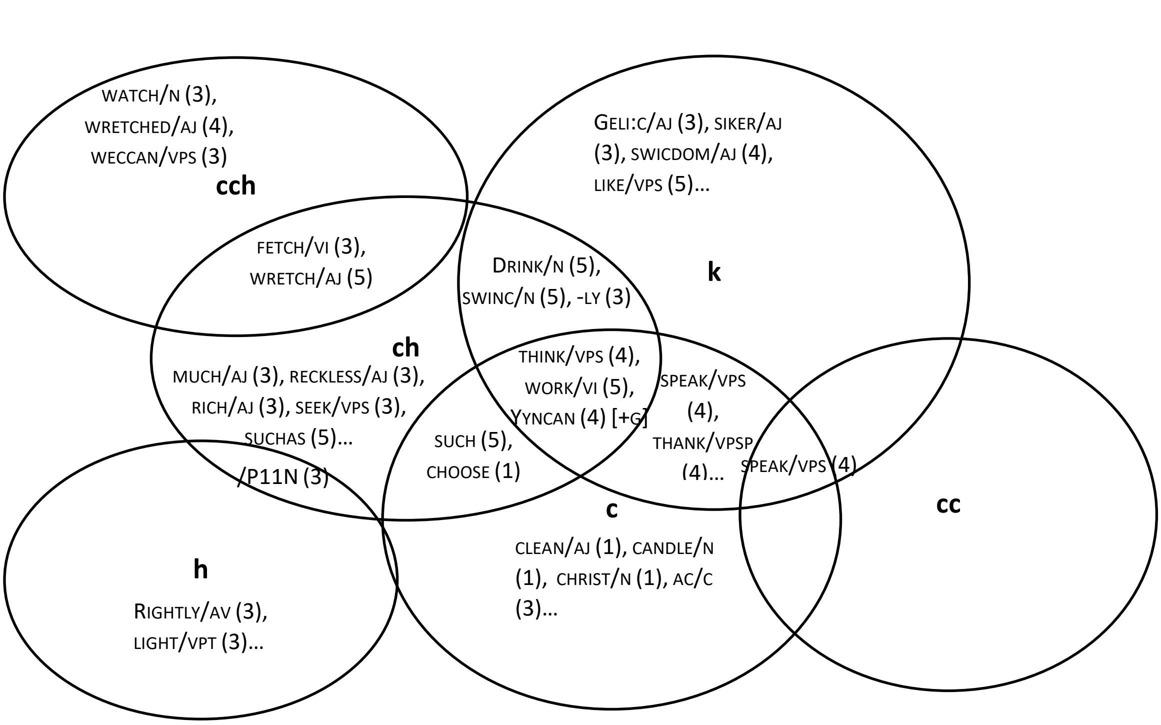


Figure 58: Visualisation of the overlapping uses of ch, k, c and cch in text #1300

The diagram shows which slots are occupied by each of the litterae and in which two or more litterae alternate. It can be used either as a basis for interpretation of sounds, including the identification of likely exemplar forms or as a material for comparison with a different text language, as long as the text shares a sufficient number of lexels. Two more diagrams are included for demonstration. The first is from #1200 (*Trinity Homilies*, scribe A):

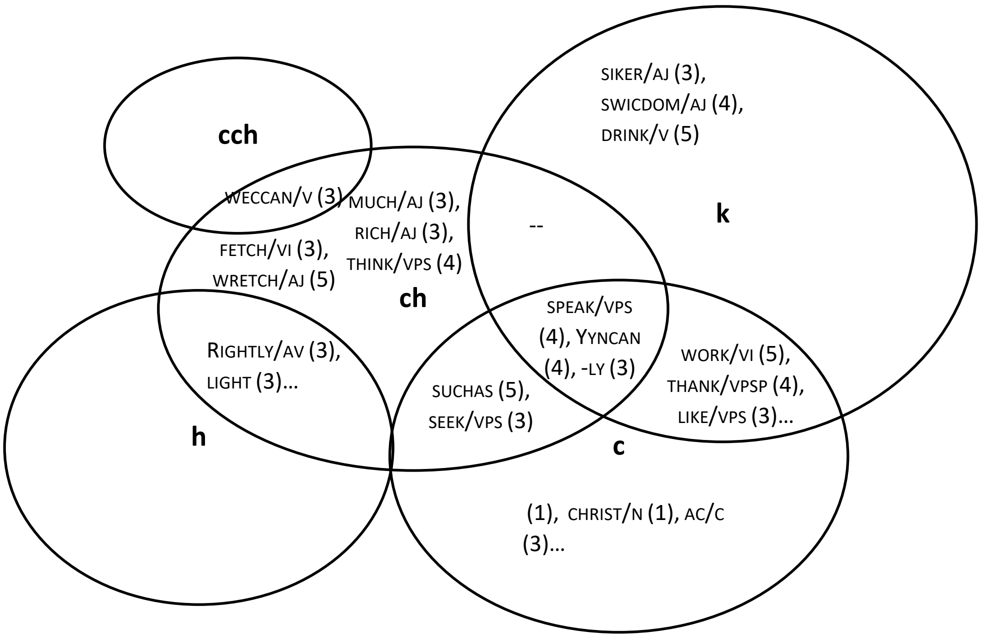


Figure 59: Visualisation of the overlapping uses of ch, k, c and cch in text #1200

Judging by the diagrams, the systems of scribe A and scribe B are very similar, which is in accordance with their locations in LAEME. The most conspicuous differences between the two is probably the interchangeability of *h*/*ch*, which is more prominent in A and the absence of *cc* from A.

The last diagram shows the corresponding litterae and items from #295 (London, British Library, Cotton Vespasian A.iii, *Cursor Mundi*), which as a northern text and as such it is geographically more distant.

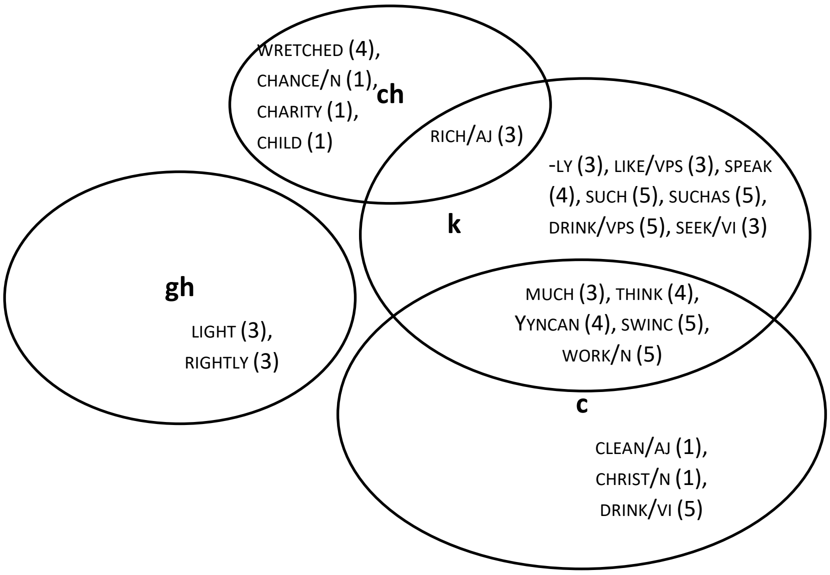


Figure 60: Visualisation of the overlapping uses of ch, k, c and gh in text #295

The diagram reflects the northern tendency to use *k* in positions where southern texts use *ch*. The *text profile* of #295 also shows that *ch* is relatively less frequent in the text compared to the *Trinity Homilies* and roughly 75 % of instances are found in the initial position. The set {*ch*, *k*} appears in one item only. Contrarily, *gh* is used much more extensively in the text. The items included in the diagram have *h* in B and {*h*, *ch*} in A.

Naturally, analyses based on the diagrams should take into account the whole forms of the items so that cases of internal consistency (e.g. *c* finally / *k* before *e*) can be revealed. Far from being the end result, the diagrams require a lot of detailed interpretative work. They merely visualise the formal aspects of a spelling system. Besides the scribe's perception of certain sounds as (dis)similar, they can reflect other phenomena, such as mixing of spelling practices from different texts, changes in progress (lexical diffusion). As the diagrams are based on the data from the spelling database, their analysis can be easily combined with statistical data from *text profiles* or maps, which should provide wider context required for interpretation. Automatic construction such diagrams is not available in the tool, however, they are highly compatible with data structure of the database – they essentially combine the data from networks with slot lists (item lists). Their potential yet needs to be properly tested, but this would be a task for a separate paper.

4.4.3. Possible upgrades

The final part of this chapter presents an overview of suggested upgrades of the tool which should either mitigate some of its current limitations or add new functionalities.

a) Search by adjacent morphemes

As the database is morpheme-based, an artificial boundary between morphemes is introduced despite the fact that segments separated by this boundary may be (and often are) relevant for phonological interpretation of the data. Querying possibilities should by improved so that filtering by context can be performed across the artificial boundary. For instance, when searching for prevocalic instances of *k*, the results should include root morphemes ending with *k* and followed by suffixes or endings beginning with a vowel e.g. *sak+e* ((FOR)SAKE/VPS).

b) Source forms

The addition of more source forms into the database would significantly improve filtering options and it would enable fast compilation of lists of items sharing a particular feature. Also, the links to CoNE could be defined not only based on sets and contextual constraints but also specific litterae attested in OE. As the database structure does not limit source forms to a single variant per item, it remains open to the inclusion of multiple reflexes from various sources, such as ON or other Germanic languages as well as PDE. Furthermore, lexels could be categorized by origin (OE/ON/Latin/French).

c) LALME

Data from LALME could in theory be parsed and coded so that it would become compatible with the spelling database. This would be especially useful for map sequences which could be extended to cover a much longer period. Although the number of items available in LALME is very limited, incorporation of LALME data in the database would significantly improve the possibilities of the tool.

d) External sources

Analogically to the links to CoNE, the tool could include links to other relatable sources. One of such sources could be the electronic version of the Bosworth-Toller dictionary hosted in Prague. An API is currently being developed for this source, which means the integration of the two databases could go beyond static links. Another relevant API is IIIF[22]. Some of the manuscripts included in LAEME are already available in the IIIF format and as such can be relatively easily loaded into the interface if required. The inclusion of images in corpora was advocated by Diemer (2012a).

e) Data storage and sharing

Most of the data in the tool including queries submitted by the researcher has a predefined standard structure. This means that queries could be easily stored for future reference or even be reused. For instance, all searches in a specific text could be re-applied to a different text and all previously generated maps could be re-displayed without the need to store them in an external file. A similar mechanism has been already implemented for item lists (which can be stored locally and reused).

f) Data export

While the data is mostly suitable for browsing and reading, it cannot be easily copied and pasted in the text of a research paper or a spreadsheet. The individual components should be extended with functions which would allow to copy their contents to the clipboard or display them in a format more suitable for export.

g) Statistical data

The possibility to retrieve global statistical data rather than lists of specific litterae or items remains underdeveloped. Examples of statistics which could be calculated based on the data in the DB include, e.g. the normalized frequency of a given littera / item calculated for the individual time periods or regions. Statistical data could also be used to check whether a specific alternation of litterae (e.g. {*s, ſ*} in #246) is rare or common (analogically to the *rare uses* statistics). Each littera in a *set* could be optionally displayed along with the periods in which it appears, so that it would be clear whether any of the litterae in the set were in use for a limited period of time.

h) Comparing contrasts

Network visualisations of correspondences between litterae in two separate texts can be used to describe contrasts in the two texts. Wherever one littera in a text corresponds to two

[22] International Image Interoperability Framework (https://iiif.io/)

different litterae in the other, it entails that the second text has two different spellings for what is likely to represent the same sound in the first text.

i) Miscellaneous ideas

- Comparison of item lists could be available in tabular format similar to the table of spelling variants in the seven texts of the *Poema Morale* described in subchapter 3.3.
- Results of searches could be filtered also by grammel.
- The machine-readable table of links between the manuscripts could be exploited in queries.
- Lexels could be grouped further, e.g. the lexel of the root morpheme LOVE in LOVE-LIKE is currently not LOVE but LOVE-LIKE, which means that there is no connection between LOVE and the corresponding part of LOVE-LIKE.

5. Conclusions

The objective of the present project is to construct a tool based on LAEME data, which would facilitate research into EME. The tool consists of a database of correspondences between segments in various spelling variants and an interface designed specifically to access the data. The concluding chapter summarizes the practical advantages and disadvantages of the tool and assesses whether its current version follows the general principles defined in the methodological chapter. It also comments on the connections between previous theoretical and methodological concepts and the features of the tool.

