”Don’t worry, it’s just noise”: quantifying the impact of files treated as single textual units when they are really collections

Thibault Clérice

“Don’t worry, it’s just noise”: quantifying the impact of files treated as single textual units when they are really collections

Thibault Clérice
Centre Jean Mabillon, École Nationale des Chartes, PSL University / 65 rue Richelieu, 75002 Paris, France
HiSoMa, Université Lyon 3
thibault.clerice@chartes.psl.eu

Abstract

Literature works may present many autonomous or semi-autonomous units, such as poems for the first or chapter for the second. We make the hypothesis that such cuts in the text’s flow, if not taken care of in the way we process text, have an impact on the application of the distributional hypothesis. We test this hypothesis with a large 20M tokens corpus of Latin works, by using text files as a single unit or multiple “autonomous” units for the analysis of selected words. For groups of rare words and words specific to heavily segmented works, the results show that their semantic space is mostly different between both versions of the corpus. For the 1000 most frequent words of the corpus, variations are important as soon as the window for defining neighborhood is larger or equal to 10 words.

1 Introduction

“You shall know a word by the company it keeps.”(Firth, 1957). Over the last decades, Firth’s sentence has seen its frequency grow as tools for analyzing both short and large corpora have found their way into the personal computer of students and researchers in other fields than linguistics, corpus linguistics and natural language processing in general (Heiden, 2010; Sinclair and Rockwell, 2016). Research using this postulate, efficiently summarizing the semantic distributional hypothesis[1], have been used in both the Latin and Ancient Greek domains on classical (Roda et al., 2019), late antiquity (Munson, 2017), medieval (Guerreau, 1989; Perreaux, 2012) and post-medieval literature (Bloem et al., 2020) with corpora spanning from few dozen thousand tokens to a few millions. As for many experiences, the set-up of this kind of information extraction[2] might have an important impact on the outcome of the analysis, as it will influence the “companionship” of the analyzed words: normalizing the text with lemmatization for example will reduce complexity and augment the signal for morphologically rich languages but reduce details found in some forms (*e.g.*, imperative in verbs); for the same reason, manipulating the size of the window defining the neighborhood of words allows for capturing more or less information. If these pre-processing steps are often known and made explicit, one is often unclear or dismissed: the way the source corpus is encoded and read is often omitted by authors[3]. But what happens when someone uses a digitized corpus made up of composite works?

If we were to take the example of the Perseus’s “Canonical Latin Literature” repository (Crane, 2021b), some files are actually composite works, such Martial’s *Epigrammata*, a collection of more than a thousand poems of varying but short lengths, while some others are “more” monolithic such as an *oratio* of Cicero. While there are (rare) studies of noise in digital humanities, and the few that exist focus on OCR quality and its impact (Eder, 2013a,b), none seems to have addressed the potential impact of treating texts as continuous strings of tokens, as they are found in digital format (plain text or the like) instead of treating them as collections of independent textual units within the same file. We note, however, the study of (Schöch, 2017) which explicitly studies two forms of segmentation, plays as a whole and arbitrary segmented plays, and their effect in a bag of words approach.

[1] Based on this hypothesis, we expect words sharing the same meaning or being semantically close to each other to be found with the same neighbor words. *e.g.*, “This morning, I drank an _ juice” is easy to fill with multiple words (*e.g.*, orange, apple) which share the same semantic trait: they are fruits.

[2] And others using bag of words such as topic modeling.

[3] *e.g.*, neither (Köntges, 2020) – in the context of a Bag-of-words approach – nor (Stringham and Izbicki, 2020) – in the context of word embeddings – address this question.

2 Corpus, concepts and methodology

2.1 Concepts for segmentation

In Classical Latin, as in most modern books, published works are usually split in various smaller textual units, which might be chapter, recipes, poems, etc. For work in prose such as novels or history books, chapters and paragraphs are usually the unit one could refer to. This segmentation is often an editorial or authorial way to indicate from light to strong topical shifts or narrative ellipsis. In poetry, most poems are published in form of collections, and, at least for Latin literature, they are not expected to be sequential: there are very few if none that connect throughout Martial's *Epigrammata* as direct sequence, and when there are connections, they probably are more echoes than the result of a progression. There are other genres and textual forms that we were able to keep over millennia, such as Apicius' *Recipes* to medicine notebooks such as Caelius Aurelianus' *Gynaeciorum Sorani* or grammatical commentaries from scholiasts such as Porphyrio: again, these are merely a single cohesive narrative sequence but rather a collection of short units, connected through a global theme.

