TOWARDS A DIGITAL PROFILE OF EARLY MODERN LATIN:

Word frequency and dispersion in some Neo-Latin historiographical texts

By Johann Ramminger

The paper examines some textual metrics commonly applied to English texts in corpus linguistics, specifically their usefulness for Latin, in particular for Neo-Latin texts. They are tested on a corpus of five Neo-Latin historiographical texts on the background of Livy's Ab urbe condita *(since it was considered a stylistic ideal by many later historiographers). The metrics concern frequency (Gini-Index and Lorenz-curve) and dispersion (inter-arrival time and dpnorm Gries). The analyses throw light on the inner structure of Latin texts in general (esp. the relative frequency of grammatical words vs. lexical words) and connect words and ideas in the political and social realities of the worlds depicted by chronologically disparate texts (e.g., dispersion of* rex *in Livy and Valla's* Gesta Ferdinandi*).*

Despite promising initiatives, quantitative linguistics research concerning Latin texts is still very much in its infancy.[1] "Traditional" Classical Philology has since the nineteenth century created one of the most nuanced philologies with grammars, lexica, and literary histories of a depth and penetration unthinkable in most modern languages. Thus the need for new philological

I would like to thank Lene Schøsler and Trine Arlund Hass for a close reading of this paper and countless improvements. Also, I discussed many points with the (unsuspecting) dedicatee of this volume.

[1] Open Access corpora with lemmatizations are offered e.g. by Perseus (http://www.perseus.tufts.edu/hopper/), Corpus Corporum (http://www.mlat.uzh.ch/), CompHistSem (https://www.comphistsem.org/home.html), and Lasciva Roma (https://github.com/lascivaroma), a.o. While these can be extremely useful in many cases, without extensive corrections they are unusable for the fine-grained studies envisaged here. Proofread Open Access lemmatizations exist, e.g., PROIEL (https://www.hf.uio.no/ifikk/english/research/projects/proiel/), the Dante Treebank (https://github.com/UniversalDependencies/UD_Latin-UDante/) and LT4HALA (https://github.com/CIRCSE/LT4HALA), but they are necessarily much smaller.

tools may seem less urgent (although new language technologies can still provide new answers to old questions).[2] The field is less cultivated as regards medieval Latin texts and even less concerning the Latin language(s) of Early Modern Europe (which in the following for convenience's sake will be called Neo-Latin), even though European writers, lawyers, clergymen, administrators, etc. produced texts at a rate which in its heyday far exceeded the contemporary production in national languages. The sheer amount of Neo-Latin texts makes this part of Latin literature a promising field for distant reading, i.e., the application of quantitative analytical methods.[3]

This paper will test the applicability of some methods of corpus linguistics for lexico-semantic research in Neo-Latin, focusing on frequency and dispersion. All the approaches I will discuss in the following have been developed for and tested on texts in English, a language with few morphological markers; some of them have also been applied to other languages. Whether parameters fitted to English are suitable for languages with rich morphology such as Latin, has seldomly been tested, since few quantitative methods have been applied to Latin texts in general, even fewer to Neo-Latin texts.[4]

Corpus

I have used a corpus of historiographical texts in Latin developed for this purpose (see Appendix). It comprises two texts from Antiquity, Livy and Ammianus,[5] and five Latin texts from Early Modern Europe,[6] from the (Italian) fifteenth century Leonardo Bruni's *Historiae*, Flavio Biondo's *Historiarum*

[2] See the extraordinary results of Field 2016 (Caesar), Stover & Kestemont 2016 (Apuleius), Vainio et al. 2019 (re-attribution of the *De optimo genere oratoris* to Cicero).

[3] As defined, e.g., by Underwood 2016.

[4] See the impressive study Bloem et al. 2020. The only field where Latin texts have played a notable role, is stylometrics, not least because texts need little preprocessing and the software developed by the Computational Stylistics Group, mainly based at Cracow University has made the method accessible also to researchers with no mathematical background (https://computationalstylistics.github.io/). For an application of stylometrics to Neo-Latin texts see Ramminger 2019/2021.

[5] It should be noted that both Livy and Ammianus are fragments; in Livy's case by far the largest part has been lost. Obviously, about the lexicon of the lost parts and lexico-semantic developments no assumptions can be made.

[6] *Token* designates the words, as they appear one after the other in the text, *form* is their morphological appearance (in Corpus Linguistics commonly called *type*), *lemma* is the form used as headword in a dictionary. E.g., in Flavio Biondo there are 75 tokens for the lemma *bombarda*; these appear in six different forms (all that are possible): *bombarda* (3x), *bombardae* (11x), *bombardam* (2x), *bombardarum* (11x), *bombardas* (15x), *bombardis* (33x). For the unlemmatized words, the Type-Token-Ratio is 75:6, for the lemmatized words 6:1. In comparison, the English lemma *cannon* has only two forms (*cannon*, *cannons*); if we hypothesize a English translation precisely corresponding to the Latin text of Biondo, the TTR for the unlemmatized English text would be 75:2, for the lemmatized version 2:1.

decades (the third decade only), and Lorenzo Valla's *Gesta Ferdinandi Regis*; from the later period Thomas More's *Historia Richardi III*, and the Danish writer Erasmus Laetus' *De nato Christiano* (for all dates see Appendix). In order to perform comparisons using frequency tests I have furthermore used three other Latin texts from different periods and genres, Cicero's *De officiis*, Thomas Aquinas' *Summa theologica* (part 4), and Thomas More's *Utopia*, as well as George Orwell's *Nineteen Eighty-Four*. All of them (except for Orwell) have been lemmatised or proofed and part-of-speech tagged with *Collatinus* and corrected manually.[7]

The lemmatization of Latin

Tokenization (the splitting of a text into words as smallest units) is generally unproblematic in Early Modern Latin. It should be remembered, however, that the division of written text into words was not always a feature of Latin; in Antiquity and the early Middle Ages *scriptura continua* (continuous script) was used widely.[8] Thus we may assume that for most Latin texts from Antiquity the tokenization found in later manuscripts and modern editions is a reconstruction, liable to misinterpretations and ambiguities; equally, the frequent problem of distinguishing collocations from compounded tokens (*quamobrem*, *postmodum*) was inexistent during a sizeable period of Latin textual transmission.[9]

No standards for defining Latin lemmas have been formulated.[10] Thus lemmatizations from different projects are usually incompatible to some

[7] *Collatinus* is a lemmatizer developed by Yves Ouvrard and Philippe Verkerk. It is published with a GNU GPL v3 license and thus can easily be adapted to specific needs (https://outils.biblissima.fr/en/collatinus/). Manual proofreading of my corpus allowed, e. g., the consistent tagging of enclictic *-que* (and), which is usually ignored in lemmatizations of Latin texts, although it is one of the most frequent words in Latin. The prepositions *a/ab* and *e/ex* have been lemmatized as *ab* and *ex, neque/nec* as *neque*. Homograph forms from different lemmata have been disambiguated (e.g. *opera* from *opus* or *opera*).

[8] See the Vatican Virgil manuscript (Vatican, *Biblioteca Apostolica*, Cod. Vat. lat. 3225), written in Rome at about 400 AD. It contains fragments of *Aeneis* and *Georgica* by the Roman poet Virgil (70 BC–19 BC). See Bischof 1990, 172. Or the Codex Florentinus of the Collection of Roman law texts, commonly called *Digesta* or *Pandectae*, made at the behest of the emperor Justinian officially issued in 533 AD; the manuscript was written between 533 and 557 AD. See Baldi 2010 (with illustrations).

