Introduction

This study offers statistical information about the chronological development of the use of Latin words throughout the history of the language, from its origins until the end of the Roman Empire.[1] Conceived to help bridge linguistic research and literary analysis, this *Computational Dictionary* contains the data derived from a digital project and is organised in two parts. The first lists lemmas alphabetically and, along with their absolute frequency in Latin literature and epigraphs, provides frequency data on their use diachronically, as well as information on the incidence of each lemma in the works of classical authors.[2] The second part, designed to support linguistic and computational research, offers a frequency-sorted list of the lemmas contained in the first volumes, giving information on their frequency and on the presence of eventual homographs (similar but etymologically or grammatically distinct word-forms). The numeric collocation of each lemma in the frequency list is repeated in the alphabetical list so that readers can easily look up the same lexical entry in the two parts of the dictionary.

This study springs from a desire to provide scholars and students working in different fields of Classical Studies with a new tool for the exploration of the historical usage of Latin words in literature and epigraphs. Although recent decades have seen the publication of many frequency dictionaries for modern languages and a growing scholarly consensus that frequency information plays a key role in both computational linguistics and literary (especially intertex-

[1] Although our corpus includes the majority of classical Latin authors, we are aware that the linguistic picture that these texts give remains ultimately provisional and imperfect due to both the difficulty of retrieving digital versions of all extant Latin works of every author at the time of the project and the often-fragmentary state of the texts. Thus, in making available the results of a digital project, this study is not exhaustive but only offers statistical data on the frequency and diachronic use of Latin words.

[2] The following volumes will be co-edited by William M. Short, Giacomo Fenzi, and Jack Leslie, who co-developed the project with Tommaso Spinelli.

Published in: Tommaso Spinelli: The Diachronic Frequency of Latin Words:
A Computational Dictionary. Volume I (A/AB–AZYMVS). Propylaeum, 2022. DOI:
https://doi.org/10.11588/propylaeum.1000

tual) studies, no comprehensive Latin frequency dictionary yet exists.[3] In fact, while Latin authors themselves often refer to the frequency of certain Latin usages in their commentaries, surprisingly few attempts have been made in the last century to provide rigorous lemmatisation and counts of Latin words (*e.g.*, Diederich 1939; Delatte/Evrard/Govaerts/Denooz 1981; and Denooz 2010) and all have relied on limited textual corpora based on 'highly representative' authors from the so-called 'Golden Age'. Moreover, in its historical approach, the still-incomplete *Thesaurus Linguae Latinae* (*ThLL*) has shown the benefits of a diachronic study of Latin, though the dictionary tends to record only the most representative occurrences of words without providing statistical frequency data on their use.[4]

Based on an analysis of a large corpus of Latin literature and epigraphy, this dictionary adopts a series of unique features to meet the needs of a new generation of researchers who are increasingly shifting their attention to the 'marginal' literature of the early Republican, late Imperial and Christian eras and pioneering new interdisciplinary approaches to different fields of Classical Studies. What follows will present, in a methodologically reflective way, the innovative features of this work, by providing details on its wide corpus, on the technologies and the methodology used to process the Latin language, and on the organisational principles deployed to present the resulting statistical data in meaningful ways. Besides discussing important practical and methodological issues, this introduction will also suggest ways to maximise the impact of our analysis on both linguistic and literary research.

[3] An example is the series of frequency dictionaries recently edited by Paul Rayson and Mark Davies for Routledge: http://www.routledge.com/Routledge-Frequency-Dictionaries/book-series/RFD. Coffee (2018, 209) has suggested that 'after type of similarity, frequency is the most important determiner' to track intertextuality. On the importance of frequency data for language teaching, see Sinclair (1991, 30) and Davies/Davies (2017, vii). On computational linguistics and language-processing, see Dee (2002) and Kornai/Halácsy/Nagy/Oravecz/Trón/Varga (2006, 1).

[4] Cf. the *Praemonenda* of the *ThLL*. New databases have been developed to support the needs of modern literary analysis and commentary writing. These tools tend to permit querying the occurrences of a stem or of a specific inflected form in literary texts with little attention to chronology. See http://www.perseus.tufts.edu/hopper/; http://tesserae.caset.buffalo.edu/; http://packhum.org/; http://mizar.unive.it/mqdq/public/. However, some important exceptions must be acknowledged. The *Corpus Corporum* (http://www.mlat.uzh.ch/MLS/), offers the possibility of querying a large chronologically-annotated corpus of Latin texts by lemma/word-form but only shows the texts in which the selected lemma/word-form appears without providing statistical data on its diachronic usage: see Roelli (2014). Leveraging the Latin corpus *LatinISE*, the premium programme *Sketch Engine* (http://www.sketchengine.eu/) offers the possibility of querying a large corpus of Latin texts divided by era (*e.g.*, '*era Romana antiqua*') in order to explore the frequency of Latin words, their synonyms, and collocations, though currently a proper diachronic analysis is not available for Latin texts. For *LatinISE* and an overview of other Latin corpora, see McGillivray/Kilgarriff (2013).

Frequential Representations of Languages

The data presented in the following volumes can be compared, at least to a certain extent, with those provided by frequency dictionaries. A frequency dictionary aims to provide statistical information on a language by studying the occurrences of words in a pre-assembled corpus of texts. The nature and aims of this kind of tool thus differ in important ways from those of standard lexica.[5] While sense-dictionaries tend to reduce the lexical complexity of a language in order to define its words on the basis of their meanings, frequency dictionaries aim to account for *all* the occurrences of the words in use in a language as attested in its sources, without, however, providing information about their semantics. Although different in scope, the data of these dictionaries are not incompatible. Frequency dictionaries provide important background information on which standard lexica are based by offering, for instance, the possibility of making frequency-based selections of the words to include in lexica as well as annotating words with tags (*e.g.*, rare, archaic, ecclesiastical) that, based on statistical considerations of the textual corpus, clarify the meanings of words as well as their stylistic and cultural significance.[6] However, while the statistical nature of frequency dictionaries might make them appear much more scientific and less biased than sense-dictionaries, a survey of the frequency-sorted dictionaries published in the last century shows that these dictionaries also potentially contribute to an idealised – and therefore biased – projection of languages.[7] Specifically, the creation of a frequency dictionary can produce remarkably different results, depending on the quantity and the nature of the texts included in the corpus, the methodologies used for its analysis, and the ways in which the data are represented.[8] By surveying how digital humanities, linguistics, and literary theories have shaped each other in the last century, the following overview aims to illuminate the nature of these tools and their potential limits.

At the core of the evolution of the computational study of languages and of its methodology is the development of digital technologies for word tagging and counting. Frequency dictionaries published in the twentieth century tended to rely on precise, manual tagging and counting of words as attested in small corpora of literary texts with no (or very little) attention paid to the spoken

[5] On the interactions between lexica and frequency dictionaries, see Kilgarriff (1997, 135–155).

[6] Cf. Bowker (2010, 155–68).

[7] On the virtual nature of every lexicon, see Bisconti (2012, 1–26).

[8] On the ways in which methodological considerations can impact on a frequency dictionary, see Rosengren (1971), Gardner (2007), and Ueda (2017).

language and to less literary texts, such as newspapers, advertisements, movie or television subtitles, and graffiti.[9] For instance, Edward Thorndike's English frequency dictionary titled *The Teacher's Word Book* (1921) proposed an alphabetically-sorted list of 10,000 words as attested in a corpus of 41 different sources and circa 625,000 words; similarly, Milton Buchanan' *Graded Spanish Word Book* (1927) was based on a 500,000-word corpus. Although partially helped by computer technologies, the dictionaries created during the second half of the twentieth century continued to be based on relatively small corpora.[10] This is the case of *The Frequency Dictionary of Spanish Words* by Alphonse Juilland and Eugenio Chang-Rodrìguez (1964), which was based on a 500,000 words corpus, and of Jean-Claude Carrière's Greek frequency dictionary titled *Tables Fréquentielles de Grec Classique* (1985), which relied on a corpus of 702,208 occurrences as attested in classical Greek texts. Such small corpora allowed scholars to tag the texts accurately and to provide users with information on the grammatical or syntactic function of each word in order to differentiate eventual homographs or ambiguous forms.[11] However, the small size of these textual corpora also made the statistical information provided by these dictionaries heavily dependent on the editors' selection of texts and scarcely representative of the language overall.

Frequency dictionaries created in the last couple of decades, such as those published by Routledge in the series edited by Paul Rayson and Mark Davies, have been more attentive to methodological issues.[12] Unlike their predecessors, these dictionaries are characterised by very large and inclusive corpora that record several million words as found in literary texts of different genres (poetry, novels, academic writing, short stories) and written in different geographical regions. These corpora also include a much greater variety of less literary texts, such as newspapers, academic or corporate conference handouts, song lyrics, and press conference transcripts, as well as samples of oral language like transcriptions of conversations, interviews on radio and television, recording of oral story-telling and conferences, or of live theatrical performances.[13] For instance, both the *Frequency Dictionary of Spanish* (2017) and the *Frequency Dictionary of French* (2009) rely on corpora of over 20 million

[9] Cf. the overview on frequency dictionaries provided by Ávila Marín (2009, 163–175).

[10] Worth mentioning are also Hayward Keniston's *Common Words in Spanish* (1920); Helen Eaton's *An English-French-German-Spanish Word Frequency Dictionary* (1940); and the famous *The Teacher's Word Book of 30,000 Words* published in 1944 by Edward Thorndike and Irving Lorge.

[11] For other examples, see the bibliography in Dee (2002, 66–67).

[12] A list of these dictionaries is available in Davies/Davies (2017).

[13] See, for instance, Davies/Davies (2017, 2–4).

words as attested in both written and oral sources. Similarly, Randall Jones and Erwin Tschirner used a 4.2-million-word corpus for their 2006 *Frequency Dictionary of German*. The *Frequency Dictionary of Contemporary American English* (2010) used an impressive corpus of over 385 million words grouped into five genres and taken from both spoken and written language. Although available only online, it is worth mentioning the frequency dictionary of English words provided by http://www.ngrams.info/. In its use of a 14-billion-word corpus (http://www.english-corpora.org/iweb/), this dictionary is representative of the modern aim of an ideally omni-comprehensive dictionary in which the role played by editorial choice is reduced to the absolutely minimum in favour of a wide and empirically inclusive corpus of texts.[14]

These multi-million-word corpora are designed to better capture the complexity of a language in its different generic, diastratic and diatopic inflections.[15] However, it would be overly optimistic to think that this inclusiveness comes without a cost. Such huge amounts of data prevent, or at least dramatically limit, the possibility of accurate, manually-curated, textual tagging.[16] In fact, modern frequency dictionaries largely rely on new digital technologies to automatically reduce all the inflected word-forms featured in a textual corpus to their respective headwords in order to count the frequency of each lexical entry.[17] This is commonly done by using pattern-matching and lemmatising technologies that can identify and count under the correct entry its inflected word-forms, yet the results of such computerised text processing are still far from perfect. While available technologies tend to query texts according to fixed and extremely general criteria, natural languages are hardly precise or perfectly consistent due to the presence of polysemy, of spelling variations and of ambiguous, homographic forms that can count as both noun and adjective, or as participle, noun, and adjective.[18] Although recent advancements in so-called ‘natural language processing’ (NLP) are making these calculi progressively more flexible and accurate by also taking into account the meaning of the words and their syntactic collocations, the computational processing of texts still produces many ambiguous or potentially incorrect results that

[14] Cf. Varga/Halácsy/Kornai/Nagy/Németh/Trón (2007).