And unlike modern literature, where we would expect chapters and paragraphs to be authorial marks on the text, the status of these marks can differ from one genre to another for ancient literature. These texts have been transmitted, reinterpreted and – as such – modified as soon as few centuries after they were first published[4]. For some of these segmentations, we know for a fact they were there originally: this is the case for the segmentation category we call "book" today which were often rolls or *volumen* published by authors at the time (Canfora, 2016, p. 13). For poetry, most of the segmentation in poems is certainly drawn from the original work with some doubts for the order of poems[5]: for rarely copied works such as the *Priapea*, some doubts can be easily instigated in how some poems can be segmented, but it remains a rare case. On the other hand, we know for a fact that some works were cut or reorganized by latter hands, such as *scholia* and commentaries in general: current hypothesis have them originating as notes to a text connected through lemma (*hypomnemata*) or *glosae*, and ended up as continuous texts in which the text was inserted (Bureau, 2012). In most other situations, the current segmentation of the text is either the effect of medieval scholars, such as for the verse numbering of the Bible, 16–17th century editors or modern ones: such is the case for the *Pro Murena*, as Fotheringham demonstrates it (Fotheringham, 2007). Not only the later text exists with two competing segmentation, but the paragraphs, when they are not numbered and identified, are sometimes not the same from one editor to the other. Of course, there are even more complex textual traditions which sometimes challenge text order, such as the one of Petron's *Satyricon* or Plaute's plays, and propose completely different forms of works, such as the *Epistola Alexandri ad Aristotelem*, an anonymous work which exists in two different *recensio*.

Whoever segmented the texts, authors or editors, they carry information about how the full work should be read by a human being. In this context, we propose to categorize the units formed by these segmentation in two types: on one hand, the ones that are clearly non-sequential – such as poems – as *autonomous textual units* (ATU), on the other, the more loosely connected elements – such as chapters – as *Semi-Autonomous Textual Units* (SATU). In this context, textual autonomy is achieved when a word from a textual unit and the word from following or preceding units cannot be classed as co-occurring, such as poems, because they are narratively, thematically or discursively unrelated. For SATU, the semi-autonomous character can be discussed, but chapters or books certainly would display a certain level of discursive autonomy with each other, while enabling discursive progression. In Latin corpora such as the ones following CapiTainS encoding guidelines for TEI (Clérice, 2017), each text has been thoroughly annotated with a citation scheme, such as Book → Poem → Line, by their corpus editorial team.

As these texts could be used within the framework of the distributional hypothesis, we propose a first metric to evaluate the potential risk of noise that would be introduced using fixed windows for context retrieval: the *theoretical window contamination rate*. For a given text t, it can be quantified as a function of the number of (S)ATU of the text

[4] And for some of the work we know under a single author's name and a single work title, we know for a fact there was either multiple authors (*e.g.*, Caesar's *De bello gallico*), multiple original works collected by later "editors" (*e.g.*, the Bible) or both such as Sulpicia whose elegies are found in the *Corpus Tibullianum*.

[5] See the difference in the edition of Leon Herrmann (Catulle and Herrmann, 1957) compared to the others such as Lafaye's one (Catulle and Lafaye, 1932).

(Poem 39)
Iliaco similem puerum, Faustine, ministro
Lusca Lycoris amat. Quam bene lusca videt!
(Poem 40)
Inserta phialae Mentoris manu **ducta**
Lacerta vivit et timetur argentum.
(Poem 41)
Mutua quod nobis ter quinquagena dedisti
Ex opibus tantis, quas gravis arca premit,
[...]

(Poem 39)
Iliaco similem puerum, Faustine, ministro
Lusca Lycoris amat. Quam bene lusca videt!
(Poem 40)
Inserta phialae Mentoris manu **ducta**
Lacerta vivit et timetur argentum.
(Poem 41)
Mutua quod nobis ter quinquagena dedisti
Ex opibus tantis, quas gravis arca premit,
[...]

Figure 1: Book 3 Poem 39–41 from Martial's *Epigrammata*. The co-occurring words of **ducta** for $W = 10$ are underlined, on the left in a segmented corpus, on right in a raw corpus.

$|U_t|$, the size of the window used for semantic information retrieval W and the number of tokens in the text $|t|$ such as

$$Rate(t) = \begin{cases} \frac{2W(|U_t|-1)}{|t|} & \text{if } |t| > 2W \\ 0 & \text{otherwise} \end{cases}$$

where each token in a (S)ATU until W has up to $2W$ co-occurring tokens drawn from neighbor (S)ATU except for the very first and last units ($|U| - 2 \times \frac{1}{2}$) of the text, which has either no following or no preceding unit (hence $\frac{1}{2}$). This rate represents the relative quantity of tokens whose window has at least one token not supposed to be counted as co-occurring. In this context, with a default window of 5-words of tools such as Gensim (Řehůřek et al., 2011), we have high rates for texts such as 21.22% for Martial's *Epigrammata*, a collection of 1,527 poems over 14 books and 71,911 tokens, and 0% for work in prose such as Sallust's *Iugurthia* (continuous which means only 1 (S)ATU while having 25,411 tokens). This would imply that up to one fourth of the tokens of Martial's work could end up polluted by non-co-occurring words, albeit at various scales (the first and last words of each unit being more polluted than the $W + 1$ one which end up with only 1 noise token). However, the theoretical window contamination rate assumes an equally distributed number of words in (S)ATU and is an efficient tool to consider the issue, the real contamination rate being dependent of the size of previous passages, following passages and size of each (S)ATU. In the case of very small poems such as Martial's *Epigrammata* 3.40 (10 tokens), not a single token window contains a clean set of co-occurring words starting with $W = 5$, with a real contamination rate reaching 1.0, and from $W \geq 10$, each token draw co-occurrences from both neighbor units at the same time (*cf.* Figure 1).

