

Glottometrics

International Quantitative Linguistics Association

55/2023

Glottometrics is an open access scientific journal for the quantitative research of language and text published twice a year by International Quantitative Linguistics Association (IQLA). The journal was established in 2001 by a pioneer of Quantitative Linguistics Gabriel Altmann.

Manuscripts can be submitted by email to glottometrics@gmail.com. Submission guideline is available at <https://glottometrics.iqla.org/>.

Editors-in-Chief

Radek Čech • University of Ostrava, Masaryk University (Czech Republic)

Ján Mačutek • Mathematical Institute of the Slovak Academy of Sciences,
Constantine the Philosopher University in Nitra (Slovakia)

Editors

Xinying Chen • University of Ostrava (Czech Republic)

Ramon Ferrer-i-Cancho • Polytechnic University of Catalonia (Spain)

Miroslav Kubát • University of Ostrava (Czech Republic)

Haitao Liu • Zhejiang University (China)

George Mikros • Hamad Bin Khalifa University (Qatar)

Petr Plecháč • Institute of Czech Literature of the Czech Academy of Sciences (Czech Republic)

Andrij Rovenchak • Ivan Franko National University of Lviv (Ukraine)

Arjuna Tuzzi • University of Padova (Italy)

International Quantitative Linguistics Association (IQLA)

Friedmangasse 50
1160 Vienna
Austria

eISSN 2625-8226

Contents

Menzerath's law: Is it just regression toward the mean?	1-16
Jiří Milička	
Applying Distributional Semantic Models to a Historical Corpus of a Highly Inflected Language: the Case of Ancient Greek	17-43
Alek Keersmaekers, Dirk Speelman	
Active or descriptive: Textual activity and its dynamic changes of Ph.D. theses across disciplines	44-58
Shuyi Amelia Sun, Wei Xiao	
Swap distance minimization in SOV languages. Cognitive and mathematical foundations.	59-88
Ramon Ferrer-i-Cancho, Savithry Namboodiripad	
Modality of verbs as stylometric feature in Czech genres	89-98
Miroslav Kubát, Xinying Chen	

Menzerath's law: Is it just regression toward the mean?

Jiří Milička^{1*} 

¹ Institute of the Czech National Corpus, Charles University

* Corresponding author's email: jiri@milicka.cz

DOI: https://doi.org/10.53482/2023_55_409

ABSTRACT

The study revisits the Menzerath's Law, which articulates the inverse relationship between the length of constructs and the mean length of their constituents. This relationship is famously modelled by Gabriel Altmann's model, which combines power and exponential relations. His formulas have been widely used to describe this relationship across linguistics and biology, however, there is no satisfactory explanation for his model. Therefore, the paper proposes shifting our perspective to examine directly the relationship between the number of constituents in a construct and the number of subconstituents in the same construct. This relationship may be explained by a simple model based on linear regression, which leads to a hyperbolic model of the Menzerath's Law. This approach is successful for several datasets, but insufficient for others.

Menzerath's Law, Menzerath-Altmann Law, MAL, regression to mean, linear model

1 Introduction

Menzerath's Law describes the relationship between the length of text segments and the mean length of their subsegments (i.e. constructs and their constituents), a principle that applies to various levels and has been confirmed in numerous languages. The law is named after Paul Menzerath, who was the first to notice the peculiar relation between the length of a syllable and its duration,¹ as well as between the length of a word and the mean length of its syllables.²

The law is also known as the Menzerath-Altmann Law (MAL), in honor of Gabriel Altmann. Altmann developed models of this relationship, popularized the concept among quantitative linguists, and most significantly, recognized that the model applied to more than just syllables and phonemes. He described the generalized form as “the longer a language construct the shorter its components (constituents)” (Altmann, 1980, p. 1).

¹“... a sound is the shorter the longer the whole in which it occurs” (Menzerath, 1928, p. 104), as translated in Altmann (1980, p. 1). Also formulated as: “... the more sounds in a syllable the smaller its relative length” (Menzerath, 1928, p. 104), translation by Altmann (1980, p. 1).

²“The relative number of sounds in the syllable decreases as the number of syllables in the word increases, or said differently: the more syllables in a word, the shorter (relatively) it is” (Menzerath, 1954, p. 100), translation by Altmann (1980, p. 1).

To be more specific, the relationship is between the number of constituents in a construct and the *mean* number of subconstituents within these constituents. Altmann's model (1980, p. 3) expresses this relationship by the equation

$$(1) \quad \bar{L}_{n-1,n-2} = aL_{n,n-1}^b e^{cL_{n,n-1}}.$$

In this equation, n refers to the level of constructs (e.g. words), while $n-1$ denotes the level of constituents (e.g. syllables), and $n-2$ denotes the level of subconstituents (e.g. phonemes). Hence the term $L_{n,n-1}$ represents the length of the construct in terms of its constituents (e.g. the number of syllables in a word). Meanwhile, $\bar{L}_{n-1,n-2}$ represents the mean length of the constituent in terms of its subconstituents, for example the mean number of phonemes in a syllable.

Many studies have found that when assuming $c = 0$, the abbreviated form of the model provides a satisfactory fit for the data:

$$(2) \quad \bar{L}_{n-1,n-2} = aL_{n,n-1}^b.$$

To give an example, the following formula describes the Menzerath-Altmann law on phoneme-syllable-word level:

$$(3) \quad \bar{L}_{\text{syllable,phoneme}} = aL_{\text{word,syllable}}^b.$$

It has been discovered that the model can be applied to almost any conceivable method of text segmentation — phonemes, morphemes (Gerlach, 1982; Milička, 2014; Pelegrinová et al., 2021), words, phrases (Mačutek et al., 2021; Mačutek et al., 2017), clauses (Buk and Rovenchak, 2008) and sentences (Milička, 2015; Motalová, 2022), but also to non-human language — geladas (Gustison et al., 2016; Semple et al., 2022) and there were also various attempts at applying MAL outside linguistics, mostly biology (Altmann, 2014; Altmann and Schwibbe, 1989; Semple et al., 2022).

Since real-world datasets of these relationships tend to be noisy, numerous models, not just Altmann's, fit the empirical data. However, very few models have actually been tested, and even those that have been published, typically bear some connection to the original Altmann models (Buk and Rovenchak, 2007; Kuřacka and Mačutek, 2007; Mačutek and Rovenchak, 2011). A hyperbolic model (4, Figure 1) has successful fit to many datasets on various levels of segmentation and languages (Milička, 2014):

$$(4) \quad \bar{L}_{n-1,n-2} = \frac{a}{L_{n,n-1}} + b.$$

Actually, already Menzerath himself modelled the relation as a hyperbolic one, however, it is a bit obfuscated. In his book from 1954 he does not analyze the MAL relationship, but the relationship of number of syllables in a word and mean number of phonemes in word. I.e. the dependent variable is not

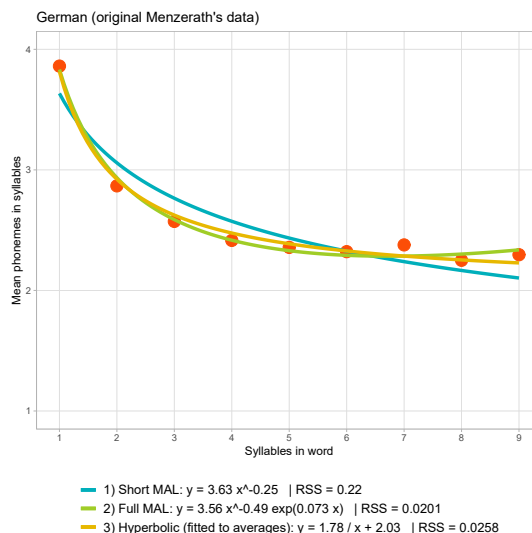


Figure 1: Three models of Menzerath’s relation fitted to the original Menzerath’s data on the phoneme-syllable-word level (Menzerath, 1954, p. 96). Residual sum of squares is reported as it scales with the main objective of the fitting function.

$\bar{L}_{n-1,n-2}$ but $\bar{L}_{n,n-2}$, in this case it means that the dependent variable is mean number of phonemes in word instead in syllables.

But it does not matter, since the mean number of phonemes in a word can be calculated as the mean number of phonemes in syllable times number of syllables. To be more general, $\bar{L}_{n,n-2} = \bar{L}_{n-1,n-2}L_{n,n-1}$. This means we get the model for this relationship by multiplying both sides of the equation 4 by length of construct $L_{n,n-1}$. This multiplication makes the hyperbolic model linear:

$$(5) \quad \bar{L}_{n-1,n-2}L_{n,n-1} = \frac{aL_{n,n-1}}{L_{n,n-1}} + bL_{n,n-1}$$

$$\bar{L}_{n,n-2} = a + bL_{n,n-1}.$$

Figure 2 shows the actual Menzerath’s data (page 108): the empirical dataset looks fairly linear, so Menzerath used linear regression to model it.

The linear model can be interpreted easily and straightforwardly: as the length of a construct increases in terms of its constituents, its length in terms of subconstituents also increases at a steady and consistent rate. The more syllables a word has, the more phonemes it contains in proportion. This relationship seems to align with our intuition, except for the parameter a . Menzerath refers to this parameter as an *inexplicable additive constant (unarklärliche additive Konstante)* and feels that it requires an explanation (Menzerath, 1954, p. 111). In order to provide this explanation, Menzerath suggests that each word contains one *core syllable (Kernsilbe)*, which is longer than the other syllables in the word. For instance, monosyllabic words are composed of only the core syllable, which is why they are relatively long.

