English Language and Linguistics (2026), **30:** 538–560
doi:10.1017/S1360674326100859

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Exploring the *MetaLing Corpus*: spelling variation and discourse in Early Modern English texts on language

Angela Andreani and Vahid Asadi

Università degli Studi di Milano, Italy
Corresponding author: Angela Andreani; email: angela.andreani@unimi.it

(Received 15 January 2026 UTC; revised 05 June 2026 UTC; accepted 09 July 2026 UTC)

Abstract

This article explores the analytical potential of the *MetaLing Corpus*, a collection of Early Modern English texts on language dating from 1500 to 1700. Combining corpus-driven and corpus-based approaches, it demonstrates both the opportunities and the challenges of working with a corpus currently under development and reflects on the methodological implications of a minimalist corpus compilation strategy adopted under project constraints. Following an outline of the principles underlying data selection and corpus construction, the article presents two exploratory case studies focusing respectively on spelling and discourse variation. The analysis of spelling practices points to the interaction between authors and typesetters as a factor shaping orthographic choices and potentially influencing the development of linguistic reflection. At the level of discourse, the analysis uncovers a spectrum of educational orientations across the texts, indicating the presence of an additional discourse domain alongside language analysis and description. This finding contributes to a preliminary understanding of the corpus' thematic organisation and informs its future development in terms of annotation. More broadly, the study highlights the interpretive possibilities offered by the corpus and the methodological challenges posed by early printed materials. By demonstrating how corpus-based methods can be combined with close reading approaches, the article contributes to ongoing discussions in historical corpus linguistics concerning the analysis, interpretation and digitisation of Early Modern English texts.

Keywords: *MetaLing Corpus*; historical corpus linguistics; spelling variation; education; Early Modern English

1. Introduction

This article stems from research undertaken as part of the *MetaLing Corpus* project (October 2023 – February 2026), dedicated to investigating the metalanguage of linguistics in Early Modern English. The goal of the project was the creation of a corpus of texts including language-related discussions from 1500 to 1700, in order to study the development of linguistic terminology and ideas, and to explore epistemic transfers across texts and genres in a period predating the establishment of philology and linguistics as academic disciplines (Andreani & Russo 2023). Over this period, English underwent substantial changes in its lexicon, phonology, spelling and morphosyntax. Equally significant were changing

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

perceptions of the language itself, as attitudes towards English evolved from Ascham's characterisation of it as dramatically inadequate: 'In the Englysh tonge … euery thinge in a maner so meanly, bothe for the matter and handelynge, that no man can do worse' (Ascham 1545), to increasingly confident assessments of its expressive potential, such as William Lisle's in his preface to *A Saxon Treatise*, published in 1623: 'our language is improued aboue all others now spoken by any nation, and became the fairest, the nimblest, the fullest … yea able to expresse any hard conceit whatsoeuer with great dexterity' (cit. in Barber 2005: 51). Combining corpus-driven and corpus-based methodological approaches, the article presents an exploratory analysis of the *MetaLing Corpus* aimed at illustrating its analytical possibilities, the types of information it can yield and its potential limitations. Focusing on spelling and discourse variation as complementary sites of linguistic standardisation, norm formation and epistemic authority, the study demonstrates how orthographic choices and discourse patterns can illuminate both the material conditions of language use and the circulation of linguistic knowledge.

2. Data selection and corpus construction

The *MetaLing Corpus* was developed using Alston's *Bibliography of the English language from the invention of printing to the year 1800* as a key reference work (Alston 1974). This multi-volume bibliography encompasses a wide range of genres and language-related phenomena, remaining a foundational resource for research in English language studies and the history of linguistics. Owing to both the project's time constraints and the limited availability of some volumes of the *Bibliography*, the selection of sources was intentionally restricted to a manageable set of texts, with a specific linguistic phenomenon chosen as the point of departure. The study initially focused on language variation, drawing on volume IX of Alston's bibliography, devoted to *English dialects, Scottish dialects: Cant and vulgar English.* Selecting works published between 1500 and 1700, we compiled an initial list of sources that formed what we termed 'Subcorpus A: Variation'. During this phase, volumes III and VI were also accessible, providing the basis for the creation of two further comparable subcorpora focusing respectively on miscellaneous works on the English language and on pronunciation (Andreani & Russo 2026).

A detailed spreadsheet was compiled to record metadata and document the various stages of the work. The information recorded for each source included author name, title, publication date, format, repository location, associated Alston volume and entry number, and provenance identifiers such as OTA (*Oxford Text Archive*) codes and EEBO (*Early English Books Online*) references. This spreadsheet laid the groundwork for metadata collection and informed subsequent steps in encoding and publication.[1] Most digitised texts were sourced from the *Oxford Text Archive*, which provides materials in various formats, including plain text and TEI-XML. Some texts unavailable in digital form were transcribed using the Transkribus platform, which combines text recognition with machine learning (Kahle *et al*. 2017). While Transkribus assisted initial transcription, the results required extensive human verification; therefore, most sources for which a digital transcription was not available were excluded in this first phase. The resulting corpus was uploaded to Sketch Engine (Kilgarriff *et al*. 2004, 2014), accessible through the library system of the University of Milan.

3. Research questions and methodology

As the corpus consists of printed texts, we expected a relatively advanced degree of spelling regularisation, which was indeed observed. Corpus creation nevertheless revealed that what

[1] See also Russo (2026).

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

may appear to the philologist as a largely regularised linguistic profile presents in fact significant challenges for the corpus linguist seeking to automate searches (Baron & Rayson 2009; Archer *et al.* 2015). In addition, because the corpus was compiled according to external criteria, we were interested in assessing the 'aboutness' of our texts on the basis of their linguistic features and in exploring the additional discourses and themes that corpus-based analysis might reveal. To this end, we employed word frequency list and keyword analysis as foundational techniques (Sinclair 1991; Tribble & Jones 1997; Baron & Rayson, 2009). The discussion that follows is organised around two case studies focusing respectively on spelling variation and discourse variation.

4. Spelling variation in the *MetaLing Corpus*

Spelling inconsistencies within large collections of texts or corpora significantly affect the results of linguistic analysis, an issue that is especially pronounced in Early Modern English texts. The existence of varied spellings has been indicated to impact the reliability of keyword analysis (Baron & Rayson, 2009), part-of-speech annotation (Rayson *et al.* 2007) and semantic analysis (Archer *et al.* 2003). One way scholars have tried to overcome this issue, particularly with Early Modern English texts, is by utilising modernised versions of a text (Culpeper 2002). However, these are not always available. Another possible technique is to manually standardise spelling variation, which is applicable for small amounts of data, but becomes time-consuming when dealing with large corpora (Schmied 1994; Marquilhas & Hendrickx 2014). While semi-automatic tools have proven valuable for spelling normalisation (Archer *et al.* 2015), resource constraints meant that no normalisation was undertaken in our project. The integration of automated and semi-automated normalisation tools in the *MetaLing Corpus* remains an important area for the future development of the corpus. In its present form, therefore, the corpus retains the original spellings of the source texts without standardisation, allowing for the investigation of emerging orthographic practices, the persistence of variation after the advent of printing, and the choices made by authors and typesetters. Within the scope of this article, we focus on a highly selective set of phenomena identified through wordlist analysis. Our research questions are fundamentally qualitative, seeking to explore the degree to which our corpus reflects linguistic conservatism and innovation, patterns of regularisation and standardisation. Exploration of these aspects is especially significant given that many of our sources can reasonably be assumed to exhibit heightened attentiveness to writing in terms of both form and normative function.