[9] See *Thesaurus Linguae Latinae* 1900-, 10.2 1225,59 s.v. *pridem* for *iam pridem* vs. *iampridem* (F. Spoth, *pridem*); 10.2 237, 73-78 (Euler, *postmodo et postmodum*) for *post modum* vs. *postmodum*, where both tokenizations are identifiable from syntactical evidence. A beautiful example of the ambiguities resulting from scriptura continua is presented in 7.2, 1807, 81-84 s.v. *lumbus* (Salvadore).

[10] See e.g. Gleim et al. 2019, Korkiakangas 2020.

degree.[11] In particular, the many words that can either be spelt as one word or a bigram (such as *prius-quam*, *uerum-tamen*, *ueri-similis*, etc.) and fixed phrases (*quo pacto*, how) present problems for which a consensus has yet to be established. The same holds true for the innumerable adjectival participles and lemmas used both as adjectives and nouns (e. g. *inimicus, contrarius*), which need to be harmonized for an effective lemmatization. Printed dictionaries have found pragmatic solutions for these (notably in the decision between different orthographies for compounds with *ab-, con-, sub- trans-* etc.), and digital databases can link such cases without deciding. In general, I have followed the lemmatizations of the *Oxford Latin Dictionary* as a standard and used analogous patterns for words not contained there.[12] The lemmatization of proper names presents different problems, not least because of the sheer mass and diversity of the material. They have in the following only been used for dispersion metrics.

For morphologically rich languages normalization either by lemmatization or stemming has long been recognized as an essential preparatory step for corpus-based research, if the inflectional properties of a word are not important for a task.[13]

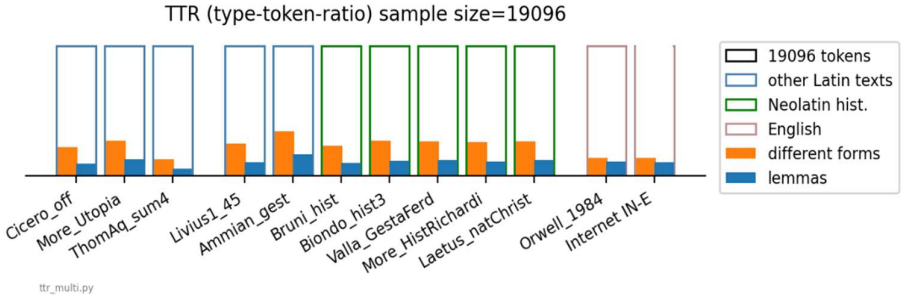

Figure 1: The lemmatization of Classical, Medieval, Neo-Latin & English texts (terminology see n.6)

[11] Cp. Mambrini & Passarotti 2019. Already in printed dictionaries some design decisions such as the numbering of homographs (e.g., the compounds of *cado* and *caedo*) are left to serendipity. The basic decision about which form to use as the lemma already decides interoperability (dictionaries of Classical Latin customarily use the first person singular of the present tense as lemma form of verbs, whereas dictionaries of Medieval Latin use the active infinitive of the present tense).

[12] The most important deviation from the *Oxford Latin Dictionary's* model is that derived adverbs are lemmatized under the adjectival form, unless the adverb has developed a significant *fortuna* of its own (i. e. *occulte* under *occultus*, but *abunde* seperate from *abundus*, since the adverb *abunde* is quite common, the adjective *abundus* rare). This follows the classification used by Gardner 1971.

[13] Kettunen 2014; Bentz et al. 2017 (for English and German); Kutuzov & Kuzmenko 2019. For Latin verbs, the morphological richness has recently been shown by Pellegrini & Passarotti 2018. For French, see Treffers-Daller 2013.

	length	sample	forms	lemmas	forms/lemmas
Cicero_off	33472	19096	5428	2292	42%
More_Utopia	27366	19096	6711	3158	47%
ThomAq_sum4	80557	19096	3126	1363	44%
Livy1_45	463349	19096	6186	2548	41%
Ammian_gest	117465	19096	8721	4159	48%
Bruni_hist	140045	19096	6054	2550	42%
Biondo_hist3	94554	19096	6931	2922	42%
Valla_GestaFerd	41290	19096	6785	3055	45%
More_HistRichardi	19096	19096	6684	2778	42%
Laetus_natChrist	42565	19096	6852	3028	44%
Orwell_1984	101866	19096	3577	2745	77%

Table 1: The lemmatization of Classical, Medieval, Neo-Latin and English texts (sample from middle of text, see n. 14)

The statistic (figure 1 and table 1) shows the type-token-ratio (TTR) between tokens and forms before, and between forms and lemmas after lemmatization.[14] The transformation of Latin texts by lemmatization is dramatic in terms of numbers. The repertory of forms in Latin seems to be about one fourth of the number of tokens (i.e., the length of the text), except for Thomas Aquinas: The formalized style of the scholastic Latin of the *Summa* needs a much smaller number of different forms to express its content. The impact of lemmatization on Latin texts is much higher than on English texts; whereas the number of forms of Latin texts are generally reduced by more than half by lemmatization, for the English texts in our comparison this is only one fifth.[15] If Orwell is any indication, English behaves very differently with a

[14] As the type-token ratio (TTR) decreases with text size and our texts are of very different size, the statistic is based on a sample equal to the shortest text in the corpus (except IN-E, see n. 15). See Baayen 2001, Malvern et al. 2004; Van Gijsel et al. 2005; Corral et al. 2015.

[15] The data for the English InternetCorpus IN-E could not be sampled (therefore the column in the graph is open at the top). For IN-E (source: http://corpus.leeds.ac.uk/list.html) the numbers are: all: 181.376.006, unique forms: 2.195.987, lemmas: 1.701.333. The ratio of numbers of forms to tokens (much smaller than in Orwell) may be due to the absolute length of the corpus (the type-token-ratio decreases with text length) as well as the repetitive nature of the texts contained. While the reduction from tokens to forms is much larger than in Orwell (1.21 forms per 100 tokens, Orwell: 19.77 per 100), the ratio of forms to lemmas is essentially the same (IN-E: 77.47 lemmas per 100 forms, Orwell: 78.16). The same proportions of forms to lemmas in English texts were observed e. g. by Toman et al. 2006. Obviously, the ratio tokens/lemmas depends largely on language typology. I have also calculated the numbers for

much lower rate of forms. Amongst the Latin texts Ammianus is an outlier. The number of forms is much higher, indicating that he either repeats the same words in different forms at a higher rate than the other authors or that he uses a larger vocabulary (which would also result in a larger pool of different forms). Since also the number of lemmas is correspondingly higher, it is his vocabulary that is more extensive than that of the other authors in this list.[16] Among the Neo-Latin texts Biondo (high number of forms) and Valla (highest ratio of lemmas) stand out – both numbers are indicators of different types of stylistic richness, the former with a more varied morphological repertory, the latter with a wider lexicon.