5.1. Strong and weak points of the tool

The presentation of the tool has focused on a basic description of the individual components and features and the links between them. Its scope is rather broad because of the range of different components and functions and at least some of the components would deserve a deeper exploration and formulation of more specific methodological recommendations. Although the current version of the tool is useable, it certainly requires further testing, fixing of errors in the database and adjustments of some of the calculations. Obtaining feedback from multiple users should make the process faster and more effective.

Arguably, the most innovative component seems to be the network visualisation which considerably facilitates analyses focused primarily on spelling systems, because it saves time and it can display multiple links between the litterae simultaneously, which is not possible in tabular form. The interactive links to item lists are useful, but it would also be convenient to be able to load a list of texts in which a given set of litterae occurs.

Similarly, to networks, one of the chief virtues of maps seems to be the possibility to pack a relatively large amount of data (all relevant spelling variants in all texts) into a single picture, which is moreover interactive. The micro analysis stressed the need to refine the maps. Maps based on larger sets of variants are almost impossible to read, on the other hand, using them as a mere point of departure seems to be a feasible strategy. As the clearest maps are naturally those based on single items, it is good that the tool enables to construct a single item map on a single click.

The inventories of litterae fulfil their basic function of a starting point for analyses of spelling systems. The visualisation of relative frequency works as expected. The precision of the statistics could be increased if the frequency was calculated relative to the number of slots, in which a given littera sometimes appears rather than to the total number of slots in a text. One weakness of the presentation of inventories is that the data concerning features like insertions, superscripts and capitalisation are difficult to access. It would be more suitable to provide a single table summarizing the frequencies of the individual features. The statistics for *rare uses* are not as reliable as intended.

Custom database searches provide access to the fundamental data types (sets, slots (items), litterae), but some pieces of the data are still difficult and slow to access. The tool should also offer the possibility to search directly for a list of texts in which a given set or littera appear. Another practical but not yet available query would be a query for sets restricted to a single text i.e. alternations of litterae normally displayed in *text profile*.

As for filtering options, the most useful one seems to be filtering by preceding and following litterae. It would be more convenient to filter by a set of litterae rather than a single littera, but this is not seen as a major drawback.

The component designed to examine the actual texts, highlight items etc. is particularly practical when used in combination with stored item lists, which help to identify and examine selected features in a specific text without necessarily reading through it. Depending on more specific needs, it might be more convenient to display all the forms matched by a search at the top of the text as clickable links which could be used to navigate the text.

As for the more experimental features, the queries for chunks, which was considered a marginal issue, were not properly tested. The links to CoNE appear to be a viable concept, although the number of potential changes displayed with sets (items) can be very high. The precision could be increased further with filters.

5.2. The perspective of principles

The first requirement was to enable identification of unusual features. The functionalities explicitly designed to fulfil this requirement are the visualisation of relative frequencies and identification of rare uses. It is also possible to identify rare alternatives of a littera because the list of alternatives provides frequency data. The testing showed that these functions are useful only to a certain extent and they have to be treated with caution, because the calculations are imperfect and statistics may become distorted, mainly in the case of low-frequency items.

The next principle was to minimize interpretative choices when parsing the forms. Parsing of complex and uncommon spelling variants required more interpretation than anticipated. In order to partly compensate for the arbitrary choices taken when dealing with ambiguous forms, chunk search was introduced. Running certain queries separately for chunks and slots could be less cumbersome if the tool automatically checked whether there are any results available for the other of the two modes, e.g. when running a query for items with alternating {*y, w*} at the level of chunk, the tool would also return the number of items with this alternation at the level of slots.

The general requirement of postponing interpretation is by definition connected with the capability to analyse more data within a certain period of time. The two features which best respond to this requirement are probably quick links to maps based on items and item lists. The links connecting sets, items forms and KWIC view are also practical in this respect. The biggest problem with postponing interpretation is that the "network" of collected data can grow so large that it becomes difficult to manage. This could be improved if the tool also included functionalities allowing to store searches in an intelligent, organised format.

The third principle concerns re-usability of the methodology and future compatibility with more data. The compatibility is satisfactory to the extent that more data can be added to the database, but automatic parsing was not efficient enough to allow fast data processing. The interface is almost data neutral. The only component adapted specifically to LAEME data are filters based on MSS metadata, but those would be relatively easy to adjust.

As far as zooming and links between different pieces of data are concerned, the interface provides links between the main data types – sets, items, maps, forms, KWIC and text profiles.

5.3. Responding to previous methodological observations

Several of the methodological concepts introduced in the theoretical chapter seem compatible with the tool. The first one is *Litteral Substitution Set* (LSS). The *sets* as defined in the tool correspond to the range of possible representations of a segment and as such, they can suggest possible sequences of literal substitutions operative in the development of spelling systems. The functionality could be improved further if it was possible to display the range of dates, indicating the order in which the individual litterae were added to the set. This information can be retrieved from the database but the corresponding function has not yet been implemented.

Sets displayed within *text profile* always reflect an occurrence of multiple litterae in the same slot. As such, they do not correspond to LSSs constituting a specific scribal lexicon as described by Laing & Lass (2013) which groups litterae by the assumed potestates. For instance, if a scribe uses *h* in MIGHT/N (3) but *ch* in NIGHT/N (3), the two litterae would not appear in the same *set* in the tool, although they would probably be included in a single LSS in a scribal lexicon. The tool is not capable of generating a scribal lexicon automatically, but it can facilitate the construction of one because the inventory of litterae and possibly also network visualisation allows to analyse the uses of different litterae in a systematic manner.

The methodological chapter also devoted some space to the problems of time and space in historical dialectology. In order to enhance the capabilities of the tool to trace the progression on changes in time as well as space, the mapping tool offers filtering by date. Moreover, some maps can be transformed into a sequence of maps for the individual time periods. Also, any queries can be filtered by date or region. Queries focusing specifically on the temporal dimension such as quantitative data on the frequency of a given littera in different time periods are not available, but they could be implemented without modifications of the data in the database.

The only feature focusing specifically on phonological changes are the links to CoNE, which are not fully developed. The linking of sets to changes is far from precise. The general patterns is that the tool suggests a number of potential changes for a single set or even a single slot and it is not possible to increase the precision without adding more data to the database. Moreover, pairing is sometimes problematic because of the wide range of possible spellings for a single sound.

Another familiar method which served as a basis for the functions in the tool is the use of item lists. The tool enables compilation of item lists based on a shared littera or a set of litterae, which can be further filtered by context of the littera, its position or occurrence in a manuscript or manuscript metadata. Items lists can be stored as repeatedly used to search for items in manuscripts or to generate maps.

Potential further development of the tool could proceed in several directions. The data could be improved by adding source forms and/or PDE forms at least for the most frequent items. Another option is to add data from LALME. Also, the interface still lacks potentially useful queries, especially queries crossing morpheme boundary and better queries for quantitative data. Also, it is incapable of exporting data (tables, query results) in a user friendly format. The possibility to store data currently limited to item lists could also be extended. The last area of improvement is the integration of data from external sources, for instance Bosworth-Toller dictionary or IIIF images.

The present project is deeply indebted to the exceptionally rich data from LAEME as well as decades of research in the field of historical linguistics represented by a number of brilliant scholars. In turn, the project can hopefully inspire thinking about potential contribution of technology to research. The final discussion of models of written language sketched out an experimental visualisation of scribal systems. Its design stems from the data structure of the spelling database but also the theoretical considerations of parallels between spoken and written language, which were repeatedly pointed out in this text.

6. References

Adams, M. (2015). “Introduction: Evidence and method in the historical study of English”. In M. Adams, L. J. Brinton, & R. D. Fulk (eds), *Studies in the History of the English Language VI* (pp. 1–12). DE GRUYTER. https://doi.org/10.1515/9783110345957.1

Aitchison, J. (2002). *Language change-progress or decay*. Cambridge: Cambridge University Press.

Alcorn, R. (2016). TIMESS grant application. Unpublished.

Benskin, M. (1997). “Texts from an English Township in Late Mediaeval Ireland.” *Collegium medievale: interdisciplinary journal of medieval research*, no. 1, pp. 91-174.

Black, M. (1999). “AB or Simply A? Reconsidering the Case for a Standard.” *Neuphilologische Mitteilungen*, vol. 100, no. 2, pp. 155–174. *JSTOR*, www.jstor.org/stable/43346192. Accessed 31 May 2021.

Brook, G. L. (1972). “A Piece of Evidence for the Study of Middle English Spelling”. *Neuphilologische Mitteilungen*, *73*(1/3), 25–28.

Browman, C. P., & Goldstein, L. (1995). “Dynamics and articulatory phonology”. *Mind as Motion*, 175–193.

Calle-Martín, J. & Moreno-Olalla, D. (2012) “Body of Evidence*:* of Middle English Annotated Corpora and Dialect Atlases”. In L. Wright and R. Dance (eds). *The Use and Development of Middle English. Proceedings of the Sixth International Conference on Middle English*, Cambridge 2008. Bern - Berlin - Bruxelles - Frankfurt am main - New York: Peter Lang AG. 17-34.

Cazal, Y., Parussa, G., Pignatelli, C., & Trachsler, R. (2003). “L’orthographe: Du manuscrit médiéval à la linguistique modern”. *Médiévales*, *45*, 99–118. https://doi.org/10.4000/medievales.969

Chomsky, C. (1971). “Invented Spelling in the Open Classroom”. *Word*, *27*(1–3), 499–518. https://doi.org/10.1080/00437956.1971.11435643

Christian, D. (1991). “The case for 'Big History'”. *Journal of World History*, 2(2), 223-238. http://www.jstor.org/stable/20078501

Corrie, M. (2006). “Middle English: Dialects and Diversity”. In L. Mugglestone, (ed.) *The Oxford History of English* [Online] (Updated ed.). Oxford: Oxford University Press. [http://site.ebrary.com/lib/cuni/Doc?id=10694614]

Daneš, F. (1966). “The Relation of Centre and Periphery as a language universal”. In *Travaux linguistiques de Prague 2*: p. 9-22.

Denison, D., Bermúdez-Otero, R., McCully, C., & Moore, E. (eds). (2011). *Analysing Older English*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139022170**.**

Diemer, S. (2012a). “Orthographic annotation of Middle English Corpora”. *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources. [Studies in Variation, Contacts and Change in English 10]*. Retrieved 7 December 2020, from https://www.academia.edu/2067888/Orthographic_annotation_of_Middle_English_Corpora

Diemer, S. (2012b). “Spelling variation in Middle English manuscripts: The case for an integrated corpus approach”. In M. Markus, Y. Iyeiri, R. Heuberger, & E. Chamson (eds), *Studies in Corpus Linguistics* (Vol. 50, pp. 31–46). John Benjamins Publishing Company. https://doi.org/10.1075/scl.50.05die

Dossena, M., & Lass, R. (2004). *Methods and Data in English Historical Dialectology*. Peter Lang.

Emerson, R. H. (1997). "English Spelling and Its Relation to Sound". *American Speech*, *72*(3), 260. https://doi.org/10.2307/455654

Faulkner, M. (2020). "Quantifying the Consistency of 'Standard' Old English Spelling". *Transactions of the Philological Society*, *118*(1), 192–205. https://doi.org/10.1111/1467-968X.12182

Fisiak, J. (1986) *A short grammar of Middle English (6. wyd.)*. Warszawa: Państwowe Wydawnictwo Naukowe.

Heggarty, P., McMahon, A., & McMahon, R. (2005). "From phonetic similarity to dialect classification: A principled approach". In N. Delbecque, J. van der Auwera, & D. Geeraerts (eds), *Perspectives on Variation*. De Gruyter Mouton. https://doi.org/10.1515/9783110909579.43

Hogg, R. (2006). "English in Britain". In R. Hogg & D. Denison (eds), *A History of the English Language* (pp. 352–383). Cambridge University Press. https://doi.org/10.1017/CBO9780511791154.008

Hopper, P. (1987). "Emergent Grammar". *Annual Meeting of the Berkeley Linguistics Society*, *13*, 139. https://doi.org/10.3765/bls.v13i0.1834

Horobin, S. (2010). *Studying the History of Early English*. Palgrave Macmillan.

Horobin, S. & Smith, J. (1999). "A database of Middle English spelling". *Literary and Linguistic Computing*, *14*(3), 359–374. https://doi.org/10.1093/llc/14.3.359

Hudson, A. (1966), "Tradition and Innovation in Some Middle English Manuscripts", *The Review of English Studies*, Vol. 17, No. 68, 359-372.

Kestemont, M. (2015). "A computational analysis of the scribal profiles in two of the oldest manuscripts of Hadewijch's letters". *Scriptorium* 69. 159–175.