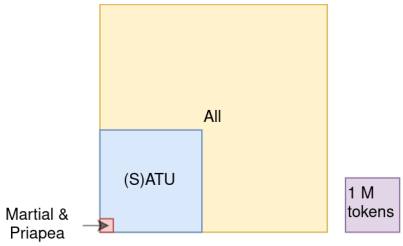

Figure 2: Scaled representation of the three sub-corpora and their relation to each other: each corpus contains the smaller one(s).

2.2 Corpora and Their Pre-Processing

In order to evaluate the effect of dismissing (S)ATU's importance, we propose a study based on the *Corpus Latin antiquité et antiquité tardive lemmatisé* (Clérice, 2020a) which aggregates original works from Perseus (Crane, 2021b), Open Greek and Latin (Crane, 2021a), Lasciva Roma (Clérice, 2020b, 2021a) and DigilibLT (Lana, 2021; Clérice et al., 2021). The lemmatization of this corpus and its pre-processing was applied with Pie-Extended's LASLA model (Manjavacas et al., 2019; Clérice, 2020c, 2021b) which has a 97.34% accuracy. Overall, the aggregated corpus spans from 254 BCE to 799 CE with 21,222,911 tokens - including punctuation, after tokenization and lemmatization - over 853 works[6], composite or not (some being collections of works of multiple authors, sometimes multiple unidentified ones, such as the *Anthologia Latina*). Each source corpus was encoded – at least partially – with CapiTainS guidelines (Clérice, 2017), allowing for machine-actionable segmentation in TEI documents and as such allowing us to retrieve or segment whole works with their editorial segmentation, such as poems, lines, books, etc.

We then divide the main corpus in three sub-corpora (*cf.* Figure 2):

1. The first corpus contains only Martial's *Epigrammata* and the anonymous *Priapea* (*Martial & Priapea* hereafter): they form a very small corpus ($|t| = 61,082$), with shared topics and vocabulary (they are full of sexual and obscene words) and are heavily composite. In fact, the first is separated in three levels (book,poem,line), the latter in only two levels (poem,line). Both text provide ATU levels (poems).

2. The second (*ATU Corpus* hereafter) consists

[6] 17,804,769, excluding punctuation.

of all works in which there is a unit level qualified as "poem", "comment" or "scholia", "letter", "speech" and "entry" (found dictionaries-like works or collection of recipes): this amounts to 125 works and 3,549,249 tokens. They do not share any specific topical unity but are massively consisting of poetry. It also is a superset of *Martial & Priapea*.

3. The last consists of the full corpus with its 17,639,626 tokens (*All*) after removing punctuation and other foreign tokens.

Each corpus displays a different value of the combined property "Number of (S)ATU-Corpus size" (*cf.* Table 1): the first is a very small corpus with a high number of autonomous textual units, the second has a high number of ATU but reaches a bigger number of tokens (nearing 1.5 million) that might be better at dealing with noise while the last is a mixture massively built of long passages in small amount per work with the exception of its contained sub-corpus *ATU Corpus* and some other texts that could not be easily fit into the latter automatically.

For each corpus, in order to quantify the effect of segmentation or its absence, each text was assigned a level at which they should be split allowing (S)ATU to be treated as non-sequential units: for texts which do not show any (S)ATU, the full text was kept as a single unit. We produce two versions of our corpora: the first one where (S)ATU are used to prevent window from overreaching, which we call segmented corpus $S(T)$, and the second where each file is treated as a single continuous unit of information, called unsegmented corpus $U(T)$.

2.3 Semantic analysis and experimental set-up

As the main objective is to analyze the impact of text segmentation on semantic analysis, we set up the experiment based on four parameters which we then combine to produce analysis using both versions of T and compare their results.

The first parameter set is composed of four groups of words we want to analyze in the corpus. These sets of words, which we call pivots, provide different distribution depending on the corpora and are composed of words appearing at least 10 times:

1. The first one, *Puer et al.*, contains words related to people and is not specific to any of the corpora. These are *dominus, mater, pater, puella, puer, uir, uxor*. They span from 2,036 occurrences (*puella*) to 48,519 (*dominus*) in *All*.

2. The second, *Carmen et al.*, might be more specific to grammarians and poetry, which are overrepresented in *ATU Corpus*. It contains *scribo, poeta, libellus, lego2, carmen1, liber1* (*cf.* Table 2)[7].

3. The third, *Puer, Carmen et al.*, is a combination of the first two and offers as such two clearly separated semantic subgroups which should be easy to cluster when time comes.

4. The last, *Futuo, Carmen, Puer et al.*, is a combination of the first two as well as crude words and words which are connected to sexuality: for some, they are heavily specific to *Martial and Priapea*, have a very low frequency compared to the first group, but are also somewhat specific to *ATU Corpus*. They are *cunnus, fello, futuo, irrumo, lasciuus, mentula, paedico2* (*cf.* Table 2 for each word frequency depending on the corpus).

5. In order to evaluate noise before analysis, we also consider a fifth word-set made of the 1000 most frequent words of the *All* dataset.

A second parameter is the size of the window, written W. To analyze words, we will only retrieve words occurring in this window. We make this window vary between four values ($[5, 10, 15, 20]$).

A third parameter is the floor-threshold frequency, noted F thereafter. Lemma co-occurring with our pivots will only be considered if they occur at least F time in the corpus of windows: if `lemma1` appears $F - 5$ times with `pivot1` and 5 times with `pivot2`, it is kept as a feature. F varies within $[1, 5, 10, 20]$ which should provide situations less prone to noise: unique co-occurrences will be ignored when $F > 1$ for example, and less important lemma will follow as we raise the value.