[15] On the modern attention to diastratic and diatopic variants of a language see, for instance, Iannàccaro/Dell’Aquila (2001).

[16] See Kornai/Halácsy/Nagy/Oravecz/Trón/Varga (2006).

[17] On the potential and the limits of this automatic processing of texts, see Héja/Takács (2012).

[18] For an overview, see Nadkarni/Ohno-Machado/Chapman (2011). On the limits of modern lemmatising technologies, see Graham (2008) and Rayson/Archer/Baron/Culpeper/Smith (2007). On the lemmatisation of the Latin language, see Cappelli/Passarotti (2003, 519–31).

need to be further refined.[19] Modern dictionaries cope with these technical and methodological issues by adopting solutions of compromise that either signal potentially ambiguous forms to readers or simply ascribe them to one of their possible categories (*i.e.*, noun, verb, or adjective) on the basis of statistical data.

Dictionaries of modern languages have also adopted different ways of presenting their data in terms of design and quantity. In addition to being organised according to their ascending or descending frequency, entries can be ordered alphabetically or thematically. In some cases, the aforementioned principles are combined (as in the case of the Routledge series) in frequency-sorted lists of words that are grouped by theme.[20] More importantly, the general enlargement of the corpora should not imply that these dictionaries are exhaustive. In fact, many of the modern frequency dictionaries are 'high frequency lists' that do not aim to provide a comprehensive treatment of a language but rather to list only its most frequently used terms, variously setting a limit between 2,000 and 5,000 entries.[21] The number of entries and their ordering principles are influenced by both methodological reasons and the dictionary's intended function and audience. Frequency dictionaries designed to support linguistic research and the creation of standard lexica, in which frequency information is used both to decide which terms to include and to mark them as common or rare forms, tend to include all the terms attested in the corpus and to order them in alphabetical and frequency-sorted lists. More often, however, modern frequency dictionaries prefer to display fewer entries organised in thematic lists. This choice is not merely the result of their pedagogical orientation, but is influenced by methodological reflections. Modern scholarship has suggested that after the first 2,000 words, the linguistic significance of frequency decreases as words have a narrower range.[22] As a result, while scholars have increasingly agreed on the fundamental role played by word frequency in language-teaching and in syllabus-design, some have also argued that only

[19] See Dereza (2018, 35–46). These programmes try to analyse texts by treating the language not as a mere sequence of symbols, but rather as complex hierarchical system in which single words are interlinked in phrases that interact to make sentences conveying ideas and meanings. This means that these technologies can do much more than simply query a text. In fact, they can operate complex operations such as automatic text summarization, topic extraction, or parts-of-speech tagging. On natural language processing for Latin, see the Classical Language Toolkit (http://cltk.org/). An overview of the many different technologies for natural language processing and textual tagging can be found in Bates (1995), Bates/Weischedel (2006), and Olsson (2009). See also Sebastiani (2002, 1–47).

[20] Baron/Rayson/Archer (2009, 41–67).

[21] See Nation (2001, 167–181) and Davies/Davies (2017, 1).

[22] Cf. Nation (2001).

‘high frequency’ really matters for language students.[23] For instance, John Read has suggested that ‘the further we move from the first 2,000 [words] or so, the less significant frequency becomes in an absolute sense’ because ‘the selection of lower-frequency words depends increasingly on the learners’ specific needs and interests’.[24]

Understanding this variety of approaches, their potential and their limits has become particularly urgent nowadays because of the role that frequency information has assumed not only in syllabus design and language teaching, but also in different fields of literary studies.[25] Following Julia Kristeva’s famous theory of ‘intertextuality’, according to which literary texts generate meanings by engaging in dialogue with and reworking other texts, recent decades have witnessed the proliferation of many digital tools for the computerised study of literary allusions.[26] Frequency data often play an important role in the ways these programmes improve the detection (and the definition) of meaningful interactions between texts.[27] For instance, the presence in two texts of the same high-frequency words (*e.g.*, ‘and’, ‘do’, ‘does’, ‘is’, ‘are’) is less significant than the reuse of rare words or word-uses to track their literary interactivity. Thus, digital programmes tend to use frequency information to refine the results of automatic intertextual search by excluding or giving less importance to the reuse of high frequency words when comparing two or more texts. At the same time, knowing the frequency by which single authors use specific words may offer important insights into the meanings that those words assume in their works as well as into the significance of their reuse. For instance, if a classical author such as Statius tends not to use a word, which is in turn very much used by another author, such as Virgil, but Statius does use that word in a narrative that thematically alludes to a Virgilian passage, this frequency information might be crucial to our understanding of that passage as an intertext, while prompting us to consider to what extent Statius’ word

[23] On the utility of frequency for language teaching, see Sinclair (1991, 30); Tribble/Jones (1997, 36); Granger/Hung/Petch-Tyson (2002). However, the pedagogical utility of frequency studies has been debated by some scholars who have also noted how the conceptualisations of ‘word’ and the lemmatising processes used to create frequency dictionaries might not match (and thus effectively support) the actual psychological ways in which the human mind processes languages: see Gardner (2007, 242–43). Furthermore, as we have anticipated, high-frequency words seem to be less relevant than low-frequency words for the determination of the meaning of a text: see Dee (2002, 62–63).

[24] Quote from Read (2000, 228).

[25] See Coffee/Koenig/Poornima/Ossewaarde/Forstall/Jacobson (2012, 383–422) and the following discussion.

[26] Cf. Allen (2011).

[27] See Coffee (2018, 209).

choice can be said to be intentionally allusive.[28] More generally, the presence in a text of a high percentage of rare words that have been intensively used by a preceding author can alert readers to the models on which the current text might be based.

Although scholarship has thus far paid surprisingly little attention to the many different ways in which modern frequency dictionaries have coped with the practical and methodological challenges of creating statistical analysis of languages, even this brief overview demonstrates that linguistic theories and digital technologies have shaped each other significantly, producing different computational presentations of languages.[29] In particular, while in recent years the use of larger corpora and of automated word-counting has helped to reduce the influence of editorial choices, most dictionaries still tend to adopt a specific angle to offer a frequential and statistical representation of a language. Thus, in order to correctly use and assess the information provided by frequency dictionaries it is fundamentally necessary to consider them as the result of the intersection between different needs (*e.g.*, research, teaching), disciplines (*e.g.*, linguistics, computer science) and, inevitably, of their practical and methodological issues (*e.g.*, the concept of word; the limits of computerised text-tagging; and issues related to the counting of ambiguous forms).

Latin Frequency Dictionaries

The nature and purpose of Latin frequency dictionaries are not radically different from those for modern languages.[30] However, the fact that classical Latin is no longer spoken by native speakers, being attested only in ancient texts that survived through the manuscript tradition, adds some specific methodological issues to the statistical analysis of this language. Firstly, classical Latin is predominantly attested in self-consciously literary texts and formal documents. Thus, while in recent years new discoveries in Pompeii and Herculaneum have prompted a reconsideration of the importance of non-literary sources (*e.g.*, graffiti), Latin textual corpora are still largely based on literary sources with no record of the spoken language and very little traces of the so-called *sermo cotidianus*, which is only occasionally attested in epistles, satires and comedies.[31]

[28] On the importance of intentionality in the study of literary allusions, see Hinds (1998).

[29] See Rosengren (1971), Kornai/Halácsy/Nagy/Oravecz/Trón/Varga (2006) and Ávila Marín (2009, 163–175).

[30] On the important role played by frequency data in the computerised detection of Latin intertextuality, see Coffee (2018). *Contra* see Dee (2002, 60).

[31] See Milnor (2014).

Contributing to this literary bias is the fact that Latin epigraphs are not usually processed with literary sources because they are difficult to lemmatise through existing technologies.[32]

The processing of Latin is further complicated by the fact that classical texts are not preserved in their original editions but have survived only through a vexed manuscript tradition. These texts present frequent gaps that have been differently resolved by scholars, resulting in many slightly different versions of the same text, as well as spelling variants, which may reflect either an author's style or the spelling conventions adopted by different generations of copyists. This means that, at least theoretically, two Latin frequency dictionaries based on apparently the same textual corpus can in fact produce different statistical data depending on the editions of Latin texts that they adopt. In addition, while we have seen that modern dictionaries for modern languages often work on corpora of millions or even billions of words, the corpus available for the Latin language is quite limited. Surviving classical Latin literature contains c. 9 million words, while the extant corpus of Latin epigraphy counts c. 5 million words.[33] Though small, this corpus still presents very rich diachronic, diastratic and diatopic variety in so far as it includes texts written in different periods and in different geographical regions of the Roman Empire by authors with different social and cultural backgrounds who wrote for different audiences.[34]

Along with the selection of a corpus, different methodological choices condition both the computational analysis of Latin and the study of its literature. The idea that frequency data can contribute significantly to the philological and literary analysis of classical texts is not linked exclusively to the modern development of digital tools for the study of intertextuality, but is well attested even in antiquity. At least from the third century BCE, Alexandrian grammarians and scholiasts started to refer in their commentaries to the variable usage of terms as a significant factor in studying an author's style and allusivity. As Folco Martinazzoli has demonstrated, ancient commentators such as Zoilus (who lived in the second half of the fourth century BCE), Zenodotus and Aristarchus of Samothrace all used frequency considerations to reflect on textual problems in the Homeric poems.[35] Whilst the first two adopted the

[32] Latin epigraphs provide important linguistic information: Kruschwitz (2014, 721–44) and Donati (2016, 21–38). Although the fragmentary nature of many epigraphs makes it difficult to lemmatise these texts correctly, we have decided to include epigraphs in our corpus because some words may be attested in epigraphs but not in literary texts.

[33] The data provided by Dee (2002, 59) are confirmed by our findings.

[34] See Adam (2013) and Kruschwitz (2014, 721–44).

[35] Martinazzoli (1953).

principle of analogy on the basis of which frequently attested forms in the Homeric corpus were more likely to be authentic, Aristarchus dealt with the same issues by introducing the concept of *hapax legomenon*, meaning literally '(something) said once'. According to this idea, particularly rare words can be considered very representative of an author's style and, therefore, more likely to be genuine in ways that can be compared with the modern philological ideas of *lectio facilior* (a more common word that is less likely to be authentic) and *lectio difficilior* (a more complex or rare word that is assumed to be more likely to be authentic).[36] Similarly, Latin authors such as Cicero, Varro, and Quintilian often referred to the *usus* of a word or to its frequency in their literary, grammatical and stylistic discussions. Joseph Denooz (2010, 1–2) has shown that the word *usus* is used to explain linguistic facts 45 times by Varro in his *De lingua Latina*, 73 times by Cicero in the *De Oratore* and the *Orator*, and 163 times in Quintilian's *Institutio Oratoria*. Moreover, Quintilian uses the adjective *frequens* and the adverb *frequenter* some 223 times in his linguistic and stylistic considerations.

Although the idea that the consideration of the quantitative use of a word can help the study of both linguistic and stylistic phenomena is well rooted in Classical Studies, only four Latin frequency dictionaries have been published in the last century, one of them (Dee 2002) being a digitalised collation of two former dictionaries (Lodge 1907; Diederich 1939).[37] These dictionaries are based on small corpora, and none has thus far attempted to study the occurrence of Latin words in a very large corpus of Latin literary and epigraphical texts, nor to adopt a diachronic approach. The first of these dictionaries was published by Gonzalez Lodge in two editions (1907; 1912) under the title *Vocabulary of High School Latin*. As the author explains, this book was intended to support the teaching of Latin in secondary schools and for this reason it is based on the authors that were considered canonical by the educational curriculum of the time, namely Caesar (*Bell. Gall.* 1–5), Cicero (*Cat.* 1–4; *Pro Arch.*; *De Imp. Pomp.*), and Virgil (*Aen.* 1–6). This 77,142-word corpus resulted in a list of 4,650 entries including every term attested at least once in the texts, the only exception being the names of people and places which were excluded from the count. This teaching-oriented dictionary of classical Latin was followed some years later by Paul Diederich's *Frequency of Latin Words and their Endings* (1939). Explicitly designed to focus on the texts left aside by Lodge, this dictionary

[36] On the notion that '*lectio difficilior preferenda est*', see Rubino (1977).

