Linguistica e Filologia

43

INDICE

Linguistica e Filologia

43

Angela Andreani
(Università degli Studi di Milano)

Daniel Russo
(Università degli Studi dell'Insubria)

Building a Corpus of the Metalanguage of English Linguistics 1500-1700: Methodological Issues[1]

Abstract
This paper focuses on the methodological issues involved in the selection of sources for a corpus-based study of the English metalanguage that was created to analyse and compare, appraise and classify, teach and learn the vernacular languages of Europe between 1500 and 1700 (MetaLing corpus). Source selection is in fact a critical aspect presenting multiple challenges, since, in a period predating the establishment of comparative philology and linguistics as academic disciplines, language-related discussions are to be found not just in grammar books and language manuals, but in works pertaining to different fields and presenting a large variety of aims. Building on previous research and combining lexicographical analysis and corpus linguistics, the aim of this paper is to explore the potential of (semi-)automated searches of online digital resources for the retrieval of underexplored or non-canonical texts. Integrating what may be conceptualised as bottom-up (lexis to subject) approaches, with top-down (subject to lexis) approaches, the results indicate that onomasiological-oriented approaches have the potential to make terminological blind spots emerge, and that the combination of different onomasiological and semasiological oriented approaches interfacing with the same corpus helps overcome the limits of each individual approach.

Key words: history of linguistics, corpus linguistics, terminology, Early Modern English, corpus methods

1. Introduction

This paper describes the initial stages of a corpus-based study of the English metalanguage that was created to analyse and compare, appraise and classify, teach and learn foreign languages, dialects and varieties between

---

1   Both authors are responsible for the overall planning and research for this paper. In particular, Angela Andreani is responsible for sections 1, 2.1, 2.2 while Daniel Russo for section 2.3.1, 2.4, 2.5, 2.6. Sections 2.3 and 3 were written jointly by the two authors.

1500 and 1700.[2] The twenty volumes of Robin Alston's monumental *A bibliography of the English language from the invention of printing to the year 1800* are the chief reference work. Studies have then been devoted to individual periods and leading figures in Britain (Gargani 1966; Polk 1989; Salmon 1988; Subbiondo 1992, 2001), traditions and movements (Beal & Iamartino 2016; Brengelman 1980; Maat 2004; Lewis 2007), language teaching (Howatt 1984; McLelland 2015), and individual genres, from grammars (Algeo 1985; Dons 2004; Robins 1986) to the large body of scholarship on English dictionaries (Considine 2022; De Witt Starnes & Noyes 1946; Hayashi 1978; Ogilvie 2020; Stein 1985). Building on this body of scholarship, the aim of our project is to build a corpus of English texts dedicated to or including discussions of vernacular languages, in order to study the dissemination of linguistic knowledge through English texts and genres and the development of the English metalanguage of linguistics. Despite much valuable scholarship on pre-modern English linguistics, in this paper we will refrain from a bibliographic approach combining the lists of primary sources of existing published scholarship and/or relevant subject bibliographies such as Alston's. In fact, the availability of large corpora of English texts in digital form, such as EEBO and ECCO, prompts us to consider how and to what extent these resources can enhance our knowledge of the discourse on language in Early Modern sources. In particular, as shall be seen, our ambition is to verify to what extent source retrieval can be "automated" with the aid of digital and computational methods. The project is therefore divided into three phases: 1) collection of texts, 2) building the corpus, 3) lexical extraction and database creation. In this paper, we focus on the methodological challenges involved in phase 1, since the selection of sources is a critical aspect that poses multiple issues.

In a period predating the establishment of comparative philology and linguistics as academic disciplines in the 19th century, language-related discussions are to be found in works with a large variety of aims and fields (Van Hal 2019; Swiggers 2010; McConchie 2012). Early Modern scholars would not have seen themselves as linguists and presented their work in these terms; rather, their primary activity may have ranged from pedagogy

---

2 An earlier version of this paper was presented at the 2022 Henry Sweet Society Symposium "Blind Spots in the History of Linguistic Thought: Forgotten or Neglected Voices, Areas, Approaches and Methods" in Leuven, Belgium. The authors wish to express their gratitude to the participants for their helpful comments and suggestions.

and history to natural philosophy. As a result, Early Modern language-related discourse is often found in contexts that may appear unexpected to modern readers, who are more accustomed to precise disciplinary boundaries.

Two such contexts in the Early Modern English landscape are chronicles and herbals. That chronicles often featured linguistic musings may not be that surprising, as Early Modern writers interested in linguistic kinship were often active as historians or antiquarians, and vice versa (e.g. Brackman 2012; Considine 2017; Parry 1995; Ruano-García 2018). Several examples of language-related discourse may be found in the pages of *Remaines Concerning Britaine* by William Camden, where the historian posited the affinity between Welsh and Gaulish and discussed the Germanic origins of English (1605: 13-14), but several others contributed to this theme (e.g. Ware 1633: 10-11). Less predictably perhaps (but the genre is well-known to historians of lexicography, see Considine 2022; De Witt 1954; Rydén 1994), debates over the names of plants in 16th-century English herbals reveal fascinating aspects of the Early Modern conception of the relationship between language, world and mind. Herbalists like William Turner and John Gerard took issue with denominations that failed to denote the observable features of plants, such as their uses or physical characteristics (Turner 1568: 99-100; Gerard 1597: 391).

