statuS Quaestionis
language text culture

Fabio Ciambella
Sapienza University of Rome

Introduction: Old Language(s), New Technologies: Corpus Linguistics and European Languages in the Renaissance, 1400s-1600s

The Renaissance[1] is universally acknowledged to have been a crucial moment in Europe for the development of vernacular national languages, which begin to establish their prestige alongside Latin. Historical linguists have focused on the many interesting peculiarities of the European vernaculars in this period, such as the high degree of spelling fluctuation, (non-)lexicalisation of words, phonological and morphological adjustments, semantic shifts, etc. To study the diachronic development of languages, historical linguists have always employed the term 'corpus' and its plural form 'corpora', as aptly suggested by Merja Kytö (2010, 418). Nevertheless, the pre-electronic idea of corpus was that of "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer 2002, xi). Historical corpus linguists generally indicated a collection of texts "intentionally created to represent and investigate past stages of a language and/or to study language change" (Claridge 2008, 242). In this sense, historical linguistics has always been based on corpora, even non-digitised corpora. One of the first examples is surely the Corpus

1 As Alessandra Petrina has pointed out, "[t]he period between the fifteenth and the seventeenth century to which we most commonly apply the label of 'Renaissance', given its trans-European validity, poses more problems, and its definition as a turning point has repeatedly been questioned and challenged, with insistent voices proposing its substitution with the locution 'early modern'" (2019, 146). In this introduction, however, I will use the terms 'Renaissance' and 'early modern' interchangeably.

Inscriptionum Latinarum (CIL), begun in 1853 under Theodor Mommsen's direction and now held at the research centre at Unter den Linden 8 in Berlin. "The CIL counts 17 volumes in folio format in about 80 parts, containing almost 200,000 inscriptions" (https://cil.bbaw.de/en/homenavigation/the-cil/history-of-the-cil) in Latin, belonging to the former area of the Roman empire. The CIL also has a searchable database of carbon copies, photos, printing blocks, and records.

Nonetheless, today linguists tend to call historical corpus linguistics a methodology grounded in the use of computational linguistics applied to historical texts, what Kytö suggested should be called more accurately 'electronic historical corpus linguistics' more than ten years ago (2010), when she triumphantly declared that the methodology "emerged as a vibrant field that [...] significantly added to the appeal felt for the study of language history and change" (418). As is well-known, diachronic corpora and archives must be machine-readable to be accessible and analysable through computational tools, as underlined by McEnery et al. (2022, 394): "Unless those words can be rendered as machine readable text, then the archive remains a source of data only for those linguists who are willing to work directly with the written records using what we might term 'hand and eye' techniques".

Between the 1970s and 1980s, electronic corpora, then available to linguists and historical datasets "which include the time dimension as a design feature" (Tognini Bonelli 2010, 22), began to be compiled. Some experiments, however, had been conducted even before the second half of the last century. Considering Latin once again, McEnery and Hardie (2012, 37) mention the case of the Italian Jesuit Roberto Busa who in 1949 began compiling the *Index Thomisticus*, an electronic corpus of 179 texts dealing with the figure of Thomas Aquinas (118 of which were written by Aquinas himself) in medieval Latin (https://www.corpusthomisticum.org/it/index.age). As for Greek, the Thesaurus Linguae Grecae (TLG) project was started in 1971 at the University of California, Irvine, from an idea of then graduate student Marianne McDonald, and on 30 October 1972 the digitisation of Ancient Greek literary texts officially began. At the moment, "[t]he TLG® Digital Library contains virtually all Greek texts surviving from the period between Homer (8 c. B.C.) and the fall of Byzantium in A.D. 1453 and a large number of texts up to the 20th century" (https://stephanus.tlg.uci.edu/history.php), searchable via an integrated search engine.

As highlighted by Tognini Bonelli, in terms of the English language, "[t]he first diachronic corpus was the Helsinki corpus [HC], which offers exemplars of English texts from c.750 to c.1700" (2010, 22). The HC was begun in the 1980s and launched in 1991. Two years later the first edited collection of essays dedicated to corpus-based and corpus-driven explorations of the HC was issued (*Early English in the Computer Age: Explorations through the Helsinki Corpus*), edited by Matti Rissanen, Merja Kytö and Minna Palander (see Bibliography).