The following workflow is then applied to the first four word sets[8] using all combinations of the W and F for a total of 16 different results per version of the corpus:

[7] When lemmas are ending with numbers such as *lego2*, it represents a disambiguation index: in Latin, the first person of indicative present is often used to represent lemma, but two verbs share the *lego* form: one is conjugated *legis* at the second person (meaning: read, *lego2*) while the other becomes *legas* (meaning: name, *lego1*)

[8] *Top1000* is not used in the whole experiment, see below.

	(S)ATU		Tokens		Texts		Distribution of Tokens / ATU						
	Count	%	Count	%	Count	%	Mean	Std	Min	25%	50%	75%	max
Martial & Priapea	1607	2.8	61,056	0.3	2	0.2	38	32	7	13	27	54	280
ATU Corpus	39,591	68.5	3,549,249	20.1	125	14.8	90	603	1	9	17	36	47,783
All	57,761	100.0	17,639,626	100.0	845	100.0	305	2,159	1	11	24	77	248,564

Table 1: Properties of the different corpora. Punctuation and foreign tokens are ignored in the token count.

1. We retrieve and store the co-occurring count in a matrix where co-occurring words constitute features (columns) and pivot classes (lines), with their number of retrieval as values. We use the output of this retrieval in section 3.1 to analyze raw variation between $U(T)$ and $S(T)$.

2. Following the work of Evert (Evert, 2005), A. Guerreau (Morsel, 2015) and N. Perreaux (Perreaux and rey, 2013), we apply a normalization algorithm called *Dice* coefficient.

3. For each pivot, we keep their 5 most correlated features (retrieved lemma in the window). If the score of the fifth word is shared by multiple words, we keep all of them. They constitute a second set of words we call *major co-occurrences* (M).

4. We retrieve and augment the original matrix in step 1 with the same retrieval and store strategy for each word in M. We use M in section 3.2 to study the impact, post-normalization, of this first step of analysis.

5. We normalize again the output with Dice coefficient: the final output here constitutes our analysis input.

This approach using bag-of-words and normalization is preferred in the context of our experiment to deep learning approaches such as Word2Vec. Given their instability in "small" corpora ($\tilde{2}$0 million)(Antoniak and Mimno, 2018) and the risk of not controlling perfectly the randomizing seed at the library (*e.g.*, *Gensim*) or Python level, we preferred non-random approaches, as any variation due to randomization, including the order in which texts are seen, might affect the results and hide the hypothetical window noise in its own randomness.

Once we have a matrix with co-occurrences count, we can then perform an analysis: while performing Dice is a first step of post-processing, we want to see how the data would react to traditional means of analysis. We preferred in this context to use Ward agglomerative clustering using Euclidean distance on pivots and major co-occurrences. In order to have the ability to compare the output of the clustering on $U(T)$ and $S(T)$, we harmonize major co-occurrences by stripping the ones which are not shared between analysis running with the same parameters over both versions of T. It produces a set of $|C|$ common classes which can finally be clustered according to a fifth parameter k, where k is the number of clusters we want to obtain. We make k vary so that $\frac{|C|}{k} < 2$ and $5 \geq k < 15$. As studying the variation is the objective of this paper, the number of clusters does not need to be fine-tuned, as we are only interested in the equality or inequality between the analysis of $U(T)$ and $S(T)$, thus varying k within a dynamic range. The output of this final step is finally studied in section 3.3.

3 Evaluation of impact

Once all combinations have been run, we want to evaluate three different kinds of differences or effects: a first raw effect on lemma co-occurrences and how the raw matrices would differ without any normalization step, the second effect on the selection of secondary classes (major co-occurrences) and finally the effect on more advanced analysis, here using clustering, through the evolution of features.

3.1 On the Co-occurrence Matrix

In order to quantify the impact on neighborhood retrieval, we propose to first analyze the impact on the most frequent words of the corpus that are either adverbs, adjectives, pronouns, nouns or verbs. Then, for each corpus, we run the step 1 described above for each combination of W and F. We compute for each lemma the Manhattan distance between its vector in the result matrices of both $S(T)$ and $U(T)$ where each absent feature is replaced by a column filled with 0.

While some lemmas do not show any variation between versions of T, a vast majority of them displays non-null distances as seen in figure 3: most of W, F, Corpus combinations have their 5% per-