In a literary and cultural landscape in which even herbals and chronicles may include digressions on matters that we would identify as linguistics today, the first challenge we are facing is that of developing a viable methodology to collect relevant sources. Our aim is to arrive at a better understanding of the genres and text types in which we could expect to find discussions on language(s), and thus potential evidence of unconventional uses of metalanguage and terminologies. Incomplete knowledge of the contexts in which linguistic ideas circulated limits our understanding of the nature and history of linguistic discourse, but it also affects the ways in which we can proceed to gather relevant texts if we want to explore writings not yet charted by scholars. Reflecting a situation in which linguistic knowledge and terminology were practised and communicated in a variety of fields, we plan to adopt a combination of approaches and methodologies in order to increase our chances of retrieving non-obvious sources.

This is in fact a familiar problem for scholars of the history of linguistic ideas and language history. Focussing on a specific "branch" of

language-related discourse, an attempt at the systematic collection of pre-modern historical sketches of English was recently undertaken by Alicia Rodríguez-Álvarez, who compiled an extremely valuable catalogue of 47 early and late modern sources (2018: 99-132). Rodríguez-Álvarez reaches very interesting conclusions regarding the circulation of these sketches, their intended readership and the dissemination of linguistic concepts from antiquarian works to later grammars, dictionaries and textbooks, thus showing the enormous potential of research in this direction. As she points out, however, "a major problem [...] was finding out which texts contain such accounts", and the retrieval of relevant texts was a "time-consuming process" which naturally entailed "much reading of primary and secondary sources" (2018: 101). Her method is essentially bibliographic, although digital and computational methods complement her searches on a component of her corpus, since the Eighteenth-Century English Grammars Database (ECEGD) compiled by María E. Rodríguez-Gil and Nuria Yáñez-Bouza (https://eceg.iatext.ulpgc.es/) allows users to refine searches by focussing on the "Subsidiary Contents" of works, in which fields such as "Origins of English language/of languages" and "Comparison of languages" can be selected. This is an essential feature that helps speeding up the identification of relevant sources significantly and is of particular relevance for the purposes of our project.

Indeed, our ambition is not only to increase our chances of retrieving sources not known to discuss language-related issues, but also to verify to what extent source retrieval can be "automated" with the aid of digital and computational methods. The retrieval of sources represents the heuristic phase of our project, which we have begun to explore testing the methods discussed in the next section of this paper.

2. *Methodological overview*

2.1 *Premise*

Considering the science of language was not yet established as a discipline before the 19th century, the challenge consists in rightly dealing with what we perceive as an overlap between "linguistics", broadly construed as reflections and observations upon language(s), and other fields, e.g. didactics, history, and theology. Taking an overly

inclusive attitude towards this discipline would make it a mere umbrella term for a variety of different practices; on the other hand, taking an overly narrow focus would miss several important connections with other fields, which in turn arguably influenced the evolution of linguistics. Thus, the challenge of this project is inherently linked to the plethora of text types involved. In addressing this difficulty, we take a pragmatic approach. By "pragmatic" we mean that we focus first on how language-related discourse was practised at the time and how this practice can be feasibly reconstructed. Instead of taking at face value today's definitions and assumptions about what Early Modern linguistics was, we attempt to reconstruct what was presented as language as a phenomenon during the 16th and 17th centuries. In particular, we aim at inventorying all those authors and works that contributed to define how languages were described and explained during the period under discussion. We intend to implement this pragmatic approach as an element of awareness in that it does not require any substantial claim about the essence and scope of linguistics in today's terms. Instead, we intend to focus on the (changing) way in which languages were presented as a subject of inquiry, appraisal, teaching and learning during the period. The availability of digital archives of Early Modern English sources, and the wide range and scope of online biographical, bibliographical and lexicographical databases prompt us to consider the application (and development) of digital and computational methods that could assist us in retrieving relevant texts in a meaningful and efficient way. In the sections that follow we illustrate the application, potential and limits of different methods for source retrieval, namely (a) the bibliographical method, (b) the biographical method, (c) corpus-based semantic method, (d) corpus-based collocate method, (e) the lexicographical method. It should be pointed out that methods (a), (c) and (d) are based on the same corpus.

2.2 *Bibliographical methods*

We have started from the digital databases of the *Universal Short Title Catalogue* (USTC) and *Early English Books Online* (EEBO) to explore their potential for the retrieval of relevant sources and of non-canonical and underexplored texts. The USTC allows searching all books printed with moveable type between 1450 and 1650. The catalogue can be searched by date, place, format or subject, and since it contains the links to the

full digital scans of many of the books, it is a convenient way not just to collect bibliographical data but also to access the texts. Of special interest to us are the subject categories within the USTC classification scheme, which identify, amongst others, "Dialectics and rhetoric", "Dictionaries, vocabularies, phrase books, instruction in foreign languages" and "Linguistics and philology". For example, a search by language "English" and subject "Linguistics and philology" retrieves 40 results that can be analysed for inclusion in the corpus. The potential of digital searches on the USTC interface however ends here, since free-text keyword searches are not possible, nor is it possible to combine different subject categories, which considerably limits the ways in which we can interact with the catalogue. Yet, the USTC subject classification has been added to EEBO records and it can be found in the Advanced Search section on the ProQuest platform (https://www.proquest.com/eebo). Here, users can select multiple subjects from the field "USTC subject classification" or from the field "Subject heading", corresponding to the full Library of Congress subject headings.[3] Results can be filtered by date, author, language, place of publication, publisher and source library. For instance, selecting the combinations "language and languages - origin - early works to 1800" OR "language and languages - grammars - early works to 1800" OR "language and languages - early works to 1800" OR "language and languages - glossaries, vocabularies, etc. - early works to 1800" OR "language and languages – grammars" OR "language and languages – origin" in the field "Subject heading", and filtering by language, we obtained seven works published between 1614 and 1668, including John Wilkins' *An essay towards a real character, and a philosophical language*, which is known to be relevant to our corpus. Another result is the treatise *Religion and Language, as they are Now in use through the Chief Regions of the World* by an anonymous "Person of Quality", of high relevance for our corpus, judging from its table of contents.[4]

Another possible way of searching EEBO is by "Document Text Subsections", which comprises the subsection "Table of Contents" (TOC). Building on the results obtained from the ECEG corpus by Rodríguez-

---

3   For a full explanation of the various fields in the advanced search function in EEBO see https://proquest.libguides.com/eebopqp/fields (accessed June 13, 2023).