[37] On the utility of frequency considerations for the study of a language, see Guiraud (1954, 5); on their impact on the study of Latin intertextuality, see Coffee (2018, 209). This overview does not consider the lists of high-frequency Latin words designed for language teaching such as those by Toner (2002) and Rook (2015).

was based on a corpus of 202,158 terms as attested in the passages provided by Maurice Avery's *Latin Prose Literature* (1931), Heathcote William Garrod's *The Oxford Book of Latin Verse* (1912), and Charles Beeson's *A Primer of Medieval Latin* (1925). In addition to mixing classical and medieval texts, Diederich's dictionary included proper names but recorded only the words that appeared at least five times in the corpus, obtaining a list of 3,800 lexical entries. This dictionary saw the first attempt to divide homographs and ambiguous forms on the basis of their meaning. However, as the author acknowledges, the principle was not applied consistently nor accurately due to the cursory hand-made tagging of many Latin texts on which the dictionary was based.

The last century has seen the creation of two modern Latin frequency dictionaries that have tried to adopt larger corpora through the help of digital technologies. The first was published online on the website of the University of California, Irvine in 2002 by James Dee and is no longer available. However, its main features can be reconstructed thanks to the description of the programme provided by Dee himself.[38] Instead of being based on a corpus of Latin texts, this dictionary collated and digitalised the frequency information provided by Lodge's and Diederich's dictionaries. This operation was complicated by the very different methodologies adopted by the two dictionaries. To cope with this issue, Dee's frequency list excluded people and place names, as Lodge did, and it included only terms that were attested at least three times in Lodge's or five times in Diederich's dictionaries, resulting in circa 3,500 lexical entries. Furthermore, Dee tended to ignore semantic differences that were impossible to maintain because the older dictionaries adopted different concepts of what counts as a word. For instance, Lodge counted '*cum*+verb' and '*cum*+name' as two different lexical entries, while Diederich treated *cum* as one single entry. Although Dee tried to normalise the information provided by Lodge and by Diederich, this database remained focused on very few entries and could not reconcile the different methods used by the former dictionaries. For this reason, it should be considered a successful digitalisation of pre-existing frequency lists rather than a new and modern Latin frequency dictionary.

An important development in the computational study of the Latin language was marked by the publication of the *Dictionnaire Fréquentiel et Index Inverse de la Langue Latine* (1981) edited by Louis Delatte, Evrard Govaerts, and Joseph Denooz within a project of the *Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA)* of the University of Liège. This dictionary used a specially designed programme to process and catalogue the terms attested in a corpus of 794,662 classical Latin words including both prose (582,411) and poetry (212,251). The corpus assembled by the *Laboratoire* between 1961

[38] Dee (2002, 59–67).

and 1981 included: Catullus, Caesar, Cicero, Horace, Juvenal, Ovid, Persius, Propertius, Quintus Curtius, Sallustius, Seneca, Tacitus, Tibullus, Livy, Virgil, and Vitruvius. The resulting dictionary, still available in PDF, features 13,077 lexical entries that are ordered alphabetically with the indication of their absolute frequency in the corpus and their relative frequency in prose and poetry. The second part of the dictionary contains an inverse lexicon that includes all the different word-forms (81,755) that appear in the corpus and their absolute frequency. The dictionary generally follows, and freely adapts, the standards of Egidio Forcellini's *Totius Latinitatis Lexicon* (1871). In particular, it normalises the Latin texts by replacing all occurrences of the letter 'u' with 'v' and by adopting, whenever possible, the forms with diphthong *ae* instead of *oe*. In addition, it counts as one entry spelling variants of the same word. Proper names are excluded from the list of lemmas, unless they are used as common names, while they appear in the inverse list of occurrences. The most important feature of the dictionary is the fact that although the data are processed electronically, the limited size of the corpus allowed the editors to curate the textual tagging quite accurately. Thus, while the dictionary does not provide a complete grammatical mark-up, some grammatical information or even the translation is provided to distinguish homographs or entries that are based on the same lemma but that can be ascribed to different grammatical categories.

In 1998, Paul Tombeur collected in the *Thesaurus Formarum Totius Latinitatis* (*TFTL*, 1998) all the Latin lexemes attested throughout the history of the language. Although this work cannot be formally considered a frequency lexicon because it records inflected and non-lemmatised forms, it represents the first attempt to provide a comprehensive analysis of Latin word-forms organised chronologically. This database theoretically contains every attested word-form, recording their frequencies in each author and work, divided into four broad periods: *antiquitas* (from Plautus to the end of the second century CE), *aetas patrum* (from Tertullian to the Venerable Bede), *medium aevum* (from 736 CE to 1499 CE) and *recentior latinitas* (from 1500 CE to the Second Vatican Council, which ended in 1965 CE). While the limits of the software used to create this ambitious work and the mistakes it inevitably produced have been discussed by Matthew Robinson, the *TFTL* certainly represents an invaluable instrument for the analysis of the Latin language and its attention to chronology has inspired the present work that, however, shares with the *Nouveau Lexique Fréquentiel de Latin* the effort to lemmatise the inflected forms attested in Latin texts.[39]

[39] See Robinson (2000, 32–34). For an overview of other Latin corpora, see footnote 4.

The Liège dictionary has been recently re-edited by Joseph Denooz (2010) with the title *Nouveau Lexique Fréquentiel de Latin.*[40] The new edition is based on an expanded and slightly different textual corpus, but it maintains the same methodology of its predecessor. However, it no longer provides the inverse lexicon, which has been replaced by a list of the entries in descending frequency order. More importantly, the dictionary gives little grammatical information to differentiate its entries, but now records their relative frequency in each of the authors analysed. The corpus of this dictionary includes nineteen authors and circa 1.7 million words, but some of the texts included in the first edition have been excluded: among them Cicero (*Tusc.* 5), Livy (extracts), Ovid's *Metamorphoses*, and Vitruvius.[41] With its 26,940 entries, this is certainly the most accurate modern Latin frequency dictionary currently available, but its corpus is still very much focused on texts of the so-called *saeculum aureum*, while it does not include any post-Virgilian epic, nor certain linguistically fundamental poems.

A Computational Study of the Diachronic Frequency of Latin Words

The following volumes offer a new computational and statistical analysis of classical Latin as attested in a wide corpus of literary and epigraphical texts ranging from the fourth century BCE to the early sixth century CE. While the present work is not designed to compete with the *Nouveau Lexique Fréquentiel de Latin*, it adopts a different perspective and innovative features that make it unique and open to bridging linguistic and literary research.

The most striking feature of this new quantitative analysis of the Latin language is its attempt to be comprehensive. Based on a textual database yielding some 14,000,000 words, which represents the majority of extant Latin classical literature, the corpus covers the works of 309 authors, fragmentary and anonymous texts, works that are usually considered 'secondary' by scholastic curricula, and 521,532 epigraphs.[42] The choice of using such a variety of

[40] See Delatte/Evrard/Govaerts/Denooz (1981).

[41] The new corpus includes: Caesar, Cato, Catullus, Curtius Rufus, Horace, Juvenal, Lucretius, Persius, Petronius, Propertius, Sallust, Tacitus, Tibullus, Vergil, Cicero (all speeches and *De Officiis*, *De Amicitia*, *De Natura Deorum*, and *De Senectute*), Ovid (*Amores*), Plautus (excluding the *Cistellaria*), Pliny the younger (*Epistles*), and Seneca (excluding the *Naturales Quaestiones*). Cf. http://www.cipl.ulg.ac.be/Lasla/tlatins.html.

[42] As I mentioned in the introduction, these texts do not represent the entirety of extant Latin literature. Therefore, while the data provided by the present analysis are statistically meaningful, they cannot be considered exhaustive. The epigraphical cor-

texts is motivated by methodological concerns. Replacing the editorial selection of a small group of supposedly 'highly representative' texts with a larger textual corpus that provides more empirical data is fundamental to ensuring that the representation of a language is scientific and unbiased. Moreover, only an extensive corpus allows the exploration of the chronological evolution of the use of a word, without the risk of prioritising the literature of the so-called *aureum saeculum*.[43] Accordingly, this dictionary not only delivers a more empirical history of the Latin language, but also offers the opportunity to explore if and to what extent the computerised analysis of a 'Big Data' corpus produces results that are significantly different from those deriving from the small and curated corpora of previous frequency dictionaries.[44]

The second stand-out feature of this analysis is its attention to diachronic development. While previous Latin frequency dictionaries record the occurrences of Latin lemmas in the textual corpus and list them in alphabetical or frequency order without providing information on their chronological development, in the first part of this work each entry (marked by the indication of its absolute frequency in the corpus) is followed by 'historical data'. This section indicates the number of authors and epigraphs in which the lemma is attested and the number of its occurrences in each century, from the fourth BCE to the sixth century CE. For each century, the dictionary records the relative frequency of the lemma in individual authors, who are named in Latin (*e.g.*, *Gaius Suetonius Tranquillus*) following the convention of the *PHI* corpus.[45]

pus was kindly provided by Manfred Clauss (cf. http://www.manfredclauss.de/?msclkid=ac574545d03a11ec8fdbafde689fb504).

[43] However, readers must be aware that our data are not representative of the Latin language as it was, but rather of the language as it attested in extant literature that, due to the vicissitudes of its transmission, is more abundant for the Imperial and late Classical periods than for the early Republican age. Moreover, fragments of earlier texts have often survived as quotations by later authors, and while our courpus encompasses the majority of classical texts, it does not contain every work of important Christian authors such as Tertullianus, Ambrosius, and Augustinus (see the appendix) because only some of their texts were available in an open source digital format when the corpus was assembled and processed.

[44] Our database leverages the open-source editions of classical texts made available by different online libraries, such as *Digital Latin Library*, *Perseus*, *Packard Humanities Institute*, and *Bibliotheca Augustana*.