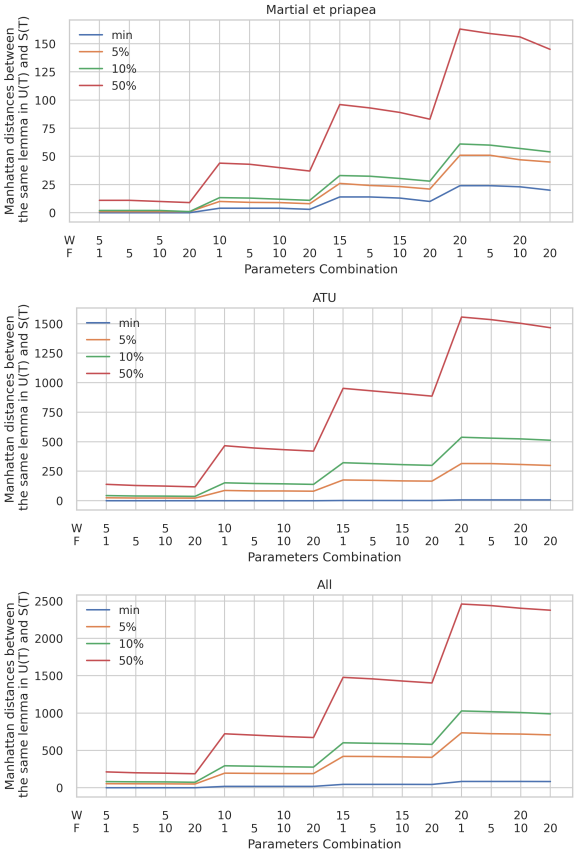

Figure 3: Variation of the distances based on W, F for each lemma's vector for the 1000 most frequent words in the full corpus, given three percentiles (5%, 10% and median) and the minimum value.

centile of distance which is not null. While each increase of W resulted in higher distances, the expected filtering effect of F also works on the noise: by ignoring less frequent co-occurring lemma, the increase of the value of F effectively lowers the distance, albeit very minimally. Indeed, F is the parameter that has the least impact on the computed distances.

3.2 On Classes

Based on this first observation, we want to evaluate what this noise can do to more advanced feature selections. To compare the effect on these major co-occurences' selection, we simply compare for each combination of parameters Word-set, W, F the set M of $U(T)$ with the set of $S(T)$. Given the differences of distances found in 3.1, this will show whether the noise accumulated through noisy windows is enough to influence the simple scoring provided by the Dice coefficient.

We first quantify the effect of a binary approach, *i.e.*, we check that $M_{S(T)}$ and $M_{U(T)}$ are equal.

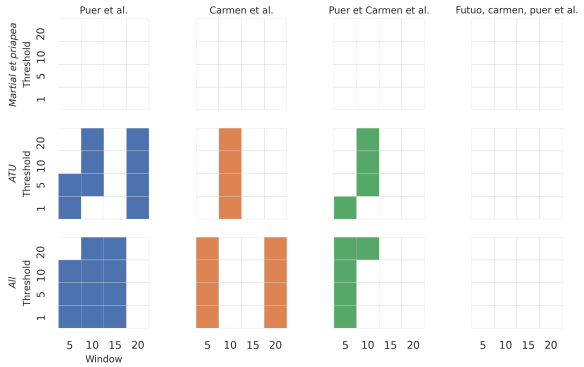

Figure 4: Binary matrix of $M_{S(T)} = M_{U(T)}$: colored cells mean the analysis on both versions of the corpus with the same parameters resulted in the same major co-occurrences.

Any situation where $M_{S(T)} \neq M_{U(T)}$ is a first proof that the absence of segmentation has an impact. In fact, with this approach, only 21.35% (41/192) of the run reaches perfect equality of their M for both versions of T, with varying results depending on the word-set and the corpus(*cf.* Figure 4):

- Overall, the word-set "Futuo et al." never results in fully similar M sets. This is probably due to very low frequencies of some of its members and would make the case, if such frequencies are acceptable from a statistical point of view, to very carefully segment the analyzed texts.

- As expected, any analysis using the corpus *Martial & Priapea* is heavily affected by its high amount of small ATU: none of them have similar output over its two versions. This corpus' size does not produce noise mitigation.

- Only the "Puer et al." regularly achieves equality over the corpus *ATU Corpus*, but it is not constant. This simply would advocate for segmentation of rich (S)ATU corpora as a prerequisite for analysis similar to the one we run here. It is also possible that the rather low frequency of "libellus" (546) in the *Carmen et al.* and *Puer et Carmen et al.* is responsible for some of the instability.

- Higher token counts do smooth the noise of features as the *All* corpora displays a higher stability between $M_{S(T)}$ and $M_{U(T)}$, specifically for *Puer et al.*, but M are still more often different than equal.

- The frequency threshold of co-occurrences F has a very small impact on major co-occurrences, while the window is irregularly affecting word sets: as an example, $W = 15$ never reaches equality for the combination *Carmen et al. + All* but it does for analysis *Puer et al.*; on the contrary, $W = 15$ fails on *Puer et al. + ATU Corpus*. This instability of the impact of W also advocates for relying on (S)ATU when doing such analysis.

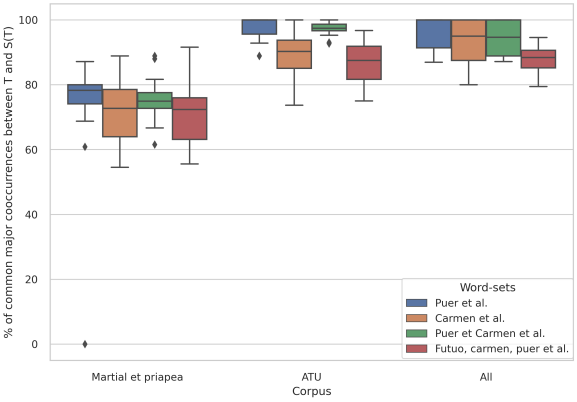

Figure 5: Dispersion of the overlap percentage or *retention rate* of M for each corpus and word-set combination. For *Puer et al.* on Martial et Priapea, major co-occurrences have a median similarity of below 80% over the 16 combinations of W, F.