If the printing press is widely regarded as a key extralinguistic factor influencing standardisation, the extent to which grammarians and schoolmasters, as well as printers, influenced standardisation remains an unresolved question (Condorelli 2023: 274–6). Spelling regularisation arguably offered practical advantages to typesetters by making printing quicker and more efficient; a degree of standardisation was also presumably favoured by the fact that early modern printers formed a tight-knit, localised community of practice (Howard-Hill 2006; Tyrkkö 2013) and networks of practice (Condorelli & De Wulf 2023). However, printers could take into account authors' own spelling idiosyncrasies and preferences, and modern standards were not firmly established until the late seventeenth century (Baron 2011: 53–4; Tyrkkö 2020: 78).

4.1. Methodological approach

Wordlist analysis was used to determine the type, range and distribution of the high-frequency vocabulary to inform the selection of terms for analysis. We examined the top 200 words to limit the number of items for closer study, and verified which amongst these

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 1. Spelling variants for the top-frequency items in the *MetaLing Corpus*

Word	Frequency	Rank	Variant	Frequency	Rank
be	8,127	16	bee	468	221
he	6,606	24	hee	741	153
they	5,859	27	theie	627	179
have	2,950	43	haue	1,485	86
hath	1,367	95	has	382	268
upon	1,173	111	vpon	817	138

presented significant spelling variants. The concordance tool in Sketch Engine was used to retrieve examples of use in context and to filter results. Spelling variants were further verified using a Python script developed for this purpose. Based on a predefined list of target forms, the script processed each corpus file individually, normalising text to lowercase, tokenising it using regular expressions and counting occurrences of each variant. The resulting data were organised by source and year to facilitate the analysis of variation patterns. This procedure provided an efficient and replicable means of tracing lexical variation across the corpus. The full code is available on GitHub for transparency and reproducibility.[2]

Table 1 provides the list of the top-frequency items displaying significant spelling variation. As may be seen, the top-frequent variants represent <u/v> distinction, the representation of long vowels – a significant source of debate amongst Early Modern English orthographers, grammarians and printers – and the ending of third person singular. In spite of its statistical significance, *they/theie* was not considered in this article because, on closer inspection, it emerged as an idiosyncratic spelling only used by Richard Mulcaster, whose *Elementarie* incidentally represents the first consistent attempt to normalise and regulate English spelling (Salmon 1999: 20). No other orthographic variants of the third-person plural pronoun are attested in the corpus.

4.2. *Results and discussion*

4.2.1. Representation of long vowels in final and medial position

The discussion of results begins with the representation of long vowels. Table 2 presents the distribution of spelling variants focusing on the sources that display alternative variants. The full names and titles of the sources cited in abbreviated form are provided in the Supplementary material Appendix.[3]

Overall, the data shows a preference for single-vowel <e> spellings, alongside variation between authors and texts. Spellings in <e> already dominate in printed works by the end of the sixteenth century. This dominance persists across all sources, even though Vaughan and Camden display comparatively higher variation with 507 and 290 occurrences of <ee> spellings which, however, do not overturn the overall preference for <e> spellings.

[2] GitHub repository: https://github.com/vahidasadi1366/METALING-keyword-analysis.git

[3] For complete metadata and bibliographical details please refer to the dataset 'Works.Tab', Work Data 1 – Metaling Corpus, accessible at https://doi.org/10.13130/RD_UNIMI/KS2PUX/WMR6IS, UNIMI Dataverse, V1, UNF:6:Ta9m7Ej+2RorDHLW2mgpDA== [fileUNF] (Andreani & Russo 2026).

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 2. Spelling variation for *he* and *be*

Source	<ee> (hee / bee)	Total	<e> (he / be)	Total
Harman 1567	21 / 24	45	300 / 189	489
Vaughan 1600	243 / 264	507	371 / 539	910
Camden 1605	257 / 33	290	508 / 402	910
Dekker 1608a	43 / 24	67	231 / 186	417
Dekker 1608b	107 / 72	179	313 / 175	488
Dekker 1616	23 / –	23	169 / 270	439
Robinson 1617	5 / 9	14	13 / 74	87
Daines 1640	2 / 8	10	18 / 268	286
Lloyd 1654	10 / 7	17	22 / 93	115
Total		1,152		4,141

Dekker texts display both intra-author and inter-textual variation: while <e> spellings outnumber <ee> forms overall, occurrences of the latter indicate that spelling was not entirely stable within a single author's production. While <e> spellings dominate in all three sources, <ee> variants rise from Dekker 1608a (13.8%) to Dekker 1608b (26.8%), both printed by Edward Allde for different booksellers; they decline in Dekker 1616 (5%), printed by Thomas Snodham. This cluster of texts constitutes a potentially interesting case study for wider scope analysis of printer-mediated variation.

The data shows that by the mid seventeenth century, <ee> spellings are erratic. Remodelled spellings of personal pronouns (such as *mee*, *hee* and *shee*) were proposed by James Howell in 1662 (Salmon 1999: 46). However, the frequency of <ee> forms in our sources does not suggest that the practice had been anticipated or applied with any consistency.

In light of the significance of this variant in the corpus, we decided to extend the analysis to the representation of long vowels in medial position. Early modern printed texts display two main approaches to represent long vowels: by using final silent <-e> or through duplication. The first attempts at regularising the spelling of long vowels were made in the later sixteenth century (Salmon 1999: 27, 34), so it is interesting to examine to what extent they are reflected in the *MetaLing Corpus.* To this end, we compared the spellings of long back and front vowels, all attested in the corpus with competing spellings; again, we relied on wordlist analysis and concordances to identify variant spellings.

Table 3 documents variation by source. After the mid seventeenth century, the sources display very little variation and a clear shift away from final <-e> spellings towards zero-final forms; for this reason, the data presented in the table focuses on earlier texts. The distribution of variant spellings indicates a marked preference for 'redundant' <VV-e> spellings. The distribution of spelling variants also shows different patterns in the representation of back and front vowels: whereas back vowels display relatively decisive shifts towards the one or the other variant spelling, front vowels exhibit variation and coexistence until the mid seventeenth century. For back vowels, duplication is dominant only in Mulcaster, whereas the 'redundant' form with duplication and final <-e> is the standard. After 1640 the picture is reversed, and simple duplication is uniformly adopted. The pattern for front vowels is less defined. Early texts select either duplication or final silent <-e> consistently (Harman and Mulcaster), but through the first half of the seventeenth century, variants with duplication

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 3. Spelling variation for final <-e> (selection 1567–1640)