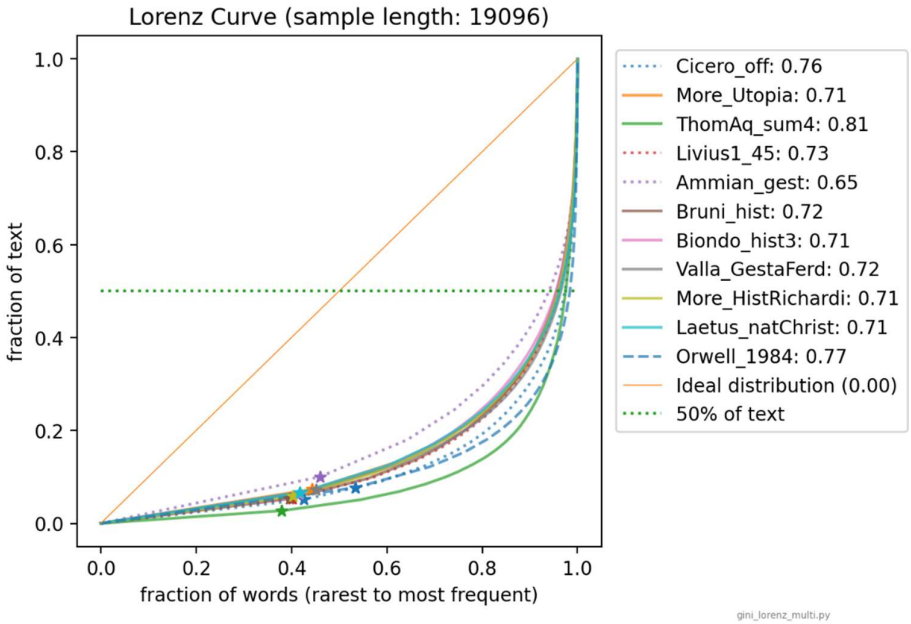

Figure 2: Distribution of lemmas within a text. The GI of the sample is given in the legend of the plot; the * indicates the point where the series of lemmas switches from frequency 1 to 2 (see Figure 2a).

samples from texts in German (Wedekind, Erdgeist: 3567 forms/2740 lemmas, 77%), Czech (the Czech translation of Orwell: 6590 forms/4058 lemmas, 62%), Bulgarian (the Bulgarian translation of Orwell: the Bulgarian translation of Orwell: 5486 forms, 3806 lemmas, 69%), and Finnish (the Finntreebank: 8344 forms, 4917 lemmas, 59%). None of them uses as many forms per lemma as Latin (numbers for these texts include proper nouns).

[16] The expansion of Ammianus's vocabulary compared to earlier Latin historians was already noted by early modern writers, who considered it a degeneration in line with the general decay of Roman culture; see Blockley 1996, 457–458. The differences are spectacular: Ammian (124345 words) 518 new words, Bruni (150944, longer than Ammianus): 201, Biondo (105922): 328, Valla (42798): 120, More (19457): 61, Laetus (43784): 205 (proportionally higher than Ammian).

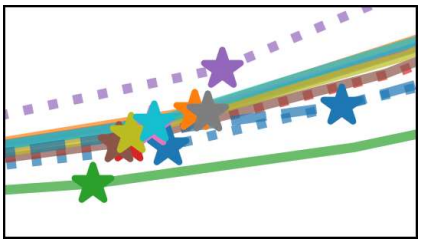

Figure. 2a: Detail from Figure 2; the * indicates the point where the series of lemmas switches from frequency 1 to 2

One reason why the TTR (in its basic form) is unreliable, lies in the internal structure of Latin (and other European languages) which organizes the information – aside from morphology – with the help of grammatical words (often also called function words or closed class words), in general comprising pronouns, prepositions, conjunctions, and some adverbs.[17] These are words of high frequency and in Latin mostly morphologically invariant. For some types of analyses these collected as stop-words and discarded. Their dominant role as the "glue" of a Latin text can be expressed by calculating the distribution of the frequencies of the lemmas present in the text.

To quantify the frequency distribution of texts Popescu in 2009 introduced the Gini index (GI) into linguistics.[18] The Gini index was formulated in 1912 by the statistician Corrado Gini as a measure of inequality and is widely used in economics to express income or wealth distribution within a population.[19] The GI is a number between 0 and 1; a GI of zero would describe a population of equally rich members, a GI of 1 would describe a population where one member possessed everything. Transferred to texts the GI describes how much of a text a word (in our case a lemma) "possesses", i. e., how frequently it occurs compared to the other lemmas in the same text. Since the GI increa-

[17] Discussions of function words/grammatical words/closed-class words in: Smith & Witten 1993, 3-4; Foolen 1996; Hartman & James 2002, 60; Argamon & Levitan 2005; Atkins & Rundel 2008, 164–165; Rosén 2009, 334-336; Kroon 2011; Rybicki & Eder 2011; Kestemont 2014. Their role in second language acquisition (note that Latin for a major part of its history was only learned as a second language): Laufer 2003; Restrepo Ramos 2015. Their role in the frequency ranking of a text (synsemantics/autosemantics): Popescu 2009, 95 (applying Hirsch 2005); Chen & Liu 2015.

[18] Popescu 2009, 54–63.

[19] Gini 1912; see also Deltas 2003 and the discussion in Ceriani & Verme 2012.

ses with text length, the following calculations are again based on a lemmatized sample equal in length to the shortest text.[20] Visually, the GI can be represented by the Lorenz curve developed by the American economist Max Lorenz.[21]

Figure 2 shows the Lorenz curves of our texts, and gives us information which the Gini index as a number alone cannot express. If none of the lemmas in a text occurred more than once, the Lorenz curve would be a straight line from the bottom left to the top right corner (here in orange) and half of the lemmas would contribute half of the text (at the intersection between orange line and the horizontal dotted line). Such a text (where every lemma occurred once), if it were possible at all, would be quite difficult to understand.

In reality, every text consists of many lemmas that occupy little of the text and a few that contribute much; the intersection of the 50%–line with the Lorenz curve shows that for most of our texts 10% or less of the lemmas contribute half of the text). Again, Orwell (in English) stands out: while the left part of the curve indicates a well-balanced variety of lemmas – more than half of the lemmas occur only once (the part of the curve to the left of the *) –, the curve rises late and is the most extreme towards the end (briefly even exceeding Thomas Aquinas), due to the high frequency of very few (grammatical) words.

Among the Latin texts Ammianus again behaves quite differently from the other texts. It is the one with the most lemmas used only once. As a further indication of the lexical richness of Ammian's text even the right tail of the curve is more spread out than that of the other texts. Thomas Aquinas' text, on the other hand, is quite repetitive – as one would expect from a medieval scholastic text; the number of lemmas used once is the smallest of all texts – though not by far (see the position of *) – since variation is not the stylistic aim of philosophical style; the internal organization seems to be dominated by few high frequent lemmas (the steep increase of the right tail). The Neo-Latin historical texts behave quite similarly to each other; the cut-off points (figure 2a) are clustered quite closely together; Valla's text is again shown to be the richest lexically.

A look at the most frequent lemmas in the historical texts (Table 2) confirms that it is indeed the grammatical words which are "in possession" of the text:

[20] For the dependence of textual richness on text length, see recently Shi & Lei 2020. The point has of course been made often before.

[21] Lorenz 1905. The Lorenz curve is construed by ordering all elements (in this case words) by frequency (rarest first, to the left) and adding their frequencies on top of each other. Since words with low frequency are at the left of the curve, it rises slowly at the beginning; words with high frequency are added last (to the right) and cause the curve to rise steeply.