4   The treatise has chapters on the origins of Italian, French and Spanish, on "Slavonish, Turkish and Arabique languages" and on the "Syriaque and Hebrew tongues" (1664: A2).

Álvarez, we have thus performed a search for the terms *language* and *tongue*,[5] and variant spellings, in the TOC of works in English between 1500 and 1700. The results are encouraging, as we have retrieved 765 works (following date range filtration) including sources known to be relevant to our corpus (i.e. Richard Mulcaster's *Elementarie*, and *The compleat French-master for ladies and gentlemen* by Abel Boyer) as well as less obvious ones. Non-canonical sources range from travel and historical writing, such as the 1626 edition of Samuel Purchas' *Purchas his Pilgrimage*, containing a section entitled "Of the diuersitie of Nations and Languages, and of the Soyle and Climate", to religious works. The method also retrieves false positives and borderline cases. An example may be the 1674 polemical tract *Symbolon Theologikon* by the bishop and theologian Jeremy Taylor (1613-1667). *Language* and *tongue* appear twice in the TOC in contexts that may or may not be relevant, i.e. section 1, chapter 2 is entitled "The notion of those words that in the Greek and Latin languages express Repentance with the definition and parts of it", indicating that this source deserves further inspection. Although religious polemic dealt with language in ways that are not necessarily connected to linguistic inquiries and practices, the significance of exegetical controversies for the evolution of linguistic thought is not in doubt.[6]

The results retrieved through the bibliographical method indicate that subject searches and key-word searches in digital corpora have the potential to make unexpected sources emerge. Yet, not all texts contained in EEBO are machine-readable or complete, and the analysis of results still requires close reading on the researcher's part in order to locate relevant sections within works, to identify false positives, and to assess the actual state of the digitised versions of the sources retrieved.[7] The EEBO advanced search function still represents an invaluable tool with inherent advantages over manual bibliographic research; in particular, subject headings gather the knowledge generated by a community of practice comprising scholars, archivists and librarians working on Early

5   They are recurrent terms in the titles of the sections dedicated to historical sketches of English identified by Rodríguez-Álvarez (2018: 103-5).

6   The bibliography is extensive, but amongst recent studies see for instance Anderson 1996; Cummings 2007; Rosendale 2001.

7   There is an ongoing debate on the advantages and limitations of EEBO for textual scholarship, a helpful overview and introduction to the history of EEBO is Gavin 2017.

Modern sources, on which researchers can rely to run searches that can be automated to some extent, considerably speeding up the process of scanning large collections of texts.

2.3 *Biographical method*

Building on the methodology discussed in Authors, we have tested the potential of another digital resource, the *Oxford Dictionary of National Biography* (ODNB). The ODNB does not contain texts but allows for searches that can be tailored to our questions. Using the refine search function, it is possible to look up authors selecting a specific "statement of occupation" and timespan.[8] For instance, typing *linguist* as a statement of occupation and 1500-1700 as our timespan, we obtained 8 results, shown in Table 1.[9]

Table 1. Biographical method: linguists 1500-1700.

Surname	Name	DoB	Death	Statement of occupation
Chamberlain	Nathaniel	b.1612	-	linguist and physician
Higgins	John	1544	1620	poet and linguist
Hollyband	Claudius	1534/5	1597	linguist and writer
Pasor	Mathias	1599	1658	linguist and philosopher
Russell	Elizabeth	1540	1609	linguist and courtier
Webbe	Joseph	-	1630	linguist and physician
Wheelocke	Abraham	1593	1653	linguist and librarian
Wotton	William	1666	1727	linguist and theologian

8   It should be pointed out that the statements of occupation are freely chosen by the authors of the entries in the ODNB; hence they do not necessarily reflect the early modern understanding of professional categories but rather the way they have been codified by modern scholarship.

9   The label 'linguist' is used as a keyword to run automated searches to help us identify intellectuals whose work has been connected with the study of language by the editors of the ODNB. It is a partially anachronistic label when applied to early modern scholars (although the term was indeed in use in the early modern period, see OED s.v. 'linguist, n'), but as we show, we merely use it as a convenient starting point to broaden the scope of the occupational labels that might lead us to identify early modern figures involved in language-related scholarship.

The names thus retrieved left out authors unanimously known to have been central to the development of the field, such as George Dalgarno, John Wallis and John Florio; therefore, we tried to expand our search keys in order to retrieve a more complete set of sources using the words *grammarian* (14 results) and *philologist* (11 results) as statements of occupation. However, obvious "linguists" were still left out, so we decided to backtrack their statement of occupation to see whether further results could be found. Dalgarno is catalogued as "writer on language", a statement of occupation which led us to include two further authors (Cave Beck, "writer on universal language and church of England clergyman" and Thomas Dyche, "schoolmaster and writer on language"). As a "mathematician and cryptographer", John Wallis remains on his own, while as a "scholar of languages" Florio is in company of Josephus Abudacnus, John Keigwin, John Palsgrave and Anthony Raymond. This led us to the statement of occupation "Anglo-Saxon scholar", which gave another 6 results, to complete our pilot list of authors which can be seen in Table 2.