For the relative rate and a more in-depth analysis of how classes vary from one version of the corpora to another, we propose to evaluate the *retention rate of classes* as a function of both sets M of major co-occurrences, computed:

$$1 - \frac{\left|M_{S(T)} - M_T\right| + \left|M_T - M_{S(T)}\right|}{\left|M_{S(T)} + M_T\right|}$$

With some exceptions on *Martial & Priapea* and *ATU Corpus*, the retention rate is generally over 60% for the smallest corpus, 80% for the two other[9] for a median number of major co-occurrences spanning from 8 (*Carmen et al.*) to 34 (*Futuo, carmen, puer et al.)* (*cf.* figure 6). However large the corpus and the word-set, the retention rate is globally high, particularly for the two biggest corpora, with one word-set being worse than the others (*Futuo et al.*).

Overall both metrics show an undisputed effect on major co-occurrences' selection. Corpus growth mitigates this effect as shown with the results of

[9]See appendix table 3

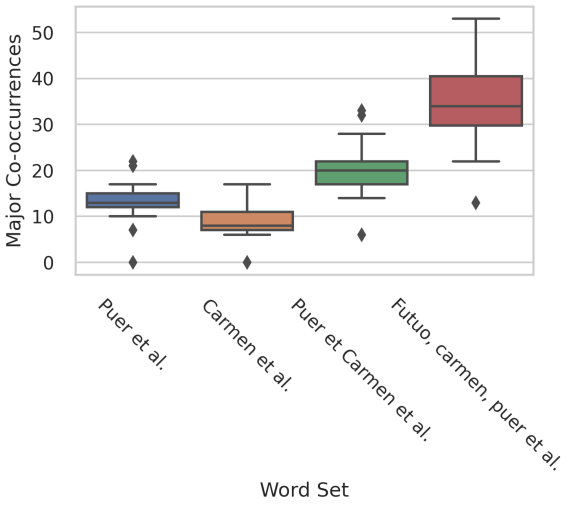

Figure 6: Number of Major Co-occurrences per Word Set

All vs. others, but it does not warranty equality of results between $S(T)$ and $U(T)$, as shown for $W = 20$ on *Puer et al.* or $W = 10$ and $W = 15$ for *Carmen et al.*. Carefully handling (S)ATU in relatively small corpora ($\leq 20M$ tokens) has to be an important step to strengthen any semantic statistical analysis.

3.3 On Features

Based on this second output, we want to identify if more advanced algorithms were as subject to variation with this normalized input. For each combination of W, F, which have the same clusters K such as $K_{S(T)} = K_T$ (*cf. figure 7*).

In this context, features have an impact that moves beyond the simple selection of classes, specifically for our first two corpora: on *Martial & Priapea*, for any word-set, there are no situations where all combinations of F, W provide the same clusters except for $k = 11$ and *Carmen et al.*, while none reaches the same clusters within the *ATU* corpus. In general, most combinations provide below 40% similar clustering for the first corpus and below 80% for the second. Similarly to the classes analysis, the size of the corpus mitigates the effect of $S(T)$ vs. $U(T)$: *All* has the biggest number of clusters which are equal in between both versions of the corpus. It, however, is still unstable and is unpredictable: while *Puer et al.* reaches 100% equality in clustering between $S(T)$ and $U(T)$ for $K = 5$ and $K = 10$, it falls down to 70% of similar results ariybd $K = 7$ and $K = 8$.