	Back vowels			Front vowels		
Source	<oo>	<oo-e>	<o-e>	<ee>	<ee-e>	<e-e>
Harman 1567	10 book	10 booke	3 boke	–	–	4 betwene
	–	27 poore	3 pore	–	–	–
	–	3 soone	3 sone	–	–	3 nede
Mulcaster 1582	47 book	–	1 boke	–	–	25 betwene
	8 poor	–	3 pore	–	–	29 greke
	34 soon	–	–	–	–	80 nede
Vaughan 1600	36 book	20 booke	–	10 between	23 betweene	–
	–	41 poore	–	–	4 greeke	–
	–	13 soone	–	5 need	–	–
Camden 1605	3 book	30 booke	–	2 between	23 betweene	–
	–	29 poore	–	4 greek	54 greeke	–
	–	6 soone	–	2 need	–	–
Dekker 1608a	–	5 booke	–	1 between	5 betweene	–
	–	24 poore	–	–	–	–
	–	4 soone	–	2 need	–	–
Dekker 1608b	–	13 booke	–	–	8 betweene	–
	–	19 poore	1 pore	1 greek	–	–
	–	2 soone	–	1 need	–	–
Dekker 1616	3 book	12 booke	–	–	2 betweene	–
	–	29 poore	1 pore	–	1 greeke	–
	–	2 soone	–	–	–	–
Robinson 1617	–	2 booke	–	–	2 betweene	–
	–	–	–	–	1 greeke	–
	–	1 soone	–	1 need	–	–
Daines 1640	–	3 booke	–	11 between	12 betweene	–
	–	2 poore	–	14 greek	22 greeke	–
	–	2 soone	–	6 need	–	–
Total	141	299	15	60	157	141

and duplication + <-e> are in competition. Up until 1640, they coexist in near-balance; after 1640 the pattern shifts towards standardisation of <ee>. In summary, the distribution of spellings shows that there was no uniform or purely chronological trajectory towards standardisation in the representation of long vowels in medial position.

4.2.2. <u/v> distinction

The grapheme <v> was not originally part of the English alphabet, but was introduced in the Middle English period by scribes trained in the French tradition, where it was present alongside its corresponding phoneme. It entered the repertoire of graphemes alongside <u>, already present, but there was no attempt at systematising their distribution; the convention became to use <u> in medial position and <v> in initial position, as continued in the sixteenth and early seventeenth century. Table 4 summarises the distribution of these graphemes in the *MetaLing Corpus.*

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 4. <u/v> distinction (extract)

Sources	Initial	Tot.	Medial	Total
Harman 1567	vpon 27 upon –	27	haue 178 have –	178
Mulcaster 1582	vpon 518 upon –	518	haue 376 have 4	380
Vaughan 1600	vpon 89 upon –	89	haue 251 have 1	252
Camden 1605	vpon 133 upon –	133	haue 41 have 317	358
Dekker 1608a	vpon 44 upon –	44	haue 151 have –	151
Dekker 1608b	vpon 95 upon 3	98	haue 121 have –	121
Dekker 1616	vpon 183 upon –	183	haue 150 have –	150
Robinson 1617	vpon 5 upon –	5	haue 38 have –	38
Daines 1640	vpon – upon 12	12	haue – have 109	109

The corpus reveals a pattern towards standardisation of <v> in medial position and <u> in initial position, which is completed by 1640, as there are no occurrences of *vpon* and *haue* after this date. Overall, medial position spellings display innovation, while initial position remains conservative. In initial position, the only deviation observed in Dekker 1608b confirms that initial <u> is exceptional in the early seventeenth century. A clear shift is observed in Daines, where *upon* is the only attested variant, reflecting the distinction introduced in the 1630s towards the separation of the functions of <u> as a vowel and <v> as a consonant (Salmon 1999: 39; Condorelli 2021: 808–9, 818).

In medial position, the distribution of the graphemes exhibits more variation. Erratic spellings with <v> intrude towards the end of the sixteenth century, and a major shift is visible in Camden, where *have* outnumbers *haue.* The innovation is not immediately generalised, however, as Dekker and Robinson revert to *haue.* Standardisation is again observable starting from Daines. The data shows that medial <v> emerges earlier but spreads unevenly, with periods of coexistence and competition of variant spellings, whereas initial <u> remains virtually unattested until the mid seventeenth century. Scattered later occurrences of *vpon* in Walker 1655 (3 occurrences), Stubbe 1657 (1 occurrence) and Newton 1677 (1 occurrence) suggest to explore residual variation after the widespread adoption of <u/v> distinction.

4.2.3. Third-person singular verb forms

The final instance of variation concerns third-person present singular verb forms. As is known, the line of development in the early modern period is the replacement of <-th> by <-s> ending. The development is known to have spread from Northumbria; by the end of the fifteenth century it is attested in London and in the early seventeenth century in East Anglia

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 5a. Third-person singular verb forms by source

Source	Verb class	<(e)th>	<(e)s>	Total
Harman 1567	HF	95	–	95
	Lex	9	5	14
Mulcaster 1582	HF	163	–	163
	Lex	27	9	36
Vaughan 1600	HF	129	1	130
	Lex	30	8	38
Camden 1605	HF	127	2	129
	Lex	17	3	20
Dekker 1608a	HF	56	21	77
	Lex	–	11	11
Dekker 1608b	HF	52	9	61
	Lex	–	19	19
Dekker 1616	HF	68	8	76
	Lex	1	22	23
Robinson 1617	HF	19	–	19
	Lex	2	–	2
Daines 1640	HF	74	4	78
	Lex	1	11	12
Anon 1644	HF	7	–	7
	Lex	–	3	3
Howell 1650	HF	3	–	3
	Lex	–	1	1
Lloyd 1654	HF	4	–	4
	Lex	1	–	1
Farnworth 1655	HF	6	1	7
	Lex	–	1	1
Walker 1655	HF	63	8	71
	Lex	–	5	5
Stubbe 1657	HF	129	14	143
	Lex	2	14	16

(*Continued*)

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 5a. *Continued*

Source	Verb class	<(e)th>	<(e)s>	Total
Reyner 1663	HF	144	–	144
	Lex	4	24	28
Head 1666	HF	90	2	92
	Lex	3	26	29
Ray 1674	HF	22	–	22
	Lex	–	9	9
Newton 1677	HF	92	9	101
	Lex	–	21	21
Mather 1681	HF	185	14	199
	Lex	15	32	47
Anon 1682	HF	15	1	16
	Lex	1	–	1
Browne 1683	HF	6	–	6
	Lex	–	–	–
Lodwick 1685	HF	2	–	2
	Lex	–	–	–
Stuart 1686	HF	–	39	39
	Lex	–	17	17
Baker 1699	HF	–	296	296
	Lex	–	16	16
HF Total		**1,467**	**430**	**1,897**
Lex Total		**195**	**265**	**460**
Total		1,662	695	2,357

(Holmqvist 1922; Kytö 1993; Nevalainen & Raumolin-Brunberg 2000; Gries & Hilpert 2010). Third-person present singular forms in <-s> are rare until the 1590s, when they begin to display variation with forms in <-th>; around 1700 <-th> is restricted to liturgical, biblical and formal contexts. The patterns of distribution of the two forms seem to be influenced by linguistic factors, such as verb frequency and phonological context, and extralinguistic factors, such as text type (Kytö 1993: 121, 130; Nevalainen & Raumolin-Brunberg 2003: 144). Table 5a presents the data for the most frequent verbs in the corpus, separating the high-frequency (HF) verbs *have* and *say* and the lexical (Lex) verbs *make* and *know.*

With 1,662 occurrences, <-(e)th> forms outnumber <-(e)s> forms (695 occurrences). Overall, the data shows a strong persistence of <-(e)th> forms well into the seventeenth century,

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 5b. Variants of *say* by source

Source	*sayth / saith*	*sayeth*	*sayes / saies*	*says*	Total
Harman 1567	49	–	–	–	49
Mulcaster 1582	–	2	–	–	2
Vaughan 1600	8	–	–	–	8
Camden 1605	40	–	–	–	40
Dekker 1608a	1	–	11	–	12
Dekker 1608b	–	–	7	–	7
Dekker 1616	2	–	6	–	8
Daines 1640	4	–	3	–	7
Lloyd 1654	1	–	–	–	1
Farnworth 1655	2	–	1	–	3
Walker 1655	–	–	1	–	1
Stubbe 1657	–	–	14	–	14
Head 1666	2	–	–	2	4
Mather 1681	–	109	–	–	109
Anon 1682	–	–	–	1	1
Stuart 1686	–	–	–	3	3
Baker 1699	–	–	–	4	4
Total	**101**	**111**	**43**	**10**	**322**

though lexical verbs show a much earlier and more consistent shift towards <-(e)s> forms. By the last quarter of the seventeenth century forms in <-th> have completely disappeared in high-frequency verbs. As expected, verb frequency conditions the pace of change, and high-frequency items resist innovation longer than less frequent lexical verbs. The verb *say* displays the greatest spelling variation, reflecting competition amongst multiple forms.