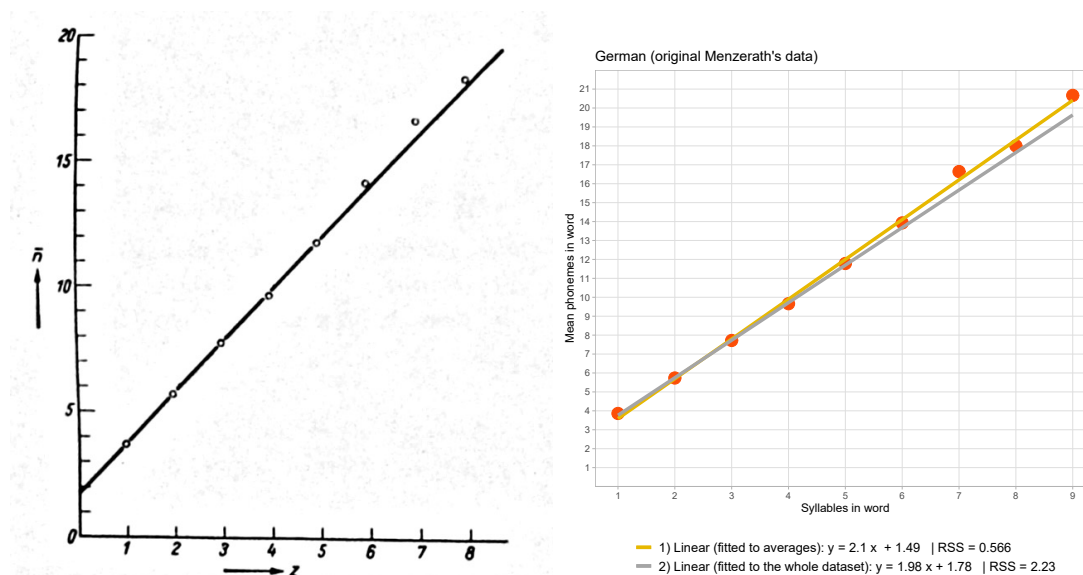


Figure 2: Original Menzerath's linear model (reprint from Menzerath (1954, p. 108)), right is recreation of the data points from his dataset (p. 96). His model $y = 2x + 1.9$ (ibid.) do not seem to match none of our linear models, presumably because he excluded the last data point. Residual sum of squares (RSS) is calculated in respect for averaged points.

Bisyllabic words, in contrast, consist of one core syllable and one ordinary syllable.

This explanation appears plausible and aligns with our experience, even when considering other units: there are core morphemes in words (root or base morphemes), and core words in clauses (the vast majority of clauses contain a predicate).³ If we really try, we would be able to find something like core clauses in sentences etc. . .

The paper by Milička (2014), which further develops the same formula, bases its explanation of the constant a upon Reinhard Köhler's idea of structure information. Köhler posits that this information is stored in constituents besides constructs (Köhler, 1984). A year later, a more generalized approach was presented in Milicka's PhD thesis (2015), in which the parameter a was examined from the perspective of the Theory of Communication.

However, it seems that no explanation is actually necessary in this instance, since the parameter a can be interpreted through the concept of Galtonian regression to the mean.

2 Paul Menzerath Meets Francis Galton

We are fortunate that Menzerath not only shared the means of the construct lengths but also provided the entire joint distribution, i.e. a table which states how many words of certain lengths he found. For

³Actually, predicates are typically short while they consist of high-frequency verbs. This observation aligns with the finding that at the syllable-word-clause level, Menzerath's law is inverted, showing an increasing function, as seen in Figure 7 of Wang and Chen, 2022, Figure 7.

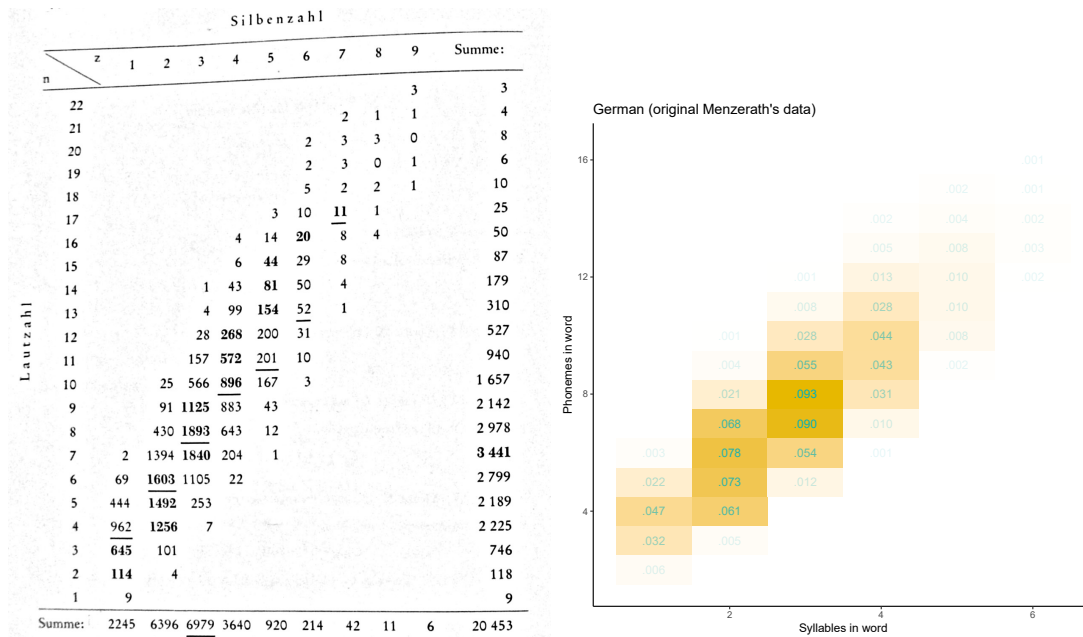


Figure 3: Original Menzerath's joint distribution of his dataset (reprint from Menzerath (1954, p. 96)). The chart on right represents the same data.

example he found 101 words that contain 2 syllables and 3 phonemes, 1893 words with 3 syllables and 8 phonemes etc., the complete distribution can be viewed in the table reprinted in Figure 3 (Menzerath, 1954, p. 96). These data make it possible to directly reanalyze his findings.

In order to get parameters of a linear model, the line is shifted and rotated until “discrepancy” between the line and the data points is as small as possible. There are several metrics of this “discrepancy”. The most favourite method for fitting is the *least squares* method, where the metric is sum of squared vertical distances between the line and the data points. Gabriel Altmann used this metric in his seminal paper on the topic (Altmann, 1980) and as far as I know everybody who fitted his model to Menzerathian relation did so.

In studies of Menzerath's law, the method of the least squares has traditionally been used to model the means, not the complete joint distribution (meaning the dataset as shown in the Figure 3), and I followed this tradition in the models presented so far in this study (Figures 1 and 2), with the exception of the yellow line in Figure 2, where the whole dataset was used. As can be observed, the two lines in the right chart of the Figure 2 are quite similar — it does not matter much, whether the model is fitted to the averages or to the whole dataset of joint distribution. This is because the least squares method inherently targets central values of the dependent variable.⁴

⁴By the way, this means we can use Galton's estimators to determine the parameters of the hyperbolic model for the Menzerath's law. To obtain these parameters, we need the correlation between the two variables as well as the mean and standard deviation of the marginal distributions. Consequently, the parameters of the hyperbolic model are straightforward to interpret.

Fitting the entire dataset with a linear model in this manner is useful for highlighting the regression toward the mean, a statistical artifact produced by averaging the values of the dependent variable (i.e., “vertically”). The line is actually called *regression* because of this. The regression toward the mean was famously discovered by Francis Galton who noticed that short people tend to have offsprings who are relatively taller than they are, and, surprisingly, also tall people have offsprings who are on average shorter than they are (Galton, 1886). This phenomenon actually resembles Paul Menzerath's observation that short words, when measured in syllables, do not appear as short when their length is counted by the number of phonemes. Conversely, words with a large number of syllables have relatively fewer phonemes on average. Such a phenomenon manifests whenever two variables are imperfectly correlated. The number of syllables in a word *is* imperfectly correlated with the number of phonemes in that word. The question is, whether the imperfect correlation can explain the parameter a completely.

Averaging the values, as we do in case of Menzerath's law, does not respect the way how the data points actually originated. We do not know which stochastic process best models the data's origin, but we can be sure that the random processes did not take place solely in the vertical direction. That is to say, it is not as if the independent variable was predetermined and all the “errors” can be attributed solely to the dependent variable. The dependent-independent dichotomy is in this case just a technical characteristic. We regard the number of phonemes in a word as being dependent on the number of syllables in the same word just for historical reasons, it is not as if some *Genius of the Language* first determined the word's length in syllables and then selected the appropriate number of phonemes to match it. The evolutionary process was presumably very chaotic and many phenomena had some effects on both variables. This situation actually mirrors Galton's data on height inheritance — there is some shared genetic material, which forms the basis for correlation, however, the actual heights of both the ancestor and their offspring are influenced by a multitude of other stochastic events, which affect both variables independently.

Therefore, let us fit the linear model using a method that accounts for errors in both vertical and horizontal directions, i.e. the method aiming to minimize the sum of squared distances between the line and the data points. We are interested in the Euclidean distance between the line and the data points. The metric is called *total least squares* (and the method is called *orthogonal fitting*).

Let us look at the difference between the blue and red lines in the [Figure 4](#). The blue line represents the classical least squares regression, while the red one shows the linear model fitted by minimizing the total least squares. The linear model represented by the blue line has a notably pronounced parameter a (commonly referred to as the intercept). This intercept nearly disappears when we fit the linear model orthogonally, diminishing to a value almost 30 times smaller. Suddenly, instead of two phonemes, the size is a mere 0.07 phonemes.

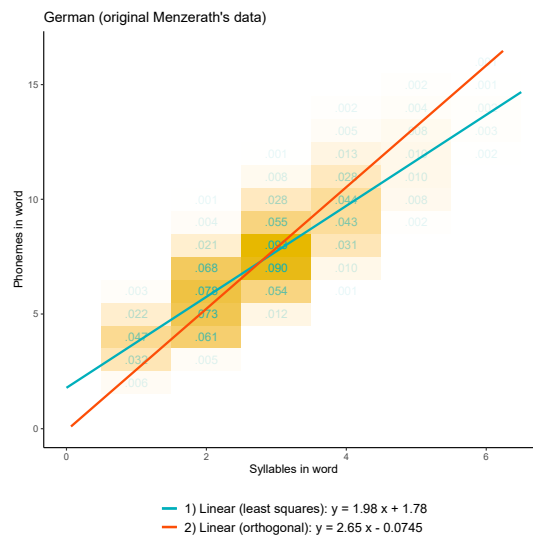


Figure 4: Linear model fitted to the original Menzerath's data (Menzerath, 1954, p. 96). Least squares fitting method and total least squares (orthogonal) method are put here in contrast.

While this result might be coincidental, we will dedicate the remainder of this study to empirically exploring this phenomenon across different texts and levels.