For languages other than English, Kytö suggests that

> [t]here is an increasing interest in historical corpora for many other modern languages, among them German and Mittelhochdeutsche Begriffsdatenbank, the Bonner Frühneuhochdeutsches Korpus and DeutschDiachronDigital, French and Textes de Français Ancien, Spanish and Corpus del Español, and Portuguese and Corpus do Português, to name just a few. (2010, 419)[2]

Of course, as one can see, not all the diachronic corpora contain (only) samples of early modern vernacular languages. What is certain is that in recent years, the interest in Renaissance European languages has risen exponentially, and for a number of reasons. Firstly, more than any other historical dataset, corpora of early modern vernaculars offer privileged observatories for the standardisation of European languages as we know them today, thanks to the invention of the printing press by Gutenberg in c.1436 and the rapid spread of printed books in the period between the 15th and the 17th centuries. Secondly, and connected to the first reason, the printed editions of books which began to circulate from the second half of the 15th century can be digitised more easily than classic/medieval manuscripts with the help of modern optical character recognition (OCR) software (see, among others, Boschetti et al. 2009; Schoen and Saretto 2022 about issues concerning OCR and classical/medieval manuscripts). As Schoen and Saretto pointed out:

> Medieval manuscripts pose significant challenges to machine learning and OCR. Unlike printed texts, medieval handwriting often contains non-discrete characters, such as the

2 For other references to digitised corpora of historical varieties of other Germanic, Romance, and Slavic languages see also Claridge 2008; Xiao 2008; McEnery et al. 2022, 394-5. Big historical corpora are also available for Chinese, but accessibility is limited (Zinin and Xu 2020). Very recently, even a small corpus of early modern Sardinian has been created (cf. Puddu and Talamo 2020).

> conjoined letters of cursive scripts or the disconnected minims of Gothic script; therefore, machines cannot simply be taught individual letterforms but must learn to transcribe larger segments of text. Compared to modern handwritten documents, medieval manuscripts feature elaborate and often cryptic handwriting systems that vary intensely across period, region, and scribe. (2022, 179)

These are the reasons why such big corpora as EEBO (Early Modern English Books Online), the Early Modern French FreEMmax corpus (Gabay et al. 2022), GerManC (Bennett et al. 2009; Scheible et al. 2011), or the HCD (Historical Corpus of Dutch; see Van De Voorde et al. 2023) are so popular today and keep expanding their number of words.

After this brief overview of the importance and popularity of historical corpora, with a focus on Renaissance European languages datasets, it is worth examining how linguists work with the large amount of data they are provided with. When it comes to diachronic approaches to corpus linguistics, scholars are sometimes sceptical about the possibilities offered by machine-readable samples of both literary and non-literary texts belonging to the Renaissance. This scepticism mainly derives from the debated issue of normalising/modernising corpora, thus eliminating, for instance, questions of variant spelling and part-of-speech (POS) tagging. Although such manipulations make examinations easier and more robust, at least from a quantitative point of view, at the same time they rule out the possibility of investigating the potentials that such variations may offer for the understanding of intra- and interlinguistic phenomena. The Hamletian (paraphrased) question 'To modernise or not to modernise' has always been a hot topic and is far beyond the scope of this introduction. Without delving into philological, almost ethical debates about the advantages and disadvantages of modernising Renaissance corpora,[3] I can only say that some attempts have been made to automatise the process of modernisation of the spelling or the morphological inflections in corpora of Renaissance texts (e.g., the VARD2 software for early modern English or the $\text{FreEM}_{\text{norm}}$ for early modern French),[4] and yet a linguistically and methodologically meaningful rationale must be developed by researchers to achieve satisfying results.

3 Moreover, not all the European languages exhibit the same degree of spelling variation in the period under consideration here.

4 Cf. Archer et al. 2015 and Bawden et al. 2022, respectively.

It goes without saying that fluctuation concerns not only spelling, but any other level of linguistic analysis, as shown in the articles in this volume. Drawing on the potentials offered by this variability, instead of considering it an obstacle, historical linguists exploit the tools offered by corpus linguistics to accelerate and broaden (both quantitatively and qualitatively) their research. The six articles included in this monographic section of *Status Quaestionis* 25 offer interesting perspectives on various levels of linguistic analysis, such as spelling fluctuation, textual pragmatics, morphosyntax, figurative language, etc., and their numerous intersections, using different corpus linguistics tools. Far for being a comprehensive overview of the state of the art on 'old languages and new technologies' (which is not the aim of this publication), the case studies presented here provide a glimpse into the potentials offered by corpus linguistics tools when dealing with Renaissance English and Romance languages such as French, Italian, Portuguese and Spanish.