Table 5b shows that variation is both diachronic and author-specific. Whereas <-(e)th> forms are standard in the earliest sources (Harman, Mulcaster, Vaughan and Camden), Dekker's texts mark the first site of variation. Dekker adopts the innovative ending earlier and more consistently than his contemporaries, although competing variants occur. By contrast, <-(e)th> forms are still dominant for the high-frequency verb *have* in all Dekker's texts: 1608a presents 55 *hath* vs 10 *has*; in 1608b *hath* occurs 52 times vs 2 *has*; in 1616 there are 66 occurrences of *hath* vs 8 occurrences of *has.*

In the mid seventeenth century authorial preferences diverge. Daines indicates ongoing competition, Lloyd and Farnworth retain <-(e)th> forms, whereas Walker and Stubbe mark a shift towards <-(e)s> forms. Later authors show increasing standardisation (but note Mather's preference for *sayeth*). By the end of the century, the innovative ending has become dominant.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

The analysis of the distribution of *say* variants demonstrates that the transition from <-(e)th> to <-(e)s> was neither linear nor uniform. Individual authors (and printers) exhibit individual preferences and idiosyncratic practices, with some maintaining conservative forms well after the introduction of innovative variants.

The group of variants analysed appears to reflect a stronger influence of printer practice than of grammarians. However, further analysis is required, particularly taking into account printers and the role of typography, and potentially of genre-specific patterns. One limitation of this study is the uneven chronological representation of the corpus, with a gap between the 1620s and the 1640s and a higher concentration of sources between 1600 and 1620. Nevertheless, one clear result is that after 1640 modern spelling standards become widespread, and for the phenomena analysed in this article they largely constitute the norm.

5. Discourse variation in the *MetaLing Corpus*

The second case study shifts focus from formal variation to discourse variation, combining corpus-driven and corpus-based research to investigate the 'aboutness' of our corpus. We begin by applying keyword analysis to help us identify units of outstanding frequency (Egbert & Biber 2019: 78). However, the analysis remains exploratory, as the current limits of our corpus regarding tokenisation and spelling normalisation prevent us from obtaining statistically robust evidence of domain coverage at this stage. For this reason, the analysis then shifts to a corpus-based approach and is primarily qualitative.

5.1. Keyword analysis

Keyword analysis was performed in Sketch Engine using the built-in keyword extraction function. The historical corpus *English Historical Book Collection* (EEBO, ECCO, Evans), preloaded in Sketch Engine, was used as a reference corpus because it provides a large and representative diachronic collection of Early Modern English texts, making it suitable for comparison with our dataset. In particular, the subcorpus EEBO-TCP Phase I (1473 – *c.*1700) was selected because it covers a comparable timespan, while Evans and ECCO focus on the later modern period (from the late seventeenth century onwards). It should be noted that the *MetaLing Corpus* and the reference subcorpus are processed using different tokenisation procedures, which affect token counts, individual word frequencies and normalised frequencies, and thus lead to keyness scores that may not reliably reflect thematic prominence in our corpus. Additionally, as already pointed out, no normalisation was applied when producing these results. Aware of these limitations, we decided to employ keyword analysis as an exploratory tool to interpret patterns across keyword sets that may indicate thematic relevance.

Informed by Jonathan Culpeper's approach (2009: 36) we tested various possibilities and chose a combination that enabled us to minimise extremely rare word types (Sketch Engine setting 10) and to retrieve a sufficient number of words (1000) with a minimum frequency threshold of 20 occurrences. The resulting keyword list was manually cleaned to remove further irrelevant items, such as contractions (e.g. *cō* for *com*) and compounds that were split in the target corpus (e.g. *them selues*), and to identify spelling variants that could be grouped as different forms of conceptually identical items. Selected keywords were further examined through concordance analysis to interpret their usage in context, and then grouped thematically. Considering the limitations of the analysis, we did not take into account frequencies, and we limited the study to clusters of semantically related keywords that could tentatively point to emerging discourse domains.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 6. Top 50 keywords and potential discourse domains

	Keyword (rank)	Type / Discourse domain
1	theie (1), themselue (5), them (8), selue (27), hee (32)	Pronouns
2	tung (2)/tongue (210), vowell (6)/vowel (36)/vowels (34), consonant (7), syllable (12)/sillable (202), verb (13), substantive (20), speche (22), pronunciation (26), sound (30), tense (38), adjective (47), language (50)	Language
3	som[e] (10), anie (28)	Determiners
4	therefor (14), bycause (16)	Linking devices
5	elementarie (15), learning (39), learner (48)	Education
6	natur (18), foren (25), waie (33)	Heterogeneous
7	Cic (4)/cicero (486), Plaut (21), Ovid (24), Caes (40), Heaut (43)	Sources/reference (Cicero, Plautus, Ovid, Terence[4])
8	maie (23), seme (46)	Verbs
9	from (29)	Prepositions
10	somtime (35), alwaie (49)	Time adverbs
11	ne (37), mihi (42)	Latin words
12	c. (11), ep. (19), sc. (45), Suss. (44)	Abbreviations (chapter, episode, scene, Sussex[5])

The results are presented in table 6. For a more synthetic overview, keywords are grouped according to type (function words are included) or discourse domains. The ranking provided in brackets is based on the keyness metric implemented in Sketch Engine, which reflects differences in relative frequency between the target and reference corpora. As noted, this is not intended to be statistically significant from a linguistic point of view, but is included for the sake of completeness.

A number of clusters of semantically related words emerge. The largest, as expected, is related to language (row 2); this can be further divided into more specific semantic fields (e.g. pronunciation, morphology and parts of speech, partly reflecting the specific areas covered by the subcorpora; see Andreani & Russo 2026) and provides a set of metalinguistic terminology for future research. Another cluster, in row 5, seems to reflect educational discourse, possibly indicating a focus on language learning. Two further potential domains are represented by rows 7 and 12, which point to textuality and intertextuality. It should be noted, however, that some of these keywords, such as 'Suss.' and 'Heut.' are not distributed across the corpus but are confined to a single text.

5.2. *Close-reading*

Given the limitations of keyword analysis, we adopted a close-reading approach to the study of discourse variation, combining corpus-driven insights with corpus-informed analysis. For

[4] Reference is to Terence's play *Heauton Timorumenos* which occurs exclusively in Walker 1655. Citations from Plautus and Caesar are also almost excusively in this text.