3 Material

It is still debatable whether the Menzerath's Law should be applied to tokens or types (Stave et al., 2021) and the difference between the two is quite pronounced (Mikros and Milička, 2014). Menzerath himself used data from a dictionary, indicating that his measurements were based on types. Gabriel Altmann (1980) also used dictionaries in his research on the topic, as did other pioneers in the field. However, many subsequent studies have applied the MAL directly to tokens. Since tokens cannot be considered independent trials, the statistical analysis and potential explanations are more complex than for types. Therefore, I prefer to use types, but for the sake of completeness, I will also present the results for tokens to illustrate the importance of this consideration.

Since we need to analyze the entire joint distribution, we can only use datasets where this distribution is available, e.g. Mikros and Milička (2014) and Milička (2014, 2015). Consequently, the number of tests for this hypothesis is limited; however, all the necessary scripts are available online so that the study can be replicated and repeated on other texts.⁵

⁵The archive is available at <http://milicka.cz/kestazeni/MenzerathRegression.zip>. The archive also includes the datasets that were used in the study. The first column in each table contains the actual forms of the given construct, for example a word tokens. The second column contains the number of its subconstituents, such as the number of phonemes. The third column contains the number of its constituents, such as the number of syllables. While the first column can be left empty, doing so will

4 Results

The first dataset used in this analysis allows for the examination of Menzerath's Law at the phoneme-syllable-word level for Greek blog posts, making it comparable to the original Menzerath's dataset discussed in previous sections. This dataset comes from Mikros and Milička (2014), although only a subset of the expansive dataset was employed. As illustrated in Figure 5, the difference between the joint distribution measured by types and tokens is relatively minor: In both cases, the intercept left by the orthogonal fitting is approximately one tenth of the parameter b , which is slightly higher than what was observed in the German data.

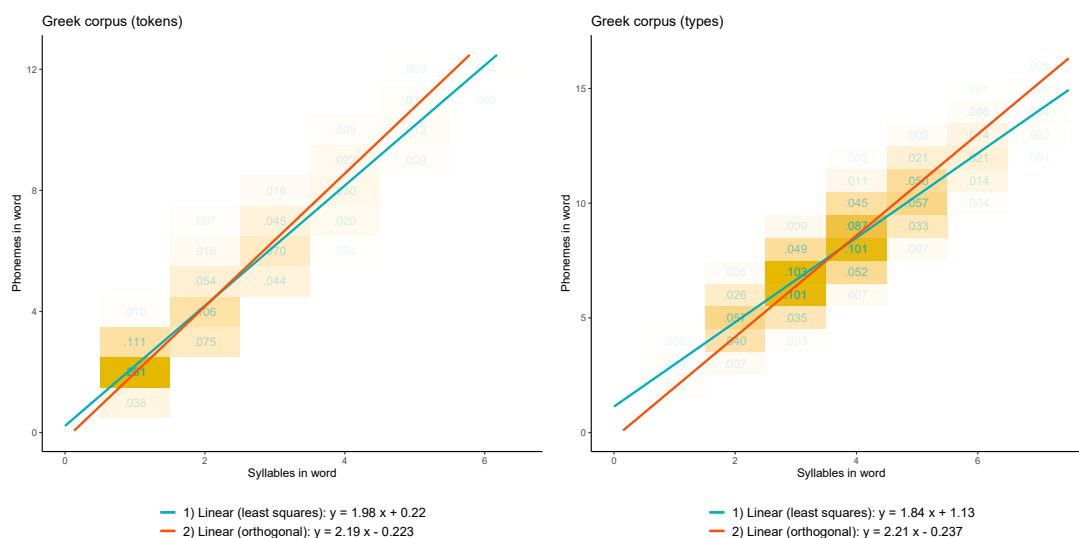


Figure 5: Phoneme — syllable — word level (Greek data, series of blog posts).

Menzerath's law can also be observed at the morpheme level, much like at the syllable level (Pelegrinová et al., 2021). Indeed, in many languages, morphemes typically equate to a single syllable. Therefore, I have incorporated several datasets that include the morpheme level, taken from the PhD dissertation by Milička (2015, Appendix C).

The first dataset allows for the exploration of the Menzerath's law at the phoneme-morpheme-word level in Czech text, specifically the novella *Krysař* by Viktor Dyk. The segmentation was done by Zuzana Komrsková and was initially published in Milička (2014). This dataset exemplifies that the hypothesis of zero intercept holds true for types rather than tokens. On tokens, the absolute intercept remains very large, but when looking at types, the absolute intercept is extremely small, even smaller than in the original Menzerath's dataset (Figure 6).

Let us stay at the phoneme-morpheme-word level. The next dataset is also taken from Milička (2014) cause the scripts to operate solely on tokens rather than on types.

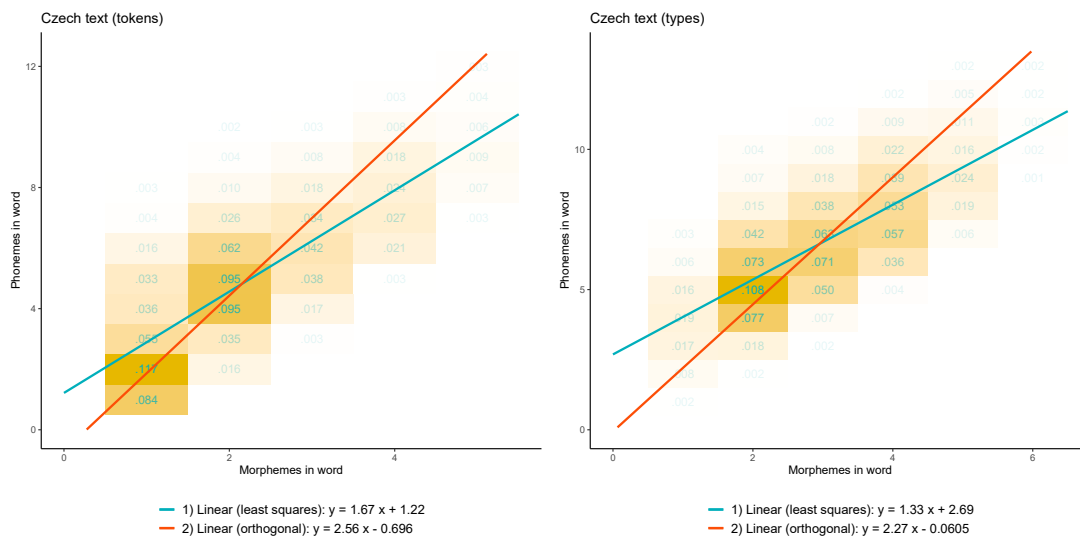


Figure 6: Phoneme — morpheme — word level (Czech data, short novel *Krysář* by Viktor Dyk).

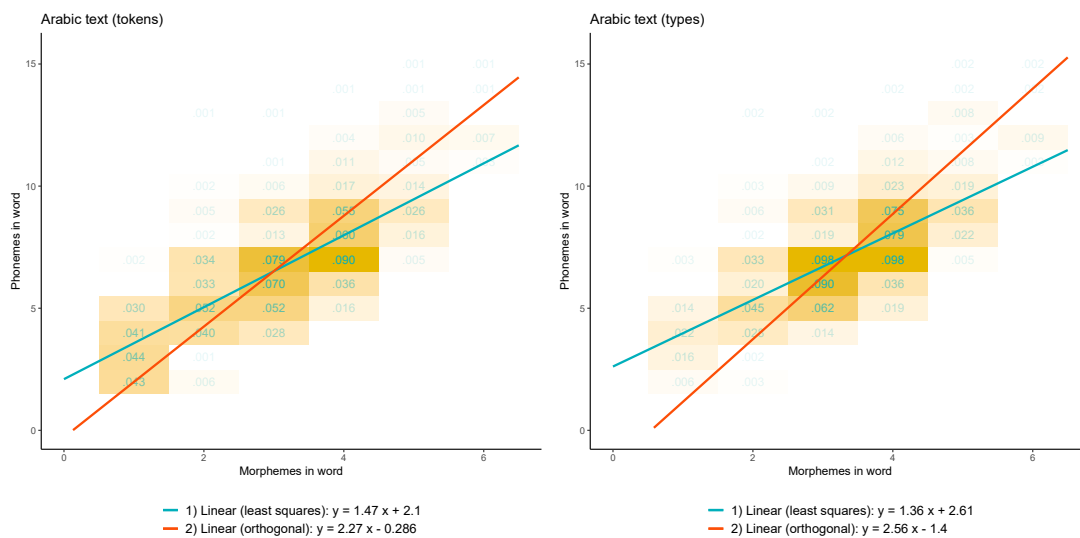


Figure 7: Phoneme — morpheme — word level (Arabic data, part of *Kalīla wa-Dimna* by Ibn al-Muqaffa').

and it is based on a chapter from the famous Arabic book *Kalīla wa Dimna* by Abdallāh ibn Muqaffa'. Interestingly, unlike the previous case, the hypothesis of zero intercept holds better for tokens than for types. I do not have a good explanation for this. However, Arabic nonconcatenative morphology differs greatly from Czech morphology, so I would not be surprised if the stochastic principles behind them also differ. The results can be seen in [Figure 7](#).

Both Czech and Arabic texts were further analyzed at the morpheme-word-clause (Figures 8 and 9) and word-clause-sentence levels [Figure 8](#). The word-clause-sentence level was not explored in the Arabic text due to insufficient data. These three datasets were only examined in terms of tokens, yet, this is likely to have a minimal impact on the results, given that clauses and sentences recur less frequently compared to words.

In all cases, the intercept was found to be negative, with its absolute value always lower than the parameter b , indicating a positive value even for the shortest construct, which makes sense. It is possible that the negative intercepts in these models are due to artifacts arising from the discrete nature of the joint distribution. Or the assumption that the type-token distinction is not necessary at the clause and sentence levels are false. However, it is also plausible that the stochastic principles underlying these datasets differ greatly from those shaping the joint distributions at the word level. Words are pre-processed units that have been shaped over the centuries of the evolution of language, while clauses and sentences are shaped by the capabilities of a single human brain.⁶ In fact, it is interesting that the data produced by these two vastly different processes are not more divergent.

Therefore it may be the case that the negative intercepts are inherent results of the stochastic processes involved. The combination of the negative intercept and the regression toward the mean explains the existence of datasets, where the Menzerath's relation manifests as an increasing function (Buk and Rovenchak, 2008).