* * *

The section dealing with early modern English begins with Marco Bagli, who explores the connection between spelling variation and grammaticalization in the Renaissance by examining the development of the pragmatic marker *harkee*. This marker originated from an imperative matrix clause with the verb *hearken/hark*. Focusing especially on the phenomenon of spelling fluctuation, Bagli demonstrates that it is evident at multiple levels in the data and the process he examined. Firstly, the matrix clause verbs exhibit alternative spelling forms, including variations with or without the digraph <ea> and with or without a final <e>. Secondly, the scholar's in-depth examination of the grammaticalization of *hearkee/harkee* reveals its emergence from a constellation of alternative spellings in the late 17th century. Bagli's essay provides a quantitative analysis of the various spelling forms of the matrix clause verbs that contributed to the pragmatic marker's development. Additionally, it offers empirical data to inform models of syntactic evolution for pragmatic markers, mapping the frequency of distinct syntactic contexts in early modern English.

Emma Pasquali discusses the creation of the *Corpus of Early Modern English Trials* (1650-1700), referred to as *EMET*, a specialised historical corpus containing 1.8 million words of trial proceedings. The primary goal of this corpus is to highlight the pragmatic aspects of early modern spoken English, as

trial proceedings offer authentic dialogues. Pasquali's *EMET* was established to investigate the influence on the choice of second-person pronouns, *thou* and *you*, as well as their various inflected forms, during the Restoration period. She begins by discussing the consultation of archives, criteria for selecting trials, and the technical process of uploading the corpus to #LancsBox for study. Her essay provides details about the *EMET*, including the number of documents, total tokens, and average tokens per text, along with the types of charges involved. She also delves into the editing, normalization, and POS tagging of trials, emphasizing the importance of proper editing for corpus linguistic analysis and comparing different normalisation methods for the *EMET*.

Closing the section dedicated to early modern English, Fabio Ciambella analyses a corpus of early modern English manuscript recipe books, denoted as *FEMER* (Folger Early Modern English Recipes), which were digitised by volunteers at the Folger Shakespeare Library. He outlines the chronological and content-based criteria for selecting the manuscripts, along with the modernization of the texts using VARD2 software, before offering a detailed corpus-driven investigation using #LancsBox and The Voyant Tools, focusing on the morphosyntactic structures found in culinary recipe texts of early modern English and their interactions with pragmatics.

Romance languages are the subject of the second section of this monographic volume. Examining late medieval Portuguese, Benjamin Fagard and José Pinto de Lima begin by affirming that a significant question in current research on adpositions is the emergence of complex prepositions in the diachronic development of Portuguese. Their essay addresses this question, concentrating on the initial centuries of available texts, to determine whether complex prepositions developed independently in Portuguese or were influenced by other European languages. Using a usage-based approach and examining texts from the 13th century, Fagard and Pinto de Lima document the presence of several complex prepositions. These findings suggest the possibility of the independent emergence of complex prepositions in early and late modern Portuguese.

Vittorio Ganfi's essay, on the other hand, analyses the structural and functional aspects of Light Verb Constructions in Italian. He focuses on texts from 1376 to 1691 extracted from the MIDIA corpus (Morfologia dell'Italiano in DIAcronia, i.e., diachronic morphology of Italian) and examines these constructions, which consist of a light verb and a noun, shedding light on their struc-

tural and functional characteristics. Ganfi also categorises these constructions based on the light verbs and nouns they use and their argument structures, while considering the semantic shifts within the constructions over time.

Lastly, adopting a comparative linguistic approach, Valentina Piunno's contribution presents a corpus-driven examination of the metaphorical and metonymic usages of the word *hand* in Spanish, French, and Italian texts from the 15th to the 17th centuries. It explores how the meaning of *hand* shifted from concrete to abstract, considering data from diachronic corpora and dictionaries. The analysis includes both qualitative and quantitative dimensions, identifying commonalities in semantic mapping, syntactic patterns, lexicalization, and functional values across these languages and gauging the level of productivity and conventionalization of each semantic shift.