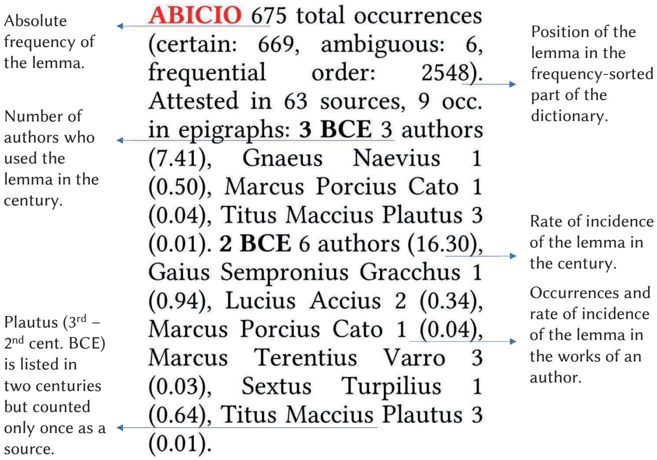
[45] The *Packard Humanities Institute* (*PHI*) corpus is one of the largest open-source Latin corpora currently available online (http://latin.packhum.org/) and, although it does not match our corpus perfectly, it can be effectively used to double check the information provided in this book and to look up the large majority of the Latin passages in which our lemmas or their inflected forms appear. However, small differences between our analysis and the data obtained through this programme are inevitable, due to the different texts included in the two databases, to the potentially different editions of the texts used and, more importantly, to the different searching tools. While our programme lemmatises inflected word forms, the *PHI* searching tool performs a simple pattern-matching query. Thus, if one searches '*ultor*'

Divisions by century have been chosen because this provides the most accurate and easy-to-read summary of the history of a word. Indeed, other organising principles such as a statistic-based division into linguistic eras (*e.g.*, archaic, classical, late) or a division on the basis of the chronology of single classical texts would provide a less consistent presentation of the diachronic data, as the precise date of composition of many classical texts is unknown and the division between different linguistic 'eras' is inevitably arbitrary. The historical data provide an invaluable help to linguistic and literary analysis by affording the possibility of obtaining frequency information for lemmas in different authors over the history of Latin literature. While the largest corpus used by former dictionaries involves 19 authors, our extensive database makes the present work an easy-to-use tool to explore the history of each Latin word. This is helpful for both high-frequency words, for which the *ThLL* does not always provide complete listings of occurrences, and for low-frequency or rare words.[46] In the latter case, use of a rare term by authors might suggest meaningful literary interactions.

More importantly, the historical data section also records the rate of incidence of a word in the works of an author and in each century. The rate of incidence per century provides information on the ratio (out of a thousand) of occurrences of the lemma in the given century to the total number of occurrences of the forms of that word. This figure gives an immediate idea of the popularity of the lemma in each century. For example, the high rate of incidence (1000) of the rare adjective *Abanteus* in the first century BCE and the first century CE shows that the lemma is attested only in the works of an author who lived between the first century BCE and the first century CE. Similarly, the low incidence of the verb *abdicare* in the third century BCE shows that this lemma is scarcely attested in that century, while it is largely attested in the sources of the first century BCE (rate of incidence 514). The rate of incidence for each author indicates the percentage of occurrences of a given word out of a thousand in the texts of that author included in our database – information crucial for understanding the significance of the statistical data. For instance, the 14,308 occurrences of the lemma *ut* in Cicero mean that the word is used around 10 times every thousand words, while the 5 occurrences of the same form in the *Laus Pisonis* mean that, on average, the word is used 2 times every thousand words. This is because the simple number of occurrences of a word in the works of an author is not necessarily significant, insofar as

the programme shows also results like '*multorum*', unless the search is made for a specific form like '#*ultor*#'. In this case, the programme displays only the occurrences of this specific graphic form and not of the lemma *ultor* and of its inflected forms.

[46] Cf. the *ThLL Praemonenda* (p. 31).



ABICIO 675 total occurrences (certain: 669, ambiguous: 6, frequential order: 2548). Attested in 63 sources, 9 occ. in epigraphs: **3 BCE** 3 authors (7.41), Gnaeus Naevius 1 (0.50), Marcus Porcius Cato 1 (0.04), Titus Maccius Plautus 3 (0.01). **2 BCE** 6 authors (16.30), Gaius Sempronius Gracchus 1 (0.94), Lucius Accius 2 (0.34), Marcus Porcius Cato 1 (0.04), Marcus Terentius Varro 3 (0.03), Sextus Turpilius 1 (0.64), Titus Maccius Plautus 3 (0.01).

Absolute frequency of the lemma.

Number of authors who used the lemma in the century.

Plautus (3rd – 2nd cent. BCE) is listed in two centuries but counted only once as a source.

Position of the lemma in the frequency-sorted part of the dictionary.

Rate of incidence of the lemma in the century.

Occurrences and rate of incidence of the lemma in the works of an author.

Fig. 1. Sample of the dictionary.

its significance largely depends on the number of words preserved from that author and contained in our database.[47] Thus, it would be misleading to think that 100 occurrences of a form in the attested corpus of the works of Cicero have the same statistical significance as 100 occurrences of the same form in a much smaller text such as the *Laus Pisonis*. The rate of incidence has thus been introduced to give an immediate indication of how often a term is used in comparison to the extent of an author’s *oeuvre*. This information could inform stylistic as well as textual-critical research, as it provides crucial information on which terms are more likely to be used by an author and, therefore, what conjectural readings might represent the *lectio facilior* or *difficilior* for a text.

A third important innovation of this study is the use of advanced digital technologies to process the Big Data deriving from an extensive corpus. While most previous Latin dictionaries relied on manual processing of texts, our analysis uses an original algorithm and the capabilities of the digital lemmatisation service offered by *LEMLAT*.[48] Lemmatisation is the process through which the variants of a term, or its inflected and graphically different forms (*e.g.*, *amas*, *amat*) are attributed to their lemma, which is the standard form of the word as it appears in a dictionary. While many programmes can perform this

[47] The authors and texts included in the corpus are listed in the appendix of this volume. Details on the editions used for each text will be published in the following volumes due to space constraints.

[48] The programme underpinning the computational dictionary is available at http://github.com/WizardOfMenlo/LatinDiachronicDatabase.

task on Latin texts quite successfully (*e.g.*, the Schinke algorithm, the Perseus lemmatiser, PROIEL, Parsley, Morpheus, Whitaker's Words, LatMor, and so on), none of these technologies yet provides entirely correct data.[49] Among them, I have chosen to use a freely adapted version of *LEMLAT*, developed by the National Research Centre (CNR) of Pisa and the *Università Cattolica del Sacro Cuore* of Milan, because it has proved to be one of the most performant open source technologies of this kind.[50] Based on a database of 40,014 lexical entries and 43,432 lemmas that also includes many late antique and medieval terms, *LEMLAT* adopts the standards of the *Oxford Latin Dictionary* (Glare 1982), thus moving beyond the *Lexicon Totius Latinitatis* by Forcellini (1871, 1887). Being able to recognise over 97% of Latin terms including many names of people and places, *LEMLAT* successfully lemmatises 319,725 lexemes into over 30,400 lexical entries.[51] Moreover, its automatic analysis is quite accurate – although not perfect – and takes into account spelling variations and even rare or archaic forms of a lemma that the former frequency dictionaries tend to neglect.[52] For instance, it recognises archaic or dialectal inflections of the same lemma such as the Faliscan variation *haba* for the standard form *faba*, or the rare form *ar* for the preposition *ad*.[53]

However, this still-developing technology does have its drawbacks. It cannot always recognise proper nouns and transliterations of Greek words, and cannot make discriminations based on the semantic or grammatical function of words.[54] As a result, the present analysis treats, for instance, *sum* as both

[49] See, for instance, *LatMor* (http://cistern.cis.lmu.de), *Words* (http://archives.nd.edu/words.html), *Parsley* (http://github.com/goldibex/parsleycore), *Morpheus* (http://github.com/tmallon/morpheus), and *PROIEL* (http://github.com/mlj/proielwebapp/tree/master/lib/morphology). On these technologies, see also Springmann/Schmid/Dietmar (2016).

[50] The programme is available at http://github.com/CIRCSE/LEMLAT3?msclkid=4dea6b71d08511ec89930f219ad58f2f and http://www.ilc.cnr.it/lemlat/lemlat/index.html. On its features and assessment see Passarotti/Budassi/Litta/Ruffolo (2017, 24–31) and Springmann/Schmid/Dietmar (2016).

[51] This figure is even more significant if one considers that the present work counts as only one lemma many forms (*e.g.*, *sum* verb and auxiliary; *cum*+verb and *cum*+name) that the Liège dictionaries divide into different lemmas.

[52] An efficient way to check the ways in which forms are lemmatised in our database through *LEMLAT* is through http://www.ilc.cnr.it/lemlat/lemlat/index.html. This programme has successfully lemmatised more than 97% of the word-forms attested in our corpus, leaving only 2.88% of the forms unrecognised. Among them many are names, Greek forms used in Latin texts, or Latin endings (*e.g.*, *-ar*; *-or*) that are mentioned by ancient grammarians in their discussions of Latin morphology but do not correspond to any lemma.

[53] Cf. Passarotti/Budassi/Litta/Ruffolo (2017, 26).

[54] The open source version of *LEMLAT* that was available when our programme was developed did not include the new Onomasticon that has recently been added to *LEMLAT3.0*. See Springmann/Schmid/Dietmar (2016, 389) and Passarotti/Budassi/Litta/Ruffolo (2017, 24).

auxiliary and main verb or *amicus* in both its adjectival and substantive uses as single entries. In addition, like its predecessors, it counts under the same lemma different spelling variants of the same word (*e.g.*, *transfero/trasfero*) while treating as different lemmas compound and derivative forms that use different suffixes or affixes.[55] The lemmatiser can also misinterpret some forms. An interesting example is provided by the form *ar*, of which the programme recognises 12 occurrences in our corpus, counting them as spelling variations of the preposition *ad* as suggested by the *OLD*.[56] However, looking at the occurrences of this form in the texts we note that in some cases *ar* is attested as the only extant part of a missing word in fragmentary texts (*e.g.*, *Carmen de Bello Aegyptiaco* 19.3, 23.8), while in other cases it is used by grammarians and commentators such as Remmius Palaemon (*Ars* 538.5, 538.10, 538.11, 545.12, 545.14) and Servius (*De Finalibus* 452.9) to refer to the ending of neuter names of the third declension in their explanations of Latin morphology. Nevertheless, the size of our corpus makes the impact of eventual imprecision negligible, providing reliable statistical data that are consistent with those obtained by smaller and hand-curated corpora, while permitting a more comprehensive history of the use of each Latin lemma by a wide range of classical authors.[57]

A final feature worth noting is the treatment of homographs, the presence of which is clearly marked in both the alphabetical and frequency lists, by providing for each lemma a double calculus of its frequency, including or excluding such ambiguous forms.[58] Homographs are alike but etymologically or grammatically distinct word-forms, and can be divided into three categories.[59]

[55] On the methodological and theoretical level, the most important issue faced by every manual or computerised lemmatisation is the definition of 'word'. For instance, we have seen that Louis Delatte's *Dictionnaire Fréquentiel* counted as homographs, and therefore as two different lemmas, the adjective *alienus* and the substantive *alienus*, but not different character-strings that are considered spelling variants of the same lemma (*e.g.*, *coelum*/*caelum*). On this issue, see the overview by Gardner (2007) and Knowles/Don (2004, 69–81). It must be noted also that the most recent developments in applied linguistics seem to head toward a derivational-morphology-based system in which words derived from the same lemma are treated as interconnected. See Passarotti/Budassi/Litta/Ruffolo (2017, 24–31).

[56] It is important to note that the programme tends to lemmatise past participles used as names or adjectives under the verb to which they belong.

[57] Some differences can also be explained by the fact that our corpus includes epigraphs. For example, the high frequency of words such as *manes* (and, consequently, *manus* that shares some homographic forms with *manes*) is due to its ritual use in inscriptions. The clear indication of the number of potentially ambiguous forms of each lemma warns readers about the statistical realiability of the historical data on lemmas. Specifically, data on lemmas with no or few homographs tend to be more reliable than data on lemmas that count numerous homographs among their inflected forms.

[58] See Passarotti/Ruffolo (2004).

[59] See Pica (1979, 154–180), Cramer (1970) and Gorfein/Viviani/Leddo (1982, 503–504).

Tab. 1. The most frequent Latin words according to four different Latin frequency dictionaries.