The last dataset examines the words-phrase-clause level and is sourced from Mačutek et al. (2017). Unlike words, clauses and sentences, the phrases were not delimited by the speakers; they were defined as chunks of text that depend on a predicate (for a more detailed definition, see the cited paper). Therefore, their segmentation relies on the linguistic annotation of the corpus they come from — the Prague Dependency Treebank 3.0 (Bejček et al., 2013). This dataset shows what the result looks like when it is negative (see [Figure 10](#)). Meanwhile, the hyperbolic model itself can be fitted well to the data.

⁶Here I describe an overall trend rather than a strict rule, there are some creative aspects in morphology — nonce words (occasionalisms) do exist, especially in some registers. On the other side large parts of clauses or even sentences can be formulaic multi-word expressions whose structure is given beforehand.

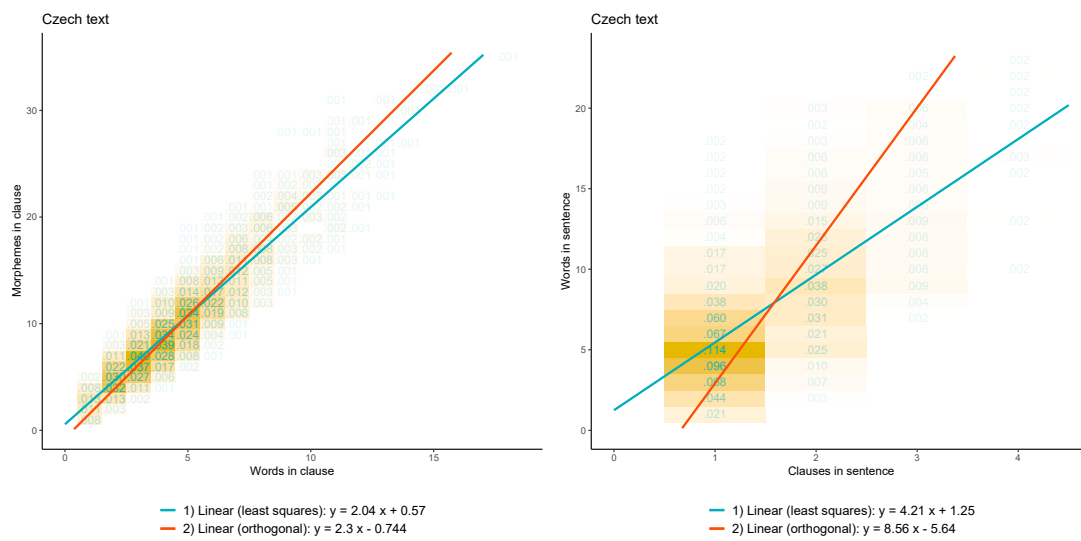


Figure 8: Morpheme — word — clause level and word — clause — sentence level (Czech data, short novel *Krysař* by Viktor Dyk).

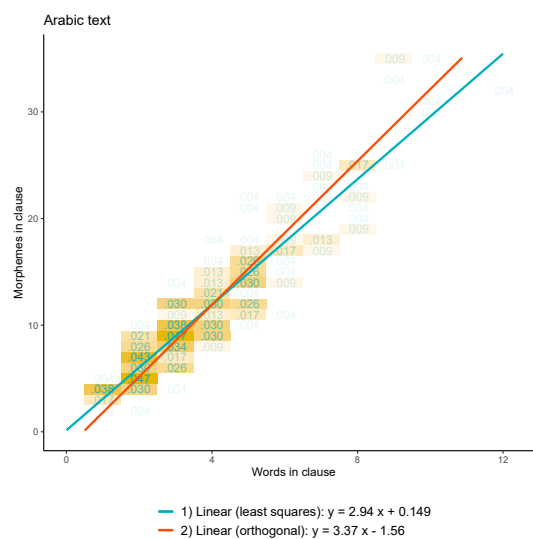


Figure 9: Morpheme — word — clause level (Arabic data, part of *Kalīla wa-Dimna* by Ibn al-Muqaffa').

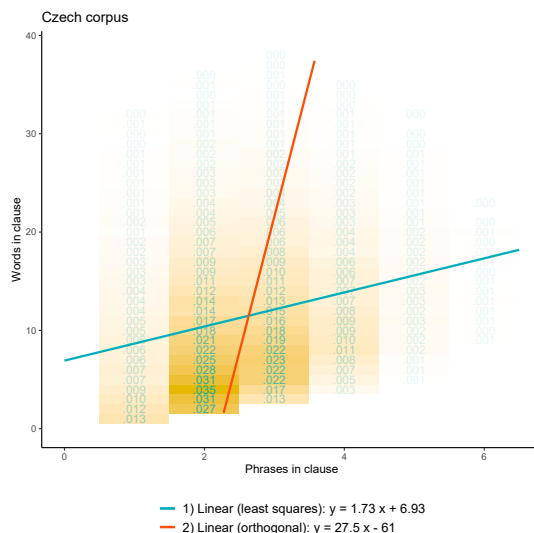


Figure 10: Word — phrase — clause level (Czech data, PDT corpus).

5 Conclusion

The main idea I propose in this study is that the Menzerath's law is a consequence of the features of the relation between the number of constituents in a construct and the number of subconstituents in the same construct. Several datasets suggest that the relation can be fairly simple with the number of constituents being directly proportional to the number of subconstituents, additionally there are some random processes scattering the data points around.

Revisiting the original question posed in the title — is the Menzerath's law just the regression toward the mean? There is a journalistic adage stating that whenever a title contains a question, the answer to that question is negative, or at least partially so. In this case I take issue with the term *just*.

I do not think the Menzerath's law is *just* a regression toward the mean in the sense that this interpretation would render the entire phenomenon obsolete and unworthy of further study.

We tend to understand a linear relation as a default but this inclination is rather magical thinking. Linearity does not mean that the process involved in the relation does not need to be studied. Besides, we need to characterize this random process. It would be worthwhile to turn our attention to the joint distribution itself, and possibly to the marginal distributions as well. When studying the phenomenon directly, the appropriate model of Menzerath's law would follow, with the advantage that this time we will have an explanation of the model and interpretation of its parameters.

Even if we find that the stochastic process involved is fairly simple, it does not mean, that the phenomenon is trivial.⁷ We need to find out which real world phenomenon works in the way corresponding with the

⁷Even not *mathematically trivial*, see Ferrer-i-Cancho et al. (2014). This stochastic process is yet to be found, Meyer

stochastic process so that it is plausible not only as a mechanism creating the data, but also as a model of reality. Moreover, one stochastic process usually does not explain the data completely, there are typically residuals left for the others to analyze.

Therefore I do not think the Menzerath's law is *just* regression toward the mean in its second sense either, meaning that it explains everything. Even some datasets presented in this paper do not fit the idea completely.

The question is, whether we can find something to generalize beyond the hyperbolic model, since the attractiveness of MAL as a research topic mostly dwells in its generality. The detailed stochastic processes on various linguistic levels may differ wildly and they may not be suitable for generalization. It may be the case that the original vague Paul Menzerath's hypothesis on decreasing function is actually the only idea that can be generalized to all the datasets on various language levels and in various human and non-human languages. And this vague idea might be adequately represented by a hyperbolic model.

In any case, for practical use for Menzerath's parameters, I suggest checking whether it might be more advantageous to use some parameters of the marginal distribution models and the correlation metric between the two variables instead. For instance instead of fitting the MAL parameters on the phoneme-syllable-word for stylometric text classification (Chen and Liu, 2022), it may be simpler to directly use the mean number of syllables in words,⁸ the mean number of phonemes in words, and the correlation coefficient between the number of syllables in a word and the number of phonemes in the same word.

Acknowledgments

I am grateful to all those who inspired me to write this paper. If I have inadvertently borrowed some of your ideas, please be generous. The notion of Menzerath's law as a linear regression took root in my mind during some discussions with Miroslav Kubát in Trier, 2013. In 2014, Damian Blasi confided in me the idea of writing several articles asserting that MAL is merely a statistical artifact, but then he moved on to more fruitful topics. I do not know what he had in mind but I guess it was something more sophisticated. In 2019 I discussed the topic with Vladimír Matlach and Tereza Motalová. Also, I probably picked up some arguments during several heated discussions with Radek Čech and Ján Mačutek.

Language models by OpenAI (GPT-3.5 and GPT-4) have been used to improve the paper stylistically.

This output was supported by the NPO "Systemic Risk Institute" number LX22NPO5101, funded by (2007) and Torre et al. (2021) found very simple stochastic models that produce a decreasing function resembling the classic Menzerathian curve, but they failed to fit their models to empirical data.

⁸If the word length distributions can be successfully modeled by 1-displaced Poisson distribution — and it often does, see Chebanow (1947) and Grzybek (2007) — then the mean word length is the only parameter needed.

European Union — Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES).

References

- Altmann, G.** (1980). Prolegomena to Menzerath's law. In R. Grotjahn (Ed.), *Glottometrika 2*, pp. 1–10. Studienverlag Dr. N. Brockmeyer.
- Altmann, G.** (2014). Bibliography: Menzerath's law. *Glottology*, 5(1), pp. 121–123. <https://doi.org/doi:10.1515/ glot-2014-0008>.
- Altmann, G., Schwibbe, M. H.** (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Georg Olms Verlag.
- Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.** (2013). Prague Dependency Treebank 3.0 [LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. <http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3>.
- Buk, S., Rovenchak, A.** (2007). Statistical parameters of Ivan Franko's novel *Perekhresni stežky* (The Cross-Paths). In: Grzybek, P., Köhler, R. (Eds.). *Exact methods in the study of language and text. dedicated to Gabriel Altmann on the occasion of his 75th birthday*, pp. 39–48. De Gruyter Mouton. <https://doi.org/doi:10.1515/ 9783110894219.39>.
- Buk, S., Rovenchak, A.** (2008). Menzerath–Altmann law for syntactic structures in Ukrainian. *Glottology*, 1(1), pp. 10–17. <https://doi.org/10.1515/glot-2008-0002>.
- Chebanow, S.** (1947). On conformity of language structures within the Indoeuropean family to poisson's law. *Comptes rendus de l'Academie de science de l'URSS*, 55(2), pp. 99–102.
- Chen, H., Liu, H.** (2022). Approaching language levels and registers in written Chinese with the Menz-erath–Altmann Law. *Digital Scholarship in the Humanities*, 37(4), pp. 934–948. <https://doi.org/10.1093/llc/fqab110>.
- Ferrer-i-Cancho, R., Hernández-Fernández, A., Baixeries, J., Dębowski, Ł., Mačutek, J.** (2014). When is Menzerath–Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, 13(6), pp. 633–644. <https://doi.org/10.1515/sagmb-2013-0034>.
- Galton, F.** (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, pp. 246–263.
- Gerlach, R.** (1982). Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In: Lehfeldt, W., Strauss, U. (Eds.). *Glottometrika 4*, pp. 95–102. Studienverlag Dr. N. Brockmeyer.
- Grzybek, P.** (2007). History and methodology of word length studies. In: Grzybek, P. (Ed.). *Contributions to the science of text and language: Word length studies and related issues*, pp. 15–90. Springer Netherlands. https://doi.org/10.1007/978-1-4020-4068-9_2.