Bibliography

Archer, Dawn, Kytö, Merja, Baron, Alistair and Rayson, Paul. 2015. "Guidelines for Normalising Early Modern English Corpora: Decisions and Justifications." *ICAME Journal* 39: 5-24.

Bawden, Rachel, Poinhos, Jonathan, Kogkitsidou, Eleni, Gambette, Philippe, Sagot, Benoît and Gabay, Simon. 2022. "Automatic Normalisation of Early Modern French." *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3354-66.

Bennett, Paul, Durrell, Martin, Scheible, Silke and Whitt, Richard J. 2009. "Annotating a Multi-genre Corpus of Early Modern German." In *Proceedings of the Fifth Corpus Linguistics Conference*, https://ucrel.lancs.ac.uk/publications/cl2009/ (accessed 4 November 2023).

Boschetti, Federico, Romanello, Matteo, Babeu, Alison, Bamman, David and Crane, Gregory. 2009. "Improving OCR Accuracy for Classical Critical Editions." *International Conference on Theory and Practice of Digital Libraries (ECDL) 2009: Research and Advanced Technology for Digital Libraries*, 156-67.

Claridge, Claudia. 2008. "Historical Corpora." In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, vol. 1, 242-59. Berlin and New York: Walter de Gruyter.

Gabay, Simon, Ortiz Suarez, Pedro, Bartz, Alexandre, Chagué, Alix, Bawden, Rachel, Gambette, Philippe and Sagot, Benoît. 2022. "From FreEM to D'AlemBERT: A Large Corpus and a Language Model for Early Modern French." In *Proceedings of the 13th Conference on Language Resources and Evaluation*, 3367-74.

Kytö, Merja. 2010. "Corpora and Historical Linguistics." *Revista brasileira de linguística aplicada* (RBLA) 11, no. 2: 417-57.

McEnery, Tony and Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony, Brookes, Gavin and Clarke, Isobelle. 2022. "Corpus Studies of Language through Time. Introduction to the Special Issue." *International Journal of Corpus Linguistics* 27, no. 4: 393-8.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Petrina, Alessandra. 2019. "All Petrarch's Fault: The Idea of a Renaissance." *Memoria di Shakespeare* 6: 145-64.

Puddu, Nicoletta and Talamo, Luigi. 2020. "EModSar: A Corpus of Early Modern Sardinian Texts." In *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*, edited by Cristina Marras, Marco Passarotti, Greta Franzini and Eleonora Litta, 210-5. Milan: Associazione per l'Informatica Umanistica e la Cultura Digitale.

Rissanen, Matti, Kytö, Merja and Palander, Minna, eds. 1993. *Early English in the Computer Age: Explorations through the Helsinki Corpus*. Berlin: Mouton.

Scheible, Silke, Whitt, Richard J., Durrell, Martin and Bennett, Paul. 2011. "A Gold Standard Corpus of Early Modern German." *Proceedings of the Fifth Law Workshop (LAW V) of the Association for Computational Linguistics*, 124-8.

Schoen, Jenna and Saretto, Gianmarco E. 2022. "Optical Character Recognition (OCR) and Medieval Manuscripts: Reconsidering Transcriptions in the Digital Age." *Digital Philology: A Journal of Medieval Cultures* 11, no.1: 174-206.

Tognini Bonelli, Elena. 2010. "The Evolution of Corpus Linguistics." In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 14-27. London and New York: Routledge.

Van De Voorde, Iris, Rutten, Gijsbert, Vosters, Rik, van der Wal, Marijke and Vandenbussche, Wim. 2023. "Historical Corpus of Dutch: A New Multi-genre Corpus of Early and Late Modern Dutch." *Taal en Tongval* 1: 114-32.

Xiao, Richard Z. 2008. "Well-known and Influential Corpora." In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, vol. 1, 383-457. Berlin and New York: Walter de Gruyter.

Zinin, Sergey and Xu, Yang. 2020. "Corpus of Chinese Dynastic Histories: Gender Analysis over Two Millenia." In *Proceedings of the 12th Language Resources and Evaluation Conference (ELRA) 2020*, 778-86.