[5] The abbreviation occurs exclusively in Ray 1674.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

exploratory purposes, we manually selected data from a single text to assess the thematic relevance of educational discourse.

As a reference, Mulcaster's *Elementarie* (1582) was chosen to help us identify further words related to the educational settings and practices in the early modern period. Mulcaster's work is considered one of the foundational works in pedagogic writing in the history of English education and, therefore, represents a prototypical example of pedagogic discourse in our corpus. The aim was to compile a representative set of terms including 'unpredictable manifestations' (Kohnen 2007: 141), or, in other words, unexpected or less intuitive words and phrases used to describe learning and teaching practices. To this end, we focused on sections 1.1 to 1.3 (pp. 1–18) of the *Elementarie* and selected the terms based on the level of confidence with which we could agree on their association with learning and teaching. The sections of the *Elementarie* were examined separately by the authors through close reading; the two wordlists were then compared to select overlapping terms, resulting in the following items: *learn*, *teach*, *trial*, *understanding*, *performance*, *comprehend*, *practice*, *principles*, *behaviour*, *precept* and *right writing*. Considering the inherent subjectivity of term identification, all selected terms were reviewed by a third annotator fluent in English to improve reliability. Given the very small number of items under evaluation, statistical measures of inter-annotator agreement such as Cohen's Kappa were not considered. Instead, agreement was assessed through direct comparison and consensus. Because these same terms were also used to identify educational discourse in the corpus, the methodology risks circularity. For this reason, in order to partially correct this limit, we complemented the qualitative selection with quantitative measures of frequency and collocational patterns, as discussed in the following section.

5.3. Results and discussions

Table 7 presents the number of occurrences of the selected terms in the *MetaLing Corpus*.

Table 7. Frequency of educational terminology in the *MetaLing Corpus*

Term	f
learn	125
teach	111
triall	14
vnderstanding / understanding	171
performance	47
comprehend	47
practice	63
principles	112
behauiour / behaviour	22
precept	71
right writing	61

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Table 8. Frequency of the terms *performance*, *principles* and *precept*

Term	Occurrences	Source	Subcorpus
performance	3	Baker 1699	B
		Reyner 1663	B
		Lloyd 1654	C
principles	3	Reyner 1663	B
		Daines 1640	C
		Lloyd 1654	C
precept	5	Baker 1699	B
		Newton 1677	B
		Reyner 1663	B
		Vaughan 1600	B
		Lloyd 1654	C

The concordance of each was examined manually to determine whether the words were really used in an educational context. For instance, the collocational profile of the word *precept*, analysed using the WordSketch function in Sketch Engine, shows it occurs in contexts marked by educational lexis, such as the verb *learn* and the nouns *practice*, *method*, *train* and *teaching.* In most cases, however, the collocational profile was not enough, so occurrences were analysed manually through concordances. As a result, three additional terms were identified as relating to educational discourse, alongside the more obvious cases of *learn* and *teach: performance*, *principles* and *precept.* Although *right writing* was also linked to educational discourse, it was excluded from the analysis because it occurs only in Mulcaster's *Elementarie.*

Table 8 summarises the terms selected for analysis, their frequencies in educational contexts following manual classification, and the corresponding subcorpora and source authors. As the table shows, occurrences in educational settings account for only a small proportion of the terms' overall frequencies. The discussion therefore moves from an examination of individual concordances to the broader textual contexts in which these terms occur.

5.3.1. Performance *in context*

The selected concordances are the following:[6]

(1) For Learners should endeavour therein exact **performance** of all circumstances required in the third and fourth, and fifth precedent Rules. (Lloyd 1654)
(2) Though better provided of Academical improvements, yet not maintaining their familiarity with their Studies, grow so negligent and loath to bestow any pains upon those **performances** which their charge exacts from them, that the crudeness of them renders them disgusted with others. (Reyner 1663)
(3) The Stores of Learning, by which means by reasoning against Learning, we must at the same time reason for it, and all our Arguments must return upon us; or if the **performance** were unlearned, it would be to no purpose. (Baker 1699)

In the first concordance, the term refers to the learners and the phrase 'should endeavour' emphasises academic discipline, likely referring to completing a task. The term 'Rule'

[6] All bold formatting has been added by the authors and does not appear in the original source.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

seems to indicate the various rules that should be followed by students. The second concordance appears to refer to the student's academic growth. In addition, the defining relative clause appears to point to tasks or duties expected of students. In the last concordance the phrase 'Stores of Learning' suggests the idea of accumulated knowledge or scholarly material within an educational setting. It appears that the author is referring to the philosophical aspect and how even criticism of learning leads to validating it, highlighting the importance of education.

5.3.2. Principles *in context*

The selected concordances are the following:

(4) Which way shall seem most reasonable and likeliest to succeed with most proficiency upon a just comparison of the **principles** in rules and practise of either methods. (Lloyd 1654)
(5) We will immediately come to treat of Letters, as the first Elements or **Principles** of speech in every Tongue or Language whatsoever. (Daines 1640)
(6) For Grammar delivers the first **Principles** of a Language, shews the Formations, and Significations, Connexions, and Dependencies of words, and how they are put together to frame a discourse, of which no sense can be made without it. (Reyner 1663)

The term *principle* has the highest frequency in the corpus compared to the other two selected terms, with varying degrees of relevance to an educational or instructional setting. In concordance (4), the lexis suggests an educational context. Although there is not sufficient contextual grounding to clearly identify and justify the pedagogical intention, the phrase 'principles in rules and practise' would seem to imply a comparison of methodologies potentially referring to a teaching or learning environment – for this reason, the concordance was selected as potentially significant. In concordance (5), the phrase 'the first Elements or Principles of speech' clearly situates the discussion within the domain of linguistics. This example offers a more direct link between principles and language instruction. Moreover, the didactic function of the text here also emerges from the metadiscourse marker 'we will immediately come to treat of', which guides the reader through the text by anticipating the goal of the discourse that will follow. The last concordance provides a strong connection between the term *principles* and formal education. The term refers explicitly to grammar as the foundational framework for understanding language, stating that principles are the underlying rules that govern language structure and usage.

5.3.3. Precept *in context*

The selected concordances are the following:

(7) Schollars that have attayned the true reason of reading rightly will be much perfected therein by reducing received **precepts** to frequent practise. (Lloyd 1654)
(8) Which may be better improved by constant practice, than by any **precepts** which can be given. (Newton 1677)
(9) Ought to choose out a good schoole-master, that will not onely instruct them in learning, but also informe them in the **precepts** of good maners. (Vaughan 1600)
(10) Therefore, though some words comply not with the **precepts** given by Grammarians. (Reyner 1663)

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

(11) Condition of Speech must be submitted to, and is no objection in plain Laws and easie **precepts**, that are intelligible enough in any Language, yet in matters of Science. (Baker 1699)

In concordances (7) and (8) teaching methodologies appear to be discussed. As can be seen, these examples show that the authors think that practice and exercise are significant and are to be preferred over rules. In concordance (9), the terms 'school-master', 'instruct' and 'learning' clearly indicate the domain of education and highlight the relevance of the school context in the extract. The text expands the function of *precept* beyond intellectual instruction; by asserting that a good schoolmaster must gain both learning and 'the precepts of good manners', *precept* is considered in a moral-pedagogical framework, reflecting early modern educational ideals where fostering good character was considered inseparable from education (Lickona 1991; Bushnell 1996). The role of grammarians in language learning is discussed in concordance (10), which implies an institutional setting or a language classroom, where grammar is being taught or discussed. The last concordance (11) discusses language in various scenarios. The sentence 'Condition of Speech must be submitted to' seems to propose that one must accept the nature or limitations of language or speech. The example would seem to reflect on the limitations of language in different types of knowledge.