- Gustison, M. L., Semple, S., Ferrer-i-Cancho, R., Bergman, T. J.** (2016). Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19), pp. E2750–E2758. <https://doi.org/10.1073/pnas.1522072113>.
- Köhler, R.** (1984). Zur Interpretation des Menzerathschen Gesetzes. In J. Boy, R. Köhler (Eds.), *Glottometrika* 6, pp. 177–183.
- Kuřacka, A., Mačutek, J.** (2007). A discrete formula for the Menzerath-Altmann law*. *Journal of Quantitative Linguistics*, 14(1), pp. 23–32. <https://doi.org/10.1080/09296170600850585>.
- Mačutek, J., Čech, R., Courtin, M.** (2021). The Menzerath-Altmann law in syntactic structure revisited. *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pp. 65–73. <https://aclanthology.org/2021.quasy-1.6>.
- Mačutek, J., Čech, R., Milička, J.** (2017). Menzerath-Altmann law in syntactic dependency structure. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 100–107. <https://aclanthology.org/W17-6513>.
- Mačutek, J., Rovenchak, A. A.** (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In: Kelih, E., Levickij, V., Matskulyak, Y. (Eds.). *Issues in quantitative linguistics* 2, pp. 136–147. RAM Verlag.
- Menzerath, P.** (1928). Über einige phonetische Probleme. *Actes du premier Congrès international de linguistes, à La Haye*, pp. 104–105.
- Menzerath, P.** (1954). *Die Architektonik des deutschen Wortschatzes* (Vol. 3). F. Dümmler.
- Meyer, P.** (2007). Two semi-mathematical asides on Menzerath-Altmann's law. In: Grzybek, P., Köhler, R. (Eds.). *Exact methods in the study of language and text. dedicated to Gabriel Altmann on the occasion of his 75th birthday*, pp. 449–460. De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110894219.449>.
- Mikros, G., Milička, J.** (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In: Altmann, G., Čech, R., Mačutek, J., Uhlřřová, L. (Eds.). *Empirical approaches to text and language analysis*, pp. 180–189. RAM-Verlag.
- Milička, J.** (2014). Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, 21(2), pp. 85–99. <https://doi.org/10.1080/09296174.2014.882187>.
- Milička, J.** (2015). Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů [The theory of communication as an explanatory principle for the natural multilevel text segmentation]. Charles University, Faculty of Arts, Prague. (PhD Thesis) <https://dspace.cuni.cz/handle/20.500.11956/77988>.
- Motalová, T.** (2022). Menzerath-Altmann law in Chinese. Palacký University, Faculty of Arts, Olomouc. (PhD Thesis) https://theses.cz/id/vqk2ml/motalova_menzerath_altmann_law_in_chinese.pdf.

- Pelegrinová, K., Mačutek, J., Čech, R.** (2021). The Menzerath-Altmann law as the relation between lengths of words and morphemes in Czech. *Journal of Linguistics / Jazykovedný časopis*, 72(2), pp. 405–414. <https://doi.org/10.2478/jazcas-2021-0037>.
- Semple, S., Ferrer-i-Cancho, R., Gustison, M. L.** (2022). Linguistic laws in biology. *Trends in Ecology & Evolution*, 37(1), pp. 53–66. <https://doi.org/10.1016/j.tree.2021.08.012>.
- Stave, M., Paschen, L., Pellegrino, F., Seifart, F.** (2021). Optimization of morpheme length: A cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard*, 7(s3: 20190076), pp. 1–11. <https://doi.org/10.1515/lingvan-2019-0076>.
- Torre, I. G., Dębowski, Ł., Hernández-Fernández, A.** (2021). Can Menzerath's law be a criterion of complexity in communication? *PLOS One*, 16(8), e0256133. <https://doi.org/10.1371/journal.pone.0256133>.
- Wang, Y., Chen, H.** (2022). The Menzerath-Altmann law on the clause level in English texts. *Linguistics Vanguard*, 8(1), pp. 331–346. <https://doi.org/doi:10.1515/lingvan-2022-0048>.

Applying Distributional Semantic Models to a Historical Corpus of a Highly Inflected Language: the Case of Ancient Greek

Alek Keersmaekers^{1*} , Dirk Speelman¹ 

¹ University of Leuven

* Corresponding author's email: alek.keersmaekers@kuleuven.be

DOI: https://doi.org/10.53482/2023_55_410

ABSTRACT

So-called “distributional” language models have become dominant in research on the computational modelling of lexical semantics. This paper investigates how well such models perform on Ancient Greek, a highly inflected historical language. It compares several ways of computing such distributional models on the basis of various context features (including both bag-of-words features and syntactic dependencies). The performance is assessed by evaluating how well these models are able to retrieve semantically similar words to a given target word, both on a benchmark we designed ourselves as well as on several independent benchmarks. It finds that dependency features are particularly useful to calculate distributional vectors for Ancient Greek (although the level of granularity that these dependency features should have is still open to discussion) and discusses possible ways for further improvement, including addressing problems related to polysemy and genre differences.

Keywords: distributional semantics, Ancient Greek, word similarity

1 Introduction

So-called “distributional” language models (also “vector space models”, “semantic spaces” or “word embeddings”) have become dominant in research on the computational modelling of lexical semantics. These techniques start from the long-held assumption that you can “know a word by the company it keeps” (Firth 1957) and try to model the semantic relatedness among different words based on their occurrence in shared contexts. While there is plenty of literature on the application of such models to modern languages, historical languages such as Ancient Greek have received less attention so far (although this is increasing, see Section 2.2). Yet there are several challenges that make Ancient Greek an interesting case study.

Many of these challenges have to do with the size and nature of the available corpus materials. First of all, we have far less data for Ancient Greek than for a modern language such as English: in the order of millions rather than billions for the whole corpus, and only on average 2 million words per century.

Since distributional language models require large amounts of data, making a selection in the already rather limited corpus material we have would inevitably lead to data sparsity. Yet the Ancient Greek corpus also spans a large period of time, and its genres are rather unevenly distributed (see Section 3), giving us a far less homogenous dataset to start from in comparison to e.g. modern language distributional models trained on Wikipedia or newspaper prose. Additionally, most of the data are of a literary or technical nature, including several genres such as epic poetry or scientific prose with a rather idiosyncratic language, while the non-literary, everyday language parts of the corpus, e.g. papyrus letters, are rather limited. But it is not just the precarious text transmission that stands in the way of a smooth application of distributional language models: the nature of the Greek language itself also presents some additional problems. We mentioned above that distributional language models measure word similarity on the basis of shared contexts: this notion of “context” typically refers to the lexical and syntactic context of a word, i.e. the words it combines with, either based on the words that precede or follow the target word (so called “bag-of-words”-models), or on more sophisticated measures such as syntactic dependency relationships. This works well for isolating languages, but it is not immediately obvious that such approaches would work equally well with a language such as Ancient Greek, which expresses much information by relying on morphological rather than syntactic means. A Greek finite verb, for instance, is inflected for person, number, tense and aspect, mood and voice. Of these features, English only expresses number and tense morphologically. Furthermore, the word and constituent order of Ancient Greek is notoriously free (see Dik 1995), which might complicate distributional bag-of-words models that only take the direct environment of a word into account.

This paper aims to test the validity of distributional semantic models on the Ancient Greek language, by evaluating how well these models are suited to retrieve semantically similar words to a given target word. While language-external issues such as genre imbalance will be addressed to some extent, the focus is first and foremost on language-internal issues, i.e. which contextual information works best to model word similarity for Ancient Greek (and other typologically related languages). It is structured as follows: Section 2 will give a broad technical background of distributional semantic models in general, and discuss previous approaches to distributional semantic modeling of Ancient Greek. Section 3 will give an overview of the corpus we used, and Section 4 will describe the specific parameters of the distributional models we compared in more detail. Section 5 will analyze the results of the word similarity task, and Section 6 will summarize and analyze the main results of this study.

2 Models of distributional semantics

2.1 Calculating distributional vectors

While it goes beyond the scope of this paper to give a full overview of the broad field of distributional semantic modelling (see Erk 2012, Lenci 2018 for some recent surveys), this section will give a concise

presentation of the terminology and techniques used in this paper. First of all, as for distributional techniques in general, a distinction can be made between so called context-counting and context-predicting models (the latter also known as “neural language models”) (Baroni et al. 2014). Both types of models represent a word as a vector of real numbers, so that the vectors of words that are semantically similar are also mathematically similar. However, they differ with respect how these vectors are calculated: the vectors of context-count models directly contain the co-occurrence frequencies (either weighted or not, see below) of the context words with which the target word occurs. The weights of context-predict models, in contrast, are calculated in such a way (on the basis of a supervised machine learning approach, using neural networks) to predict the contexts in which the target word tends to appear. Such an approach has been found to outperform context-count models on a wide range of tasks (Baroni et al. 2014). However, one of the main advantages of using context-count models is their greater transparency: the individual elements of these vectors directly refer to the contexts in which the target word appears, while the elements of vectors calculated with a context-predict approach do not have any obvious meaning. This paper aims to compare and understand the underlying reasons why certain models are better suited to perform a number of specific tasks than others. Since the focus is not on achieving state-of-the-art performance for these tasks, we will stick to a context-count approach, although a comparison with context-predict models is certainly a desideratum for the future.