Table 2. Biographical method: list of authors.

Surname	Name	DoB	Death	Statement of occupation
Abudacnus	Josephus	fl.1595	1643	scholar of Oriental languages
Beck	Cave	b.1622	1706	writer on universal language and church of England clergyman
Bentley	Richard	1662	1742	philologist and classical scholar
Burton	William	1609	1657	antiquary and philologist
Butler	Charles	1560	1647	philologist and apiarist
Chamberlain	Nathaniel	b.1612		linguist and physician
Dalgarno	George	1616	1687	writer on language
Dyche	Thomas	-	1722x7	schoolmaster and writer on language
Elstob	Elizabeth	1683	1756	Anglo-Saxon scholar

Elstob	William	1674	1715	Anglo-Saxon scholar and Church of England clergyman
Florio	John	1553	1625	author and teacher of languages
Higgins	John	1544	1620	poet and linguist
Hollyband	Claudius	1534/5	1597	linguist and writer
Jacob	Henry	1608	1652	philologist
Junius	Franciscus	1591	1677	philologist and writer on art
Keigwin	John	b.1642	1716	scholar of the Cornish language
Lhuyd	Edward	1659/60	1709	naturalist and philologist
Lisle	William	1569	1637	translator and An-glo-Saxon scholar
Littleton	Adam	1627	1694	Church of England clergyman and philologist
Malcolm	David		1748	philologist
Marshall	Thomas	1621	1685	Dean of Gloucester and philologist
Pakington	John	1621	1680	politician and Anglo-Saxon scholar
Palsgrave	John	-	1554	teacher and scholar of languages
Pasor	Mathias	1599	1658	linguist and philosopher
Raymond	Anthony	1675	1726	Church of Ireland clergyman and Irish language scholar
Rhys	Siôn Dafydd	1533/4	1620	grammarian
Rightwise	John	1490	1533	grammarian
Robert	Gruffydd	fl.1558	1598	Roman Catholic priest, humanist, grammarian
Robertson	William	fl.1651	1685	grammarian and lexicographer

Russell	Elizabeth	1540	1609	linguist and courtier
Salesbury	Henry	1560/61	1632	grammarian and lexicographer
Sanford	John	1565	1629	poet and grammarian
Simson	Andrew	1526	1591	grammarian and church of Scotland minister
Skinner	Stephen	b.1623	1667	physician and philologist
Somner	William	b.1598	1669	antiquary and Anglo-Saxon scholar
Stanbridge	John	1463	1510	schoolmaster and grammarian
Thomson	Richard	-	1613	philologist and Church of England clergyman
Thwaites	Edward	b.1671	1711	Anglo-Saxon and Greek scholar
Tonneys	John	-	1514	prior of the Austinfriars and grammarian
Vaus	John	1484	1539	grammarian
Vossius	Isaac	1618	1689	philologist and author
Wallis	John	1616	1703	mathematician and cryptographer
Webbe	Joseph	-	1630	linguist and physician
Wedderburn	David	b.1580	1646	poet and Latin grammarian Latin scholar
Wharton	Jeremiah	fl.1654		grammarian
Wheelocke	Abraham	1593	1653	linguist and librarian
Whittington	Robert	1480	1553	schoolmaster and grammarian
Willymott	William	b.1672	1737	grammarian
Wotton	William	1666	1727	linguist and theologian

Following this procedure, it has indeed been possible to expand our set of relevant authors, although it is evident that this is not a very productive method given that it can hardly be automated; moreover, with respect to our ambitions this approach shows inherent deficiencies.

Crucially, some authors that we would have expected to encounter due to their fundamental contribution to the history of linguistic thought were left out. It is the case of John Wilkins, catalogued as a "theologian and natural philosopher"; of Francis Bacon, listed as "Lord Chancellor, politician and philosopher"; and of John Locke, defined "simply" as a "philosopher". This means their statements of occupation include their historically more established definitions, which makes the retrieval of authors barely viable from a strictly computational point of view, because this type of information is hardly quantifiable or codifiable, as we do not have precise in-text or metatextual cues that can be scraped in an automated fashion. The method must in fact rely on occupational labels that do not attempt to be uniform or systematic, but are created individually by the authors of the biographical essays of the ODNB. Neither did we come across any historians or religious controversialists who wrote on language; instead, the procedure yields several false positives, such as authors writing in Latin, not to be included in our corpus according to our premises. Finally, one obvious fundamental issue of this procedure is that we get the authors, but certainly not the texts, even though the ODNB entries provide titles of works of potential relevance to expand our corpus. For these reasons, the method does not seem to be helpful to lead us to non-obvious sources, though it can assist us in expanding our knowledge of potential writers for our corpus.

Even so, there is still something that we believe is valuable in this method. First, the diverse occupational labels associated with authors known to have contributed to the development of linguistic ideas and discourse effectively show the extent to which language-related concerns in the early modern period transcended present-day disciplinary boundaries. Additionally, although it does not supersede traditional bibliographic research, the method may be helpful in identifying the range of intellectual occupations that involved an interest in language(s), as well as the names of the people and their connections, which provides helpful background information for archival research.