Livy	ab	ad	atque	et		in	is		que	qui	sum	ut	
Ammian		ad		et	hic	in	is		que	qui	sum	ut	per
Bruni	ab	ad	atque	et	hic	in	is		que	qui	sum	14	
Biondo	ab	ad		et		in	is		que	qui	sum	ut	pontifex
Valla	14	ad	atque	et	hic	in	is	non	que	qui	sum	11	
More	20	17	atque	et	hic	in	is	non	que	qui	sum	ut	
Laetus	13		atque	et	hic	in	is	12	que	qui	sum	ut	cum *cj.*

Table 2: The most frequent words (in alphabetical order) shared in the historical texts. Numbers indicate a rank higher than 10.

The list in table 2 gives an overview over the ten most frequent words in our texts. When a word (i.e., a lemma) ranked under the first ten in most texts is not under the first ten, but under rank 11 to 20, the number in the table indicates the rank between 11 and 20; empty spots indicate that a word is not even among the twenty most frequent words. Part of the inequalities may be due to the vagaries of sampling.

The list shows the remarkable homogeneity of the texts. While the ten most frequent words are nearly all grammatical words (with the exception of *pontifex* in Biondo), up to rank 20 some semantic lemmas make an appearance (11–20 are not shown in table 2): *urbs* and *consul* in Livy, *hostis* and *urbs* in Bruni, in addition to *pontifex* also *hostis* in Biondo, *rex* in Valla and More.

Dispersion

Frequency alone is a poor indicator of Lexical Richness (metrics of the quality of the vocabulary), Keyness (the importance and distinctiveness of terms used) etc. An additional and equally important metric is dispersion. Dispersion is in corpus linguistics commonly understood as the pattern of recurrence of a word (or any other phenomenon) in a corpus of texts.[22] Equally, dispersion measures can be used to discover structures and patterns within a single text (i.e., the regularity of a word's recurrence).

The following will focus on the latter, i. e., on intra-textual dispersion measures. I will explore two strategies: first I will explore the distance between individual occurrences of a lemma (interarrival time), secondly, I will test regularity by observing the occurrence within segments of the text (deviation of dispersion). For reasons of space the following analysis will focus on one text, Valla's *Gesta Ferdinandi*.

[22] See the example in Gries 2020..

Interarrival time

The analysis of interarrival time is a common analytical process when a sequence of events recurring at uneven intervals is considered (arrival of flights, entry of clients into a shop, etc.), e. g., to predict waiting times. It was first used in corpus linguistics by Lijfjit in 2011[23] and is an attractive exploratory method because it replicates the natural reading process.

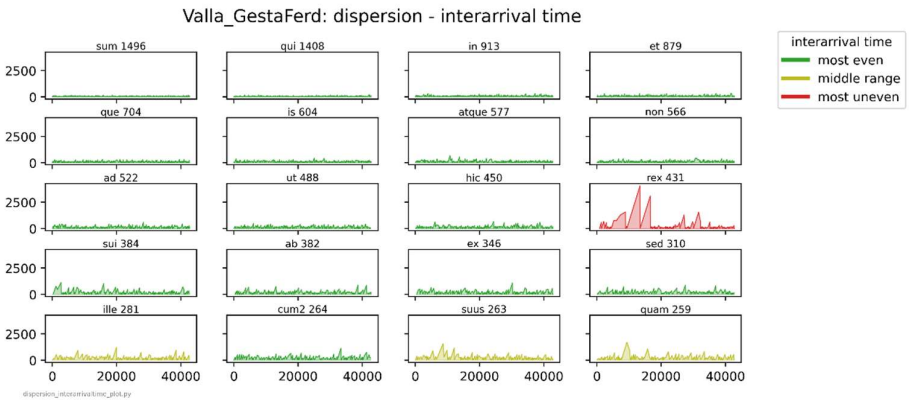

Figure 3: Interarrival time of the 20 most frequent lemmas in Valla (with frequency)

Interpreting figure 3 we see that for the majority of the most frequent lemmas interarrival times (or usage intervals) stay low continuously. The only exception is *rex* – also the only lexical lemma among them – which is only used intermittently in approximately the first third of the text.

The boxplot in figure 4 shows a more fine-grained view of the intervals between recurring instances of a lemma. The graph for every lemma consists of a core block at the bottom and some outliers (small horizontal lines above) connected with a vertical line. The core block at the bottom contains the most frequent intervals. With lemmas such as *sum*, *qui*, and *in* the core is very compact because the intervals are uniform and uniformly short (i. e., the lemmas are frequent and evenly spaced out). The cores are somewhat higher towards the right, indicating that there is a bigger variation in the distance even where the lemma is used regularly. Also, many of the grammatical words (coloured in grey) are very evenly dispersed, there are hardly any passages in the text where they do not occur (few outliers).

[23] Lijffijt et al. 2011, later also applied by Lijffijt et al. 2016.

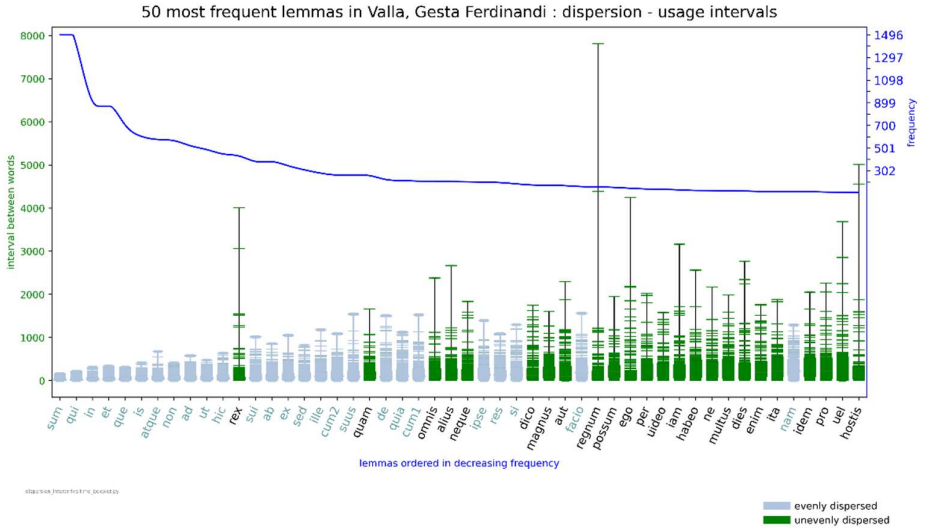

Figure 4: Dispersion of the 50 most frequent lemmas in Valla[24]

The most frequent unevely dispersed lemma is *rex*, with maximum intervals between examples of ca. 4000 and 3000 words), and *regnum*, which has one interval of nearly 8000 words with no occurrence (see below). The blue line (with the scale on the right) indicates the frequency of the lemmas in Valla. Except for the most frequent words there is no correlation between frequency and dispersion (note *nam*, which is relatively rare, but used evenly).[25]