5.4. Educational terminology in context

In this section, we examine whether the concordances identified are embedded in larger educational sections and consider how these relate to the rationale of the work in which they appear. In doing so, it explores the extent to which corpus-based analysis can illuminate aspects of the corpus 'aboutness' that are not immediately apparent from its compilation criteria. The concordances analysed suggest that Lloyd 1654, Reyner 1663 and Baker 1699 are potential sources for investigating pedagogical discourse in the corpus. Significantly, these texts originate from different sections of the corpus: Lloyd 1654 is included in volume VI of Alston's *Bibliography*, devoted to works on pronunciation, whereas Reyner and Baker are catalogued in the miscellaneous section of volume III, lacking any explicit association with a particular branch of linguistic knowledge. Educational discourse thus emerges as a domain that cuts across the corpus's original classification and contributes to a more nuanced understanding of its thematic profile.

5.4.1. The practice of teaching (Lloyd 1654)

Lloyd discusses *principles* in a section entitled 'Observations upon the Letters tending to right Syllabizing', in which he contrasts the conventional syllabising method with his own approach. The section begins by underlying rules of sound–letter relations:

> This did move the Author to propound a new Abcedary, with a Preface, declaring the true reasons of Innovation ... And for evidencing the truth thereof, hee doth willingly referre himselfe to any impartiall Judgements to determine which way shall seem most reasonable and likeliest to succeed with most proficiency upon a just comparison of the **principles** in rules and practise of either methods, to wit, of the vulgar way formerly accustomed, and of the new form here propounded, desiring first these premonitions to be remembred. (Lloyd 1654: 22–3)

Lloyd's practice substitutes 'the vulgar way formerly accustomed' with a 'new form' that is presented as more systematic and logical. The passage functions as both a beginning to

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

procedural instruction and a rationale for pedagogical reform. Here, the pedagogical intent of the author is clear: Lloyd seeks not only to instruct learners in correct syllabising but also to provide a reasoned framework that allows teachers and students to evaluate methods critically through 'just comparison of the principles in rules and practise of either methods' (23). The next passage is from the section 'Of Schollars communicating knowledg', which underscores the pedagogical value of applying *precepts* through active practice, especially in peer-to-peer instruction:

> Schollars that have attayned the true reason of reading rightly will be much perfected therein by reducing received **precepts** to frequent practise, which is done to the best advantage when such do teach their fellowes; for as the more ingenious are thereby more exercised, so the duller sort are more capable of instruction from them, having among themselves an easier method in a proportion neer their own capacity, by a more familiar discourse, fuller of freedome for ingenuity to expresse it selfe…besides that Scholers made super-visors of their fellows, are far more studious and considerate in teaching others, then when they learn themselves. (Lloyd 1654: 30)

Lloyd emphasises that when scholars teach their fellows, both the advanced and the less capable learners benefit; the former refine their knowledge through teaching, while the latter gain instruction in a more accessible and familiar discourse. The passage situates *precept* at the intersection of theory and practice, demonstrating its role in fostering collaborative and reflective modes of education. This may be compared with concordance (8) in section 5.3.3; the fuller context is as follows:

> And thus much concerning Elocution, as for Memory and Pronounciation, which are the other two parts of Rhetorick, I purposely omit them, as being natural Endowments, which may be better improved by constant practice, than by any **Precepts** which can be given. (Newton 1577: 172)

In this discussion of memory and pronunciation, *precept* is contrasted with the value of practice. Unlike Lloyd, who acknowledges the usefulness of precepts when applied, Newton appears to believe in practice as superior to perceptive instruction in certain respects, particularly for memory and pronunciation, which he considers natural endowments that can be enhanced through repetition and practice rather than precepts. The pedagogical implication of this text is that not all aspects of learning can be governed by rules; rather, repetition and experience may have greater effectiveness.

In contrast, Lloyd returns to 'rules' in the final passage, from a section of his work aptly entitled 'The rules of Art in writing'. In it, the author lists a series of 'grounds' or 'principles' which 'Such as will learn to write with good proficiency' should observe (Lloyd 1654, p. 40). Our concordance appears in point 12:

> 12. When Learners are well skilled in the draughts of the smaller letters, it will be expedient, before they enter upon the Capitall, to teach them upon fresh memory how to joyne the same together, […] and when that is performed regularly, to proceed unto the third succeeding president: both which will require much time to gain therein perfection; for Learners should endeavour therein exact performance of all circumstances required in the third and fourth, and fifth precedent Rules. (Lloyd 1654: 46)

This is clearly embedded within a larger educational framework, in which pedagogic content is fundamental. Lloyd's work is not simply a treatise on pronunciation, but also a pedagogical

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

manual for teachers: *The Schoole-Masters Auxiliaries*, a 'guide' in two volumes, the first devoted to how to teach to read and write English 'dexterously', and the second devoted to 'asserting the LATINE TONGUE in Prose and Verse, to its just Inlargement, Splendor, and Elegancy' (Lloyd 1654).

5.4.2. *Disciplined learning (Reyner 1663)*

John Reyner's *A Treatise of the Necessity of Humane Learning for a Gospel Preacher* was published from the papers inherited from his father Edward, and represents an earnest justification of the necessity of a classical and academic education for ministers. The occurrences of the terms *principles* and *performance* appear in chapter I, which discusses the value of a knowledge of languages for preachers:

> Nor are those to be excused, who, though better provided of Academical improvements, yet not maintaining their familiarity with their Studies, grow so negligent and loath to bestow any pains upon those **performances** which their charge exacts from them, that the crudeness of them renders them disgusted of others, and their conceptions become as jejune and void of any sprightful relish, as if their minds had never received any Artificial seasoning, nor were tinctured with any **principles** besides those which are ordinarily found among the vulgar throng of men. (Reyner 1663: 31)

This passage is situated within a larger discussion about the responsibilities of learners and the consequences of neglecting academic work. Unlike Lloyd's prescriptive attitude, Reyner's emphasis is not on explicit procedural instruction; instead, he emphasises the qualitative results of scholarly effort. He claims that even those having access to academic resources may fail in improving proficiency if they do not actively engage in their studies. This emphasises the potential loss of meaningful learning. In addition, the author highlights that developing scholarly skills hinges upon consistent practice and regular effort. From a pedagogical perspective, this supports the idea that learners' advancement depends not only on the exposure to materials, but also on the disciplined engagement and the development of intellectual habits, which are central to effective instruction in any field of study. It shows how neglecting scholarly work can hinder learning, which in turn highlights the responsibilities of learners and the importance of instruction in developing capable practitioners. Thus, it can be concluded that the section deals with teaching implications, even if not directly. The next occurrence appears in section IX of chapter I, 'Of Grammar':

> For Grammar delivers the first **Principles** of a Language, shews the Formations, and Significations, Connexions, and Dependencies of words, and how they are put together to frame a discourse, of which no sense can be made without it. (Reyner 1663: 36)

According to the author, grammar is considered as the tool through which the foundational principles of language are conveyed. The knowledge of a language is not innate but should be systematically gained through grammar, which elucidates the forms, meanings, relationships and dependencies of words, ultimately enabling a coherent discourse. Unlike Daines, who emphasises letters as the fundamental principles, Reyner highlights the role of grammar as the basis that shapes these elements into a meaningful structure. Training in grammar is presented as indispensable for developing linguistic competence by establishing a bridge between elementary knowledge, such as letters, and functional language use. The discussion is within a section that supports structured learning and the rational acquisition of linguistic skills, aligning with early modern views on the centrality of grammar in

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

education (Hedge 2000; Leblanc 2020). Just like Lloyd's educational approach, Reyner's too starts from principles and then moves to practice.