2.3.1 *Further applications of the biographical methods*

Besides using metatextual tags such as "statement of occupation", there is another corpus-based approach that can be experimented with within this bio-bibliographical method. To further expand the range of

sources, the biographical profiles of ODNB can be queried as a corpus by using the advanced search tool to extract relevant keywords in the full-text mode. However, this procedure involving in-text queries in the ODNB raises various issues that compromise its viability. For example, using the advanced search tool, we carried out a full-text search for *linguist* or *linguistic* in the date range under investigation, which enabled us to expand our set. However, a large number of false positives were also retrieved, which prevented a fully automated process. The examples listed below show the nature of these false positives. In (1), the description of the conspirator and informer Charles Bailey shows that he was a polyglot, but since he did not write anything about language, he cannot be included in our sources. Similarly, in (2), navigator William Adams is presented as a multilingual official with no record of written reflections about language, and in (3), government official William Blathwayt's proficiency in Dutch is described as a promoting factor in his career. In (4), royal army officer Henry Bard also moved up the career ladder thanks to his linguistic skills, which did not produce any relevant textual sources. Interestingly, a large number of entries obtained through this method returned results such as (5): in the army officer William North's entry, his sister is indicated as a linguist proficient in classical languages, but neither produced any significant writings in the field of linguistics.

(1) His use of the IHS monogram underlines his allegiance to Catholicism, while his use of French and Italian phrases points to his linguistic skills. (Charles Bailey)

(2) His services, especially his linguistic ones in official negotiations, were valuable but not indispensable to either the English or the Dutch. (William Adams)

(3) His linguistic skill (he was the only official at the embassy who knew Dutch) rendered him a most useful member of the staff. (William Blathwayt)

(4) Bard returned to England about 1642, and on the outbreak of the civil war his reputation as a traveller and linguist secured him a colonel's commission, possibly through the influence of the queen, Henrietta Maria. (Henry Bard)

(5) Educated privately along with her brothers by their tutors, her [Dudleya North's] natural facility as a linguist was obvious and she quickly gained fluency in Latin and Greek. (William North)

2.4 *Corpus-based semantic method*

The next approach that was considered for the purposes of this paper is called corpus-based semantic method**.** The previous approach involves retrieving texts by selecting authors through a top-down approach; conversely, this bottom-up method starts directly from annotated texts. Semantic Early English Books Online (EEBO)[10], a semantically annotated version of the above discussed EEBO, is the ideal tool for this task. Large textual datasets, such as linguistic corpora, can be annotated with semantic tags to create new, effective ways to explore the material they contain. Users can search Semantic EEBO not only for words but also for concepts and can quickly and precisely investigate how these concepts relate to one another, a task that might be lengthy and laborious when using conventional resources. This tool can be used for both semasiological and onomasiological queries, as it allows browsing of both vocabulary and semantic fields. Furthermore, this tool promises that its semantic annotation removes the need to laboriously filter away irrelevant results from search results to find the desired meaning of a word that has several definitions; however, we will demonstrate that this is not always the case.

One of the advantages of Semantic EEBO is that lexical items extracted from the EEBO Corpus are already categorised. The terms can be browsed in a highly intuitive interface divided into macro- and microfields: Figure 1 shows Language as a macrocategory and the subcategories "languages of the world", "speech", "narration", and so on.

Figure 2 presents a sample search that was carried out by selecting a random word (*mutations*) from within the semantic category "Language: Phonetic and Phonology". Given the semantic label under which it is listed, this should undoubtedly be related to the contents of our corpus. Selecting the 1550s as a timespan, the interface returns the list of the first occurrences of the word in context, and through this, we can retrieve the texts in which the term was first used. This is a purely semasiological approach based on previously classified lexical items. This appears to be extremely promising for the purposes of the corpus; nevertheless, the results clearly show that there is an inherent problem in this semantic

---

10   Available at https://www.english-corpora.org/ (accessed June 13, 2023).

Figure 1. Linguistics-related macro and micro categories in Semantic EEBO
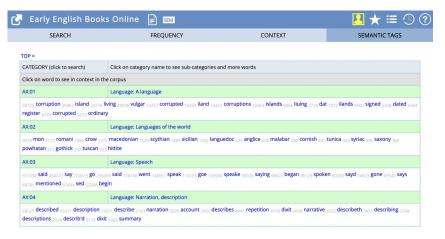(source: https://www.english-corpora.org/eebo/ accessed June 13, 2023).

Figure 2. Results of the semantic search of mutations in Semantic EEBO
(source: https://www.english-corpora.org/eebo/ accessed June 13, 2023).

categorisation. For instance, the first occurrences of *mutations* in Figure 1 show that they have no semantic reference to linguistics; in fact, the linguistic acceptation of the term is only first attested in 1808 (OED s.v. 'mutation, n.'). Terms like this are listed because they later acquire a language-related meaning, but there is no way to filter them out. Therefore, this classification makes no onomasiological distinction between the meanings of the same term in different areas.

This brings us to the significant shortcomings of Semantic EEBO. As shown above, the earliest occurrences of terms do not necessarily relate to the semantic field under investigation. However, an extra corollary of interest to this is that this presence of unrelated lexical items allows us to gain evidence of linguistic terminology in development so that we can distinguish which words started off as mainly "linguistic" and which were subject to semantic shift and acquired specific/specialised meanings.

2.5 *Corpus-based collocate method*

In his review of different approaches to collocational theory for lexicographic practice, Siepmann (2006) states that "the traditional dictionary-making process should be turned on its head: rather than starting from an alphabetical framework, it should proceed from a bilingual or multilingual onomasiological research base". A similar corpus-based approach was adopted by Price (2015) in his lexicographic study of New Testament Greek. Various studies in the last two decades (e.g. Grzega 2002; Geeraerts 2009; Fernández-Domínguez 2019; Keersmaekers and Van Hal 2022) have shown the potential of syntactic relations as a way to expand our lexicological knowledge starting from context rather than from lexical items intended as lemmas.