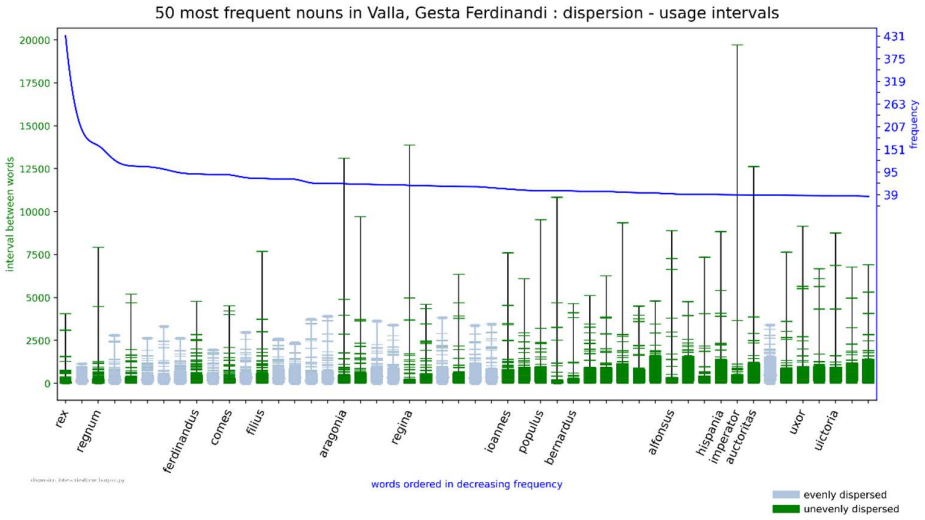

Figure 5: The most frequent nouns in Valla

[24] *cum1* is the preposition, *cum2* the conjunction.
[25] See Gries 2020, 117–118.

Figure 5, a list of the fifty most frequent nouns in Valla, brings us closer to the actual contents of Valla's work.[26] Greyed out are the nouns that are evenly dispersed throughout the *Gesta*.[27] Also I have not labelled the unevenly dispersed nouns with a more general meaning and no specific relation to the contents of the *Gesta*.[28] While the importance of the kings Ferdinando (I) (*ferdinandus*) and Alfonso (V) (*alfonsus*) is hardly surprising, *bernardus* brings another person to the foreground: most of the examples refer to the Spanish-Sicilian nobleman Bernardo de Centelles, a leading figure at the courts of both Alfonso and his successor.[29] I have left *ioannes* in the list to emphasize the need to disambiguate proper names before analysis; it refers to a number of different personages from the Aragonese orbit, notably kings John I and John II of Castile; thus its frequency has no analytical value. A further caveat relates to the dispersion of proper nouns: In passages with high occurrence the proper noun would often be substituted by a pronoun or in Latin just be implicit in the verb. Thus, neither frequency nor dispersion alone represent the semantic presence of the persons within the text.

dpnorm (Gries)

An alternative approach to measuring dispersion was proposed by Stefan Gries; it is based on registering the presence or absence of a word in contiguous segments of text.[30] If the dispersion within a corpus is measured, the segment is usually a work. The method can equally well be used for individual texts; in this case the segments can either be extracted from the structure of the text (chapters etc.) or arbitrarily established (if there are no "natural" segments).[31] An advantage of this method is that it allows us to put a number to the dispersion and thus compare different texts.

---

[26] Cp. Gries 2021. Words plotted (50): *alfonsus, animus, aragonia, arma, auctoritas, auxilium, bellum, bernardus, caput, castra, causa, comes, consilium, corpus, dies, domus, dux, eques, equus, ferdinandus, filius, hispania, homo, hostis, imperator, ioannes, ius, locus, manus, miles, mors, murus, nomen, oppidum, pars, pater, populus, regina, regnum, res, rex, socius, spes, tempus, turris, uictoria, uir, uita, urbs, uxor*.

[27] Words evenly dispersed, not labelled (17): *animus, bellum, causa, dies, domus, dux, homo, locus, manus, nomen, pars, pater, res, spes, tempus, uir, urbs*.

[28] Words unevenly dispersed, not labelled (17): *arma, auxilium, caput, castra, consilium, corpus, eques, equus, hostis, ius, miles, mors, murus, oppidum, socius, turris, uita.*

[29] See Putzulu 1979.

[30] See Gries 2008, Gries 2010; Lijffijt & Gries 2012.

[31] If used for individual texts with arbitrary segment length, the result will be influenced by word clusters occurring at segment borders; if a cluster is divided between two segments, the dispersion will appear more even than it actually is. To compensate for such cases I have calculated the dispersion of overlapping segments by moving the segment border one word

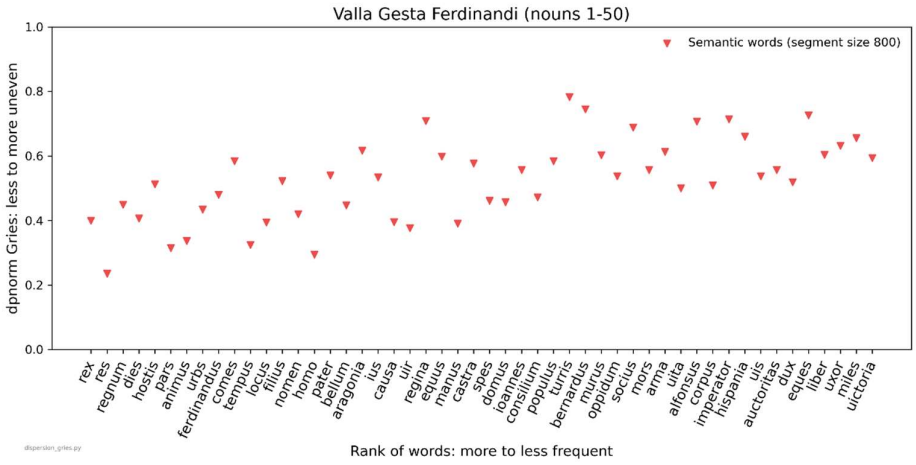

Fig. 6: The 50 most frequent nouns (incl. proper nouns) and their dispersion

The results plotted in figure 6 are essentially the same; again, it is quite clear that frequency and dispersion are independent from each other (even if we discern a slight tendency towards more uneven dispersion at the right). The dispersion metric becomes extremely useful for comparing texts. In our case I have compared Valla's *Gesta* to the authoritative example of historiography in Latin, Livy's *Ab urbe condita*.

The comparison plotted in figure 7 allows us several observations.[32] Some differences in dispersion are clearly connected with the contents: *rex* and *regina* are simply more pervasive factors in Valla's narrative than in Livy's. The same holds true for *pater* and *filius,* in many cases connected to dynastic considerations or family politics – again less prominent in Livy. On the other hand, words and concepts such as *hostis* and *bellum* are much more unevenly dispersed in Valla; this may have to do with changing political concepts, but also with the fact that a large part of Livy is occupied with warfare (not least with Hannibal, the public enemy number one). Certainly, the numbers for dispersion and frequency support each other in this case, *bellum* occurs in Livy 2452 times, in Valla a mere 67; this is significant, even allowing for the difference in length.

In other cases, the difference can be explained by a semantic development (for *imperator* and *socius* see below). It needs to be emphasized, however, that not in all cases the differences of dispersion can easily be

further through the whole text and calculating the mean. The segment length has been arbitrarily established at 800 words.

[32] The following words have the same dispersion in both texts (threshhold 0.05) and are not plotted in figure 7: *animus, caput, causa, consilium, dies, dux, equus, locus, nomen, pars, populus, res, tempus, urbs.*

explained; the differences in dispersion between the two texts of *vita* and *victoria* would probably merit a closer analysis.