An appropriate starting point for explaining the procedure behind the creation of context-count vectors is Turney and Pantel (2010). The first step consists in counting for each target word how often certain other words occur in its context, for example a window of N preceding and following words (see Section 4 for alternative ways of determining the context).¹ Next, the elements on the matrix are typically weighted to give more weight to more “surprising”² co-occurrences. This paper will use the Pointwise Positive Mutual Information (PPMI) measure to do so, which has been shown to outperform other weighting approaches (Bullinaria and Levy 2007).³ Function words and/or stop words are often removed from the matrix. However, as their removal has been shown to have no significant positive or negative effect on performance for English data (Bullinaria and Levy 2012), we refrained from removing them in the context of this paper (although we left out tokens indicating punctuation or “gaps” in the text): our early experiments suggested that removing them does not have an effect for Ancient Greek either.

¹ The target and context words can be either lemmas or word forms. Since Greek is a highly inflectional language (a Greek participle, for instance, has more than 150 possible forms), using word forms would lead to data sparsity, so all the models described in this paper are based on word forms.

² The term “surprising” is used here in a statistical context, to refer to co-occurrences that appear more than we would expect from random chance.

³ The PPMI is calculated by first log-transforming the observed frequency of a co-occurrence pattern divided by its expected frequency (i.e. the PMI measure), which has a negative value when the observed frequency is lower than the expected frequency and a positive value when it is higher than the expected frequency. Subsequently, all negative PMIs are set to 0 (i.e. all patterns with an observed frequency that is lower than the expected frequency). See Turney and Pantel (2010: 157-158) for more information.

Subsequently, a dimension reduction technique such as Singular Value Decomposition (SVD) is often applied to the co-occurrence matrix in order to reduce the context information to a smaller number of latent dimensions, which often improves the performance of context-count models (Bullinaria and Levy 2012). However, we will refrain from doing so in the context of this paper, in order to gain a better insight in the specific context features that cause semantic similarity (see Section 5).

To detect semantic similarity, we next need to calculate by some measure how similar the vectors of the different target words are. We will use the cosine similarity measure for this purpose, which has been found to outperform other measures to detect semantic similarity in the vector space (Bullinaria and Levy 2007, Lapesa and Evert 2014). The cosine similarity (as is obvious from its name) captures the cosine of the angle between the two vectors that are compared, and is 1 when they are completely similar and 0 when they are completely dissimilar (see Turney and Pantel 2010: 160-161 for the calculation).

2.2 Related work

This section will give an overview of the relevant literature: more details about the model parameters of the main studies discussed here can be found in Table 1. Most studies investigating distributional models for Ancient Greek are applied in nature, in particular using them in order to track lexical semantic change. As for context-count models, the first study was Boschetti (2010), who used a context-count model to examine the Greek lexicon in various ways, including the diachronic development of specific words, their polysemy structure in different genres and the taxonomical relations among them. Additionally, Boschetti argues that such models can also be used for text-critical ends, i.e. to evaluate the appropriateness of editorial conjectures. Rodda et al. (2017) use distributional models trained on a part of the *TLG* corpus (36 million tokens in total) to evaluate the hypothesis whether Christianity had a significant effect on the Greek lexicon. Their results confirm the crucial role of Christianity on lexical semantic change in Greek, and also show that distributional models can bring unexpected patterns of change to light. Rodda et al. (2019) have developed distributional models in order to study linguistic variation in Ancient Greek epic formulae. They are one of the only studies that compare several (context-count, SVD-reduced) distributional models against independent benchmarks from various sources (ancient scholarship – the *Onomasticon* by Julius Pollux – modern lexicography – Schmidt’s dictionary of synonyms – and an NLP resource – the *Open Ancient Greek WordNet*). These models vary with regard to the context window (1, 5 and 10 words to the left and right) and frequency threshold (including all words, words that occur at least 20, 50 and 100 times in the corpus). They find that context windows of 5 words and frequency thresholds of 20 or 50 words achieve the best results on their benchmarks (with the *Onomasticon* and Schmidt’s dictionary matching the semantic spaces of the distributional models better than *Ancient Greek WordNet*).

There have also been some studies on context-predict models for Ancient Greek: an experimental *word2vec* model has been implemented in the Python Classical Language Toolkit (Burns 2019), although their results

have not been evaluated yet. Perrone et al. (2021) compare the results of two context-predict models to a dynamic Bayesian mixture model for the task of detecting semantic change, but conclude that the latter approach delivers superior results over the context-predict models. List (2022) investigates how Word2Vec models can be used for lexicographic purposes. Finally, recently various transformer models have been trained for Ancient Greek, including Singh et al. (2021) (BERT), Yamshchikov et al. (2022) (BERT) and Spanopoulos (2022) (RoBERTa). These studies did not evaluate their models for semantic purposes, however, making them less relevant for this paper. In contrast, Riemenschneider and Frank (2023) evaluate RoBERTa models on various non-semantic and semantic tasks, including their ability to distinguish synonyms from antonyms. Additionally, Mercelis et al. (Forthcoming) evaluate the results of an ELECTRA model for word sense disambiguation, comparing both unsupervised and supervised techniques.

Stopponi et al. (2023) compare the performance of both context-count and context-predict models trained on the Diorisis corpus. The evaluation is done on the AGREE benchmark, containing evaluations of word similarity rated by experts. For the context-count models, the authors compare both dimensionally reduced vectors (with SVD) and non-reduced vectors, while for the context-predict models, they compare a SGNS model to a Continuous Bag-of-Words model (CBOW), in all cases using a window size of 5 words. They conclude that context-count models perform better than context-predict models against this benchmark, with the non-reduced vectors performing the best of all 4 models.

While interest in distributional models for Ancient Greek is clearly increasing, in all of these studies only bag-of-words models are investigated,⁴ and dimension reduction or neural networks are generally employed, making the resulting vectors difficult to interpret. The main contribution of this paper is therefore the following: it will compare various ways to incorporate syntactic context as well (see Section 4), and offer a thorough investigation of the resulting semantic spaces and the various context features that cause semantic similarity. Additionally, it will employ the GLAUx corpus (see Section 3), the largest openly available Greek corpus so far, allowing for higher quality semantic spaces than the previous studies.

Table 1: Previous studies on distributional semantic modeling for Ancient Greek.

Study	Architecture	Application	Corpus
Boschetti (2010)	Count, SVD (window 100)	Describing the lexicon	TLG
Rodda et al. (2017)	Count, SVD (window 5)	Lexical semantic change	TLG
Rodda et al. (2019)	Count, SVD (varying window)	Model comparison, epic formulae	Diorisis
Perrone et al. (2021)	Predict, word2vec (SGNS/TR)	Lexical semantic change	Diorisis
List (2022)	Predict, word2vec (SGNS)	Lexicography	Diorisis
Riemenschneider & Frank (2023)	Predict, transformer (RoBERTa)	Model comparison	Custom
Mercelis et al. (forthcoming)	Predict, transformer (ELECTRA)	Word Sense Disambiguation	GLAUx
Stopponi et al. (2023)	Count, SVD/Non-SVD; Predict, word2vec (SGNS/CBOW)	Model comparison	Diorisis

⁴ However, a future investigation into the performance of syntactic models has been announced by Stopponi et al. (2023).

3 The corpus

As mentioned in the introduction of this paper, the Ancient Greek corpus is quite small as compared to some modern language corpora. What is more, the largest collection of Greek text – the corpus of the *Thesaurus Linguae Graecae* (TLG) – has not made its data publicly available. However, there have been some recent large-scale open initiatives: the Diorisis corpus (Vatri and McGillivray 2018), containing 10.2M tokens from the 8th century BC to the 5th century AD, and the GLAUx corpus (Keersmaekers 2020), containing 27.7M tokens from the 8th century BC to the 8th century AD. Since the Diorisis corpus is much smaller and does not contain syntactic annotation, which was essential for the experiments described in the next sections, we made use of the latter corpus. More specifically, we used an earlier version of GLAUx, which was larger (37.2M tokens) but also noisier, containing several texts with OCR problems. The accuracy is about 0.95 for part-of-speech/morphological tagging and 0.98 for lemmatization, while syntactic parsing accuracy (Labeled Attachment Score) ranges between 0.75 and 0.88 depending on text genre (see Keersmaekers 2020).

The literary texts are quite diverse with respect to texts genre, ranging from epic poetry to drama, philosophy, historical narrative, scientific prose and so on. Previous studies have already indicated that text genre has an important effect for the computational modelling of semantics for Ancient Greek (Boschetti 2010, McGillivray et al. 2019). Since we did not want to further reduce the corpus to avoid data sparsity, we used the full corpus for the construction of distributional vectors. However, in our analysis we will also consider how genre and diachrony may influence the resulting semantic spaces.

4 Construction of context models

As mentioned in Section 2, all techniques discussed in this paper make use of some notion of “context”. In traditional collocational and distributional semantic approaches, this context is simply defined as a window of preceding and/or following words – a so-called “bag-of-words” approach. This context window can be as wide or small as the researcher wants to define it, but in general it has been found that larger context windows leads to a more associative, topical similarity (e.g. “soldier”/“war”) while smaller context windows lead to cosine similarities that indicate relationships that are more taxonomic (e.g. “soldier”/“warrior”) (Peirsman et al. 2008; Kolb 2009).

Another way to define “context” is to use the *syntactic* context of a word as features, in particular involving syntactic dependencies (Lin 1998, Padó and Lapata 2007). This approach has been shown to return even tighter taxonomic syntactic relationships than small-window bag-of-words approaches (e.g. Heylen et al. 2008, see also Levy and Goldberg 2014 for context-predict models). In such an approach context features typically look like *child/OBJ* (as in *child* is the object of the target word X, e.g. of *raise* in *he raised the child*), although it is in principle possible to include less or more detailed information (see below).

Finally, in the context of a highly inflectional language such as Ancient Greek, it also makes sense to consider the *morphological* context of a word. Greek dictionaries such as Liddell-Scott-Jones (Jones et al. 1996), for instance, typically list what cases, moods etc. a given word frequently combines with. In fact, one could wonder whether language-internal categories such as case are in fact not better suited to model the semantics of Ancient Greek than categories that are considered to be more language-general such as “object” (i.e. by replacing “child is the object of X” by e.g. “child is a dative dependent on X”) – see in this context Croft’s (2013) skepticism on defining such language-general categories. Particularly with context-predict models, there have been several approaches that integrated morphological or other formal characteristics of the target word itself in its vector embedding, i.e. to assign similar vectors to formally similarly looking words (e.g. Luong et al. 2013; Botha and Blunsom 2014, Bojanowski et al. 2017), but the use of morphological features as context features has, to the best of our knowledge, not been explored yet.