One way to apply this approach is to browse an existing corpus of historical texts of the period under investigation using corpus linguistic tools to find collocates of the terminology that have emerged from other methods listed in this paper. EEBO is available on the corpus manager and text analysis software Sketch Engine (Kilgariff et al. 2004, Kilgariff et al. 2014). By analysing the word "sketches" (collocates) and the concordances of terms, we reached less common, less obvious or obsolete language-related terminology. For example, by searching the collocates of *noun* in EEBO through Sketch Engine (Figure 3), especially those with fewer occurrences, we can see that there are five unexpected occurrences with the verb *to betoken* in the list of verbs collocating with *noun*. These appear in only two works written in 1653 (Lloyd's *Latine Grammar*) and 1669 (Milton's *Accedence commenc't grammar*), in which this collocation is associated mainly with gender expressions or noun classes, thus meaning *be assigned with*, *carry* or *signify*.

Figure 3. KWIC concordances of noun + to betoken *in EEBO*
(source: https://www.sketchengine.eu/ accessed June 13, 2023).

This onomasiological method has two advantages: it does not rely on previously selected (and potentially biased) sources and it has the potential to make more terminological blind spots emerge. It is mainly automatised in the outputs; on the other hand, it requires close reading by the researcher. The potential of this method can be further expanded combining it with the use of a historical thesaurus in order to include searches of historical variants of linguistic terminology (eg. *name*, *common name*, or *appellative*, all attested alongside *noun* in the period under scrutiny).

2.6 *Lexicographical method*

Following what Kipfer (1986) postulated, online dictionaries can be useful tools for extracting not only semasiological results but also onomasiological data, provided that the textual references that can be retrieved in lemmas are sufficient. In this methodological review, a similar approach was applied by using the advanced research of the Oxford English Dictionary (OED) online to retrieve the first citations in the its "Categories" section. Through this procedure, it is possible to create a database of first occurrences, extract language-related lemmas and record relevant metadata such as author, work title, date and definition. More specifically, in the OED online, this is done by selecting Categories > language > linguistics and by setting the date of the first citation. This produces helpful results that are operationally more direct than those obtained through the bio-bibliographical method. Because of the visual properties of the website, the portal tends to list authors or sources with a significant number of results and automatically hides sources with fewer results (especially those with one result). The only solution to obtaining this information from the OED website is to further break down the timespan, which means searching within 1501–1510, 1511–1520, 1521–

1530, and so on. This method reveals the sources with only one result. This is potentially tedious because the information we require can be retrieved in a large number of queries; however, this also has the potential to become an automated process that we are trying to investigate in this project. In a merely explorative fashion, a quick Python script was implemented to extract the first citations of the linguistic lemmas from 1500 to 1509 and to copy them into an Excel spreadsheet (see Figure 4). This has proven successful; however, the script must be launched for every decade (which can possibly be further automatised), and the resulting spreadsheet misses important data, such as actual citations.

Figure 4. Explorative extraction of first occurrences of linguistic terminology in the timespan 1501-1510 from the OED online.

	author	lemma	work-title	date
1	author	lemma	work-title	date
2	Holt J.	accident	Lac Puerorum	1504
3	Holt J.	active	Lac Puerorum	1504
4	Holt J.	adversative	Lac Puerorum	1504
5	Holt J.	impersonally	Lac Puerorum	1504
6	Holt J.	potential	Lac Puerorum	1504
7	Holt J.	reciprocation	Lac Puerorum	1504
8	Holt J.	regiment	Lac Puerorum	1504
9	Holt J.	subcontinuative	Lac Puerorum	1504
10	Holt J.	subjunctive	Lac Puerorum	1504
11	Holt J.	verbal	Lac Puerorum	1504
12	Holt J.	verbal noun	Lac Puerorum	1504
13	Barclay Alexander	out-	Brant's Shyp of Folys	1509
14	Thomson D.	agree	Middle Eng. Grammatica	1500
15	Longe Paruula	presentens	Longe Paruula	1509
16	Parliamentary papers	verbatim	Parliamentary papers	1503
17	Henryson Robert	write	Poems	1505
18	Hawes Stephen	declinal	Pastime of Pleasure	1509

The best way to obtain raw data in a more systematic and automated fashion is to use the OED Research Application Programming Interface (API). By employing a range of advanced endpoints to access data, this tool facilitates quick and effortless data retrieval and manipulation from the OED, thereby paving the way for innovative research avenues, as exemplified in this project. The API entails some level of command of HTML-based coding language to perform advanced queries quickly. This tool is not yet available to general OED users; however, those who

are interested can apply to gain access to it. At the time of writing this paper, we are still waiting for this access, which has required long-running correspondence with the OED support team[11]. Obviously, we are aware of one limit of this approach at the structural level, which has to do with the biases associated with source selection. At present, the OED is attempting to overcome and correct it, but representativeness is still slanted (as can be seen in other studies such as Brewer 2010; Considine 2009).

3. *Conclusions*

This paper poses the question of how to build a corpus of texts containing linguistics terminology and concepts representative of Early Modern linguistic metalanguage in English. Various approaches were described with their strengths and weaknesses (Table 3).

Table 3. Strengths and weaknesses of the methods described in this paper.

Approach	Strengths	Weaknesses
a. Bibliographical method	Finds unexpected sources Speeds up bibliographic research	Can hardly be automatised False positives
b. Biographical method	Finds unexpected authors	Does not include texts Can hardly be automatised Large number of false positives
c. Corpus-based semantic method	Finds earliest occurrences Study linguistic terminology in development (resemantisation)	Failures in in-built categorisation
d. Corpus-based collocate method	Collocation-based Finds blind spots and previously unobserved terminology	Only retrieval of lexical items can be automatised
e. Lexicographical method	Potentially completely automatised Selection bias of the sources	Unlikely to find blind spots

11   At the time of writing this article, the online OED was still using the previous graphical user interface, which hindered automatic data extraction. However, as of late August 2023, the new OED user interface now facilitates the automatic extraction of lexical queries in the form of spreadsheet files.