Diederich (1939)	Delatte/Evrard/ Govaerts/Denooz (1981)	Denooz (2010)	*The Diachronic Frequency of Latin Words* (2022)
1. Que 6771	1. Et 22059	1. Et Coord. 42251	1. Qui 300995
2. Qui/Quis 6553	2. Sum Verb 19001	2. Sum Verb 41781	2. Sum 284423
3. Et 6412	3. Qui Adj. 18215	3. Qui Pron. Rel. 41396	3. Et 269373
4. Sum 4853	4. In 13409	4. In 29813	4. Eo 258913
5. In 3481	5. Que 13113	5. Que 27318	5. Quis 216180
6. Is 1998	6. Non 10013	6. Non 21393	6. In 161672
7. Hic 1995	7. Hic Adj. 8174	7. Hic Pron. 20206	7. Is 161385
8. Non 1827	8. Is 6990	8. Is 18209	8. Hic 118783
9. Cum 1567	9. Ille 6460	9. Ego 14878	9. Facio 94736
10. Ad 1563	10. Ad Prep. 5837	10. Ille 14329	10. Non 88534

The first category is represented by forms that belong to the same lemma while expressing a different grammatical function (*e.g.*, *rosa* can be sing. nominative, vocative, and ablative of the first declension). The second is that of adjectives that are also used as names or vice versa (*e.g.*, *amicus; augustus*). The third category involves forms that, despite being graphically identical, belong to different lemmas. This is the case for instance of *făcĭēs* and *făcĭes*, the first being a feminine noun of the fifth declension (meaning 'face') and the other the future indicative of the verb *facio* ('you will make'). The disambiguation of homographs is not a problem for native speakers who can instinctively perform morpho-syntactic lemmatisation, by matching the grammatical information provided by the morphology of a word with the semantic and phonetic information provided by the context in which the word is used.[60] However, existing lemmatisers are only able to perform so-called 'morphological lemmatisation'. By analysing word-forms as character-strings independently of their meaning and in isolation from their syntactical context, this second form of lemmatisation is unable to resolve the ambiguity created by homographic forms.[61] Thus, modern dictionaries have generally coped with this issue by manually intervening on the corpus to disambiguate homographs, or by using

[60] See Dereza (2018, 35–47).

[61] Some programmes have recently taken the first steps toward such technologies by offering a seminal syntactically annotated Latin treebank. See the *Latin Dependency Treebank* by Bamman and Crane (2011), and the *Index Thomisticus* by Passarotti (2019). See also the LiLa project (http://lila-erc.eu/?msclkid=d44adb1cd08911ec85e6acdfd9ff8d64#page-top). However, as Mark Davies and Katy Davies (2017, 4) have recently suggested, despite the development of NLP technologies, the automated semantic tagging of languages is still very far from being perfect and requires substantial refining of the results.

'probabilistic tagging' that automatically assigns these ambiguous forms to a lemma according to a maximum likelihood principle on the basis of statistical calculi made on a sample corpus.[62]

The programme underpinning this work instead counts eventual homographic lexemes under all the lemmas to which they could belong (in order to maintain mathematical consistency of the statistical figure), while marking those entries as ambiguous, in addition to providing for each lemma the number of its certain and potentially ambiguous forms.[63] This choice considers several factors. Firstly, studies have shown that besides being very time consuming, manual tagging of relatively large corpora still tends to result in an approximative disambiguation of homographs.[64] Secondly, 'probabilistic' disambiguation is inappropriate for a Latin diachronic frequency dictionary that aims to record the occurrences of each lemma in different authors in order to show the evolution of its use through time, as this method risks deforming the statistical data by privileging common lemmas over rare words and, consequently, depriving scholars of the possibility to see the use of rare forms by authors. In particular, this would affect the dictionary's ability to support intertextual analysis, as the reuse of rare forms can be particularly indicative of meaningful interconnections between texts.[65] Thirdly, provided that homography affects less than 8.5% of Latin word-forms and that in 88% of cases the homographic form can be attributed to no more than two lemmas, the impact of this phenomenon on the statistical data provided by the present analysis is negligible. By clearly marking the resolution of eventual ambiguity resulting from homography, our programme opens new avenues for studying the phenomenon of Latin homography, which remains largely unexplored in modern scholarship.[66]

Overall, while the dictionary's broad coverage and attention to archaic or rare lexemes make it a valuable tool for both computational and historical linguistics, its 'historical data' provide commentary writers and philologists with a concise story on the use of each word by Latin authors, also offering statistical data on the influence played by each lemma in the works of an author. Our wish is that this tool will further support new interdisciplinary approaches to Classical Studies by providing researchers in linguistics, philology, intertextuality, and digital humanities with statistical data that they can

[62] See Merialdo (1991, 809–812); Tufis (2000); Davies/Davies (2017, 4).

[63] On the negligible impact of homographs on the data produced by *LEMLAT*, see Passarotti/Ruffolo (2004, 106).

[64] See Dee (2002, 60 n. 3).

[65] Coffee (2018).

[66] See the seminal discussion and scholarly review by Passarotti/Ruffolo (2004, 100 and n. 1).

use as a starting point for their research as well as in syllabus-design and in the creation of new teaching resources.[67]

Tommaso Spinelli
University of Manchester
tommaso.spinelli@manchester.ac.uk

Appendix

List of Latin authors and texts included in the corpus:

Ablabius (4 CE; Epigramma)
Aemilius Asper (2 CE; Commentarii in Terentium Sallustium Vergilium, Grammatica Vergiliana)
Aemilius Macer (1 BCE; Carmina)
Aemilius Sura (2 BCE; De Annis Populi Romani)
Albinovanus Pedo (1 CE; Carmina)
Albinus (?; De Metris, Rerum Romanarum)
Albius Tibullus (1 BCE; Carmina Tibulliana, Elegiae)
Ambrosius Mediolanensis (4 CE; Epistolarum Classis I et II, Expositio Psalmi CXVIII)
Ammianus Marcellinus (4 CE; Res Gestae)
Annius Florus (1 CE, 2CE; Carmina, Epitomae de Tito Livio Bellorum Omnium Annorum DCC, Ex Epistulis ad Divum Hadrianum, Vergilius Orator an Poeta)
Anonymi Comici et Tragici (?; Ribbeck's Scaenicae Romanorum Poesis Fragmenta)
Anonymi Epici et Lyrici (?; Morel's Fragmenta Poetarum Latinorum Epicorum et Lyricorum praeter Ennium et Lucilium)
Anonymi Fragmenta de Iure Fisci (2 CE, 3 CE)
Anonymi Grammatici (?; Funaioli's Grammaticae Romanae Fragmenta)
Anonymi de Differentiis (2 CE, Keil's De Differentiis)
Antonius Panurgus (?; Grammatica)
Appendix Vergiliana (1 CE; Aetna, Catalepton, Ciris, Copa, Culex, De Est et Non, De Institutione Viri Boni, De Rosis, Dirae, Elegiae in Maecenatem, Lydia, Moretum, Priapea, Priapeum)

[67] On the utility of frequency for language teaching, see: Sinclair (1991, 30); Tribble/Jones (1997, 36); Granger/Hung/Petch-Tyson (2002); and Davies/Davies (2017, vii). On their use for experiment design and language-processing technologies, see: Kornai/Halácsy/Nagy/Oravecz/Trón/Varga (2006, 1).

Appius Claudius Caecus (4 BCE, 3 BCE; Sententiae)
Aprissius (2 BCE, 1 BCE; Fragmentum)
Apuleius Madaurensis (2 CE; Anechomenos, Apologia, Carmina, De Deo Socratis, De Mundo, De Platone, Florida, Fragmenta, Metamorphoses)
Aquilius (2 BCE, 1 BCE; Palliata)
Arbonius Silo (1 BCE; Carmen)
Argumenta Aeneidis et Tetrasticha in Vergilii Aeneida, Bucolica et Georgica (?)
Atilius (?; Palliatae)
Attius Labeo (1 CE; Versio Latina Iliados)
Aufidius Bassus (1 CE; Historiae)
Aufustius (1 BCE; Grammatica)
Augustinus Hipponensis (4 CE, 5 CE; Confessiones, De Civitate Dei, Epistulae, Laus Cerei)
Aulus Caecina (1 CE; Fragmentum)
Aulus Cascellius (1 BCE; Liber Bene Dictorum)
Aulus Cremutius Cordus (1 BCE, 1 CE; Annales)
Aulus Furius Antias (1 BCE; Carmina)
Aulus Gellius (2 CE; Noctes Atticae)
Aulus Hirtius (1 BCE; C. Iulii Caesaris Commentariorum de Bello Gallico Liber VIII, Epistulae)
Aulus Persius Flaccus (1 CE; Saturae)
Aulus Postumius Albinus (2 BCE, 1 BCE; Annales)
Aurelius Opillus (2 BCE, 1 BCE; Grammatica)

Balbus (1 CE; Expositio et Ratio Omnium Formarum)
Bellum Africum (1 BCE)
Bellum Alexandrinum (1 BCE)
Bellum Hispaniense (1 BCE)
Bruttedius Niger (1 CE; Historiae)
Bucolica Einsidlensia (?)

C. Iul. Caes. Augustus Octavianus (1 BCE, 1 CE; Malcovati’s Imperatoris Caesaris Augusti Operum Fragmenta)
C. Plinius Caecilius Secundus (1 CE, 2 CE; Epistulae, Fragmenta, Panegyricus)
Caecilius Statius (3 BCE, 2 BCE; Palliatae)
Caelius Apicius (1 CE; Brevis Ciborum, Brevis Pimentorum, De Re Coquinaria)
Caelius Aurelianus (5 CE; E Parmenide De Natura)
Caesellius Vindex (2 CE; Grammatica)
Calpurnius Flaccus (2 CE; Declamationes)
Carmen Arvale (3 BCE)
Carmen Devotionis (?)

Carmen Evocationis (?)
Carmen de Bello Aegyptiaco (1 CE)
Chalcidius (4 CE; Ex Graecis Conversiones)
Claudius Caesar Germanicus (1 CE; Aratea)
Cloatius Verus (1 BCE, 1 CE; Grammatica)
Cn. Arulenus Caelius Sabinus (1 CE; Iurisprudentia)
Cn. Cornel. Lentulus Gaetulicus (1 BCE; Carmen)
Cn. Cornel. Lentulus Marcellinus (1 BCE; Orationes)
Cornelia, mater Gracchorum (2 BCE; Epistula)
Cornelius Nepos (1 BCE; Fragmenta, Vitae)
Cornelius Severus (1 BCE, 1 CE; Carmina)
Cornelius Tacitus (1 CE, 2 CE; Annales, De Origine et Situ Germanorum, De Vita Iulii Agricolae, Dialogus de Oratoribus, Historiae)

Decimus Iunius Iuvenalis (1 CE, 2 CE; Saturae)
Decimus Iunius Silanus (1 BCE; Versio Latina Magonis)
Decimus Laberius (1 BCE; Mimi)
Didascaliae et Argumenta in Plautum (?)
Didascaliae in Terentium (?)
Domitius Marsus (1 BCE, 1 CE; Epigrammata ex Bobiensibus)
Dorcatius (1 BCE, 1 CE; Carmen)

Fabius Dossennus (?; Carmina)
Fabius Pictor (3 BCE; Annales, Iuris Pontificis Libri)
Favorinus (1 CE, 2 CE; Oratio)
Fenestella (1 BCE, 1 CE; Annales)
Flavius Caper (2 CE; De Orthographia, De Verbis Dubiis)