Kestemont, M. & Karina van Dalen-Oskam (2009). "Predicting the Past: Memory Based Copyist and Author Discrimination in Medieval Epics". In *Proceedings of the twenty-first Benelux conference on artificial intelligence (BNAIC 2009)*. ResearchGate. Retrieved 9 June 2020, from https://www.researchgate.net/publication/237349804_Predicting_the_Past_Memory_Based_Copyist_and_Author_Discrimination_in_Medieval_Epics

Kohnen, Thomas (2014) *Textbooks in English Language and Linguistics (TELL), Volume 6 : Introduction to the History of English*. Frankfurt am Main, DEU: Peter Lang AG. ProQuest ebrary. Web. 10 February 2016.

Kopaczyk, J., Molineaux, B., Karaiskos, V., Alcorn, R., Los, B., & Maguire, W. (2018). "Towards a grapho-phonologically parsed corpus of medieval Scots: Database design and technical solutions". *Corpora*, *13*(2), 255–269. https://doi.org/10.3366/cor.2018.0146

Kretzschmar, Jr, W. (2015). "Complex systems and the history of the English language". In *Language and Complex Systems* (pp. 105-130). Cambridge: Cambridge University Press. doi:10.1017/CBO9781316179017.006 Laing, M. (1988).

Laing, M. (2015). "Some illustration of useful ways to compare and contrast the maps of early Middle English data in LAEME with those of late Middle English data in eLALME". Unpublished. University of Edinburgh.

Laing, M (2004) “Multidimensionality: Time, Space and Stratigraphy in Historical Dialectology”. in M. Dossena & R. Lass (eds), *Methods and Data in English Historical Dialectology: Linguistic Insights 16*. Peter Lang Publishing Group, Bern, 49-96.

Laing, M. (1999). “Confusion "wrs” Confounded: Litteral Substitution Sets in Early Middle English Writing Systems”. *Neuphilologische Mitteilungen, 100(3*), 251-270. Retrieved May 31, 2021, from http://www.jstor.org/stable/43346203

Laing, M. (1993). *Catalogue of Sources for a Linguistic Atlas of Early Medieval English*, Cambridge: Brewer.

Laing, M. (1988). “Dialectal Analysis and Linguistically Composite Texts in Middle English”. *Speculum*, *63*(1), 83–103. https://doi.org/10.2307/2854323

Laing, M. (1992). “A Linguistic Atlas of Early Middle English: The Value of Texts surviving in more than one Version”. *History of Englishes: New Methods and Interpretations in Historical Linguistics: Topics in English Linguistics 10*, 566–581.

Laing, M. & Lass, R. (2013). *Introduction to the Linguistic Atlas of early Middle English.* [http://www.lel.ed.ac.uk/ihd/laeme2/laeme_intro_ch1.html; http://www.lel.ed.ac.uk/ihd/laeme2/laeme_intro_ch2.html]. Edinburgh: © The University of Edinburgh.

Laing, M., & Lass, R. (2009). “Shape-shifting, sound-change and the genesis of prodigal writing systems”. *English Language and Linguistics*, *13*(01), 1. https://doi.org/10.1017/S1360674308002840

Laing, M., & Lass, R. (2003). “Tales of the 1001 nists: The phonological implications of litteral substitution sets in some thirteenth-century South-West Midland texts”. *English Language and Linguistics*, *7*(2), 257–278. https://doi.org/10.1017/S1360674303001102

Laing, M., & Williamson, K. (2004). “The Archaeology of Medieval Texts”. In C. Kay & J. Smith (eds). *Categorization in the History of English*, p. 85- 145.

Lass, R. (2015). “Interpreting Alphabetic Orthographies”. In P. Honeybone & J. Salmons (eds), *The Oxford Handbook of Historical Phonology.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199232819.013.024

Lass, R. (2006). “The end of linear narrative? Reflections on the historiography of English”. In N. Love, (ed.), *Language and History: Integrationist Perspectives* (1st ed.). Routledge. https://doi.org/10.4324/9780203592588

Linell, P. (2019). “The Written Language Bias (WLB) in linguistics 40 years after”. *Language Sciences*, 76. Online [https://www.sciencedirect.com/science/article/pii/S0388000118303875] https://doi.org/10.1016/j.langsci.2019.05.003

McIntosh, A., Samuels, M. L., & Laing, M. (1989). *Middle English dialectology: Essays on some principles and problems*. Aberdeen University Press.

McMahon, A. M. S. (1994). *Understanding Language Change* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139166591

McMahon, A., Foulkes, P., & Tollfree, L. (1994). “Gestural Representation and Lexical Phonology”. *Phonology*, *11*(2), 277–316.

Millward, C. M. & Hayes, M. (2012), *A Biography of the English Language*, Wadsworth: Cengage Learning.

Minkova, D. (2015). “Establishing Phonemic Contrast in Written Sources”. In P. Honeybone & J. Salmons (eds), *The Oxford Handbook of Historical Phonology.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199232819.013.024

Minkova, D. (2013). *A Historical Phonology of English*. 441. Edinburgh: Edinburgh University Press.

Minkova, D. (2003). *Alliteration and sound change in early English*. Cambridge: Cambridge University Press.

Minkova, D., & Stockwell, R. P. (eds). (2002). *Studies in the history of the English language: A millennial perspective*. Mouton de Gruyter.

Minkova, D., & Stockwell, R. (1998). “Are Diphthongs Neglected?”. *Publication of the American Dialect Society*, *80*(1), 34–49. https://doi.org/10.1215/-80-1-34

Mossé, F. (1968). *A handbook of Middle English (5th printing, corrected and augmented.)*. Baltimore: Johns Hopkins University Press.

Ogura, M. & William S-Y. Wang. (2004). “Dynamic Dialectology and Complex Adaptive System”. In M. Dossena & R. Lass (eds), *Methods and Data in English Historical Dialectology: Linguistic Insights 16*. Peter Lang Publishing Group, Bern, 137-170.

Pandey, P. K. (1997). “Optionality, Lexicality and Sound Change”. *Journal of Linguistics*, *33*(1), 91–130.

Read, C. (1971). “Pre-School Children’s Knowledge of English Phonology”. *Harvard Educational Review*, *41 (1)I*, p. 1-34.

Sebba, M. (Ed.). (2007). “Between language and dialect: Orthography in unstandardised and standardising vernaculars”. In *Spelling and Society: The Culture and Politics of Orthography around the World* (pp. 102–131). Cambridge University Press. https://doi.org/10.1017/CBO9780511486739.010

Smith, J. (2020). “On Scriptae: Correlating Spelling and Script in Late Middle English”. *Revista Canaria de Estudios Ingleses*, *80*, 13–27. https://doi.org/10.25145/j.recaesin.2020.80.02

Smith, J. J. (2007). *Sound change and the history of English*. Oxford University Press.

Smith, J. J., Black, M., & Horobin, S. (2002). “Towards a new history of Middle English spelling”. In P. J. Lucas & A. M. Lucas (eds), *Middle English from Tongue to Text: Selected Papers from the Third International Conference on Middle English: Language and Text, Held at Dublin, Ireland, 1-4 July 1999* (No. 4; Issue 4, pp. 9–20). Peter Lang. http://eprints.gla.ac.uk/8961/

Stenroos, M. (2004). “Regional Dialects and Spelling Conventions in Late Middle English”. In M. Dossena & R. Lass (eds), *Methods and Data in English Historical Dialectology: Linguistic Insights 16*. Peter Lang Publishing Group, Bern, 257-286.

Teresi, L., & University of Manchester (1998). *A computer-assisted analysis of spellings in two vernacular manuscripts of the transition period: MS Cambridge, Corpus Christi College 302 and MS London, British Library, Cotton Faustina A. ix.* Manchester: University of Manchester.

Browman, C. P., & Goldstein, L. (1992). “Articulatory Phonology: An Overview”. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

Tolkien, J. R. R. (1929). “Ancrene wisse and Hali meiðhad”. *Essays and Studies, 14.*

Upward, C., & Davidson, G. (2011). *The history of English spelling* [Online]. Malden, Mass.: Wiley-Blackwell. [http://site.ebrary.com/lib/cuni/Doc?id=10483263]

Vachek, J. (1978) *A brief survey of the historical development of English* (4. ed.). Praha: Státní pedagogické nakladatelství.

Vachek (1966). "On the Integration of the Peripheral Elements into the System of language". *Travaux linguistiques de Prague 2*. p. 23—37.

Vachek, J. (1942). "Písmo a transkripce ve světle strukturálního jazykozpytu". *ČMF*, *28*, p. 403-408.

Vachek, J., & Luelsdorff, P. A. (1989). *Written language revisited*. Amsterdam: John Benjamins.

Vaňková, M. (2021, forthcoming). "Testing a Spelling Database Created from The Linguistic Atlas of Early Middle English".

Vaňková, M. (2016). *Localisation of version D of "The Poema Morale" based on "The Linguistic Atlas of Early Middle English*. Unpublished MA thesis.

Venezky, R. L. (2011). *The Structure of English Orthography*. Walter de Gruyter.

Wiggins, A. (2007). "Middle English Romance and the West Midlands". In Scase, W. (ed.), *Essays in manuscript geography: Vernacular manuscripts of the English West Midlands from the Conquest to the sixteenth century*. Brepols.

Williamson, K. (2004). "On Chronicity and Space(s) in Historical Dialectology". In M. Dossena & R. Lass (eds), *Methods and Data in English Historical Dialectology: Linguistic Insights 16*. Peter Lang Publishing Group, Bern, 97-136.

Wood, M. (1982). "Invented Spelling". *Language Arts*, 59(7), 707-717. Retrieved May 31, 2021, from http://www.jstor.org/stable/41405102

6.1. Online Resources

A Linguistic Atlas of Older Scots, Phase 1: 1380-1500 [http://www.lel.ed.ac.uk/ihd/laos1/laos1.html] (Edinburgh: © 2008- The University of Edinburgh).

Angus McIntosh Centre for Historical Linguistics [http://www.amc.lel.ed.ac.uk/]

Benskin, M., Laing, M., Karaiskos, V. & K. Williamson. An Electronic Version of A Linguistic Atlas of Late Mediaeval English [http://www.lel.ed.ac.uk/ihd/elalme/elalme.html] (Edinburgh: © 2013- The Authors and The University of Edinburgh).

Bosworth, J. *An Anglo-Saxon Dictionary Online*., Ed. Thomas Northcote Toller and Others. Comp. Sean Christ and Ondřej Tichý. Faculty of Arts, Charles University in Prague, 21 Mar. 2010. Web. 24 Feb. 2015. [http://bosworth.ff.cuni.cz/]

Heggarty, Paul, Aviva Shimelman, Giovanni Abete, Cormac Anderson, Scott Sadowsky, Ludger Paschen, Warren Maguire, Lechoslaw Jocz, María José Aninao, Laura Wägerle, Darja Dërmaku-Appelganz, Ariel Pheula do Couto e Silva, Lewis C. Lawyer, Jan Michalsky, Ana Suelly Arruda Câmara Cabral, Mary Walworth, Ezequiel Koile, Jakob Runge & Hans-Jörg Bibiko. 2019. Sound Comparisons: Exploring Diversity in Phonetics across Language Families. (Available online at https://soundcomparisons.com, Accessed on 2021-05-31.)

Laing, M. (2013) *A Linguistic Atlas of early Middle English, 1150–1325*, Version 3.2 [http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html]. Edinburgh: © The University of Edinburgh.