It may be noted that the bibliographical (a), corpus-based semantic (c) and corpus-based collocate (d) methods represent three ways of interrogating the same electronic corpus. While (c) and (d) adopt what may be conceptualised as a bottom-up (lexis to subject) approach, (a) can be described as top-down (subject to lexis). Combining different onomasiological and semasiological approaches interfacing with the same corpus is essential to transcend the limitations inherent to each individual approach.

To create our corpus, our initial intention was to rely on an approach that may be characterised by some level of automation, given the availability of online digital resources and in order to complement traditional bibliographical research, since it was obvious that the few researchers involved in the project could not read a potentially enormous number of texts to avoid blind spots during text collection. This approach, in fact, represents a novel endeavour, prompting us to systematically evaluate diverse resources to gauge the degree of automation achievable with each. The analysis of the different approaches indicated that there was only partial room for automation unless we relied on one source of preselected texts (such as the OED). As such, our conclusion is that a combination of onomasiological and semasiological approaches is recommended for this task. We conclude that the starting point can be an integration of the lexicographical method – to retrieve the bulk of terminology that has already been collected by the OED (possibly facilitated by the advanced user interface of the new online OED, which enables data extraction, see note nr. 10 above) – and the onomasiology-oriented corpus-based collocate method – to explore potential lexical blind spots emerging from the terms collected in the former approach.

Angela Andreani
Università degli Studi di Milano
piazza S. Alessandro 1, 20123 Milano (MI)
angela.andreani@unimi.it

Daniel Russo
Università degli Studi dell'Insubria
via Dunant 7, 21100 Varese (VA)
daniel.russo@uninsubria.it

References

1 *Primary sources (all consulted via EEBO)*

[Anonymous]. 1664. *Religion and Language, as they are Now in use through the Chief Regions of the World*. London: Printed for Samuel Speed.

Camden, William. 1605. *Remaines of a Greater Worke, Concerning Britaine*. London: Printed by G[eorge] E[ld] for Simon Waterson.

Gerard, John. 1597. *The Herball Or Generall Historie of Plantes*. London: Iohn Norton.

Hanmer, Meredith. 1633. 'The Chronicle of Ireland'. *Two histories of Ireland*, edited by James Ware. Dublin: The Society of Stationers and London: Thomas Harper.

Lloyd, Richard. 1653. *The Latine grammar*. London: Thomas Roycroft.

Milton, John. 1669. *Accedence commenc't grammar*. London: Printed for S.S., and are to be sold by John Starkey.

Turner, William. 1568. *The First and Seconde Partes of the Herbal*. Cologne: Birckman.

2 *Secondary sources*

Aarsleff, Hans. 1967. *The Study of Language in England, 1780-1860*. Princeton: Princeton UP.

Algeo, John. 1985. The Earliest English Grammars. In Arn, Mary-Jo & Wirtjes, Hanneke & Jansen, Hans (eds.), *Historical and Editorial Studies in Medieval and Early Modern English for Johan Gerritsen*. Groningen: Wolters-Noordhoff.

Alston, Robin. 1974-. *A bibliography of the English language from the invention of printing to the year 1800*. Ilkley, Yorkshire: Janus Press..

Anderson, Judith H. 1996. *Words that Matter: Linguistic Perception in Renaissance English*. Stanford: Stanford University Press.

Beal, Joan C. & Iamartino, Giovanni, eds. 2016. *Towards a History of the English Normative Tradition*, monograph issue of Language and History, 59(1).

Brackmann, Rebecca. 2012. *The Elizabethan invention of Anglo-Saxon England: Laurence Nowell, William Lambarde, and the study of Old English*. Suffolk: Boydell & Brewer.

Brengelman, Fred H. 1980. Orthoepists, Printers, and the Rationalization of English Spelling. *The Journal of English and Germanic Philology* 79. 332–354.

Brewer, Charlotte. 2010. The use of literary quotations in the Oxford English Dictionary. *The Review of English Studies* 61(248). 93–125. (doi:10.1093/res/hgp020)

Considine, John. 2009. Literary Classics in *OED* Quotation Evidence. *The Review of English Studies* 60(246). 620–38. (doi:10.1093/res/hgp021)

Cummings, Brian. 2003. *The Literary Culture of the Reformation: Grammar and Grace*. Oxford: Oxford University Press.

De Witt, Starnes T. & Gertrude E. Noyes. 1946. *The English Dictionary from Cawdrey to Johnson 1604-1755*. Amsterdam/Philadelphia: John Benjamins.

De Witt, Starnes T. 1954. *Renaissance Dictionaries, English-Latin and Latin-English*. Austin: University of Texas Press.

Dons, Ute. 2004. *Descriptive adequacy of Early Modern English grammars*. Berlin: Mouton de Gruyter.

Fernández-Domínguez, Jesús. 2019. The Onomasiological Approach. In *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. (doi:10.1093/acrefore/9780199384655.013.5 79)

Formigari, Lia. 1988. *Language and experience in 17th century British philosophy*. Amsterdam: Benjamins.

Gargani, Aldo G. 1966. Idea, Mondo E Linguaggio in T. Hobbes E J. Locke. *Annali della Scuola normale superiore di Pisa. Lettere, storia e filosofia* 35(3/4). 251–292.