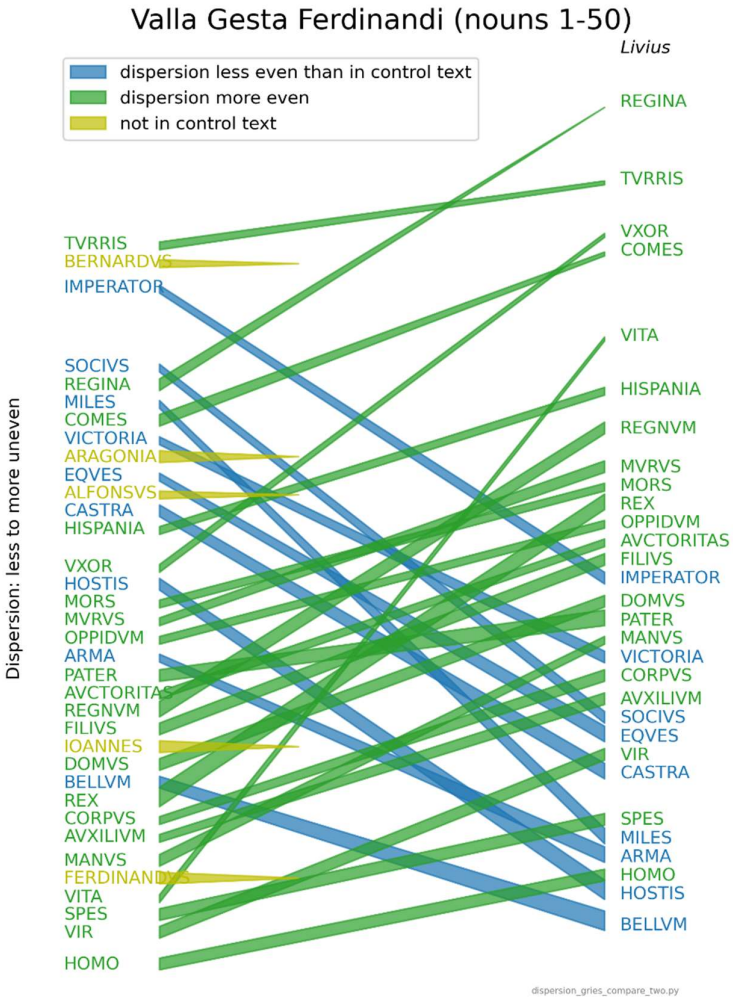

Figure 7: Comparison of dispersion between Valla and Livy. The fifty most frequent nouns in Valla ordered according to their dispersion.

Obviously, the proper nouns important for Valla do not occur in Livy (marked in yellow), except for *Hispania* which shows an interesting dispersion both in Valla and in Livy (figure 8):

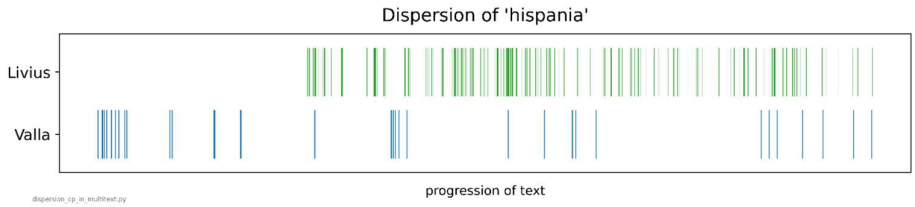

Figure 8: Dispersion of *Hispania* in Valla and Livy

In Valla there is a cluster at the very beginning of the work, where Valla gives a brief outline of the geographic and political parameters of Spain.[33] A second cluster corresponds to the speech Fernando gives upon the death of king Martin I in 1410, in which he claims the throne of Aragon. In Livy mention of Spain begins with Book 21, treating the beginning of the Second Punic War which is triggered by warfare between Roman and Carthaginian military in Spain. Other differences in dispersion indicate semantic change: *imperator* in Livy is the commander-in-chief of an army and as such pervasive from early on, while in Valla it is nearly exclusively the *imperator Romanorum*, i. e., the Emperor, and thus limited to parts of his narration where the Emperor plays a role. With *rex* and *regnum* (figure 9), the realities referred to are different in both authors:

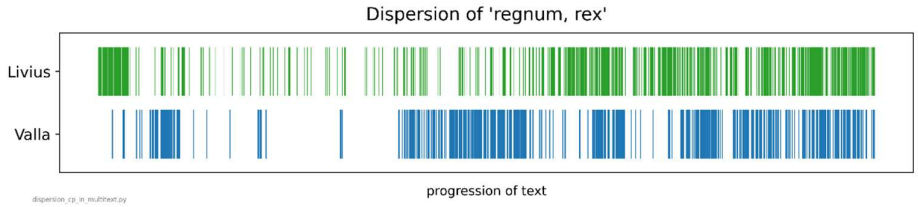

Fig. 9: Dispersion of *rex* and *regnum* in Valla and Livy

In Valla, after the initial discussion of the Spanish political situation, *rex/regnum* become key terms in the narration after the accession of Fernando to the throne. In Livy, there is an initial cluster, dealing with the early era of the Roman kings; in the second half the graph reflects the ever-increasing presence of the kings of the East in Livy's narration, with the description of

[33] As Valla announces at the end of the preface: "Sed quoniam de duobus Hispanis regibus locuturus sum, Ferdinando qui primus e Castella regno Aragonie, Alfonso eius filio qui primus ex Aragonia regno Italie potitus est, aliquid de ipsa Hispania altius repetam" (But since I am going to talk about two Spanish kings, Ferdinando who as the first from Castile gained the kingdom of Aragon, and his son Alfonso who as the first from Aragon gained the kingdom of Italy, I will present Spain itself in more detail).

the Eastern wars starting in Book 31 and filling the remainder of the preserved part of Livy's *History*.

We find an analogous semantic development with *socius* ('ally'). Whereas allies are part of the political fabric of Roman warfare from early on, *socius* is hardly ever used in this sense in Valla's narration (*socius* in Valla mainly refers to an individual in a group connected to someone(?) by some circumstance, i. e., 'companion').

Conclusion

This paper has hardly scratched the surface as regards the possibilities of quantitative research in our authors. The vast fields of Lexical Richness and Keyness have only been marginally touched upon. Nevertheless, we could establish the usefulness of several text metrics for research in Latin and specifically Neo-Latin. The Gini Index (and the Lorenz-curve as its graphical representation) analysed the very fabric of Latin, drawing attention to the importance of grammatical words for the construction and organization of information in the text. Differences due to the chronological dispersion of the texts were hardly visible; outliers in the lemmatization metrics such as Ammianus and Thomas Aquinas are due to the individual style of the former and the genre of the latter. Metrics of dispersion allowed us to connect the information contained in the narratives to specific features (i.e., words) of the texts; the comparison between Valla's *Gesta Ferdinandi* and Livy's *Ab urbe condita* expressed the differences in political structure between the worlds depicted in the two texts by connecting it to specific words in the texts. Thus, through a combination of quantitative analysis (distant reading) and 'traditional' close reading we can highlight important aspects of our texts.