To test the role of the type of context model in detecting Ancient Greek word similarity, we have constructed five types of context models, as summarized in Table 2 below. All models use PPMI weighting and require a context feature to occur at least 150 times, so as to avoid features that are too infrequent as well as noise in the data. The first context model is a simple bag-of-words model (model *BOW*). We used a context of 4 preceding and following words, since this window size turned out to be the most optimal to detect word similarity for Ancient Greek without bringing in too much noise in some early (unpublished) experiments. The four other models make use of syntactic information, using the automatically parsed data described in Section 3. The first (which we will style *DepMinimal*) simply states the frequency of lemmas that have a direct dependency link with the target word, i.e. when the context word occurs as the head or as a child of the target word, without adding information about syntactic relation or whether the context word occurs as the head or child (i.e. the direction of the arc). The second (*DepHeadChild*) enhances this with the information whether the given context word occurs as the target word’s head or child, i.e. in ἡ θυγάτηρ τῆς μητρὸς “the mother’s daughter”, the relevant feature for μήτηρ “mother” would be θυγάτηρ/head (“daughter”), while in ἡ μήτηρ τῆς θυγατρὸς “the daughter’s mother” the feature would be θυγάτηρ/child. In the third model (*DepSyntRel*) a syntactic label is added, e.g. θυγάτηρ/head/ATR for “μήτηρ is an attribute of θυγάτηρ” or θυγάτηρ/child/ATR for “θυγάτηρ is an attribute of μήτηρ”. Finally, in a fourth model (*DepMorph*) we use morphological information instead of syntactic labels. Instead of using the full morphology of the context words (which can be quite extensive for words such as participles and as a result increases data sparsity) we only include two features that we considered to be most relevant in a word’s combinatorial behavior (and are therefore often mentioned in dictionaries such as Jones et al. 1996): case (nominative, accusative, dative, genitive, vocative) and mood (indicative, subjunctive, optative, imperative, infinitive, participle). In such a case a feature would look like θυγάτηρ/child/gen for “θυγάτηρ is a genitive with μήτηρ” (see Table 2 below for an illustration).

These syntactic models required us to implement a special treatment of prepositions and conjunctions on the one hand, and coordination structures on the other hand. In a sentence such as ἔρχομαι εἰς πόλιν “I go to a city”, εἰς (“to”) is treated in our syntactic corpus as a prepositional group with ἔρχομαι (“I go”) and πόλιν (“city”, accusative of πόλις) as the “object” of εἰς (which is in fact the relation that εἰς πόλιν has to ἔρχομαι). When it comes to determining the syntactic context of ἔρχομαι, one has four options: (1) εἰς, (2) πόλις, (3) both εἰς and πόλις, or (4) a single feature “εἰς πόλιν”. Since we considered both εἰς and πόλις to be relevant for the meaning of ἔρχομαι, and since adding a single feature “εἰς πόλιν” would considerably reduce the influence of πόλις to the vector — there are many other prepositional groups with the same noun possible, such as ἀπὸ πόλεως “from the city”, ἐκ πόλεως “out of the city” etc. — we preferred to count two context features in such a case, respectively “εἰς” and “πόλις”. Secondly, the use of dependencies implies that coordination structures are somewhat awkwardly annotated: in a hierarchical representation it is much more straightforward to annotate subordination than coordination. In our representation, one coordinate has been made dependent of the other: i.e. in a sentence such as ἀκούω φωνὴν καὶ βοήν “I hear a voice and a scream” φωνή (“voice”) is annotated as the object of ἀκούω “to hear”, while βοή (“scream”) is annotated as a conjunct of φωνή. Since we considered both the fact that βοή is an object of ἀκούω and that φωνή is coordinating with βοή to be relevant for the meaning of βοή, we again added two features for βοή in such a case, its technical head “φωνή” and the head of the whole group “ἀκούω”.

Finally, since our corpus contains many proper names which would be less useful as either context features (the specific name would not matter except for some rare cases such as “Zeno’s paradox”) or target words (a vector for specific names, which are shared by several people who have little in common, would make little sense) we chose to replace all words starting with a capital letter simply by the lemma “NAME” (although in the future, it would be preferable to distinguish personal names such as “Socrates” from place names such as “Greece”).

Table 2: Distributional models constructed for this study.

	Context	Head/child	Extra info	Example features
BOW	Window (size 4)	N/A	NO	μήτηρ, δίδωμι
DepMinimal	Dependencies	NO	NO	μήτηρ, δίδωμι
DepHeadChild	Dependencies	YES	NO	μήτηρ/child, δίδωμι/head
DepSyntRel	Dependencies	YES	Syntactic label	μήτηρ/child/ATR, δίδωμι/head/OBJ
DepMorph	Dependencies	YES	Morphology	μήτηρ/child/genitive, δίδωμι/head/dative

5 Evaluation of the context models

5.1 Main benchmark

Various benchmarks exist for the evaluation of distributional semantic models for Ancient Greek, described in Section 2.2. However, since they did not exist when the main research for this paper was carried out, and generally only contain lists of semantically related words without specifying in which way they are related, we decided to create our own benchmark, offering more detailed information about semantic relatedness (nevertheless, we will also offer results evaluated on these other benchmarks in Section 5.6). More concretely, we examined a sample of 100 lemmas – 50 nouns and verbs each – divided into 5 frequency bands, with 10 randomly chosen verbs or nouns in each band. We only selected lemmas with a frequency of at least 50 and chose to divide the frequency ranges for each band in such a way that the first group contains the 50% most frequent noun or verb tokens, the second group the next 25% most frequent tokens, the third group the next 12.5%, the fourth group the next 6.7% and the final group the remaining 6.7% tokens.⁵ This resulted in the randomly chosen lemmas in Table 3.

Table 3: Words evaluated for the similarity task.

Band	Type	Freq.	Lemmas
1	Nouns	3600+	ἀλήθεια “truth”, πέρασ “boundary”, ὄνομα “name”, πόλις “city”, ἀπορία “difficulty”, μάχη “battle”, ἀδελφός “brother”, αἰτία “cause”, ἡδονή “pleasure”, καρδία “heart”
1	Verbs	8000+	δοκέω “seem”, συμβαίνο “agree”, καλέω “call”, φημί “say”, δρᾶω “see”, μένω “stay”, ἵστημι “stand”, πάρεμι “be present”, κρίνω “judge”, μανθάνω “learn”
2	Nouns	850-3600	συμφορά “accident”, ὀδούς “tooth”, κῦμα “wave”, σιωπή “silence”, ἔρις “strife”, ἀγαλμα “statue”, πλοῖον “ship”, ὄς “pig”, νεανίσκος “young man”, οὐλή “scar”
2	Verbs	1900-8000	ἀπαντάω “meet”, ἀφήμι “let go”, κατασκευάζω “equip”, ἀποκρίνω “answer”, τέμνω “cut”, συντίθημι “put together”, οἴχομαι “be gone”, γαμέω “marry”, βιάζω “force”, φιλέω “love”
3	Nouns	300-850	λοχαγός “commander”, ἄχος “distress”, ἴρις “iris”, ψάμμος “sand”, ἀνάμνησις “remembrance”, προσευχή “prayer”, κωμωδία “comedy”, ταμειῶν “treasury”, ἡῶν “shore”, δελφίς “dolphin”
3	Verbs	650-1900	χαρίζω “please”, ἀποστερέω “rob”, δανείζω “lend”, φορέω “wear”, ἀεῖρω “lift up”, ἀποτίθημι “put away”, μετέρχομαι “pursue”, ἀποτίνω “pay”, περιαιρέω “remove”, ἀπελαύνω “expel”
4	Nouns	150-300	παραφυλακή “guard”, ἵππόδρομος “chariot-road”, οἶστρος “frenzy”, ῥαφή “seam”, καλοκάγαθία “nobleness”, πολεμιστής “warrior”, θήκη “case”, ἐστίασις “feasting”, σκοπιά “hill-top”, πέδιλον “sandal”
4	Verbs	250-650	εὐδαιμονέω “be prosperous”, ἀνασκευάζω “remove”, εὐθύνω “make straight”, κρούω “strike”, λήζομαι “carry off as booty”, σκεπάζω “cover”, κατακρύπτω “hide”, ποιμαίνω “herd”, ἀναδείκνυμι “display”, δεξιόομαι “greet”
5	Nouns	50-150	ἀκρόαμα “anything heard”, ἄρπαγμα “booty”, στρύχνον “winter cherry”, γάρως “sauce”, πρόβασις “advance”, ἔλασις “driving away”, εὐπλοία “fair voyage”, εἰδωλολατρία “idolatry”, ὀποβάλασαμον “balsam”, ἰμάσθλη “whip”
5	Verbs	50-250	ἐναπολαμβάνω “intercept”, αὖω “shout”, προλείπω “abandon”, ἐπιβοηθέω “come to aid”, προκατασκευάζω “prepare beforehand”, ἐξισόω “make equal”, προαπαντάω “go forth to meet”, ἐπισυντίθημι “add successively”, ἐκθειάζω “deify”, ἐξοδιάζω “scatter”

⁵ This seemed a good compromise to us instead of dividing the groups into five groups of an equal number of types (which would result in a first group consisting of several highly frequent and averagely frequent words, and the other groups consisting of only lowly frequent words), or an equal number of tokens (which would result in the first groups containing only a few very frequent items and the other groups containing all other items).

For each lemma, we calculated the cosine distance with all other remaining nouns/verbs of the full dataset, using the PPMI vectors of the models described in Section 4. Next, we examined the 10 nearest neighbors (i.e. the lemmas with the highest cosine similarity) of each lemma and annotated them with the following labels, which we considered to be useful to distinguish some very basic distinctions of semantic relatedness:

- **Synonym:** has a synonymous or near-synonymous meaning with the target lemma. E.g. *νεανίσκος* – *νεανίας* (both “young man”) or *κρούω* – *τύπτω* (both “strike, knock”).
- **Related:** while the words are not strictly synonymous, they are closely semantically and syntactically related, for instance because they share a hypernym or one word is the hypernym of the other (i.e. there is a taxonomical relationship between the two words). E.g. *νεανίσκος* – *παρθένος* (“young man” – “young woman”) or *κρούω* – *ώθέω* (“strike” – “thrust”).
- **Distantly-related:** there is a vague resemblance between the two words, but they share a hypernym higher up in the ladder, and as a result they will still frequently occur in the same syntactic environments. E.g. *νεανίσκος* – *στρατιώτης* (“young man” – “soldier”) or *κρούω* – *αείδω* (“strike (often musically)” – “sing”).
- **Same domain:** while there is no shared hypernym between the two words, they still often occur in the same thematic contexts (the relation is more associative). E.g. *νεανίσκος* – *ἡλικία* (“young man” – “youth”) or *κρούω* – *ὀρχέομαι* (“strike (often musically)” – “dance”).
- **Unrelated:** there is no overlap in syntactic or thematic contexts. E.g. *νεανίσκος* – *δῆμος* (“young man” – “populace”) or *κρούω* – *ἵστημι* (“strike” – “stand”).