Lass, R., Laing, M., Alcorn, R. & K. Williamson (2013). A Corpus of Narrative Etymologies from Proto-Old English to Early Middle English and accompanying Corpus of Changes, Version 1.1 [http://www.lel.ed.ac.uk/ihd/CoNE/CoNE.html]. Edinburgh: © The University of Edinburgh. https://soundcomparisons.com/

Treharne, E., Cambridge, Trinity College, B. 14. 52., in *The Production and Use of English Manuscripts 1060 to 1220*, edited by Orietta Da Rold, Takako Kato, Mary Swan and Elaine Treharne (University of Leicester, 2010), accessed 8 March 2016. [http://www.le.ac.uk/english/em1060to1220/mss/EM.CTC.B.14.52.htm]

7. Appendices

7.1. LAEME files referenced in the thesis

text id	manuscript
3	London, British Library, Cotton Caligula A ix (*The Owl and the Nightingale*)
4	Cambridge, Trinity College B.14.52 (*Poema Morale T*)
7	London, British Library, Egerton 613 (*Poema Morale E*)
8	Oxford, Bodley Digby 4 (*Poema Morale D*)
9	Oxford, Jesus College 29 (*Poema Morale J*)
10	Cambridge, Fitzwilliam Museum, McClean 123 (*Poema Morale M*)
64	London, British Library, Stowe 34, Hand A (*Vices and Virtues*)
129	Cambridge University Library Ff.VI.15 (*The Ten Comandements*)
136	London, Lambeth Palace Library 499 (lyrics)
137	London, British Library Arundel 248 (short pieces)
149	Oxford, Bodleian Library, Laud Misc 636 (*The Peterborough Chronicle*)
150	London, BL Arundel 292 (*The Bestiary*)
155	Cambridge, Corpus Christi College 444 (*Exodus*, *Genesis*)
158	Oxford, Bodleian Library, Bodley 652 (*Iacob and Iosep*)
160	Oxford, Bodleian Library Add E.6, roll (*Sayings of St Bernard*)
161	Oxford, Bodleian Library, Additional E.6, roll (*An Exposition of the Pater Noster I*, *The XV signs before Doomsday*)
169	Oxford, Merton College 248 (short pieces)
170	Worcester Cathedral, Chapter Library Q 29 (*A sermon on the Nativity*)
171	Oxford, Bodleian Library, Junius 121 (*Nicene Creed*)
172	Worcester Cathedral, Chapter Library F 174 (short rhythmic prose text, *The Debate between the Body and Soul (theme)*)
173	Worcester Cathedral, Chapter Library F 174 (*Ælfric's Grammar and Glossary*)
214	Oxford, Bodleian Library, Digby 86 (*Iesu dulcis memoria, The XI Pains of Hell*)
218	Oxford, Bodleian Library, Digby 86 (*The Proverbs of Alfred, The Proverbs of Hending*)
227	Oxford, New College 88 (*religious pieces*)
242	London, British Library, Cotton Caligula A ix (*The Latemest Day*)
246	Cambridge, Trinity College B.14.39, hand A
247	Cambridge, Trinity College B.14.39, hand B
248	Cambridge, Trinity College B.14.39, hand C
249	Cambridge, Trinity College B.14.39, hand D
261	London, British Library, Royal 17 A xxvii (*On Lofsong of Ure Lefdi / Oreisun of Seinte Marie, Sawles Warde, St Juliana*)
263	London, British Library, Royal 2.F.viii (*religious pieces*)
273	London, British Library, Cotton Cleopatra C.vi (*Ancrene Riwle*)
276	Cambridge, Gonville and Caius 234/120, pp. 1-185 (*Ancrene Riwle*)
277	London, British Library, Cotton Caligula A.ix, part 1 (*Layamon A I*)
280	London, British Library, Cotton Otho C xiii (*Laȝamon B*)
282	Oxford, Bodleian Library, Laud Misc 108 (*The Debate between the Body and Soul (theme)*)
285	Oxford, Bodleian Library, Laud Misc 108 (*Havelok*)
291	London, British Library, Arundel 57 (containing the *Ayenbyte of Inwyt*)
295	London, British Library, Cotton Vespasian A.iii (*Cursor Mundi*)
297	Edinburgh, Royal College of Physicians (*Cursor Mundi*)
300	London, British Library, Arundel 292 (miscellaneous religious pieces)
301	Oxford, Bodleian Library, Junius 1 (*The Orrmulum*)

304	London, British Library, Cotton Claudius D iii (*Benedictine Rule*)
1100	Oxford, Jesus College 29
1200	Cambridge, Trinity College B.14.52, hand A (*Trinity Homilies*)
1300	Cambridge, Trinity College B.14.52, hand B (*Trinity Homilies*)
1400	Cambridge University Library Ff.II.33 (*Bury documents*)
1600	Oxford, Bodleian Library Laud Misc 108, part 1 (South English Legendary)
1800	London, British Library, Cotton Nero A xiv (*miscellaneous religious pieces*)
2000	London, Lambeth Palace Library 487 (*Lambeth Homilies A*)
2001	London, Lambeth Palace Library 487 (Lambeth Homilies B)
2002	Oxford, Bodleian Library, Digby 86

7.2. Anchor texts

text id	total tokens	manuscript	hand	anchor type
16	110	Oxford, Bodleian Library, Rawlinson C 317		L
124	745	Oxford, Bodleian Library, Tanner 169*, p. 175		L
125	403	Herefordshire Record Office AL 19/2, Registrum Ricardi de Swinfield		D
126	198	Stratford-upon-Avon, Shakespeare Birthplace Library, DR 10/1408, pp. 23-24		D
128	303	London, Lincoln´s Inn Hale 135		L
130	50	Oxford, Bodleian Library, Rawlinson C 510		L
131	530	London, BL Cotton Galba E ii		D
132	437	Carlisle, Cumbria RO, D/Lons/L Medieval Deeds C1		D
133	5238	London, PRO, E 164/28	A	D
134	366	London, PRO E 164/28	B	D
135	2304	London, BL Cotton Otho B xiv		D
140	2479	Cambridge, Emmanuel College 27		L
143	1446	London, British Library, Add 15340		D
144	205	London, British Library, Harley 978		L
147	1431	London, British Library, Cotton Roll ii.11		D
148	423	London, British LibraryL Cotton Roll ii.11		D
149	6812	Oxford, Bodleian Library, Laud Misc 636		A
156	1642	Wells Cathedral Library, Liber Albus I		D
157	1315	Wells Cathedral Library, Liber Albus I		D
160	3007	Oxford, Bodleian Library Add E.6, roll	A	L
163	364	Aberdeen University Library 154		L
170	3303	Worcester Cathedral, Chapter Library Q 29		L
171	595	Oxford, Bodleian Library, Junius 121		L
172	8114	Worcester Cathedral, Chapter Library F 174		L
173	47031	Worcester Cathedral, Chapter Library F 174		L
177	151	Oxford, Bodleian Library, Bodley 57		L
183	356	Private		L
184	2328	London, British Library, Cotton Vitellius A xiii, Chertsey Cartulary		D
185	487	Cambridge University Library, Add 3020, Red Book of Thorney 1		D
186	432	Cambridge University Library, Add 3021, Red Book of Thorney 2		D
187	136	Worcester, Herefordshire and Worcestershire Record Office, BA 3814		D
188	4485	London, British Library, Cotton Julius A v		L

229	2689	Oxford, Corpus Christi College 59		L
230	1272	London, British Library, Cotton Charter iv 18		L
256	417	London, British Library, Cotton Faustina A.v fols. 10r-v	A	L
257	362	London, British Library, Cotton Faustina A.v	B	L
266	340	Cambridge University Library Hh.6.11		L
279	725	London, British Library, Add 46487, Sherborne Cartulary		D
291	93603	London, British Library, Arundel 57		L

7.3. Database statistical overview

Query	**result**	**note**
Total number of LAEME tags (words)	651823	Total rows in the table *laeme_tags*
Total number of LAEME tags (morphemes)	834398	Total rows in the table *laeme_morphemes*
Total number of unique LAEME lexels (morphemes)	11019	
Total number of processed items	8955	
Total number of unique forms	55508	
Total number of unique segments (litterae)	361	Includes single occurrences
Total number of unique slots	40028	i.e. combinations of item id and position number
Total number of supersets	813	Smaller sets are subsumed under larger sets, e.g. {*c*, *k*} and {*c*, *h*, *k*, *q*} are counted as one set

7.4. Tags excluded from processing

lexel	**tokens**
GRAMMATICAL WORDS	52
-ed	1
-els	2
-en	1
-en{d}	1
-en{i}	1
-er	5
-est	1
-fast	1
-ig	16
-in	2
-ing	1
-isc	1
-less	1
-ly	1
-self	2
-some	1
-th	6
-uY	1
-ward	2
&	2
1000	1
600000	1
7night	1

a-	2
a:nle:pig	1
accord	2
account	2
ade:adian	1
affraien	1
againcerran	1
allinge	1
almighty	3
amend	2
among	1
amount	1
an-	1
andaful	1
anent	2
angel	1
annoy	4
anonright	12
anonso	8
anonthat	1
anupon{p}	1
anykin	1
apostle	1
aready	1
arch-	1
as	1
as-sum	6
assail	2
assoilen	1
assoonas	3
astound	1
athome	1
aturnen	1
avi:len	1
avow	1
await	7
be	1
be-	2
beaufrere	2
becatch	1
befall	1
begitan	2

bethink	2
bewinnan	1
byrdan	1
come	1
confound	1
cumber	1
dearworthly	1
declension	1
decli:nigendli:c	1
defoul	1
deserve	1
dismay	1
drunken	1
e:adig	3
e:admo:dig	1
e:aYele:te	5
eachone	5
eachonedeal	1
elYe:odigli:ce	4
encounter	1
endebyrdli:ce	1
enough	6
enoughhraDe	2
envenom	1
evereach	2
evereachdeal	4
everywhere	1
fell	1
foe	1
forcu:Y	1
forlose	1
forthat	1
forthgewi:tan	1
forthright	1
forthythat	1
frumbyrdling	1
further	3
ge-	1
gebe:gedness	1
geli:c	1
gemyndig	2
gewis	2

godalmighty	1
gospel	1
gospeller	1
half	1
handle	1
hardi	1
harrow	1
have	3
he:rsumian	1
he:rsumness	1
hereupon{re}	2
hie	1
holy	2
honour	5
christian	1
ilk	123
in	2
last	1
lecherous	2
linen	1
man	3
manifoldly	12
manya	1
menen	1
mighty	3
mildsian	1
mis-	1
morning	1
n-	1
narrow	1
nevermore	1
niman	1
nowhere	1
onsi:gan	1
onufeward	1
oppose	2
other	3
out-	1
over	1
pay	1
perceive	1
perform	1
pursue	1
racente:ah	1
ready	1
red	1
right	2
righteous	1
ruthfully	1
sae:lig	1
sagol	1
sainthood?	1
say	1
scourge	1
see	1
shall	1
smell	1
so	2
so:cn	1
sorriness	1
sorry	1
sosum	2
sosumthat	1
strangle	6
strength	1
suchas	1
sum	2
sweotolli:ce	1
that	1
thereteke	1
thereupon{p}	1
thing	1
ti:Da	1
tintregian	2
to	2
to-	1
toflowedness	1
tooth	1
toward	1
turn	1
under-	1
unsae:lig	1
up-	1
wan-	2

weep	1
welcome	1
whereupon{p}	1
whilethat	5
whitsuntide	1
why	1
wi:sely	1
wi:seness	3
will	2
willnot	1
winnan	1
within	1
withmetenli:c	1
Yan	2
Ye:aw	1
Ye:od	1
Ye:ostrian	1
ymbee:ode	1
youth	1

7.5. Manually defined word classes

Label: manually created label

Total tokens: total number of tokens

LAEME grammels: the full list of original LAEME grammels subsumed under the label

Label	**total tokens**	**LAEME grammels**
A-dat-acc	1279	{A-av,A-av-k,A-av+H,A-av+V,A<pr,A<pr-k,A<pr-k{rh}",A<pr+H,A<pr+V,A>pr,A>pr+H,A>pr+V}"
av	1	{pr<}
c	29	{av,av>=}
Dat-acc	726	{DatOd,DatOd-ad,DatOd-as,DatOd{rh}",DatpnOd,DatpnOd-ad,"DatpnOd{rh}",DatpnOd>=,DatpnOdRTA,DatpnOdRTA>pr,DatpnOdRTI,DatpnOdRTIOd}"
Dat-dat	26	{DatOi,DatpnOi,DatpnOi>=,DatpnOiRTA,DatpnOiRTAOd,DatpnOiRTI}
Dat-dat-acc	649	{Dat-av,Dat-av-as,Dat<pr,Dat<pr-ad,Dat>pr}
Dat-gen	41	{DatG,DatG-ad,DatpnG,DatpnG{rh}",DatpnGRTI,DatpnGRTIOd}"
Dat-nom	2591	{DatN,DatN-ad,DatN-as,Datpn,Datpn-ad,Datpn-as,Datpn-as{rh}","Datpn{rh}",Datpn<pr,Datpn<pr-ad,"Datpn<pr{rh}",Datpn<pr>=,Datpn<prRTA,Datpn<prRTAG,Datpn<prRTAOd,Datpn<prRTI,Datpn<prRTI-ad,Datpn<prRTI<pr,Datpn<prRTIOd,Datpn<prRTIOd-ad,Datpn>=,Datpn>pr,Datpnpl,Datpnpl<pr,Datpnpl<prRTIpl,Datpnpl>=,DatpnRTA,DatpnRTA-ad,DatpnRTA>pr,DatpnRTAOd,DatpnRTAOi,DatpnRTI,DatpnRTI-ad,DatpnRTI>pr,DatpnRTIOd,DatpnRTIOd-ad}"
Des-acc	277	{DesOd,DesOd-ad,DesOd<{rh}",DespnOd}"
Des-dat	13	{DesOi,DespnOi}
Des-dat-acc	282	{Des-av,Des<pr,Des<pr-ad}
Des-gen	18	{DesG,DespnG}
Des-nom	719	{DesN,DesN-ad,Despn,Despn-ad,Despn<pr,Despn<pr{rh}"}"
Dis-acc	1014	{DisOd,DisOd-ad,DisOd-as,DispnOd,DispnOd-ad,DispnOd{rh}"}"