Gaius Aelius Gallus (1 BCE; De Verbis ad Ius Civile, Iurisprudentia)
Gaius Aquilius Gallus (1 BCE; Iurisprudentia)
Gaius Asinius Gallus (1 BCE, 1 CE; Carmen, Grammatica)
Gaius Asinius Pollio (1 BCE, 1 CE; Carmina, Grammatica, Historiae, Orationes)
Gaius Ateius Capito (1 BCE, 1 CE; Iurisprudentia)
Gaius Aurelius Cotta (2 BCE, 1 BCE; Oratio)
Gaius Caesius Bassus (1 CE; Breviatio Pedum, Carmen, De Metris, De Metris Horatii, De Compositionibus, Genera Versuum, Poeticae Species)
Gaius Calpurnius Piso (2 BCE, 1 BCE; Oratio)
Gaius Cassius Hemina (2 BCE; Annales)
Gaius Cilnius Maecenas (1 BCE; Carmina)
Gaius Cornelius Gallus (1 BCE; Elegiae)
Gaius Erucius (2 CE; Oratio)

Gaius Fannius (2 BCE, 1 BCE; Historiae, Orationes)
Gaius Helvius Cinna (1 BCE; Carmina)
Gaius Iulius Caesar (1 BCE; Anticatones, Bellum Civile, Carmina, De Analogia, De Bello Gallico, Epistulae ad Ciceronem, Epistulae ad Familiares, Orationes)
Gaius Iulius Caesar Strabo (2 BCE, 1 BCE; Orationes, Tragoediae)
Gaius Iulius Hyginus (1 BCE, 1 CE; Grammatica, Historiae)
Gaius Laelius Sapiens (2 BCE; Orationes)
Gaius Licinius Macer (1 BCE; Annales, Oratio)
Gaius Licinius Macer Calvus (1 BCE; Carmina, Orationes)
Gaius Licinius Mucianus (1 CE; Historiae)
Gaius Lucilius (2 BCE; Saturae)
Gaius Memmius (1 BCE; Carmina, Orationes)
Gaius Oppius (1 BCE; De Silvestribus Arboribus, Vitae)
Gaius Papirius Carbo (2 BCE; Oratio)
Gaius Papirius Carbo Arvina (2 BCE, 1 BCE; Oratio)
Gaius Plinius Secundus (1 CE; Dubius Sermo, Naturalis Historia)
Gaius Sallustius Crispus (1 BCE; Ad Caesarem de Re Publica, Bellum Iugurthinum, Catilinae Coniuratio, Historiae, Historiarum Fragmenta, In M. Tullium Ciceronem)
Gaius Scribonius Curio, avus (2 BCE, 1 BCE; Oratio)
Gaius Scribonius Curio (1 BCE; Orationes)
Gaius Sempronius Gracchus (2 BCE; Orationes)
Gaius Servilius Glaucia (2 BCE; Orationes)
Gaius Suetonius Tranquillus (1 CE, 2 CE; De Grammaticis et Rhetoribus, De Historicis, De Poetis, De Vita Caesarum, Fragmenta, Prata)
Gaius Titius (2 BCE; Oratio)
Gaius Trebatius Testa (1 BCE, 1 CE; Iurisprudentia)
Gaius Valerius Catullus (1 BCE; Carmina, Carminum Fragmenta)
Gaius Valerius Flaccus (1 CE; Argonautica)
Gaius Valgius Rufus (1 BCE; Carmina)
Gaius (2 CE; Fragmenta Aegyptia, Fragmenta Oxyrhynchitica, Gai Institutionum Epitome, Institutiones)
Gavius Bassus (1 BCE; De Origine Vocabulorum, Grammaticae Fragmentum)
Gnaeus Domitius Ahenobarbus (1 BCE, 1 CE; Oratio)
Gnaeus Gellius (2 BCE; Annales)
Gnaeus Marcius, vates (3 BCE; Praecepta)
Gnaeus Matius (1 BCE; Carmina)
Gnaeus Naevius (3 BCE; Alia Carmina Epica, Bellum Punicum, Carmina, Palliatae, Praetextae, Tragoediae, Versus in Metellos)
Gnaeus Tremelius Scrofa (1 BCE; De Re Rustica)
Granius Flaccus (1 BCE; Iurisprudentia)

Granius Licinianus (1 BCE; Annales)
Grattius (1 BCE, 1 CE; Cynegetica)

Hadrianus (1 CE, 2 CE; Carmina, Orationes)
Helvius Mancia (1 BCE; Oratio)
Hilarius Arelatensis (5 CE; Carmina)
Hieronymus Stridonensis (4 CE, 5 CE; Actus Apostolorum, Apocalypsis Ioannis, Canticum Canticorum, Commentaria in Ezechielem, Epistola Iacobi, Epistola Ioannis I, Epistola Ioannis II, Epistola Ioannis III, Epistola Iudae, Epistola Pauli ad Colossenses, Epistola Pauli ad Corinthios I, Epistola Pauli ad Corinthios II, Epistola Pauli ad Ephesios, Epistola Pauli ad Galatas, Epistola Pauli ad Hebraeos, Epistola Pauli ad Philemonem, Epistola Pauli ad Philippenses, Epistola Pauli ad Romanos, Epistola Pauli ad Thessalonicenses I, Epistola Pauli ad Thessalonicenses II, Epistola Pauli ad Timotheum I, Epistola Pauli ad Timotheum II, Epistola Pauli ad Titum, Epistola Petri I, Epistola Petri II, Epistolae, Evangelium secundum Ioannem, Evangelium secundum Lucam, Evangelium secundum Marcum, Evangelium secundum Matthaeum, Liber Danielis, Liber Deuteronomii, Liber Ecclesiastes, Liber Esther, Liber Exodi, Liber Ezechielis, Liber Ezrae, Liber Genesis, Liber Ieremiae, Liber Iob I, Liber Iob II, Liber Iosue, Liber Isaiae, Liber Iudicum, Liber Iudith, Liber Levitici, Liber Numerorum, Liber Proverbiorum, Liber Psalmorum, Liber Ruth, Liber Tobiae, Libri Samuelis et Malachim, Libri XII Prophetarum)
Homerus Latinus (1 CE; Ilias Latina)
Hostius (2 BCE; Bellum Histricum)
Hyginus (2 CE; Fabulae)
Hyginus Astronomus (2 CE; Astronomica)
Hyginus Gromaticus (2 CE; Constitutio Limitum, De Condicionibus Agrorum, De Generibus, De Limitibus, De Munitionibus Castrorum)

Iulius Africanus (2 CE, 3 CE; Oratio)
Iulius Montanus (1 BCE, 1 CE; Carmina)
Iulius Valerius (3 CE, 4 CE; Carmina)
Iustinianus (5 CE, 6 CE; Digesta Iustiniani)
Iuventius, comoed. (1 BCE; Palliatae)

L. Aelius Praeconinus Stilo (2 BCE, 1 BCE; Grammatica)
L. Aemilius Paulus (2 BCE; Oratio)
L. Aurel. Avianius Symmachus (4 CE, 5 CE; Carmina)
L. Iunius Moderatus Columella (1 CE; De Arboribus, De Re Rustica)
Lactantius Placidus (4CE, 5 CE; In Statii Thebaida Commentum)
Laelius Felix (2 CE, Iurisprudentia)

Laevius (2 BCE, 1 BCE; Carmina)
Laus Pisonis (1 CE)
Lentulus, mimus (1 CE; Mimus)
Licinius Imbrex (3 BCE, 2 BCE; Palliatae)
Lucilius minor (1 CE; Carmina)
Lucius Accius (2 BCE, 1 BCE; Carmina, Praetextae, Tragoediae)
Lucius Afranius (2 BCE, 1 BCE; Togatae)
Lucius Ampelius (3 CE, 4 CE; Liber Memorialis)
Lucius Annaeus Cornutus (1 CE; Grammatica)
Lucius Annaeus Seneca minor (1 CE; Agamemnon, Apocolocyntosis, De Beneficiis, De Clementia, De Vita Patris, Dialogi, E Cleanthe Versus, Epistulae Morales ad Lucilium, Hercules Furens, Hercules Oetaeus, Medea, Naturales Quaestiones, Octavia, Oedipus, Phaedra, Phoenissae, Thyestes, Troades)
Lucius Annaeus Seneca maior (1 BCE; Controversiae, Fragmenta, Suasoriae)
Lucius Arruntius (1 BCE, 1 CE; Historiae Belli Punici)
Lucius Ateius Praetextatus (1 BCE; Grammatica)
Lucius Calpurnius Piso Frugi (2 BCE; Annales)
Lucius Cincius (3 BCE, 2 BCE; Grammatica, Iurisprudentia)
Lucius Coelius Antipater (2 BCE; Annales)
Lucius Cornelius Sisenna (2 BCE, 1 BCE; Historiae, Milesiae)
Lucius Cornelius Sulla (2 BCE, 1 BCE; Commentarii Rerum Gestarum)
Lucius Iulius Caesar (2 BCE, 1 BCE; Auspiciorum Liber)
Lucius Licinius Crassus (2 BCE, 1 BCE; Orationes)
Lucius Livius Andronicus (3 BCE; Odyssia, Palliatae, Tragoediae)
Lucius Marcius Philippus (1 BCE; Orationes)
Lucius Neratius Priscus (1 CE, 2 CE; Fragmenta Vaticana)
Lucius Orbilius Pupillus (2 BCE, 1 BCE; Grammatica)
Lucius Pomponius Bononiensis (2 BCE, 1 BCE; Atellanae)
Lucius Quinctius (?; Orationes)
Lucius Varius Rufus (1 BCE; Carmina, Tragoediae)
Lucius Verginius Rufus (1 CE; Epigramma)
Lucius Volusius Maecianus (2 CE; Assis Distributio)
Luscius Lanuvinus (3 BCE, 2 BCE; Palliatae)

M. Aemilius Lepidus Porcina (2 BCE; Orationes)
M. Valerius Messalla Corvinus (1 BCE, 1 CE; Commentarii de Bello Civili, Orationes)
M. Valerius Messalla Rufus (1 BCE; De Auspiciis, De Familiis Romanis)
Manilius (1 CE; Carmina)
Manius Manilius (2 BCE; Iurisprudentia)
Marcus Aemilius Scaurus (2 BCE, 1 BCE; De Vita Sua, Orationes)