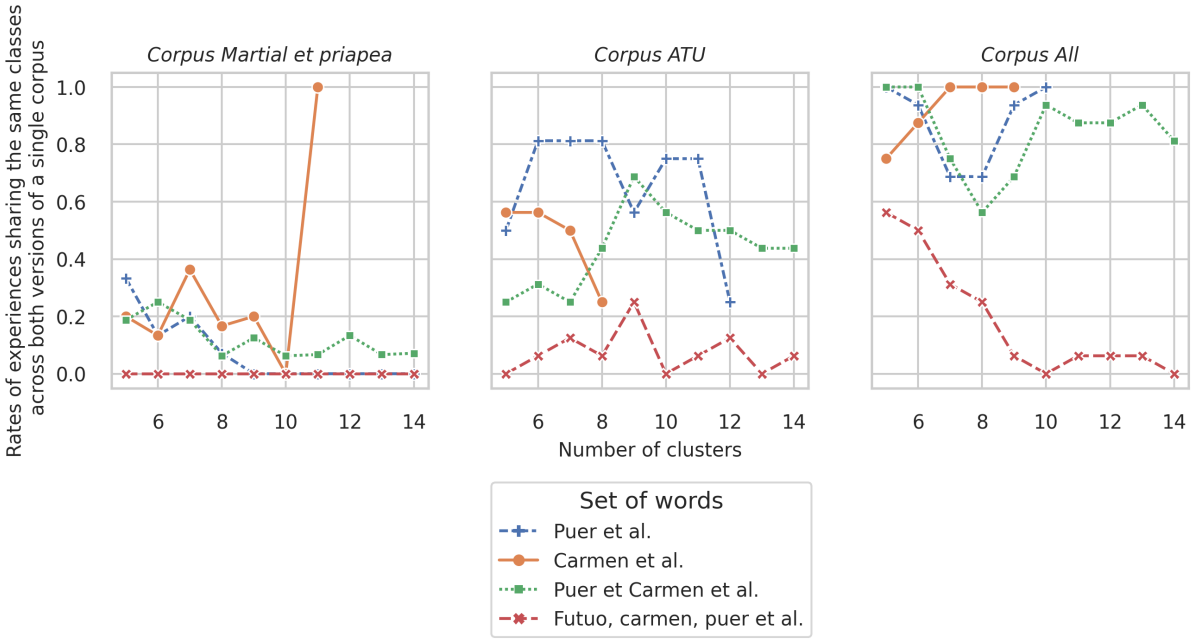

Figure 7: Ratio of experiments where $K_{S(T)} = K_T$ given the number of clusters for each corpus and word-set

4 Conclusion

When J. R. Firth used "company", which in our experiment becomes neighborhood and windows, what is meant is definitely more than having simply two words that follow each other: should two tweets in a timeline be treated as a single unit just because they appear in the same HTML "context"? The answer should be clear to any linguist, yet, texts and composite works have been treated this way by studies in DH or NLP[10].

In this experiment, we evaluated the impact of not taking into account the composition of texts. While working with statistics on such small corpora inherently requires caution, we demonstrated that treating digitized works as single cohesive units rather than as a patchwork of smaller units is altering the results of distributional analysis with corpora around 15 million tokens. In our experiments, only few combinations of the various parameters common to these applications (window size, frequency threshold, cluster size) yielded consistent results between an edited corpus of texts ($S(T)$) and a raw version of it ($U(T)$). With frequent words accross the corpus, with a small window ($W = 5$) and large corpora, the effect of noise is mitigated. But, if these words are more frequent in highly segmented works such as poetry compilation, the size of the overall corpus will have less impact on the issue.

These results do strengthen the necessity of metadata-enriched texts which allow for post-processing such as the one allowed by CapiTainS and the original XML TEI environment. It does definitely advocate for using declarations of segmentation with metadata such as CiteStructure (Cayless and Clérice, 2020) in order to make these corpora usable in machine actionable ways.

The current work was voluntarly limited in scope to both the Latin language and the use of deterministic methods. The Latin corpus is limited in size and does not provide a testing field for bigger corpora. Any experiment on larger corpus will have to deal with the annotation of larger corpora for segmentation purposes. The use of deterministic methods to represent clusters or distances between words allowed us for an easy and reproducible experiment. Applying the same approach to non-deterministic methods such as the one found in Word2Vec (Mikolov et al., 2013) and evaluating these results would provide a second testing field. However, the size of the corpus might already be a constraint difficult to overcome according to (Antoniak and Mimno, 2018).

Acknowledgments

We want to thank Florian Cafiero, Jean-Baptiste Camps, Marie Puren and Simon Gabay for their feedback on this article.

[10] See (Köntges, 2020) for example.

References

Maria Antoniak and David Mimno. 2018. Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, 6(0):107–119.

Jelke Bloem, Maria Chiara Parisi, Martin Reynaert, Yvette Oortwijn, and Arianna Betti. 2020. Distributional semantics for neo-Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 84–93, Marseille, France. European Language Resources Association (ELRA).

Bruno Bureau. 2012. Quelques réflexions sur la notion de littérarité à partir de l'édition numérique de commentateurs anciens. *Interférences. Ars scribendi*, (6).

Luciano Canfora. 2016. *Conservazione e perdita dei classici*. Stilo editrice, Bari, Italie.

Catulle and Léon Herrmann. 1957. *Les deux livres de Catulle*. Latomus Revue d'Etudes Latines, Bruxelles (Berchem), Belgique. ISSN: 1378-8760.

Catulle and Georges Lafaye. 1932. *Poésies*. les Belles Lettres, Paris, France. ISSN: 0184-7155.

Hugh Cayless and Thibault Clérice. 2020. FR: Declarative Citation Structure · Issue #1957 · TEIC/TEI.

Thibault Clérice. 2017. Les outils CapiTainS, l'édition numérique et l'exploitation des textes. *Médiévales. Langues, Textes, Histoire*, 73(73):115–131.

Thibault Clérice. 2020a. Corpus latin antiquité et antiquité tardive lemmatisé.

Thibault Clérice. 2020b. Lasciva roma, priapea.

Thibault Clérice. 2020c. Pie extended, an extension for pie with pre-processing and post-processing.

Thibault Clérice. 2021a. Lasciva roma, additional texts.

Thibault Clérice. 2021b. Latin lasla model.

Thibault Clérice, Hippolyte Souvay, Etienne Ferrandi, Vincent Giovannangeli, Akim Ouchen, Léa Maronet, Émilien Arnaud, Krister Kruusmaa, and Jean Barré. 2021. lascivaroma/digiliblt: Release 0.0.64.

Gregory R. Crane. 2021a. Open greek and latin, corpus scriptorum ecclesiasticorum latinorum.

Gregory R. Crane. 2021b. Perseusdl/canonical-latinlit 0.0.752.

Maciej Eder. 2013a. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, page fqt066.