Dis-dat	10	{DisOi}
Dis-dat-acc	1323	{Dis-av,Dis<pr,Dis<pr-ad,Dis>pr}
Dis-gen	145	{DisG,DisG-as,DisG-av,DispnG}
Dis-nom	1574	{DisN,DisN-ad,DisN-as,Dispn,Dispn-ad,Dispn-as,Dispn{rh}",Dispn<pr,"Dispn<pr{rh}"}"
Dos-acc	140	{DosOd,DosOd{rh}",DospnOd,DospnOd>=,"DospnOd>={rh}",DospnOdRTApl,DospnOdRTApl>pr,DospnOdRTAplOd,DospnOdRTAplOi,DospnOdRTIpl,DospnOdRTIplOd}"
Dos-dat	38	{DospnOi,DospnOi>=,DospnOiRTApl,DospnOiRTApl>pr,DospnOiRTAplOi}
Dos-dat-acc	38	{Dos-av,Dos<pr}
Dos-gen	23	{DosG,DospnG,DospnGRTApl,DospnGRTIpl,DospnGRTIplOd}
Dos-nom	688	{DosN,DosN-ad,Dospn,Dospn-ad,Dospn-as,Dospn{rh}",Dospn<pr,"Dospn<pr{rh}",Dospn<pr>=,"Dospn<pr>={rh}",Dospn<prRTApl,Dospn<prRTAplOd,Dospn<prRTAplOi,Dospn<prRTIpl,Dospn<prRTIpl>pr,Dospn<prRTIplOd,Dospn>=,DospnRTApl,DospnRTApl-ad,DospnRTApl-as,DospnRTApl+in,DospnRTApl>pr,DospnRTAplOd,DospnRTAplOi,DospnRTIpl,DospnRTIpl>pr}"
n	1300	{indef,int,int{rh}",pr}"
T-acc	3756	{T-xp-ajOd,T-xp-av,T-xp-av-ad,T-xp-cj,T-xp-pnOd,TOd,TOd-ad,TOd-as}
T-accPl	1249	{T-xp-ajplOd,T-xp-pnplOd,TplOd,TplOd-ad,TplOd-as}
T-dat	319	{TOi,TOi-ad,TOi-as}
T-dat-acc	6839	{T-av,T-av-ad,T-av-as,T-xp-pn>pr,T<pr,T<pr-ad,T<pr-as,T>pr}
T-datPl	90	{TplOi,TplOi-ad}
T-datPl-accPl	1172	{Tpl-av,Tpl<pr,Tpl<pr-ad,Tpl>pr}
T-gen	825	{TG,TG-ad,TG-as}
T-genPl	119	{TplG,TplG-ad}
T-nom	7386	{T-int,T-inv,T-voc,T-xp-aj,T-xp-pn,TN,TN-ad,TN-as}
T-nomPl	1717	{T-xp-pnpl,TplN,TplN-ad,TplN-as}

7.6. Grammatical items

Word_class	Total forms	Processed forms
A-dat-acc	21	21
AG	18	0
AN	7	7
AOd	19	19
AOi	6	0
Apl	3	0
Dat	3	0
Dat-acc	63	0
Dat-dat	19	0
Dat-dat-acc	36	0
Dat-gen	18	0
Dat-nom	147	0
Des	1	0
Des-acc	37	0
Des-dat	6	0
Des-dat-acc	39	0

Des-gen	9	0
Des-nom	62	0
Dis	2	0
Dis-acc	44	0
Dis-dat	5	0
Dis-dat-acc	58	0
Dis-gen	27	0
Dis-nom	58	0
Dos	1	0
Dos-acc	36	0
Dos-dat	19	0
Dos-dat-acc	12	0
Dos-gen	7	0
Dos-nom	118	0
P11<pr	3	3
P11<prX	3	0
P11>pr	1	0
P11G	21	21
P11MX	1	0
P11N	33	34
P11NX	4	0
P11O	4	4
P11OdX	5	0
P11OiX	4	0
P11X	5	0
P12<pr	8	8
P12<prX	4	0
P12>pr	2	2
P12G	53	53
P12MX	4	0
P12N	25	25
P12O	10	10
P12OdX	6	6
P12OiX	2	2
P12X	2	2
P13>prF	5	5
P13>prI	1	1
P13>prM	9	9
P13>prXF	1	0
P13>prXM	4	4
P13GF	40	0

P13GI	7	7
P13GM	4	4
P13MXF	3	3
P13MXI	2	2
P13MXM	6	6
P13NF	35	35
P13NI	24	24
P13NM	23	23
P13NXM	2	2
P13OdF	14	14
P13OdI	65	0
P13OdM	22	22
P13OdXF	5	5
P13OdXI	3	3
P13OdXM	9	9
P13OiF	10	10
P13OiI	1	1
P13OiM	14	14
P13OiXF	2	0
P13OiXM	6	6
P13XF	5	0
P13XI	2	2
P13XM	4	4
P21<pr	5	5
P21<prX	3	0
P21>pr	3	3
P21G	61	0
P21MX	3	3
P21N	16	16
P21O	14	14
P21OdX	5	5
P21OiX	2	2
P21X	2	2
P22<pr	10	10
P22<prX	2	2
P22>pr	2	2
P22G	87	0
P22MX	4	4
P22N	20	20
P22O	35	35
P22OdX	15	15

P22OiX	7	7
P22X	6	0
P23<pr	26	26
P23<prX	13	13
P23>pr	18	18
P23G	32	32
P23MX	8	8
P23N	29	29
P23O	35	35
P23OdX	17	17
P23OiX	11	11
P23X	5	5
RTA	6	6
RTA<pr	3	0
RTA>pr	8	8
RTAG	2	2
RTAOd	11	11
RTAOi	6	6
RTApl	0	0
RTApl<pr	2	2
RTApl>pr	5	5
RTAplOd	11	11
RTAplOi	5	5
RTI	4	4
RTI<pr	2	2
RTI>pr	13	13
RTIG	1	0
RTIOd	10	10
RTIOi	4	0

RTIpl	0	0
RTIpl>pr	6	6
RTIplOd	13	13
RTIplOi	2	2
T-acc	96	0
T-accPl	30	0
T-dat	30	0
T-dat-acc	128	0
T-datPl	13	0
T-datPl-accPl	42	0
T-gen	50	0
T-genPl	21	0
T-nom	88	0
T-nomPl	36	0
vi	0	0
vn	0	0
vpp	0	0
vps11	7	7
vps12	18	18
vps13	57	57
vps21	25	25
vps22	25	25
vps23	23	23
vpt11	24	24
vpt12	32	32
vpt13	41	41
vpt21	20	20
vpt22	17	17
vpt23	50	50

7.7. Text groups (manual processing)

label	**LAEME files Ids**	**Description**
Poema Morale	4-10	The seven version of *The Poema Morale*
Owl and Nightingale	2, 3, 1100	The files containing copies of The Owl and the Nightingale. #1100 includes also other texts.
Vices and Virtues	64, 65, 302, 303	The text of Vices and Virtues and corrections thereof from London, British Library, Stowe 34 written in different hands.

Laȝamon	174, 175, 271, 277, 278, 280, 286	The MSs containing Laȝamon A, Laȝamon B plus files related to the by similar language or localisation.
The Homilies	1200, 1300, 2001, 2000, 189, 63	The MSs containing Trinity Homilies, Lambeth Homilies plus other texts contained in the Trinity and Lambeth MSs.
AB language	118-121, 123, 245, 262, 263, 272, 273, 275, 276, 1000	A large group of texts directly or indirectly associated with AB language. Most of the MSs contain versions of *Ancrene Riwle* and texts from the Katherine group.
Digby 86	161, 214, 218, 220, 222, 227, 2002	Texts found in MS Digby 86 plus texts related to the by similar language.
Cusor Mundi	295-298	The MSs of *Cursor Mundi*.
Worcester Tremulous Hand	170-173, 1800	The work of the Worcester scribe plus texts in similar language.
Trinity B,14.39	246-249, 169	Texts from MS Trinity College B.14.52 copied by four different scribes.
Kent	291, 142	The texts localised in Kent - anchor MS Arundel 57 plus the MS of Kentish Sermons.
Documents	a) 156, 157, b) 147, 148, c) 133-135	Documentary texts further grouped by MS or scribe.
Cleopatra C	1701, 1702, 1400, 146	Texts from MS Cotton Cleopatra C vi plus two texts related by shared content or localisation.
Laud Misc 108	282, 285	The two parts of MS Laud Misc 108.
Arundel 292	150, 155, 300	The two files covering MS Arundel 292 plus MS Cambridge, Corpus Christi College 444 related by similar language.

7.8. Conversion of LAEME conventions

feature	**LAEME notation**	**Replacement in text**
yogh	z	ȝ
Insular g	g	ᵹ
wynn	w	ƿ
ash	ae	æ
edh	d	ð
thorn	y	þ
Capital letter	*+letter	Capital letter
Expanded abbreviation	Lowercase letters	Uppercase letters
r+superscript, u+superscript	r^, u^	none
Stacked letters	letter^letter	none
insertion	>string>	none
deletion	<string<	none
reconstruction	[string]	none
De nexus		none
Flourished s	^S	none

7.9. Litterae metadata

Category	value	tag
Type	consonant	C
	vowel	V
	diphthong	DP
Vowel length	short	ST
	long	LN
Vowel height	low	1
	Low-mid	2
	mid	3
	High-mid	4
	high	5
Consonants – place of articulation	labial	l
	Labio-dental	ld
	Labio-velar	lv
	dental	d
	alveolar	a
	palatal	p
	velar	v
	glottal	g
Consonants – manner of articulation	plosive	P
	fricative	F
	affricate	A
	Approximant - coronal	Xc
	Approximant - lateral	Xl
	nasal	N
	liquid	L
	spirant	S
Consonants - voicing	voiced	V1
	voiceless	V0

7.10. Table samples

7.10.1. Litterae statistics

rank	id	littera	average frequency	length	category	tags	types	tokens
1	141	e	0.14167283202712390902	1	V	{ST,3,f,V}	9212	287170
2	139	i	0.07374576942468126585	1	A	{A}	3099	141358
3	140	o	0.07031025797868838466	1	V	{ST,2,b,r,V}	2757	131486
4	138	n	0.06999597429039487117	1	C	{a,N,+,C}	2805	124719
5	134	a	0.06127317575010526898	1	V	{ST,1,c,V}	3082	122844
6	137	r	0.05648360102641355267	1	C	{a,Xc,+,C}	3163	111190
7	136	t	0.04969679719876167214	1	C	{P,a,-,C}	2521	99332
8	135	d	0.05120171578784676458	1	C	{a,P,+,C}	1889	92554
9	131	s	0.04593911411889760746	1	C	{a,F,-,C}	2307	89030

10	133	l	0.04502288004807719923	1	C	{a,Xl,+,C}	2283	85775
11	132	u	0.03820322311172977532	1	A	{A}	3283	75850
12	129	h	0.03036787226519378978	1	C	{g,F,C}	1481	72461
13	130	m	0.02935575727996485189	1	C	{l,N,+,C}	1059	58020
14	128	f	0.02927744444320442604	1	C	{ld,F,-,C}	1035	55379
15	127	þ	0.024281208087719698090681	1	C	{d,F,C}	874	52207
16	125	b	0.02068487935220337729	1	C	{P,l,+,C}	701	33961
17	124	ƿ	0.011462889509624006125721	1	C	{C}	1373	32028
18	126	w	0.019923131084937519538454	1	C	{C}	1258	28896
19	123	g	0.01328477594795075435	1	C	{C}	1139	26411
20	120	ð	0.006354146800349953585375	1	C	{d,F,C}	624	23031
21	119	c	0.00988788396231620780	1	C	{C}	1158	21969
22	118	ch	0.00846486491243506495	2	C	{C}	615	17929
23	116	eo	0.005189050687119595106236	2	V	{DP,3-2,V}	912	16214
24	117	k	0.00802259842864759934	1	C	{P,v,-,C}	896	14174
25	115	p	0.00744720079111580617	1	C	{l,P,-,C}	679	13372
26	122	y	0.010670002276399628142439	1	A	{A}	1220	11882
27	107	ll	0.00510669915915231179	2	C	{a,Xl,+,C}	327	10526
28	114	ȝ	0.003015111691962910070293	1	C	{C}	673	9511
29	111	nn	0.00471217077269121581	2	C	{a,N,+,C}	347	8451
30	178	E	0.002397537261921340798682	1		{M}	461	6537
31	110	ea	0.002564302784002089840068	2	V	{DP,3-1,V}	860	6387
32	264	N	0.002559399542833497938879	1		{M}	664	6044
33	105	ss	0.00263571507560788614	2	C	{a,F,-,C}	273	5108
34	88	ei	0.001895265391052778771543	2	V	{DP,3-5,V}	400	4703
35	109	ou	0.002419509683794441477327	2	V	{DP,2-5,V}	402	4631
36	121	ᵹ	0.00250490917537898113	1	C	{C}	466	4356
37	103	v	0.002482075594304959803555	1	A	{A}	516	4060
38	106	sch	0.001952622410735526712408	3	C	{p,F,-,C}	219	3805
39	98	ie	0.001355831091357274650619	2	V	{DP,5-3,V}	453	3332
40	112	th	0.003286811523509589857346	2	C	{d,F,C}	383	3123
41	87	dd	0.001082128869400179928582	2	C	{P,a,+,C}	160	3100
42	260	M	0.000899089133282436431768	1		{M}	179	3070
43	102	hƿ	0.000892555890653343775937	2	C	{l,F,C}	76	2577