Marcus Annaeus Lucanus (1 CE; Bellum Civile, Carmina)
Marcus Antistius Labeo (1 BCE, 1 CE; Iurisprudentia)
Marcus Antonius (1 BCE; Orationes)
Marcus Aurelius (2 CE; Carmen)
Marcus Caelius Rufus (1 BCE; Orationes)
Marcus Calidius (1 BCE; Oratio)
Marcus Cornelius Fronto (2 CE; Ad Amicos, Ad Antoninum Pium, Ad M. Antoninum Imp., Ad M. Caesarem et Invicem, Ad Verum Imp., Additamentum Epist. Aceph., Arion, Carmina, De Bello Parthico, De Eloquentia, De Feriis Alsiensibus, De Nepote Amisso, De Orationibus, Fragmenta, Laudes Fumi et Pulveris, Laudes Neglegentiae, Principia Historiae)
Marcus Duronius (2 BCE, 1 BCE; Oratio)
Marcus Fabius Quintilianus (1 CE; Declamationes Maiores, Declamationes Minores, Institutio Oratoria)
Marcus Furius Bibaculus (1 BCE; Carmina)
Marcus Iuventius Laterensis (1 BCE; Oratio)
Marcus Manilius (1 CE; Astronomica)
Marcus Pacuvius (3 BCE, 2 BCE; Praetextae, Tragoediae)
Marcus Porcius Cato (3 BCE, 2 BCE; Carmen De Moribus, De Agri Cultura, De Medicina, De Re Militari, De Rhetorica, Dicta Memorabilia, Epistulae, Incertorum Librorum Fragmenta, Iurisprudentia, Orationes, Origines)
Marcus Porcius Cato Uticensis (1 BCE; Orationes)
Marcus Terentius Varro (2 BCE, 1 BCE; Annales, Antiquitates Rerum Divinarum, Antiquitates Rerum Humanarum, Carmina, De Gente Populi Romani, De Lingua Latina, De Vita Populi Romani, Epistulae, Grammaticae Fragmenta, Logistorici, Menippeae, Res Rusticae, Res Urbanae)
Marcus Tullius Cicero (1 BCE; Academica, Arati Phaenomena, Arati Prognostica, Brutus, Carmina, Cato Maior de Senectute, Commentarii Causarum, De Divinitatione, De Domo Sua, De Fato, De Finibus, De Haruspicum, De Inventione, De Iure, De Lege Agraria, De Legibus, De Natura Deorum, De Officiis, De Optimo Genere Oratorum, De Oratore, De Partitione Oratoria, De Provinciis Consularibus, De Republica, Epistula ad Octavianum, Epistulae ad Atticum, Epistulae ad Brutum, Epistulae ad Familiares, Epistulae ad Quintum Fratrem, Epistulae, Facete Dicta, Hortensius, In Catilinam, In Pisonem, In Q. Caecilium, In Sallustium, In Vatinium, In Verrem, Laelius de Amicitia, Orationum Incertarum Fragmenta, Incertorum Librorum Fragmenta, Lucullus, Orator, Paradoxa Stoicorum, Philippicae, Philosophicorum Librorum, Post Reditum ad Populum, Post Reditum in Senatu, Pro Archia, Pro Balbo, Pro Caecina, Pro Caelio, Pro Cluentio, Pro Flacco, Pro Fonteio, Pro Lege Manilia, Pro Ligario, Pro Marcello, Pro Milone, Pro Murena, Pro Plancio, Pro Q. Roscio, Pro Quinctio, Pro Rabirio Perduellionis Reo, Pro Rabirio Postumo, Pro Rege Deiotaro, Pro Roscio Amer-

ino, Pro Scauro, Pro Sestio, Pro Sulla, Pro Tullio, Rhetorica ad Herennium, Timaeus, Topica, Tusculanae Disputationes)
Marcus Tullius Tiro (1 BCE; Grammatica)
Marcus Ulpius Traianus (1 CE, 2 CE; Dacica)
Marcus Valerius Martialis (1 CE, 2 CE; Epigrammata, Liber de Spectaculis)
Marcus Valerius Probus (1 CE, 2 CE; De Notis Iuris, Fragmenta, Vita Persii)
Marcus Verrius Flaccus (1 BCE, 1 CE; Etruscarum Rerum Libri, Grammatica)
Maurus Servius Honoratus (4 CE, 5 CE; Commentarius in Artem Donati, De Centum Metris, De Finalibus, De Metris Horatianis, In Vergilii Aeneidos Libros, In Vergilii Bucolicon Librum, In Vergilii Georgicon Libros)
Mimi Poetarum Incertorum (?; Bonaria's Mimi Poetarum Incertorum)
Mummius (2 BCE; Atellanae)

Naevius (?; Cypria Ilias)
Nero (1 CE; Carmina)
Ninnius Crassus (1 BCE; Ilias)
Novius (1 BCE; Atellanae)

P. Cornel. Scipio Aem. Afr. (2 BCE; Orationes)
P. Cornel. Scipio Afr. ma. (3 BCE, 2 BCE; Oratio)
P. Cornel. Scipio Nasica (3 BCE, 2 BCE; Orationes)
P. Terentius Varro Atacinus (1 BCE; Carmina)
Parthenius Presbyter (4 CE; Carmina)
Passienus Crispus (1 CE; Oratio)
Paulus Quaestor (4 CE; Carmina)
Petronius (1 CE; Fragmenta, Satyrica)
Phaedrus (1 BCE, 1 CE; Fabulae Aesopiae, Fabularum Appendix)
Pompeius Trogus (1 BCE, 1 CE; De Animalibus, Historiae Philippicae)
Pomponius Mela (1 CE; De Chorographia)
Pomponius Porphyrio (2 CE, 3 CE; Commentum in Horatium, Vita Horati)
Porcius Licinus (2 BCE, 1 BCE; Carmina)
Precatio Omnium Herbarum (?)
Precatio Terrae (?)
Priapea (1 CE, 2 CE; Vollmer's Poetae Latini Minores)
Pseudo-Varro (?; Sententiae)
Publilius Optatianus Porfyrius (3 CE, 4 CE; Epistola ad Constantinum, Panegyricus)
Publilius Syrus (1 BCE; Mimi, Sententiae)
Publius Alfenus Varus (1 BCE; Iurisprudentia)
Publius Aufidius Namusa (1 BCE; Iurisprudentia)
Publius Cannutius (1 BCE; Oratio)

Publius Clodius Pulcher (1 BCE; Orationes)
Publius Ovidius Naso (1 BCE, 1 CE; Ars Amatoria, Carmina, Epicedion Drusi, Heroides, Epistulae ex Ponto, Fasti, Halieutica, Ibis, Medea, Medicamina Faciei Femineae, Metamorphoses, Nux, Remedia Amoris, Tristia)
Publius Papinius Statius (1 CE; Achilleis, Belli Germanici Fragmenta, Silvae, Thebais)
Publius Pomponius Secundus (1 CE; Praetextae, Tragoediae)
Publius Rutilius Lupus (1 CE; Schemata Lexeos)
Publius Terentius Afer (2 BCE; Adelphoe, Andria, Eunuchus, Heauton Timorumenos, Hecyra, Phormio)
Publius Vergilius Maro (1 BCE; Aeneis, Eclogae, Georgica)
Pupius (?; Fragmenta)

Q. Aurelius Symmachus (4 CE, 5 CE; Carmina)
Q. Caecilius Metellus Macedonicus (2 BCE; Oratio)
Q. Caecilius Metellus Numidicus (2 BCE, 1 BCE; Orationes)
Q. Fabius Maximus Servilianus (2 BCE; Annales)
Q. Lutatius Catulus minor (2 BCE, 1 BCE; Oratio)
Q. Mucius Scaevola, pontifex (2 BCE, 1 BCE; Iurisprudentia)
Q. Pompeius Rufus (2 BCE, 1 BCE; Orationes)
Quintus Aelius Tubero (1 BCE; Historiae)
Quintus Asconius Pedianus (1 CE; In Senatu, In Toga, Pro Cornelio, Pro Milone, Pro Scauro)
Quintus Claudius Quadrigarius (1 BCE; Annales)
Quintus Cornificius (1 BCE; Carmina)
Quintus Curtius Rufus (1 CE, 2 CE; Historiae Alexandri Magni)
Quintus Ennius (3 BCE, 2 BCE; Annales, Fragmenta, Palliatae, Praetextae, Saturae, Tragoediae)
Quintus Horatius Flaccus (1 BCE; Ars Poetica, Carmina, Epistulae, Epodi, Sermones, Carmen Saeculare)
Quintus Hortensius Hortalus (2BCE, 1 BCE; Carmina, Orationes)
Quintus Lutatius Catulus (2 BCE, 1 BCE; Communes Historiae, Epigrammata)
Quintus Remmius Palaemon (1 CE; Ars Grammatica)
Quintus Serenus Sammonicus (2 CE, 3 CE; Liber Medicinalis)
Quintus Servilius Caepio (2 BCE, 1 BCE; Oratio)
Quintus Terentius Scaurus (2 CE; De Adverbio, De Ordinatione Partium Orationis, De Orthographia)
Quintus Tullius Cicero (1 BCE; Carmina, Commentariolum Petitionis)
Quintus Valerius Soranus (2 BCE, 1 BCE; Carmina)

Rabirius (1 BCE, 1 CE; Carmina)

Santra (1 BCE; Grammatica, Tragoediae)
Saserna (1 BCE; De Agri Cultura)
Scaevus Memor (1 CE; Tragoediae)
Scribonius Largus (1 CE; Compositiones)
Scriptores Historiae Augustae (4 CE; Hohl's Scriptores Historiae Augustae)
Sempronius Asellio (2 BCE, 1 BCE; Rerum Gestarum Libri)
Sentius Augurinus (1 CE; Carmen)
Servius Clodius (1 BCE; Grammatica)
Sevius Nicanor (2 BCE, 1 BCE; Carmen)
Sextilius Ena (1 BCE; Carmen)
Sextus Iulius Frontinus (1 CE; De Agrorum Qualitate, De Aquis Urbis Romae, De Arte Mensoria, De Controversiis, De Limitibus, Strategemata)
Sextus Pompeius Festus (2 CE; De Verborum Significatione)
Sextus Pomponius (2 CE; Liber Regularum)
Sextus Propertius (1 BCE; Elegiae)
Sextus Turpilius (2 BCE; Palliatae)
Siculus Flaccus (?; De Condicionibus Agrorum)
Silius Italicus (1 CE; Punica)
Sinnius Capito (1 BCE; Grammatica)
Staberius Eros (1 BCE; Grammatica)
Sueius (1 BCE; Carmina)
Sulpicia, Caleni uxor (1 CE; Carmen, De Statu Rei Publicae)

Tarquitius Priscus (1 BCE; De Disciplina Etrusca)
Terentianus Maurus (2 CE, 3 CE; De Litteris Syllabis et Metris Horatii)
Tertullianus (2 CE, 3 CE; Ad Marcionem, De Anima, De Carnis Resurrectione)
Ticidas (1 BCE; Carmina)
Titinius (2 BCE; Togatae)
Titus Annius Luscus (2 BCE; Oratio)
Titus Calpurnius Siculus (1 CE; Eclogae)
Titus Labienus (1 BCE; Oratio)
Titus Livius (1 BCE, 1 CE; Ab Urbe Condita, Periochae)
Titus Lucretius Carus (1 BCE; De Rerum Natura)
Titus Maccius Plautus (3 BCE, 2 BCE; Amphitruo, Asinaria, Aulularia, Bacchides, Captivi, Casina, Cistellaria, Curculio, Epidicus, Fragmenta, Menaechmi, Mercator, Miles Gloriosus, Mostellaria, Persa, Poenulus, Pseudolus, Rudens, Stichus, Trinummus, Truculentus, Vidularia)
Titus Quinctius Atta (2 BCE, 1 BCE; Epigramma, Togatae)
Tullius Laurea (1 BCE; Epigramma in Ciceronis Obitum)

Vagellius (1 CE; Carmen)
Valerius Aedituus (2BCE, 1 BCE; Epigrammata)
Valerius Antias (1 BCE; Annales)
Valerius Maximus (1 CE; Facta et Dicta Memorabilia)
Velius Longus (2 CE; De Orthographia)
Velleius Paterculus (1 BCE, 1 CE; Historia Romana)
Veranius (1 CE; Libri de Rebus Sacris)
Vibius Crispus (1 CE; Orationes)
Vita Iuvenalis (?)
Vitruvius (1 BCE; De Architectura)
Volcacius Sedigitus (2 BCE, 1 BCE; Liber De Poetis)

Zeno Veronensis (4 CE; Tractatus)

References

Adams, J. N. 2013. *Social variation and the Latin language.* Cambridge University Press.

Allen, G. 2011. *Intertextuality.* Routledge.