Appendix: Sources

Titus Livy (59 BC–17 AD), *Ab urbe condita* (From the foundation of Rome). Digital source: http://data.perseus.org/texts/urn:cts:latinLit:phi0914.phi0011. Length: 463349 tokens. Abbreviation: Livy1_45

Ammianus Marcellinus (fl. 390), *Rerum gestarum quae exstant* (Deeds [of Emperor Julian]) (books 14–31). Digital source: http://data.perseus.org/texts/urn:cts:latinLit:stoa0023.stoa001. Length: 117465 tokens. Abbreviation: Ammian_gest

Leonardo Bruni (c. 1370–1444), *Historiae Florentini populi* (History of the Florentine People) (1442). Edition used: Leonardi Aretini *Historiarum Florentini populi libri XII* [Dalle origini all'anno 1404], a cura di Emilio Santini, *Rerum italicarum Scriptores* XIX,3. Città di Castello 1914, 3–288. Length: 140045 tokens. Abbreviation: Bruni_hist

Flavius Blondus, Biondo Biondi (1388–1463), *Historiarum ab inclinatione Romanorum imperii decades* (History for the Decline of the Roman Empire) (1439–1453). Edition used: *Historiarum ab inclinatione Romanorum libri XXXI* (Basileae 1531). Only the third decade was lemmatized. Length: 94554 tokens. Abbreviation: Biondo_hist3

Lorenzo Valla (1407–1457), *Gesta Ferdinandi Regis Aragonum* (Deeds of King Ferdinand of Aragon) (ca. 1445/1446). Edition used: Laurentii Valle *Gesta Ferdinandis Regis Aragonum* edidit Ottavio Besomi. *Thesaurus mundi* 10. Patavii 1973. Length: 41290 tokens. Abbreviation: Valla_GestaFerd

Thomas More (1478–1535) *Historia Richardi III* (History of Richard III.) (ca. 1513). Digital Source: *Lemmatized Concordance of Historia Richardi Tertii*, CW 2 (Thomas More Studies 10.1, 2015). URL: https://thomasmorestudies.org/concordance/. Length: 19096 tokens. Abbreviation: More_HistRichardi

Erasmus Laetus, Rasmus Glad (1526–1582), *De nato baptisatoque primo Friderici II filio Christiano* (The birth and baptism of Christian, the first son of Frederik II.). Edition used: Erasmi Michaelii Laeti *de nato baptisatoque primo Friderici II potentissimi Danorum regis filio Christiano, duce Holtzatiae, deque istius inaugurationis magnificentia, plausu et solennitate, historiarum libri IIII*. Hafniae 1577. URL: http://renaessancesprog.dk/tekstbase/Laetus_De_Nato_1577/1/. Length: 42565 tokens. Abbreviation: Laetus_natChrist

**

M. Tullius Cicero (106–43 BC), *De officiis* (Duties) (44 BC). Digital Source: PROIEL. https://github.com/proiel/proiel-treebank/blob/master/cic-off.conll. Length: 33472 tokens. Abbreviation: Cicero_off

Thomas Aquinas (1225–1274), *Summa theologiae* (The Sum of Theology; unfinished). Only part 4 was used. Digital source: Index Thomisticus Treebank, https://github.com/Universal Dependencies/UD_Latin-ITTB. Length: 80557 tokens. Abbreviation: ThomAq_sum4

Thomas More, *Utopia* (1516). Digital source: *Major Latin Terms in Thomas More's Utopia, CW 4: A Lemmatized Concordance.* Thomas More Studies 11.1. 2016. URL: https://thomasmorestudies.org/wp-content/uploads/2020/09/tms11.1-1.pdf. Length: 27366 tokens. Abbreviation: More_Utopia

George Orwell (1903–1950), *Nineteen Eighty-Four* (1949). Digital source: MULTEXT-East "1984" annotated corpus 4.0. URL: https://www.clarin.si/repository/xmlui/handle/11356/1043. Length: 101866 tokens. Abbreviation: Orwell_1984. The corpus includes Czech and Bulgarian translations (see n.15).

Finntreebank ftb1u (sentences or sentence fragments used as linguistic examples in a descriptive grammar of Finnish) (2014). URL: http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/sources/ftb1u-v1.zip

The texts of Biondo and Laetus were rendered machine-readable with *OCR4all* (https://github.com/OCR4all). In all cases (except for Orwell), the lemmatizations were extensively controlled and revised. All sources will be available in the context of the *Danish Center for Neo-Latin*, URL: www.dcnl.dk.

Bibliography

(All internet addresses were last visited on 6 January 2022).

Argamon, Shlomo & Shlomo Levitan 2005, "Measuring the usefulness of function words for authorship attribution", *Proceedings of the 2005 ACH/ALLC Conference, 4-7 June*. URL: https://www.researchgate.net/publication/227400638

Atkins, B. T. Sue & Michael Rundell 2008, *The Oxford Guide to Practical Lexicography*. Oxford.

Baldi, Davide 2010, "Il Codex Florentinus del Digesto e il Fondo Pandette della Biblioteca Laurenziana (con un'appendice dei documenti inediti)", *Segno e Testo* 8, 99–186. URL: http://www.bml.firenze.sbn.it/it/PDF/pandette.pdf

Baayen, Harald 2001, *Word Frequency Distributions*, Dordrecht.

Bentz, Christian et al. 2017, "Variation in Word Frequency Distributions: Definitions Measures and Implications for a Corpus Based Language Typology", *Journal of Quantitative Linguistics* 24, 128–162.

Bischoff, Bernhard 1990, *Latin Palaeography Antiquity and the Middle Ages*, trans.: Daibhm O. Cróinin & David Ganz, Cambridge. Originally 1979; 2nd rev. ed. 1986.

Bloem, Jelke et al. 2021, "Distributional Semantics for Neo-Latin", *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, 84–93. URL: https://aclanthology.org/2020.lt4hala-1.13

Blockley, Roger 1996, "Ammianus Marcellinus and His Classical Background: Changing Perspectives", *International Journal of the Classical Tradition* 2, 455–466.

Ceriani, Lidia & Paolo Verme 2012, "The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini", *The Journal of Economic Inequality* 10, 421–443.

Chen, Ruina & Haitao Liu 2015, "Ideologies of Supreme Court Justices: Quantitative Thematic Analysis of Multiple Opinions of 'Bush v. Gore 2000'", *Glottotheory* 6.2, 299–322.

Corral, Álvaro, Gemma Boleda & Ramon Ferrer-i-Cancho 2015, "Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts", *PLoS One* 10.7, e0129031. doi: 10.1371/journal.pone.0129031. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4497678/#pone.0129031.ref027

Deltas, George 2003, "The Small-Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research", *The Review of Economics and Statistics* 85.1, 226–234.

Field, Anjalie 2016, "An Automated Approach to Syntax-based Analysis of Classical Latin", *Digital Classics Online* 2, 3, 57–78. URL: https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/32315.

Foolen, Ad 1996, "Pragmatic particles", In *Handbook of Pragmatics*, eds.: Jef Verschueren, Jan Ala Ostmann, Jan Blommaert & Chris Bulcaen, Amsterdam, 1–24.

Gardner, David Dixon 1971, *A Frequency Dictionary of Classical Latin Words*, Ph.D. dissertation, Stanford University.

Gini, Corrado 1912, *Variabilità e Mutuabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*, Bologna.

Gleim, Rüdiger et al. 2019, "A practitioner's view: a survey and comparison of lemmatizationand morphological tagging in German and Latin", *Journal of Language Modelling* 7.1, 1–52.

Gries, Stefan Th. 2008, “Dispersions and adjusted frequencies in corpora”, *International Journal of Corpus Linguistics* 13. 4, 403–437.

Gries, Stefan Th. 2010, “Dispersions and adjusted frequencies in corpora: further explorations”, *Corpus-linguistic applications. Current studies, new directions,* eds.: Stefan Th. Gries, Stefanie Wulff, & Mark Davies (*Language and Computers* 71), Leiden, 197–212.