44	85	tt	0.0011060795548963246664993	2	C	{a,P,-,C}	287	2568
45	90	sc	0.0014334804075045961115758	2	C	{C}	250	2562
46	99	bb	0.0013941591369066249204090	2	C	{C}	44	2498
47	113	æ	0.0012243984966626043658980	1	V	{ST,2,f,V}	694	2496
48	285	R	0.0007381186260919536566130	1		{M}	200	2287
49	91	ai	0.0010398104771704340157540	2	V	{DP,1-5,V}	216	2236
50	80	z	0.0004774410147826387608620	1	C	{a,F,+,C}	203	1839
51	63	rr	0.0003671752122665743565070	2	C	{a,Xc,+,C}	215	1425
52	94	sh	0.0005554839447885546762550	2	C	{p,F,-,C}	136	1387
53	65	q	0.0006245839919201942145850	1	C	{P,v,-,C}	91	1333
54	57	gg	0.0004188640752187078261680	2	C	{C}	70	1104
55	64	mm	0.0004848365095031090266300	2	C	{l,N,+,C}	86	1058
56	71	qu	0.0002341884376047060133020	2	C	{l,F,C}	61	895
57	78	ye	0.0001395414210205943044700	2	V	{DP,5-3,V}	105	819
58	41	au	0.0002642314693260608545180	2	V	{DP,1-5,V}	139	710
59	83	ey	0.0006518957136058969972510	2	V	{DP,3-5,V}	131	705
60	49	x	0.0003451456237405519109130	1	C	{C}	60	704
61	69	ff	0.0002647559831408229876480	2	C	{ld,F,-,C}	65	685
62	66	pp	0.0003371244941821975245300	2	C	{l,P,-,C}	67	639
63	79	ay	0.0004476952044365885761800	2	V	{DP,1-5,V}	117	572
64	55	cch	0.0002304963865045209790000	3	C	{C}	77	531
65	100	gh	0.0005774911425533115100410	2	C	{C}	122	531
66	27	hu	0.0000501018005977395731530	2	C	{l,F,C}	38	469
67	329	U	0.0002075077223035165275150	1		{M}	73	436
68	86	hw	0.0001666179187625678396380	2	C	{l,F,C}	33	436
69	28	eu	0.0001371092062552332079800	2	V	{DP,3-5,V}	56	400
70	70	þþ	0.0000687218801471423402270	2	C	{C}	25	371
71	56	ng	0.0001738409263822202699940	2	C	{C}	86	367
72	67	hh	0.0000653497066747652723850	2	C	{g,F,+,C}	57	362
73	30	oi	0.0001732115924379866517360	2	V	{DP,2-1,V}	66	348
74	45	ck	0.0001759477316428574575530	2	C	{C}	90	316
75	97	ȝw	0.0000532847904921767728820	2	C	{l,F,C}	29	302
76	332	uo	0.0000369082216906864000010	2		{M}	26	279
77	59	ee	0.0003095247078992034757420	2	V	{LN,3,f,V}	118	251

78	298	sk	0.000121657203650391932404	2		{M}	20	240
79	11	ph	0.000052029748277512582018	2	C	{ld,F,-,C}	10	236
80	84	ȝh	0.000037235050837113298407	2	C	{C}	73	227
81	73	ow	0.000156639928826990312758	2	V	{DP,2-5,V}	57	210
82	82	é	0.000120206353230048530756	1	V	{LN,3,V}	107	207
83	68	oe	0.000250409022243097630023	2	V	{DP,2-3,V}	54	191
84	43	ui	0.000122780542256955842622	2	V	{V}	60	180
85	35	ue	0.000108074093476606825581	2	V	{DP,5-3,V}	71	171
86	29	op	0.000039749270962297558700	2	V	{DP,2-4,V}	22	168
87	46	cc	0.000370124978359262811985	2	C	{C}	43	166
88	62	ᵹh	0.000066079357909527483336	2	C	{v,F,+,C}	50	162
89	39	gu	0.000011053412762253053406	2	C	{C}	13	159
90	36	ia	0.000055925127646252117208	2	V	{DP,5-1,V}	45	157
91	108	í	0.000078986806267774709529	1	A	{LN,5,A}	63	156
92	58	þh	0.000145780646520117174246	2	C	{l,F,C}	30	152
93	214	ᵹᵹ	0.000025249098675882301412	2		{M}	12	147
94	31	uy	0.000013645610460095821347	2	V	{V}	37	145
95	16	aw	0.000021282814810993298429	2	V	{DP,1-5,V}	18	142
96	24	ðð	0.000048691651653961753962	2	C	{d,F,+,C}	21	139
97	72	ó	0.000063913272646838375163	1	V	{LN,2,V}	44	136
98	271	O	0.000055531848507913938229	1		{M}	19	136
99	101	wh	0.000228002683095605863292	2	C	{l,Xc,+,C}	20	116
100	48	á	0.000054813884124534452991	1	V	{LN,1,V}	40	105
150	53	ǣ	0.000007967385657188035016	1	V	{LN,f,2,V}	12	16
151	21	cu	0.000005203254960740603318	2	C	{C}	9	16
152	89	ssch	0.000044832635534181054934	4	C	{p,F,-,C}	10	15
153	197	eou	0.000001926224771429112837	3		{M}	10	15
154	38	tþ	0.000037697235744531578880	2	C	{C}	7	14
155	219	hg	0.000001249958086754308332	2		{M}	9	13
156	268	NN	0.000002232913488343332765	2		{M}	8	13
157	18	iw	0.000002621254173317332547	2	V	{DP,V}	4	12
158	218	hȝ	0.000006728966026705936179	2		{M}	8	11
159	244	ih	0.000005019194975054107294	2		{M}	8	11
200	7	ua	0.000011883437406506082137	2	V	{DP,5-1,V}	4	4

201	336	vu	0.000001726512901987524925	2		{M}	4	4
202	352	yh	0.000018338121648031862703	2		{M}	4	4
203	344	ƿu	0.000004243485474584323569	2		{M}	2	4
204	346	ww	0.000000589015739973111412	2		{M}	2	4
205	267	nm	0.000002961909839464486765	2		{M}	1	3
206	243	ig	0.000000686031429124415695	2		{M}	2	3
207	321	þw	0.000025480547950405458647	2		{M}	2	3
208	337	vy	0.000000751128507299646458	2		{M}	2	3
209	333	uw	0.000001945008136617371529	2		{M}	2	3
240	198	et	0.000032144005143040822882	2		{M}	1	2
241	331	uh	0.000000272115494738909888	2		{M}	2	2
242	196	eoo	0.000000320033514477660953	3		{M}	2	2
243	188	eha	0.000014817080288432127941	3		{M}	2	2
244	195	eoie	0.000000287455857559873462	4		{M}	1	2
245	186	eeo	0.000000278678839358369840	3		{M}	1	2
246	334	uƿ	0.000000342291623365668646	2		{M}	2	2
247	3	eaa	0.000000272963104665908304	3	V	{DP,3-1,V}	2	2
248	176	ðt	0.000000323019847954557568	2		{M}	2	2
249	311	tð	0.000010086353474744569430	2		{M}	2	2

7.10.2. Texts-litterae statistics

Norm tokens: the number of tokens divided by the number of slots

Potential types: the number of slots present in the given text, in which the littera can be expected to appear (i.e. it has at least one occurrence in the slot in the whole corpus)

Types ratio: potential types divided by the number of types

text id	littera	tokens	Norm tokens	types	potential types	Types ratio
8	hƿ	27	0,00236	7	14	0,50000
8	ƿ	407	0,03555	122	184	0,66304
280	hw	1	0,00002	1	19	0,05263
280	w	1544	0,03665	276	391	0,70588
280	ƿ	1	0,00002	1	307	0,00326
295	qu	105	0,00361	18	29	0,62069
295	w	914	0,03140	166	317	0,52366

301	qu	4	0,00012	1	16	0,06250
301	w	8	0,00023	6	234	0,02564
301	p	1368	0,03995	162	241	0,67220
301	ph	68	0,00199	16	18	0,88889
1100	hw	214	0,00483	25	29	0,86207
1100	qu	1	0,00002	1	26	0,03846
1100	w	1277	0,02881	331	470	0,70426
1100	p	6	0,00014	6	414	0,01449

7.10.3. Rare uses

mss: LAEME ids of text in which instances of the rare use are found

mss total: the number of texts in which the combination of lexel/word class occurs

mss ratio: the ratio of texts in which the rare use occurs

rank	littera	pos	mss		mss_total	mss_ratio	lexel	word_class
1	f	1	{1100}		107	0.00934579439252336	if	c
2	f	5	{7}		88	0.0113636363636364	-self	xs
3	f	1	{222}		79	0.0126582278481013	soul	n
4	f	1	{296}		75	0.0133333333333333	see	vi
5	f	1	{249}		73	0.0136986301369863	shall	vps12
50	f	3	{173,280,301,2000}		49	0.0816326530612245	above	{av,pr}
54	f	3	{301}		12	0.0833333333333333	lording	n
55	f	2	{64}		12	0.0833333333333333	offer	vi
56	f	3	{159,185,298,301,304,2000,2001}		83	0.0843373493975904	heaven	n

57	f	3	{65,277,278,291,301,2000}		71	0.0845070422535211	woman	n
100	f	3	{4,1300,1800}		20	0.15	efning	n
101	f	3	{184,1400,2000}		19	0.157894736842105	reeve	n
102	f	3	{3,159,280,295,297,298,301,1400,1800,2000}		61	0.163934426229508	give	vi
103	f	3	{173}		6	0.166666666666667	beaver	n
104	f	4	{245}		6	0.166666666666667	clifer	n
1	ea	1	{173}		139	0.00719424460431655	all	{aj,av,cj}
2	ea	2	{246}		123	0.00813008130081301	that	{av,cj}
3	ea	2	{276}		121	0.00826446280991736	have	vps
4	ea	2	{277}		103	0.00970873786407767	not	n
5	ea	2	{173}		93	0.010752688172043	when	{av,cj,RT}
50	ea	1	{173}		40	0.025	alder-	xp
54	ea	2	{142,155,172}		119	0.0252100840336134	day	n
55	ea	1	{6,64}		79	0.0253164556962025	each	{aj,pn}
56	ea	1	{280}		39	0.0256410256410256	benot	vsjpt
57	ea	2	{278}		38	0.0263157894736842	sit	vSpt
100	ea	2	{260,1000}		50	0.04	faran	vi
101	ea	1	{6,131,143,156}		97	0.0412371134020619	as	{av,cj,pr,RT}
102	ea	2	{65}		24	0.0416666666666667	elsewhere	av

103	ea	2	{118}		24	0.0416666666666667	low	av
104	ea	2	{272}		24	0.0416666666666667	last	n

7.10.4. N-grams

formid	morphid	pos	pre	pre_tags	main_tags	main	main_pos_tags	post	post_tags
113	525	1			{C}	ᵹ	{mI}	o	{ST,2,b,r,V}
113	525	2	ᵹ	{C}	{ST,2,b,r,V}	o	{}	d	{a,P,+,C}
113	525	3	o	{ST,2,b,r,V}	{a,P,+,C}	d	{mF}	_	{A}
112	525	1			{C}	gh	{mI}	o	{ST,2,b,r,V}
112	525	2	gh	{C}	{ST,2,b,r,V}	o	{}	d	{a,P,+,C}
112	525	3	o	{ST,2,b,r,V}	{a,P,+,C}	d	{mF}	_	{A}
117	525	1			{C}	g	{mI}	oe	{DP,2-3,V}
117	525	2	g	{C}	{DP,2-3,V}	oe	{}	d	{a,P,+,C}
117	525	3	oe	{DP,2-3,V}	{a,P,+,C}	d	{mF}	_	{A}
117	525	4	d	{a,P,+,C}	{A}	_	{}	_	{A}