The final occurrence appears in an additional section appended at the end of chapter I, which is entitled 'Concerning the usefulness of the Arabic'. The language is compared with Hebrew:

> The matter would come all to one, and the sense remain inviolate. Therefore, though some words comply not with the **Precepts** given by Grammarians, which are conformed onely to the examples of the Bible, and not to the whole extent of the Hebrew Tongue, of which since we have no more monuments, we may well be ignorant of much of the ancient Hebraism: yet such words are not presently to be rejected as faults and mistakes. (Reyner 1663: 45)

Here *precept* is explicitly connected to the grammatical rules. The author ties these rules' limitations when applied exclusively to biblical Hebrew rather than to languages more broadly. In contrast to Lloyd's discussion, where *precept* carries a more didactic or moral connotation, Reyner's use of the word highlights its role within linguistic description. The pedagogical implication of it is that *precepts* provide a framework for order and instruction.

5.4.3. Critical pedagogy (Baker 1699)

Thomas Baker's *Reflections upon learning* was a successful work that saw numerous editions; it asserted the primacy of revealed knowledge over human reason and countered criticism of the Scriptures. Contrasting with the prescriptive approaches of Reyner and Lloyd, Baker's writing is predominantly reflective and critically engaged, examining the boundaries of human learning. In the preface to his work he writes:

> An attempt of this nature would be utterly impracticable, for either it would be well perform'd, and then it must be done by reasons borrow'd from the Stores of Learning, by which means by reasoning against Learning, we must at the same time reason for it, and all our Arguments must return upon us; or if the **performance** were unlearned, it would be to no purpose, and might as well be let alone. (Baker 1699: 1)

The passage exemplifies his sophisticated perspective: while he critiques an over reliance on traditional sources of knowledge, he concurrently maintains that engagement with the subject matter necessarily entails interaction with them. The text serves as both a teaching and philosophical guide, preparing the reader to question knowledge rather than just delivering it. Baker presents himself as a guide who points out the flaws and misuses of traditional learning while still recognising its value. Here, teaching isn't about step-by-step instructions emphasized in Lloyd's approach, but rather about encouraging critical thinking and reflection on how we learn and how we reason. The teaching purpose in Baker's text is both subtler and more nuanced, as it highlights the reflective aspect of pedagogy in early modern scholarly practice. This refers to encouraging learners to critically examine their own learning processes, reasoning and assumptions, rather than only following instructions. The other 'pedagogical' term used by Baker occurs in the second chapter entitled 'Of Languages':

> Languages being the Chanels by which most of our Learning is convey'd, it is necessary to the attaining of Knowledge, that these should be kept clear and open; if the Streams in these run muddy, or are corrupted, all the knowledge that is convey'd by them, must be obscure: words at the best are no very certain signs of things, they are liable to ambiguity, and under that ambiguity are often subject to very different meanings; and tho' this, as far as it is the common condition of Speech must be submitted to, and is no

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

> objection in plain Laws and easie **Precepts**, that are intelligible enough in any Language, yet in matters of Science, it is much otherwise; these are nice things; the strict meaning is to be observ'd in them; nor can we mistake a word without losing the Notion. (Baker 1699: 7)

The author broadens the scope of *precept* by applying it to the scientific domain, distinguishing between 'plain Laws and easie Precepts' – which are comprehensible in everyday language – and the heightened precision necessary in scientific discourse. In this context, 'precepts' are to be intended as simplified and accessible instructional tools (such as 'prescriptions' or 'rules') that would seem to govern moral and behaviours, similarly to laws. The passage is not properly pedagogical but is interesting in that it highlights the epistemological tension between conventional speech and the demands of scientific knowledge, illustrating the inherent limitations of language in conveying meaning unambiguously. The themes are the nature of language itself, and the development of field-specific terminology. In fact, later in his work, Baker expresses reservations about John Wilkins' philosophical language project, and adopts a rather pessimistic view of the possibility of human mutual understanding: 'The Divisions of Tongues was inflicted by God as a Curse upon humane Ambition, and may have been continu'd since for the same reason; and as no Remedy has been yet found' (17–18).

Although these findings are preliminary and should be interpreted with caution, given the limited number of concordance data, they point to both the presence of educational discourse in the corpus and its multifaceted nature. The discourse appears to include not only intellectual instruction but also the elements of moral guidance, reflective thinking and organised pedagogical frameworks. Authors such as Lloyd appear to highlight prescriptive and systematic approaches, focusing on step-by-step methods and procedural rules, while Baker encourages critical reflection and the development of reasoning, by pointing out the importance of practice, repetition and active participation in learning.

6. Concluding remarks

Through two exploratory case studies, this article has assessed the analytical potential of the *MetaLing Corpus* and the extent to which linguistic evidence can illuminate both orthographic practices and the aboutness of the texts it contains. By focusing on spelling and discourse variation, the study has demonstrated how the corpus can contribute to the study of both orthographic practices and the discursive domains represented within its texts.

The first case study has shown that spelling variation is generally limited in the corpus, confirming trends towards standardisation already identified by previous scholarship. Nevertheless, significant variation is observed in certain forms. The variation observed is individual and uneven, while it also points to author- and printer-specific practices. In fact, the inclusion of multiple works by individual authors makes it possible to explore authors' choices and the role of typesetters in shaping orthographic choices. This pilot survey provides valuable data for future developments, such as closer analysis of the interaction between authorial preferences and printers' practices, which might be further illuminated by recent re-evaluations of the role of printers and their historical context (Condorelli 2021). Another avenue for further research may involve comparative analyses of spelling variation in print and manuscript sources identified by the project.

In the second case study we have focused on the investigation of a potentially emerging discourse field combining corpus-driven insights with corpus-based close-reading approaches. Although the relatively low number of concordances retrieved suggests that the selected terms may not be highly productive in early modern pedagogical contexts, their distribution

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

points to distinct discursive uses across texts. Three sources, in particular, illustrate a spectrum of educational approaches, ranging from prescriptive modes of instruction to more practice-based and experiential forms of learning. The findings suggest that the type of educational discourse represented in the corpus integrates intellectual, moral and reflective dimensions, pointing to a possible interconnection of knowledge acquisition and ethical judgement. More broadly, the analysis suggests the presence of an educational discourse domain alongside language analysis and description that is not immediately apparent from its bibliographic classification. As such, it contributes to a preliminary understanding of the range of discourse domains represented in the corpus and provides a basis for its future semantic and discourse annotation.

As anticipated, one limitation of this research lies in the absence of spelling normalisation in the corpus. While this was balanced by the analytical potential afforded by retaining original forms, orthographic variation has a clear impact on keyword analysis, dispersing frequencies across variant forms. Although some manual correction was undertaken, this inevitably affects the results. In addition, differences in tokenisation further affect frequency counts and limit the reliability of comparisons between the target corpus and the reference corpus. These constraints are not only methodological limitations but also serve to identify priorities for corpus development. In particular, they highlight how variation in spelling and annotation can obscure thematic patterns and reduce the effectiveness of keyword analysis, which, for this reason, has been treated here as an exploratory and heuristic tool, rather than a measure of thematic importance.