Gries, Stefan Th. 2020, “Analyzing Dispersion”, *A Practical Handbook of Corpus Linguistics*, eds.: Magali Paquot & Stephan Th. Gries, Cham, 99–118.

Gries, Stefan Th. 2021 “A new approach to (key) keywords analysis: Using frequency, and now also dispersion”, *Research in Corpus Linguistics* 9.2: 1–33. URL: http://ricl.aelinco.es/first-view/150-Article%20Text-1031-3-10-20210224.pdf

Hartmann, Reinhardt Rudolf Karl & Gregory James 2002, *Dictionary of Lexicography*, London & New York (first published 1998).

Hirsch, Jorge Eduardo 2005, “An index to quantify an individual’s scientific research output”, *Proceedings of the National Academy of Sciences* 102, 16569–16572.

Kestemont, Mike 2014, “Function Words in Authorship Attribution. From Black Magic to Theory?” *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, eds.: Anna Feldman, Anna Kazantseva & Stan Szpakowicz, Gothenburg, 59–66.

Kestemont, Mike & Jeroen de Gussem 2016, “Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning”, *Journal of Data Mining and Digital Humanities: Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, eds.: Marco Büchler et al. URL: https://jdmdh.episciences.org/3835

Korkiakangas, Timo 2020, “Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters”, *SSL* 58.1, 67–94.

Kroon, C. 2011, “Latin Particles and the Grammar of Discourse”, *A Companion to the Latin Language*, ed.: James Clackson, Chichester & Malden, 176–195.

Kettunen, Kimmo 2014, “Can Type-Token Ratio be Used to Show Morphological Complexity of Languages?”, *Journal of Quantitative Linguistics*, 21.3, 223–245. DOI:10.1080/09296174.2014.911506.

Kutuzov, Andrey & Elizaveta Kuzmenko 2019, “To lemmatize or not to lemmatize: how word normalisation affects ELMo performance in word sense disambiguation”, *NODALIDA2019 Deep Learning for Natural Language Processing workshop*. URL: https://aclanthology.org/W19-6203/

Laufer, Batia 2003, "Vocabulary Acquisition in a Second Language: Do Learners Really Acquire Most Vocabulary by Reading? Some Empirical Evidence", *Canadian Modern Language Review / La Revue canadienne des langues vivantes* 59, 4, 567–587. URL: https://doi.org/10.3138/cmlr.59.4.567

Lijffijt, Jefrey et al. 2011, "Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping". *Proceedings of ECML-PKDD 2011—Part II*, eds.: D. Gunopulos et al., Berlin, 341–357.

Lijffijt, Jefrey & Stefan Th. Gries 2012, "Correction to Stefan Gries' 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics*, 13, 4 (2008), 403–437", *International Journal of Corpus Linguistics* 17.1, 147-149.

Lijffijt, Jefrey et al. 2016, "Significance testing of word frequencies in corpora", *Digital Scholarship in the Humanities* 31.2, 374–397. URL: https://doi.org/10.1093/llc/fqu064

Lorenz, Max O. 1905, "Methods of measuring the concentration of wealth", Publications of the American Statistical Association 9.70, 209–219.

Malvern, David et al. 2004, *Lexical Diversity and Language Development: Quantification and Assessment*, London.

Mambrini, Francesco & Marco Passarotti 2019, "Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin", *Proceedings of the 13th Linguistic Annotation Workshop*, 71–80.

Oxford Latin Dictionary 1968, eds.: Peter G. W. Glare et al., Oxford [1]1968. [2]2012.

Pellegrini, Matteo & Marco Passarotti 2018, "LatInfLexi: an Inflected Lexicon of Latin Verbs", *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018): CEUR Workshop Proceedings* 2253. URL: ceur-ws.org/Vol-2253/

Popescu, Ioan-Iovitz 2009, *Word Frequency Studies*, Berlin & New York.

Putzulu, Evandro 1979, "Centelles, Bernardo de", *Dizionario Biografico degli Italiani* 23. URL: https://www.treccani.it/enciclopedia/bernardo-de-centelles_%28Dizionario-Biografico%29/

Ramminger, Johann 2019/2021, "Stylometry in a Language without Native Speakers: A Test Case from Early Modern Latin", *Philology Then and Now. Proceedings of the Conference held at The Danish Academy in Rome, 16 July 2019*, eds.: Trine Arlund Hass & Marianne Pade (*Analecta Romana Instituti Danici* 44), 151-167.

Restrepo Ramos, F. D. 2015, "Incidental vocabulary learning in second language acquisition: A literature review", *PROFILE Issues in Teachers'*

Professional Development 17.1, 157–166. URL: http://dx.doi.org/10.15446/profile.v17n1.43957.

Rosén, Hannah 2009, "Coherence, sentence modification, and sentence-part modification – the contribution of particles", *New perspectives on historical Latin syntax*, eds.: Philip Baldi & Pierluigi Cuzzolin, Berlin, 317–442.

Rybicki, Jan & Maciej Eder 2011, "Deeper Delta across genres and languages: do we really need the most frequent words?", *Literary and Linguistic Computing* 26.3, 315–321. URL: https://doi.org/10.1093/llc/fqr031

Shi, Yaqian & Lei Lei 2020, "Lexical Richness and Text Length: An Entropy-based Perspective", *Journal of Quantitative Linguistics*. URL: https://doi.org/10.1080/09296174.2020.1766346

Smith, Tony & Ian H. Witten 1993, *Language inference from function words*. (*Working Paper Series* 93.3), Hamilton. URL: https://core.ac.uk/download/pdf/44290163.pdf

Stover, Justin & Mike Kestemont 2016, "Reassessing the Apuleian Corpus: A Computational Approach to Authenticity", *The Classical Quarterly* 66.2, 645–672.

Thesaurus Linguae Latinae 1900–. Leipzig & Stuttgart.

Toman Michal, Roman Tesar, Karel Jezek 2006, "Influence of word normalization on text classification", *Proceedings of InSciT*, Mérida, 354–358. URL: http://textmining.zcu.cz/publications/inscit20060710.pdf

Treffers-Daller, Jeanine 2013, "Measuring lexical diversity among L2 learners of French. An exploration of the validity of D, MTLD and HD-D as measures of language ability", *Vocabulary Knowledge, Human ratings and automated measures*, eds.: Scott Jarvis & Michael Daller (*Studies in Bilingualism* 47), Amsterdam & Philadelphia, 79–104.

Underwood, Ted 2016, "Distant Reading and Recent Intellectual History", *Debates in the digital humanities 2016*, eds.: Matthew K. Gold & Lauren F. Klein, Minneapolis & London, 530–533. URL: https://doi.org/10.5749/j.ctt1cn6thb.47

Vainio, Raija et al. 2019, "Reconsidering Authorship in the Ciceronian Corpus Through Computational Authorship Attribution", *Ciceroniana on line* 3.1, 15–48. URL: https://doi.org/10.13135/2532-5353/3518. Data are on https://github.com/propreau/reconsideringciceronianauthorship

Van Gijsel, Sofie, Dirk Speelman & Dirk Geeraerts 2005, "A Variationist, Corpus Linguistic Analysis of Lexical Richness", *Proceedings from the Corpus Linguistics Conference Series* 1.1, 1–16.