7.10.5. Chunks

morphid	formid	pos	char	id
3	32578	1_2	æ	10646
3	32578	2_3	æh	10976
3	32578	3_4	ht	71071
3	32578	4_5	t	146108
78	44233	1_2	bea	20114
78	44233	2_3	ear	38916
78	44233	3_4	re	126793
119	1895	1_2	bl	20613

119	1895	2_3	ly	94273
119	1895	3_4	yn	173597
119	1895	4_5	nd	102966
119	1895	5_6	d	27005
140	37208	1_2	br	21382
140	37208	2_3	ri	130771
140	37208	3_4	ih	77298
140	37208	4_5	ht	71276
140	37208	5_6	te	148911

7.10.6. Special features

id	**char**	**text**	**morpheme_id**	**feature**
48912	f	163	216896	reconstruction
132635	h	163	216897	capital
167729	a	286	217699	u+superscript
164871	a	286	226954	r+superscript
146095	s	286	226957	capital
48093	c	272	271520	insertion
48774	u	272	260754	reconstruction
163662	i	272	261596	r+superscript
48195	e	276	305713	insertion
48258	i	276	305770	insertion
163144	i	276	305801	r+superscript

7.10.7. Source forms

id	**period**	**dialect**	**lexel**	**form**	**word_class**	**souce**
86	OE		beaver	befer	n	DOE
185	OE		blood	blód	n	DOE
92	OE		boat	bát	n	DOE
48	OE		boneless	bán	aj	DOE
210	OE		carbunclestone	carbunculus	n	DOE

197	OE		crafty	cræftig	aj	DOE
425	PDE		dark	dark	aj	OED
73	OE		dark	deorc	aj	DOE
42	OE	-	earl	eorl	n	
41	OE	-	earl	heorl	n	
147	OE		father	fæder	n	DOE
204	OE		fiend	féond	n	DOE
13	OE	-	fight	feohtan	v	
2	OE	Kentish	fire	fur	n	
158	OE		fish	fisc	n	DOE
44	OE	-	flee	fléon	v	
369	OE		foam	fám	n	DOE
368	OE		folk	folc	n	DOE
314	OE		friend	fréond	n	DOE
75	OE		house	hús	n	DOE
78	OE		child	cild	n	DOE

7.11. JSON data samples (#170, *A Sermon on the Nativity*)

7.11.1. Inventory of litterae

The sample comprises the litterae *a*, *cch* and *ea*.

```
[{
        "str": "a",
        "tokens": 214,
        "types": 71,
        "normTokens": 0.06465256797583081571,
        "rareSlots": 4,
        "mssRatio": 0.14225504542405932727,
        "litAvg": [{
            "label": "global",
            "normTokens": 0.06127317575010526898
        }],
        "specialFeatures": [{
            "str": null,
            "tokens": 210
        }, {
            "str": "capital",
            "tokens": 2
        }, {
            "str": "r+superscript",
            "tokens": 1
        }, {
            "str": "reconstruction",
            "tokens": 1
        }]
    },
{
        "str": "cch",
        "tokens": 7,
        "types": 6,
        "normTokens": 0.00211480362537764350,
        "rareSlots": 5,
        "mssRatio": 0.01633993531060058000,
        "litAvg": [{
            "label": "global",
            "normTokens": 0.000230496386504520979000
        }],
        "specialFeatures": [{
            "str": null,
            "tokens": 7
        }]
    },
{
        "str": "ea",
        "tokens": 3,
        "types": 3,
        "normTokens": 0.00090634441087613293,
        "rareSlots": 1,
        "mssRatio": 0.01282051282051280000,
        "litAvg": [{
            "label": "global",
            "normTokens": 0.002564302784002089840068
        }],
        "specialFeatures": [{
            "str": null,
            "tokens": 3
        }]
    }]
```

7.11.2. Sets

The sample comprises two sets, namely {*ea*, *eo*} and {*c*, *cch*}

```
[{
        "simple": ["ea", "eo"],
        "types": 1,
        "tokens": 2,
        "members": [{
            "str": "ea",
            "tokens": 1
        }, {
            "str": "eo",
            "tokens": 1
        }]
    },
{
        "simple": ["c", "cch"],
        "types": 1,
        "tokens": 3,
        "members": [{
            "str": "cch",
            "tokens": 2
        }, {
            "str": "c",
            "tokens": 1
        }]
    }]
```

7.11.3. Items

The item list for {q} in text #170 comprising QUEEN/N (1), KNOW/VSPT (1) and CWEÞAN/VPSP (1).

```
[{
        "morphid": 6439,
        "pos": 1,
        "lexel": "queen",
        "wordClass": "n",
        "litterae": [{
            "str": "q",
            "tokens": 1
        }]
    }, {
        "morphid": 24292,
        "pos": 1,
        "lexel": "know",
        "wordClass": "vSpt",
        "litterae": [{
            "str": "q",
            "tokens": 1
        }]
    }, {
        "morphid": 26605,
        "pos": 1,
        "lexel": "cweYan",
        "wordClass": "vpsp",
        "litterae": [{
            "str": "q",
            "tokens": 2
        }]
    }]
```

7.11.4. Map data

The map of EARTH/N (1), only data for texts #261, #214 and #2001 is included.

```
[{
        "id": 261,
        "litterae": [{
            "str": "eo",
            "tokens": 6
        }],
        "tokens": 6
    }, {
        "id": 214,
        "litterae": [{
            "str": "e",
            "tokens": 1
        }],
        "tokens": 1
    }, {
        "id": 2001,
        "litterae": [{
            "str": "o",
            "tokens": 7
        }, {
            "str": "eo",
            "tokens": 6
        }, {
            "str": "e",
            "tokens": 1
        }],
        "tokens": 14
    }]
```

7.12. Statistics

7.12.1. Mixed slots ratios

The table shows the ratio of slots which in which two or more litterae are used interchangeably, calculated for each text in the table. The texts with the highest ratio appear on top. The table does not covers only texts with 1000 slots or longer.

rank	text id	total	mixed	ratio	manuscript
1	278	6141	1387	0.225858980622049	London, British Library, Cotton Caligula A.ix, part 1 (*Laȝamon A II*)
3	1400	2725	507	0.18605504587156	Cambridge University Library Ff.II.33 (*Bury documents*)
5	277	6432	1107	0.172108208955224	London, British Library, Cotton Caligula A.ix, part 1 (*Layamon A I*)
6	246	5492	901	0.164056809905317	Cambridge, Trinity College B.14.39, hand A
7	2000	8324	1346	0.161701105237866	London, Lambeth Palace Library 487 (*Lambeth Homilies A*)
8	64	7221	1162	0.160919540229885	London, British Library, Stowe 34, Hand A (*Vices and Virtues*)
9	285	6710	1056	0.157377049180328	Oxford, Bodleian Library, Laud Misc 108 (*Havelok*)
10	298	7856	1223	0.155677189409369	Edinburgh, Royal College of Physicians, MS of Cursor Mundi (*Northern Homily Collection*)
12	1300	8757	1286	0.146853945415097	Cambridge, Trinity College B.14.52, hand B (*Trinity Homilies*)
14	173	7317	1054	0.144048107147738	Worcester Cathedral, Chapter Library F 174 (*Ælfric's Grammar and Glossary*)
15	280	6074	862	0.141916364833717	London, British Library, Cotton Otho C xiii (*Laȝamon B*)
16	291	8720	1219	0.139793577981651	London, British Library, Arundel 57 (containing the *Ayenbyte of Inwyt*)
17	296	7347	1022	0.139104396352253	Edinburgh, Royal College of Physicians (*Cursor Mundi*)
18	2002	7803	1057	0.135460720235807	Oxford, Bodleian Library, Digby 86
19	249	3026	405	0.133840052875083	Cambridge, Trinity College B.14.39, hand D

20	304	1856	245	0.132004310344828	London, British Library, Cotton Claudius D iii (*Benedictine Rule*)
21	6	3329	439	0.131871432862722	London, British Library, Egerton 613 (*Poema Morale e*)
22	7	3428	448	0.130688448074679	London, British Library, Egerton 613 (*Poema Morale E*)
23	1600	9612	1247	0.12973366625052	Oxford, Bodleian Library Laud Misc 108, part 1 (*South English Legendary*)
24	1100	8237	1051	0.127594998178949	Oxford, Jesus College 29
25	248	1609	203	0.12616532007458	Cambridge, Trinity College B.14.39, hand C
26	2001	4927	621	0.126040186726203	London, Lambeth Palace Library 487 (Lambeth Homilies B)
27	286	8537	1067	0.124985357854047	Cambridge, Corpus Christi College 145 (*South English Legendary*)
28	155	5917	734	0.124049349332432	Cambridge, Corpus Christi College 444 (*Exodus*, *Genesis*)
29	295	6065	747	0.123165704863974	London, British Library, Cotton Vespasian A.iii (*Cursor Mundi*)
30	3	3787	466	0.12305254819118	London, British Library, Cotton Caligula A ix (*The Owl and the Nightingale*)
31	169	1945	238	0.122365038560411	Oxford, Merton College 248 (short pieces)
32	1200	5744	702	0.122214484679666	Cambridge, Trinity College B.14.52, hand A (*Trinity Homilies*)
33	149	1958	239	0.122063329928498	Oxford, Bodleian Library, Laud Misc 636 (*The Peterborough Chronicle*)
34	297	7507	910	0.121220194485147	Edinburgh, Royal College of Physicians (*Cursor Mundi*)
35	65	3758	449	0.119478445981905	London, British Library, Stowe 34, hand B (Vices and Virtues)
36	142	2396	285	0.118948247078464	Oxford, Bodleian Library, Laud Misc 471 (*The Kentish Sermons*)
37	276	6250	718	0.11488	Cambridge, Gonville and Caius 234/120, pp. 1-185 (*Ancrene Riwle*)
38	247	3953	445	0.112572729572477	Cambridge, Trinity College B.14.39, hand B
40	182	2667	298	0.111736032995876	London, Dulwich College MS XXII (*La Estorie del Euangelie*)

42	2	5267	574	0.108980444275679	London, British Library, Cotton Caligula A ix (*The Owl and the Nightingale*)
43	8	3208	339	0.105673316708229	Oxford, Bodley Digby 4 (*Poema Morale D*)
45	161	1838	193	0.105005440696409	Oxford, Bodleian Library, Additional E.6, roll (*An Exposition of the Pater Noster I*, *The XV signs before Doomsday*)
46	5	2698	279	0.103409933283914	London, Lambeth Palace Library 487 (*Poema Morale L*)
47	245	8505	865	0.101704879482657	London, British Library, Cotton Nero A xiv (*Ancrene Riwle*)
48	9	3364	338	0.100475624256837	Oxford, Jesus College 29 (*Poema Morale J*)
49	271	2268	227	0.100088183421517	London, British Library, Cotton Vitellius D iii (*Floriz and Blauncheflur*)
51	170	1321	130	0.0984102952308857	Worcester Cathedral, Chapter Library Q 29 (*A sermon on the Nativity*)
53	301	4379	417	0.0952272208266728	Oxford, Bodleian Library, Junius 1 (*The Orrmulum*)
54	273	8356	790	0.0945428434657731	London, British Library, Cotton Cleopatra C.vi (*Ancrene Riwle*)
55	188	1871	175	0.0935328701229289	London, British Library, Cotton Julius A v (*A Ballad on the Scottish Wars*)
57	123	6710	606	0.0903129657228018	London, British Library, Cotton Titus D xviii (*St Katherine*)
58	282	3255	292	0.0897081413210445	Oxford, Bodleian Library, Laud Misc 108 (*The Debate between the Body and Soul (theme)*)
61	1701	2260	201	0.0889380530973451	Cambridge, Trinity College 43 (B.1.45) and BL Cotton Cleopatra C vi (short pieces)
62	4	3356	296	0.0882002383790226	Cambridge, Trinity College B.14.52 (*Poema Morale T*)
63	260	8208	722	0.087962962962963	London, British Library, Royal 17 A xxvii (*Sawles Warde*, *St Katherine*, *St Margaret*)
64	172	2617	227	0.0867405426060374	Worcester Cathedral, Chapter Library F 174 (short rhythmic prose text, *The Debate between the Body and Soul (theme)*)
65	1800	4371	376	0.0860215053763441	London, British Library, Cotton Nero A xiv (*miscellaneous religious pieces*)
66	118	8314	713	0.0857589607890305	BL Cotton Titus D xviii (*Ancrene Riwle*)