In conclusion, the findings highlight both the potential and the current limits of the *MetaLing Corpus.* The two case studies demonstrate that the corpus can support investigations of orthographic practices, linguistic standardisation and discourse variation, while also revealing aspects of the corpus's aboutness that are not immediately apparent from its bibliographic classification. At the same time, the study highlights the challenges posed by orthographic variation and text-processing inconsistencies. Addressing these issues through spelling normalisation, improved tokenisation and lemma-based annotation represents a key priority for the corpus's future development and analytical robustness.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1360674326100859

Acknowledgements. The research leading to these results has received funding from MUR, PRIN2022 project 'MetaLing Corpus: Creating a corpus of English linguistics metalanguage from the 16th to the 18th century', ref.: 202233C93X, funded by the European Union under the programme NextGenerationEU, PNRR Missione 4 – Componente 2 – Investimento 1.1 (October 2023 – February 2026). Angela Andreani was responsible for sections 1 to 5.1; Vahid Asadi was responsible for the remaining sections. OpenAI's ChatGPT (version 5.2) was used to support language revision and the design of table 5. Data selection, analysis and interpretation of the results remain the responsibility of the authors.

References

Alston, Robin. 1974. *A bibliography of the English language from the invention of printing to the year 1800*. A corrected reprint of volumes I–X. Ilkley, Yorks.: Janus Press.

Andreani, Angela & Daniel Russo. 2026. Work Data 1 – Metaling Corpus. UNIMI Dataverse. https://doi.org/10.13130/RD_UNIMI/KS2PUX

Andreani, Angela & Daniel Russo. 2023. Building a corpus of the metalanguage of English linguistics 1500–1700: Methodological issues. *Linguistica & Filologia* 43, 151–74.

Archer, Dawn, Paul Rayson, Tony McEnery & Andrew Wilson (eds.). 2003. *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL Technical Papers. Centre for Computer Corpus Research on Language Technical Papers. Lancaster: University of Lancaster.

Archer, Dawn, Merja Kytö, Alistair Baron & Paul Rayson. 2015. Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal, The International Computer Archive of Modern and Medieval English* 39(1), 5–24.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Ascham, Roger. 1545. *Toxophilus the schole of shootinge contayned in two bookes*. London.
Barber, Charles Laurence. 2005. *Early Modern English*. Edinburgh: Edinburgh University Press.
Baron, Alistair. 2011. Dealing with spelling variation in Early Modern English texts. PhD dissertation, Lancaster University.
Baron, Alistair, Paul Rayson & Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik* 20, 41–67.
Bushnell, Rebecca W. 1996. *A culture of teaching: Early modern humanism in theory and practice*. Ithaca, NY: Cornell University Press.
Condorelli, Marco. 2021. Positional spelling redistribution: Word-Initial <u>/<v> and <i>/<j> in Early Modern English (1500–1700). *English Language and Linguistics* 25(4), 799–823.
Condorelli, Marco. 2023. What is spelling standardization? In Condorelli & Rutkowska (eds.), 265–82.
Condorelli, Marco & Chris De Wulf. 2023. Networks of practice across English and Dutch corpora. In Condorelli & Rutkowska (eds.), 457–76.
Condorelli, Marco & Hanna Rutkowska (eds.). 2023. *The Cambridge handbook of historical orthography*. Cambridge: Cambridge University Press.
Culpeper, Jonathan. 2002. Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*. In *Conversation in life and in literature: Papers from the ASLA Symposium* 15, 11–30. Uppsala: Universitestryckeriet.
Culpeper, Jonathan. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1), 29–59.
Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1), 77–104.
Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(3), 293–320.
Hedge, Tricia. 2000. *Teaching and learning in the language classroom*. Oxford: Oxford University Press.
Holmqvist, Erik Ossian. 1922. *On the history of the English present inflections: Particularly -th and -s*. Heidelberg: Carl Winter.
Howard-Hill, T. H. 2006. Early modern printers and the standardization of English spelling. *Modern Language Review* 101(1), 16–29.
Kahle, Philip, Sebastian Colutto, Günter Hackl & Günter Mühlberger. 2017. Transkribus – a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 4, 19–24. IEEE.
Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž & David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, 105–16.
Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1, 7–36.
Kohnen, Thomas. 2007. Text types and the methodology of diachronic speech act analysis. In Susan Fitzmaurice & Irma Taavitsainen (eds.), *Methods in historical pragmatics*, 139–66. Berlin and New York: De Gruyter Mouton.
Kytö, Merja. 1993. Third-person singular verb inflection in early British and American English. *Language Variation and Change* 5(2), 113–39.
Leblanc, Hélène. 2020. Grammar in the early modern period. In Dana Jalobeanu & Charles T. Wolfe (eds.), *Encyclopedia of early modern philosophy and the sciences*, 781–84. Cham: Springer International.
Lickona, Thomas. 1991. *Educating for character: How our schools can teach respect and responsibility*. New York: Bantam Books.
Marquilhas, Rita & Iris Hendrickx. 2014. Manuscripts and machines: The automatic replacement of spelling variants in a Portuguese historical corpus. *International Journal of Humanities and Arts Computing* 8(1), 65–80.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2000. The third-person singular -(E)S and -(E)TH revisited: The morphophonemic hypothesis. In Christiane Dalton-Puffer & Nikolaus Ritt (eds.), *Words: Structure, meaning, function: A festschrift for Dieter Kastovsky*, 235–48. Berlin and New York: Mouton de Gruyter.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Pearson Education.
Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper & Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics Conference: CL 2007*. University of Birmingham: UCREL.
Russo, Daniel. 2026. Metadata and annotation in historical text corpora: Standards, challenges and best practices. *English Language and Linguistics* 30(3) (special issue, Angela Andreani & Daniel Russo (eds.), *Investigating the metalanguage of English: Texts, corpora and models*). https://doi.org/10.1017/S1360674326100847
Salmon, Vivian. 1999. Orthography and punctuation. In Roger Lass (ed.), *The Cambridge history of the English language*, vol. 3, 13—55. Cambridge: Cambridge University Press.
Schmied, Josef. 1994. The Corpus of Early Modern English. *Anglistik und Englischunterricht* 53(1), 75–93.
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press

Tribble, Christopher & Glyn Jones. 1997. *Concordances in the classroom*. Houston, TX: Athelstan.
Tyrkkö, Jukka. 2013. Printing houses as communities of practice: Orthography in early modern medical books. In Joanna Kopaczyk & Andreas H. Jucker (eds.), *Communities of practice in the history of English*, 151–76. Amsterdam: John Benjamins.
Tyrkkö, Jukka. 2020. Early modern medicine in manuscript and print: A triangulation approach to analysing spelling standardisation. *International Journal of English Studies* 20(2), 67–93.

Cite this article: Andreani, Angela and Vahid Asadi. 2026. Exploring the *MetaLing Corpus*: spelling variation and discourse in Early Modern English texts on language. *English Language and Linguistics* 30, no. 3: 538–560. https://doi.org/10.1017/S1360674326100859

https://doi.org/10.1017/S1360674326100859 Published online by Cambridge University Press