Avery, M. W., (Ed.) 1931. *Latin prose literature: Cato to Suetonius.* Little, Brown & Co.

Ávila Marín, M. D. C. 2009. "Estadística y lingüística de corpus: Implicaciones pedagógicas en la enseñanza y el aprendizaje del léxico". *Cauce, Revista Internacional de Filología*, 33: 163–175.

Bamman D., Crane G. 2011. "The Ancient Greek and Latin Dependency Treebanks". In Sporleder C., Van Den Bosch A., Zervanou K. (Eds.), *Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing*, pp. 5–22. Springer.

Baron, A., Rayson, P., Archer, D. 2009. "Word frequency and key word statistics in corpus linguistics". *Anglistik*, 20: 41–67.

Bates, M. 1995. "Models of natural language understanding". *Proceedings of the National Academy of Sciences*, 92: 977–982.

Bates, M., Weischedel, R. M., (Eds.) 2006. *Challenges in natural language processing.* Cambridge University Press.

Beeson, C. H., (Ed.) 1986. *A primer of medieval Latin: An anthology of prose and poetry.* Catholic University of America Press.

Bisconti, V. 2012. "La svolta lessicografica di Tullio De Mauro e i dizionari contemporanei". *Chroniques Italiennes*, 23: 1–26.

Bowker, L. 2010. "The contribution of corpus linguistics to the development of specialised dictionaries for learners". In Olivera, P. A. F. (Ed.) *Specialised Dictionaries for Learners*, pp. 155–168. Walter de Gruyter.

Buchanan, M. A. 1927. *A graded Spanish word book.* The University of Toronto Press.

Cappelli, G., Passarotti, M. C. 2003. "*LEMLAT*: uno strumento computazionale per l'analisi linguistica del latino. Sviluppo e prospettive". *Euphrosyne*, 31: 519–531.

Carrière J. C. 1985. *Tables Fréquentielles de Grec Classique.* Groupe de recherches et d'action pédagogiques en langues anciennes (Besançon) and Centre national de la recherche scientifique (France).

Coffe, N., Koenig, J. P., Poornima, S., Ossewaarde, R., Forstall, C., Jacobson S. 2012. “Intertextuality in the digital age”. *Transactions of the American Philological Association*, 142: 383–422.

Coffee, N. 2018. “An Agenda for the Study of Intertextuality”. *Transactions of the American Philological Association*, 148: 205–223.

Cramer, P. 1970. “A study of homographs”. In L. Postman, G. Keppel, (Eds.), *Norms of Word Association*, pp. 361–82. New York: Academic Press.

Davies, M. and Gardner, D. 2013. *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. Routledge.

Davies, M., Davies, K. H. 2017. *A frequency dictionary of Spanish: Core vocabulary for learners*. Routledge.

Dee, J. H. 2002. “The first downloadable word-frequency database for classical and medieval Latin”. *The Classical Journal*, 98: 59–67.

Delatte, L., Evrard, É., Govaerts, S., Denooz, J. 1981. *Dictionnaire fréquentiel et index inverse de la langue latine*. L.A.S.L.A.

Denooz, J. 2010. *Nouveau lexique fréquentiel de latin. Alpha-Omega*. Georg Olms Verlag.

Dereza, O. 2018. “Lemmatization for Ancient Languages: Rules or Neural Networks?”. *Communications in Computer and Information Science*, 930: 35–47.

Diederich, P. B. 1939. *The frequency of Latin words and their endings*. University of Chicago Press.

Donati, M. 2016. “Variation and Text Type in epigraphic corpus classes I”. *Studi e Saggi Linguistici*, 53: 21–38.

Eaton, H. S., 1940. *An English-French-German-Spanish Word Frequency Dictionary: A Correlation of the First Six Thousand Words in Four Single-Language Frequency Lists*. Dover Publications.

Forcellini, E. 1871. *Totius Latinitatis Lexicon*. Typis Aldinianis.

Forcellini, E., Furlanetto, G., De Vit, V. 1887. *Totius Latinitatis Lexicon: Totius Latinitatis Onomasticon*. Typis Aldinianis.

Gardner, D. 2007. “Validating the construct of word in applied corpus-based vocabulary research: A critical survey”. *Applied Linguistics*, 28: 241–265.

Garrod, H. W., (Ed.) 1912. *The Oxford book of Latin verse: from the earliest fragments to the end of the V century AD*. Clarendon Press.

Glare, P. G. 2012. *Oxford Latin Dictionary*. Oxford University Press.

Gorfein, D. S., Viviani, J. M., Leddo, J. 1982. “Norms as a tool for the study of homography”. *Memory & Cognition*, 10: 503–509.

Graham, A. 2008. "The Effects of Homography on Computer-generated High Frequency Word Lists". *All Theses and Dissertations*. 1617. (http://scholarsarchive.byu.edu/etd/1617).

Granger S., Hung, J., Petch-Tyson, S. (Eds.) 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. John Benjamins Publishing.

Guiraud, P. 1954. *Les caractères statistiques du vocabulaire*. Presses universitaires de France.

Héja, E., Takács, D. 2012. “Automatically generated customizable online dictionaries”. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012: 51–57. Association for Computational Linguistics.

Hinds, S. 1998. *Allusion and intertext: dynamics of appropriation in Roman poetry*. Cambridge University Press.

Iannàccaro, G., Dell'Aquila, V. 2001. "Mapping languages from inside: notes on perceptual dialectology". *Social & Cultural Geography*, 2: 265–280.

Jones, R., Tschirner, E. 2015. *A frequency Dictionary of German: Core vocabulary for learners*. Routledge.

Juilland, A., Chang-Rodríguez, E. 1964. *Frequency dictionary of Spanish words*. De Gruyter.

Keniston, H. 1920. "Common words in Spanish". *Hispania*, 3: 85–108.

Kilgarriff, A. 1997. "Putting frequencies in the dictionary". *International Journal of Lexicography*, 10: 135–155.

Knowles, G., Don, Z. M. 2004. "The notion of a lemma: Headwords, roots and lexical sets". *International Journal of Corpus Linguistics*, 9: 69–81.

Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V., Varga, D. 2006. "Web-based frequency dictionaries for medium density languages". *Proceedings of the 2nd International Workshop on Web as Corpus* (http://www.aclweb.org/anthology/W06-1701), pp. 1–9.

Kruschwitz, P. 2014. "Linguistic variation, language change, and Latin inscriptions". In Bruun, C. and Edmondson, J. (Eds.), *The Oxford Handbook of Roman Epigraphy*, pp. 721–744. Oxford University Press.

Lodge, G. 1907. *The Vocabulary of High School Latin. Being the Vocabulary of: Caesar's Gallic War, Books IV; Cicero Against Catiline, on Pompey's Command, for the Poet Archias; Vergil's Æneid, Books I-VI; Arranged Alphabetically and in the Order of Occurrence (No. 9)*. Columbia University Press.

Lonsdale, D. and Le Bras, Y. 2009. *A frequency dictionary of French: Core vocabulary for learners*. Routledge.

Martinazzoli, F. 1953. *Hapax legomenon: Parte prima*. Gismondi Editore.

McGillivray, B., Kilgarriff, A. 2013. "Tools for historical corpus research, and a corpus of Latin". In Bennett, P., Durrell, M., Scheible, S., Whitt, R. J. (Eds.), *New Methods in Historical Corpus Linguistics*. Narr.

Merialdo, B. 1991. "Tagging text with a probabilistic model". *ICASSP*, 91: 809–812.

Milnor, K. 2014. *Graffiti and the literary landscape in Roman Pompeii*. Oxford University Press.

Nadkarni, P. M., Ohno-Machado, L., Chapman, W. W. 2011. "Natural language processing: an introduction". *Journal of the American Medical Informatics Association*, 18: 544–551.

Nation, I. S. P. 2001. "How many high frequency words are there in English?". In Gill, M., Johnson, A. W. Kiski, L. M., Sell, R. D., Warvik, B., (Eds.), *Language, learning, literature: Studies presented to Hakan Kingdom*, pp. 167–181. English Department Publications, Abo Akademi University.

Olsson, F. 2009. "A literature survey of active machine learning in the context of natural language processing". *SICS Technical Report*, 6: 1–59.

Passarotti, M. 2019. "The Project of the Index Thomisticus Treebank". In Monica Berti (Ed.), *Digital Classical Philology*, pp. 299–320. De Gruyter Saur.

Passarotti, M., Budassi, M., Litta, E., Ruffolo, P. 2017. "The *LEMLAT3.0* Package for Morphological Analysis of Latin". *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 133: 24–31. Linköping University Electronic Press.

Passarotti, M., Mambrini, F. 2012. "First Steps towards the semi-automatic development of a word formation-based lexicon of Latin". In *LREC*, 2012: 852–859.

Passarotti, M., Ruffolo, P. 2004. “L’utilizzo del lemmatizzatore *LEMLAT* per una sistematizzazione dell’omografia in latino”. *Euphrosyne*, 32: 99–110.

Pica, P. 1979. “Homography or Polysemy”. *ITL-International Journal of Applied Linguistics*, 45: 154–180.

Rayson, P., Archer, D., Baron, A., Culpeper, J., Smith, N. 2007. “Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora”. *Proceedings of Corpus Linguistics*. University of Birmingham Press.

Read, J. 2000. *Assessing vocabulary*. Cambridge University Press.

Robinson, M. 2000. “Cetedoc’s *Thesaurus Formarum Totius Latinitatis*”. *Computers & Texts*, 19: 32–35.

Roelli, P. 2014. “*The Corpus Corporum*, a new open Latin text repository and tool”. *Archivum Latinitatis Medii Aevi-Bulletin du Cange*, 72: 289–304.

Rook, N. 2015. *The 750 most frequently used Latin adjectives*. Amazon.

Rosengren, I. 1971. “The quantitative concept of language and its relation to the structure of frequency dictionaries”. *Etudes de linguistique appliquée*, 1: 103–27.

Rubino, C. A. 1977. “*Lectio Difficilior Praeferenda Est*: Some Remarks on Contemporary French Thought and the Study of Classical Literature”. *Arethusa*, 10: 63–84.

Sebastiani, F. 2002. “Machine learning in automated text categorization”. *ACM computing surveys (CSUR)*, 34.1: 1–47.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Springmann, U., Schmid, H., Dietmar N. 2016. “LatMor: A Latin finite-state morphology encoding vowel quantity”. *Open Linguistics - Topical Issue on Treebanking and Ancient Languages: Current and Prospective Research*, 2: 386–392.

Thorndike, E. L. 1921. *The teacher’s word book*. Columbia University press.

Thorndike, E. L., Lorge, I., 1944. *The Teacher’s Word Book of 30,000 Words*. Columbia University Press.

Toner, J. 2002. *Latin Key Words: The Basic 2000 Word Vocabulary Arranged by Frequency*. Oleander Press.

Tribble, C., Jones, G. 1997. *Concordances in the classroom: A resource guide for teachers*. Athelstan Press.

Tufis, D. 2000. “Using a Large Set of EAGLES-compliant Morpho-syntactic Descriptors as a Tagset for Probabilistic Tagging”. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhauer, (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Ueda, H. 2017. “Two statistical treatments of Spanish vocabulary: composite indices of frequency and dispersion and principal component analysis applied to ordinal frequencies”. *Dialectologia: revista electrònica*, 7: 187–227.

Varga D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V. 2007. “Parallel corpora for medium density languages”. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, Ruslan Mitkov, (Eds.), *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4, pp. 247–58. John Benjamins Publishing Company.