67	158	3257	277	0.0850475898065705	Oxford, Bodleian Library, Bodley 652 (*Iacob and Iosep*)
68	121	6007	504	0.0839021142000999	London, British Library Cotton Titus D xviii (*Hali Meiðhad*)
71	120	3934	321	0.081596339603457	London, British Library Cotton Titus D xviii (*Sawles Warde*)
72	122	3255	264	0.0811059907834101	London, British Library, Cotton Titus D.xviii (*Þe Wohunge of Ure Lauerd*)
73	10	3204	257	0.0802122347066167	Cambridge, Fitzwilliam Museum, McClean 123 (*Poema Morale M*)
74	1000	7630	606	0.0794233289646134	Oxford, Bodleian Library, Bodley 34 (*Andrene Riwle B*)
75	119	5775	437	0.0756709956709957	London, British Library Cotton Titus D xviii (*Ancrene Riwle*)
76	66	1581	118	0.0746363061353574	Maidstone Museum A.13 (*Proverbs of Alfred*, *The names of the Old English letters*)
77	262	4960	368	0.0741935483870968	London, British Library, Royal 17 A xxvii (*St Juliana, St Margaret*)
78	220	2552	189	0.0740595611285266	Oxford, Bodleian Library, Digby 86 (*Dame Sirith*)
79	261	5607	409	0.0729445336186909	London, British Library, Royal 17 A xxvii (*On Lofsong of Ure Lefdi / Oreisun of Seinte Marie, Sawles Warde, St Juliana*)
81	137	1649	118	0.0715585203153426	London, British Library Arundel 248 (short pieces)
83	222	1460	104	0.0712328767123288	Oxford, Bodleian Library, Digby 86 (*The Debate between the Body and Soul (theme)*)
84	214	1980	141	0.0712121212121212	Oxford, Bodleian Library, Digby 86 (*Iesu dulcis memoria, The XI Pains of Hell*)
85	150	3763	263	0.0698910443794845	London, BL Arundel 292 (*The Bestiary*)
86	140	1182	81	0.0685279187817259	Cambridge, Emmanuel College 27 (miscellaneous religious pieces)
87	189	1729	118	0.0682475419317525	London, Lambeth Palace 487 (*On Ureisun of Ure Loverde*)
88	272	8798	598	0.0679699931802682	Cambridge, Corpus Christi College 402 (*Ancrene Wisse*)

93	242	1592	102	0.064070351758794	London, British Library, Cotton Caligula A ix (*The Latemest Day*)
95	229	1279	81	0.0633307271305708	Oxford, Corpus Christi College 59 (prayers)
96	160	1537	95	0.0618087182823683	Oxford, Bodleian Library Add E.6, roll (*Sayings of St Bernard*)
97	218	2359	145	0.0614667231877914	Oxford, Bodleian Library, Digby 86 (*The Proverbs of Alfred, The Proverbs of Hending*)
104	275	1492	85	0.056970509383378	London, British Library, Cotton Cleopatra C.vi (*Ancrene Riwle*)

7.12.2. Alternatives

This table presents global statistics showing the degree of interchangeability of individual litterae.

average set size: the average number od litterae found in one set containing the described littera ("1" means that the littera is never used interchangeably)

min: the minimum set size

max: the maximum set size

rank	**littera**	**average set size**	**min**	**max**	**Slot types**
1	p	1.1424375917767988	1	4	681
2	b	1.1626248216833096	1	10	701
3	r	1.2651515151515152	1	7	3168
4	l	1.2822862129144852	1	10	2292
5	m	1.4547169811320755	1	11	1060
20	rr	2.2798165137614679	1	16	218
21	e	2.3954862976894143	1	18	9305
22	k	2.4044198895027624	1	16	905
23	c	2.4243986254295533	1	17	1164
24	bb	2.4772727272727273	1	10	44
25	ă	2.5000000000000000	2	3	2
50	x	3.0655737704918033	1	16	61
51	mn	3.2500000000000000	2	5	4
52	gg	3.2857142857142857	1	16	70

53	ss	3.3321428571428571	1	18	280
54	sc	3.4488188976377953	1	17	254
55	ð	3.4960254372019078	1	17	629
100	tth	4.5000000000000000	3	8	4
101	ui	4.5081967213114754	1	16	61
102	ó	4.6000000000000000	1	11	45
103	ȝ	4.6523668639053254	1	21	676
104	ä	4.6666666666666667	4	5	3
105	sck	4.6666666666666667	4	6	3
200	hw	8.6666666666666667	2	21	33
201	eoi	8.8000000000000000	5	18	5
202	ph	8.8666666666666667	2	21	30
203	uh	9.0000000000000000	9	9	2
204	vy	9.0000000000000000	4	14	2
205	opp	9.0000000000000000	2	16	2
225	fw	12.0000000000000000	11	13	2
226	chs	12.2500000000000000	7	17	4
227	hv	12.5000000000000000	8	17	2
228	ðd	12.6666666666666667	5	17	3
229	ȝth	14.5000000000000000	14	15	2
230	wȝ	14.5000000000000000	11	21	4
231	shs	16.0000000000000000	15	17	2

7.12.3. Slot variability

This table presents selected rows from an overview of slots ordered by the number of litterae appearing in them along with the sets of specific litterae.

rank	**morphid**	**pos**	**no of litterae**	**lexel**	**word class**	**set**
1	8356	1	21	when	{av,cj,RT}	{_,ȝ,ȝw,h,hh,hu,hw,hƿ,q,qu,qv,þƿ,uu,v,vu,vv,w,ƿ,wȝ, wh,ƿh}
2	29378	2	18	be	vps23	{e,ea,ee,ei,eo,eoi,i,ie,o,oe,oei,oi,ss,u,ue,uo,y,ye}

3	4293	4	17	flesh	n	{_,c,cs,hc,hs,ch,chs,s,sc,sh,shs,sch,schs,ss,ssc,ssch,xs}
4	28884	1	17	-er	xs	{_,a,æ,e,E,ea,ee,eo,eou,i,iu,o,ou,u,U,y,ye}
5	8351	1	17	what	{aj,av,cj,in,pn,pr,RT}	{ʒ,ʒw,h,hu,hv,hw,hƿ,q,qu,þ,v,w,ƿ,wh,ƿh,ƿv,ww}
6	29387	3	17	may	vpt13	{_,c,cch,ʒ,ʒh,g,gh,ᵹ,h,hh,hs,ch,chc,s,þ,xis,y}
7	28921	2	17	-th	xs	{_,cþ,d,ð,ðd,ðð,ðh,h,hð,ht,hþ,t,th,þ,þh,tt,y}
8	8244	2	16	either	{aj,cj,pn}	{_,a,æi,ai,aie,au,ay,e,ei,ey,i,o,oi,ou,oƿ,oƿƿ}
9	29163	1	16		P13NF	{_,ʒ,g,gg,gh,ᵹ,ᵹh,h,ch,s,sc,sg,sh,sch,þ,y}
10	8293	3	16	nigh	{aj,av,pr}	{_,cs,ʒ,ʒh,g,gh,ᵹ,h,hg,hᵹ,ch,ks,rr,þ,x,xs}
300	28893	2	8	-hood	xs	{_,a,e,ea,ee,ei,i,o}
301	4795	5	8	ha:lga	n	{ʒ,g,gh,ᵹ,h,ch,w,ƿ}
302	26733	1	8	give	vSpp	{_,ʒ,g,gh,ᵹ,ih,þ,y}
303	4912	2	8	heart	n	{e,E,eo,i,ie,o,oe,u}
304	26581	2	8	burn	vpsp	{_,a,e,E,ea,eo,i,u}
305	21555	5	8	follow	vpt	{_,ʒ,g,ᵹ,h,ch,w,ƿ}
306	28960	2	8	un-	xp	{a,i,o,ou,ow,u,v,w}
307	21794	2	8	go	vps	{_,a,aa,e,ea,o,oi,u}
308	29146	1	8		P22N	{_,ʒ,g,gh,ᵹ,h,þ,y}
309	363	2	8	fair	aj	{æi,ai,ay,e,eai,eay,ei,ey}
310	24339	2	8	run	vSpt	{_,æ,e,eo,o,ou,u,v}
900	27022	3	5	swinge	vSpp	{_,e,i,o,u}
901	6842	1	5	ship	n	{s,sc,sh,sch,ss}
902	28960	3	5	un-	xp	{_,m,n,N,nn}
903	6851	1	5	shire	n	{s,sc,sh,shc,sch}
904	29133	2	5		P11G	{_,e,i,í,y}
905	6864	1	5	shower	n	{s,sc,sh,sch,ss}
906	27392	5	5	bletsian	vpp	{c,cc,s,sc,ss}
907	1859	2	5	e:aYe	av	{a,e,ea,i,ie}
908	22913	2	5	teach	vps	{a,æ,e,ea,eæ}

909	6872	1	5	shroud	n	{s,sc,sh,sch,ss}
910	22119	2	5	live	vps	{e,eo,i,u,y}
1800	3863	4	4	devil	n	{_,e,i,o}
1801	23076	2	4	turn	vpt	{e,o,u,U}
1802	3868	3	4	dew	n	{u,v,w,p}
1803	24712	2	4	listen	v-imp	{_,e,i,u}
1804	7386	4	4	thane	n	{_,g,ȝ,N}
1805	29177	2	4		P12<pr	{e,é,ee,i}
1806	7389	2	4	thank	n	{a,e,eo,o}
1807	22656	1	4	ship	vpt	{sc,sh,sch,ss}
1808	7393	1	4	thigh	n	{th,þ,þh,z}
1809	29383	4	4	may	vps13	{_,g,ȝ,ȝȝ}
1810	7393	2	4	thigh	n	{e,ei,eo,i}

7.13. Littera correspondences between texts #301 and #246

The Ormulum (#301)	**The corresponding litterae in text #246**
_	{_,a,au,c,cg,ck,d,dd,e,E,ea,ei,eo,f,ff,g,gk,h,i,ie,ii,k,l,m,M,n,N,nn,o,oe,oo,r,R,rr,s,t,þ,u,v,w,p,y,z}
a	{_,a,ai,au,d,e,E,ea,ei,eo,ey,i,l,o,oi,oo,u}
á	{_,o,oo}
ă	{a}
ä	{e}
æ	{_,a,ai,e,E,ea,ee,ei,eo,o}
ǣ	{a,e,E,ei}

b	{b}
bb	{b,bb,u}
c	{_,c,ch,k,q,s}
cc	{_,c,ch}
cch	{ch}
d	{_,d,dd,g,gk,hit,ht,k,l,t,þ}
ð	{_,þ}
dd	{_,d,dd,t}
e	{_,a,ai,au,e,E,ei,eo,ey,h,i,ie,l,n,N,o,oe,oi,s,ss,u,y}
E	{e,E,ei,i}
é	{e}
ĕ	{eo}
ȅ	{e,ei}
eo	{e,E,ei,eo,o,oe,oi,u}
eoo	{E}
f	{_,b,bb,f,ff,o,ph,u,v,w,ƿ}
ff	{_,f,ff,s,u,w,ƿ}
g	{_,c,cg,ck,ȝ,g,gh,gk,h,k,þ}
gg	{_,g}
gh	{_}
ᵹ	{_,ȝ,g,h,ch,i,þ,u}
ᵹᵹ	{_,g,gg,þ}
ᵹh	{_,h,þ,u}
h	{_,c,h,w,ƿ,y}
hᵹh	{_,þ}
hh	{_,c,cs,ch,s,th,þ}
ch	{ch,k}
i	{_,a,e,ei,eo,i,ie,ii,l,o,u,ui,v,y}
í	{_,eo,i,o,u}
ĭ	{i}
k	{_,c,ch,k}

l	{_,e,l,ll}
ll	{_,a,l,ll,rl}
lll	{l,ll}
m	{m,M,mm,n}
M	{m,M,mm,n}
mm	{m,M,mm,n}
n	{_,e,i,n,N,nn,r}
N	{n,N}
ng	{n}
nn	{_,g,n,N,nn}
NN	{_,n,N,nn}
o	{_,a,au,e,ei,eo,i,o,oe,ohi,oi,oo,ou,u}
ó	{o,ohi,oi}
ő	{e,ehi,ei,o}
oo	{e,eo,o}
op	{e,o}
p	{p}
pp	{p,pp}
r	{_,E,r,R,rr}
rr	{E,r,R,rr}
s	{_,a,f,s,sc,ss}
sh	{c,f,ch,s,sc,sl}
ss	{_,e,i,s,ss}
t	{_,d,dd,N,s,t,th,þ,tt}
þ	{_,cþ,d,st,t,þ}
þþ	{_,d,t,þ}
tt	{_,d,N,t,tt}
u	{_,a,o,ou,u,U,v}
ú	{ou,u,v}
v	{o,u,v}
w	{v,w,p}

p	{_,e,u,v,w,p}
ph	{hu,v,w,p}
pp	{_}
x	{cc,x}
y	{i}

7.14. Programming languages and resources

Segmentation script	Python 3 (https://www.python.org/)
Database	Postgres 9.2 (https://www.postgresql.org/)
Interface	Angular 7 (https://angular.io/)
Maps	OpenLayers library (https://openlayers.org/)
Networks	Vis.js library (https://visjs.org)