Maciej Eder. 2013b. Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4):603–614.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

John Rupert Firth. 1957. *Papers in linguistics, 1934-1951*. Oxford University Press.

Lynn S. Fotheringham. 2007. The Numbers in the Margins and the Structure of Cicero's "Pro Murena". *Greece & Rome*, 54(1):40–60. Publisher: Cambridge University Press.

Alain Guerreau. 1989. Pourquoi (et comment) l'historien doit-il compter les mots ? *Histoire & Mesure*, 4(1):81–105. Publisher: Persée - Portail des revues scientifiques en SHS.

Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, volume 2/3, page 389-398, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Thomas Köntges. 2020. Measuring Philosophy in the First Thousand Years of Greek Literature. *Digital Classics Online*, pages 1–23.

Maurizio Lana. 2021. Metodologie e problematiche per una biblioteca digitale. il caso di digiliblt.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Joseph Morsel. 2015. Quelques propositions pour l'étude de la noblesse européenne à la fin du Moyen Âge. In *Discurso, memoria y representación : la nobleza peninsular en la Baja Edad Media (XLII Semana de Estudios Medievales de Estella)*, Discurso, memoria y representación : la nobleza peninsular en la Baja Edad Media (Actas de la XLII Semana de Estudios Medievales de Estella, 21 al 24 de julio 2015), pages 449–499, Estella, Spain. Gobierno de Navarra, Pamplona, Gobierno de Navarra, Pamplona.

Matthew Munson. 2017. *Biblical Semantics: Applying Digital Methods for Semantic Information Extraction to Current Problems in New Testament Studies*, 1 edition. Shaker, Aachen.

Nicolas Perreaux. 2012. Mesurer un système de représentation ? Approche statistique du champ lexical de l'eau dans la Patrologie Latine. In *Mesure et histoire médiévale*, pages 365–374, Tours, France. Publications de la Sorbonne.

Nicolas Perreaux and coraline rey. 2013. CBMA. Chartae Burgundiae Medii Aevi VII.“ Le “vocabulaire courant” en diplomatique : techniques et approches comparées ”. *Bulletin du Centre d'études médiévales d'Auxerre*, (17.1).

Martina Astric Roda, Philomen Probert, and Barbara McGillivray. 2019. Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique des Langues*, 60(3/2019):63–87.

Christof Schöch. 2017. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, 011(2).

Stéfan Sinclair and Geoffrey Rockwell. 2016. Voyant Tools.

Nathan Stringham and Mike Izbicki. 2020. Evaluating Word Embeddings on Low-Resource Languages. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 176–186, Online. Association for Computational Linguistics.

Radim Řehůřek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

A Online Resources

The code and output of this article can be found at https://github.com/lascivaroma/dont-worry-its-just-noise .

B Appendices

	Martial & priapea	ATU Corpus	All	Puer et al.	Carmen et al.	Puer et Carmen et al.	4 Fields
cunnus	33	42	43				✓
fello	11	14	28				✓
futuo	46	52	52				✓
irrumo	10	16	16				✓
lasciuus	35	155	300				✓
mentula	68	75	75				✓
paedico2	17	20	22				✓
carmen1	90	1753	3101		✓	✓	✓
lego2	95	3030	10252		✓	✓	✓
libellus	119	546	1190		✓	✓	✓
liber1	36	1773	21015		✓	✓	✓
poeta	53	1366	2944		✓	✓	✓
scribo	71	6282	20501		✓	✓	✓
dominus	112	7420	48519	✓		✓	✓
mater	43	2132	9271	✓		✓	✓
pater	73	5154	29927	✓		✓	✓
puella	120	989	2036	✓		✓	✓
puer	159	1812	5824	✓		✓	✓
uir	94	3908	20062	✓		✓	✓
uxor	63	1306	7832	✓		✓	✓

Table 2: Frequency distribution over the 3 corpora

Corpus	Word-Set	Experiments	Common classes mean	std	min	25%	50%	75%	max
Martial et priapea	Puer et al.	16.0	72.9	20.5	0.0	74.1	78.3	80.0	87.2
	Carmen et al.	15.0	71.8	11.6	54.5	64.0	72.7	78.6	88.9
	Puer et Carmen et al.	16.0	75.7	6.9	61.5	72.7	74.9	77.6	88.9
	Futuo, carmen, puer et al.	16.0	71.3	9.7	55.6	63.2	72.4	75.9	91.6
ATU	Puer et al.	16.0	96.9	4.4	88.9	95.6	100.0	100.0	100.0
	Carmen et al.	16.0	88.6	9.8	73.7	85.1	90.3	93.8	100.0
	Puer et Carmen et al.	16.0	97.3	2.3	92.7	96.7	97.4	98.7	100.0
	Futuo, carmen, puer et al.	16.0	86.6	6.7	75.0	81.7	87.5	91.9	96.7
All	Puer et al.	16.0	96.3	5.8	87.0	91.4	100.0	100.0	100.0
	Carmen et al.	16.0	93.5	7.1	80.0	87.5	95.0	100.0	100.0
	Puer et Carmen et al.	16.0	94.3	5.1	87.2	88.9	94.7	100.0	100.0
	Futuo, carmen, puer et al.	16.0	87.9	4.4	79.5	85.2	88.4	90.6	94.5

Table 3: Ratio of common classes over experiments