Gavin, Michael. 2017. How to Think about EEBO. *Textual Cultures* 11(1/2). 70–105. (doi: 10.14434/textual.v11i1-2.23570)

Geeraerts, Dirk. 2009. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Grzega, Joachim. 2002. Some Aspects of Modern Diachronic Onomasiology. *Linguistics* 40(5). 1021–1045. (doi:10.1515/ling.2002.035)

Hayashi, Tetsuro. 1978. *The Theory of English Lexicography 1530-1791*. Amsterdam: John Benjamins.

Howatt, Anthony P.R. 1984. *A History of English Language Teaching*. Oxford: Oxford University Press.

Kilgarriff, Adam & Baisa, Vít & Bušta, Jan & Jakubíček, Miloš & Kovář, Vojtěch & Michelfeit, Jan & Rychlý, Pavel & Suchomel, Vít. 2014. The Sketch Engine: ten years on. *Lexicography* 1. 7–36. (doi: 10.1007/s40607-014-0009-9)

Kilgarriff, Adam & Rychlý, Pavel & Smrž, Pavel & Tugwell, David. 2004. The sketch engine. *Proceedings of the 11th EURALEX International Congress*. 105–116.

Kipfer, Barbara Anne. 1986. Investigating an Onomasiological Approach to Dictionary Material. *Dictionaries: Journal of the Dictionary Society of North America* 8. 55–64.

Kohonen, Viljo. 1978. On the Development of an Awareness of English Syntax in Early (1550-1660). Descriptions of Word Order by English Grammarians, Logicians and Rhetoricians. *Neuphilologische Mitteilungen* 79. 44–58.

Lewis, Rhodri. 2007. *Language, Mind and Nature: Artificial Languages in England from Bacon to Locke*. Cambridge: Cambridge University Press.

Maat, Jaap. 2004. *Philosophical Languages in the Seventeenth Century. Dalgarno, Wilkins, Leibniz*. Dordrecht: Kluwer.

McConchie, Roderick. 2012. Doctors and Dictionaries in Sixteenth-Century England. In McConchie, Roderick (ed.), *Ashgate Critical Essays on Early English Lexicographers. Volume 3. The Sixteenth Century*. Farnham: Ashgate.

McDiarmid, John. F. 1980. *Language: System to Act to Report. The development of English linguistic thought in the sixteenth century*. Ann Arbor, Michigan-London: University Microfilms International.

McLelland, Nicola. 2015. *German Through English Eyes: A History of Language Teaching and Learning in Britain 1500-2000*. Wiesbaden: Harrassowitz Verlag. (doi:10.2307/j.ctvc2rjrp)

Ogilvie, Sarah. 2020. *The Cambridge Companion to English Dictionaries*. Cambridge: Cambridge University Press.

Parry, Graham. 1995. *The trophies of time: English antiquarians of the seventeenth century*. Oxford: Oxford University Press.

Price, Todd. 2015. *Structural Lexicology and the Greek New Testament: Applying Corpus Linguistics for Word Sense Possibility Delimitation Using Collocational Indicators*. Piscataway, NJ, USA: Gorgias Press.

Robins, Robert. H. 1986. The Evolution of English Grammar Books since the Renaissance. In Leitner, Gerard (ed.), *The English Reference Grammar. Language and Linguistics, Writers and Readers*, 292–306. Tübingen: Max Niemeyer Verlag.

Rodríguez-Álvarez, Alicia. 2018. Cataloguing the First Histories of the English Language Written from the Late 16th to the End of the 18th Century. *Historiographia Linguistica* 45(1). 99–132. (doi: 10.1075/hl.00017.rod)

Rosendale, Timothy. 2001. 'Fiery toungues:' Language, Liturgy, and the Paradox of the English Reformation. *Renaissance Quarterly* 54(4). 1142–1164.

Ruano-García, Javier. 2018. *Etymological Collections of English Words and Provincial Expressions*. Oxford: Oxford University Press.

Rydén, Mats. 1994. William Turner and the English plant names. In Kastovsky, Dieter (ed.), *Studies in Early Modern English*, 349–370. Berlin/Boston: De Gruyter Mouton.

Salmon, Vivian. 1988. *The study of language in 17th-century England*. Second edition. Amsterdam Philadelphia: John Benjamins.

Schäfer, Jürgen. 1980. *Documentation in the O.E.D.: Shakespeare and Nashe as Test Cases*. Oxford: Clarendon Press.

Siepmann, Dirk. 2006. Collocation, colligation and encoding dictionaries. Part II: Lexicographical aspects. *International Journal of Lexicography* 19(1). 1–39. (doi:10.1093/ijl/eci051)

Stein, Gabriele. 1985. *The English Dictionary before Cawdrey*. Tübingen: Max Niemeyer.

Stillman, Robert. 1995. *The New Philosophy and Universal Languages in Seventeenth-Century England: Bacon, Hobbes, and Wilkins*. Lewisburg: Bucknell University Press.

Subbiondo, Joseph L. (ed.). 1992. *John Wilkins and 17th-century British Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Subbiondo, Joseph L. 2001. Educational reform in seventeenth-century England and John Wilkins' philosophical language. *Language & communication* 21(3). 273–284.

Swiggers, Pierre. 2010. History and Historiography of Linguistics: Status, Standards and Standing. *EUTOMIA. Revista Online De Literatura e Linguística* 3(2). 1–18.

Van Hal, Toon. 2019. Early Modern Views on Language and Languages (ca. 1450-1800). In *Oxford Research Encyclopaedia of Linguistics*. Oxford: Oxford University Press.