The data were annotated by an independent researcher on Ancient Greek linguistics, starting from the meanings described in the LSJ lexicon of Greek (Jones et al. 1996). Since in most cases there is only partial overlap in meaning between words, overlap with any meaning was checked, e.g. when there was synonymy with at least one meaning (even though the two words might not be synonymous in all meanings) the label “synonym” was used (and similarly for “related” and so on).⁶ Since the training data of the distributional model contains a very long time span (16 centuries) and various text genres, polysemy was considered for the full ranges of uses of a word over time and genre: i.e. two words were also called ‘synonymous’ if they had one meaning that was synonymous, even if this meaning was only present in certain periods or text genres.

⁶ For comparative purposes, we also annotated the data ourselves to evaluate how much of the differences described in this section are simply due to the subjectivity of the annotation. Our labeling only overlapped with the independent one in 45.5% of all cases (1012/2226), Cohen’s kappa = 0.312 (although in an additional 36% of cases the difference was only one level). Nevertheless, the general tendencies described in this section still hold, although the effect of frequency (see 5.3) was a little stronger in our annotation.

5.2 Main results

The following tables detail the general results we found with each syntactic model. For the top 10 we looked at 500 nearest neighbors in total for each model (the 10 nearest neighbors of 10 verbs per frequency band, with 5 frequency bands in total) and for the top 5 the 250 nearest neighbors.

Table 4: Classification of 10 nearest neighbors among verb distributional models.

Top 10 - Verbs	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.142	0.184	0.178	0.192	0.304
DepMinimal	0.160	0.192	0.214	0.186	0.248
DepHeadChild	0.162	0.188	0.216	0.200	0.234
DepSyntRel	0.140	0.192	0.222	0.214	0.232
DepMorph	0.164	0.192	0.226	0.176	0.242

Table 5: Classification of 10 nearest neighbors among noun distributional models.

Top 10 - Nouns	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.088	0.255	0.335	0.244	0.078
DepMinimal	0.108	0.296	0.356	0.166	0.074
DepHeadChild	0.104	0.318	0.336	0.166	0.076
DepSyntRel	0.102	0.324	0.324	0.160	0.090
DepMorph	0.090	0.326	0.316	0.170	0.098

Table 6: Classification of 5 nearest neighbors among verb distributional models.

Top 5 - Verbs	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.180	0.212	0.180	0.176	0.252
DepMinimal	0.212	0.228	0.204	0.164	0.192
DepHeadChild	0.180	0.208	0.244	0.172	0.196
DepSyntRel	0.188	0.212	0.224	0.212	0.164
DepMorph	0.212	0.232	0.192	0.176	0.188

Table 7: Classification of 5 nearest neighbors among noun distributional models.

Top 5 - Nouns	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.104	0.284	0.356	0.180	0.076
DepMinimal	0.148	0.312	0.356	0.120	0.064
DepHeadChild	0.148	0.356	0.300	0.140	0.056
DepSyntRel	0.148	0.380	0.276	0.124	0.072
DepMorph	0.120	0.384	0.304	0.112	0.080

These data first and foremost reveal that there is a clear difference between the bag-of-words model on the one hand and the syntactic models on the other hand: syntactic models prove to be better suited to return synonyms and closely related words than the former. Although the number of totally unrelated words does not differ that much for nouns, the bag-of-words model returns several more words that are only tangentially or associatively related (“same domain”), which corroborates the findings mentioned in

Section 4. For verbs there were no real differences for the “same domain” label, but it is more difficult to say when a verb belongs to the same domain as another verb (since the meaning of a verb tends to be more abstract and/or vague than that of a noun). Consequently, this might simply be an effect of the annotation: the annotator might have been more disposed to say that two nouns belong to the same domain than in the case of verbs. On the other hand, the number of totally unrelated words is clearly higher for BOW in the verb category than for the syntactic models. Within the four syntactic models, however, there is far less differentiation, with only a one or two percent difference for most categories, and no consistent best performing model. We will analyze the reason for this lack of clear differences below.

5.3 Effect of frequency

The following plots detail the effect of frequency by counting the percentage of **synonymous** and **related** words in the 10 nearest neighbors (N=100 per frequency band) – since many words do not have direct synonyms, it makes more sense to consider both in the evaluation of the performance of the different models.

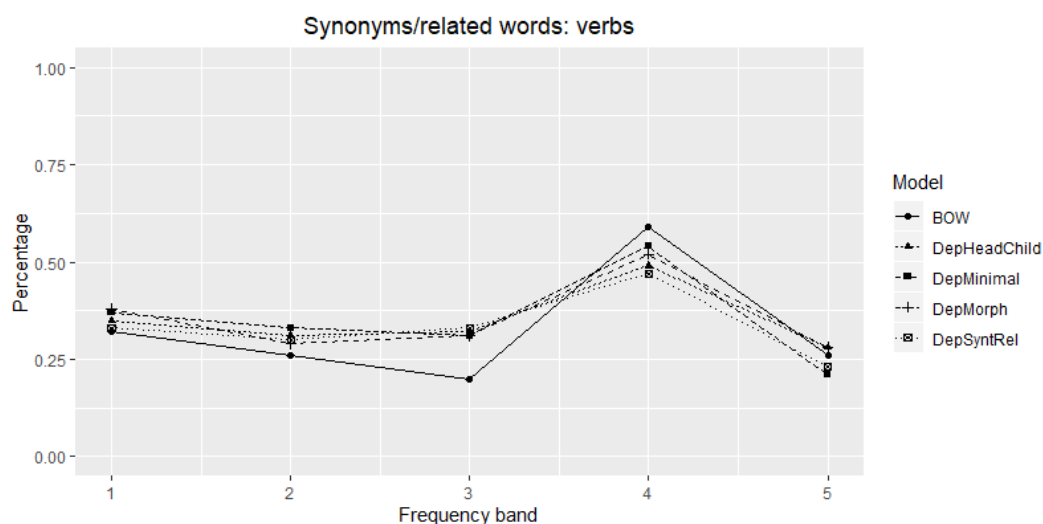


Figure 1: Percentage of synonyms/related words in 10 nearest neighbors by frequency band (verbs).

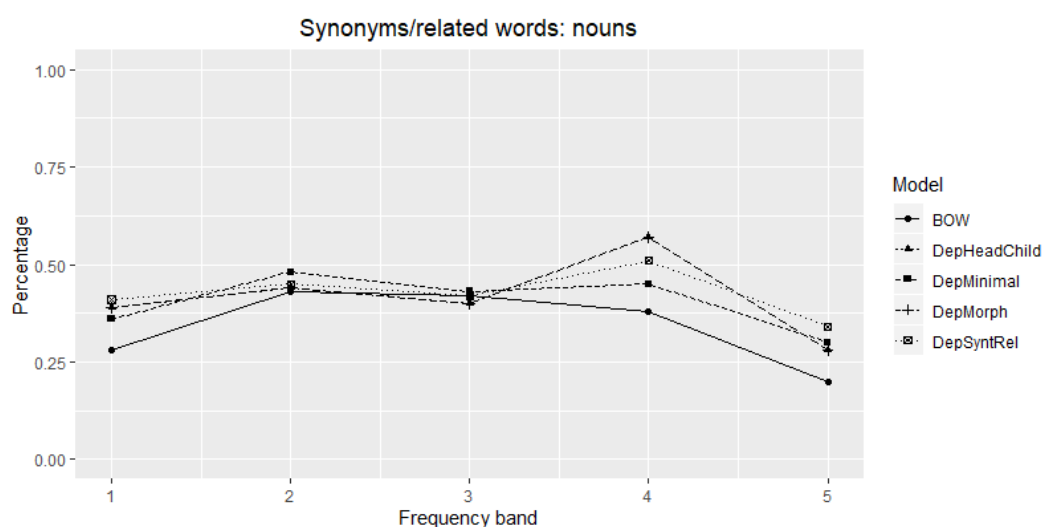


Figure 2: Percentage of synonyms/related words in 10 nearest neighbors by frequency band (nouns).

For almost each model (except the verbs *BOW* model) frequency band 5, containing the lexical items with the lowest frequencies, returns the least number of synonymous/related words in the nearest neighbors. Interestingly, however, the words in the highest frequency band do not seem to substantially outperform the ones in the second to fourth frequency band (or perform even worse, in the case of the nouns). This might possibly suggest a diminishing effect of frequency, i.e. as long as the distributional vectors contain enough observations, adding more data would not have a large effect anymore. Another factor to take in mind is that the highest frequency band contains several words with a quite general and/or abstract meaning, which makes their meaning more difficult to model (see below). These frequency effects seem to be relatively consistent across all 5 distributional models, and any differences are probably caused by random fluctuations.

5.4 Causes of the differences between the various context models

There are two reasons that may explain the limited differentiation between the syntactic models: either these models return the same types of words, or they do not, but the drawbacks of a certain model cancel out its benefits. In order to establish which of these two situations applies, we investigated the degree of overlap of the words that are in the 10 nearest neighbors, as shown in Table 8 (since the numbers for nouns and verbs were almost identical, we did not separate them).

Table 8: Degree of overlap between 10 nearest neighbors returned by each model.

	BOW	DepMinimal	DepHeadChild	DepSyntRel	DepMorph
BOW		54%	52%	43%	42%
DepMinimal	54%		73%	53%	51%
DepHeadChild	52%	73%		61%	56%
DepSyntRel	43%	53%	61%		64%
DepMorph	42%	51%	56%	64%	

This table demonstrates that there is not a high degree of overlap between the nearest neighbors returned by the bag-of-words models on the one hand and the syntactic models on the other hand, with especially the models with syntactic or morphological specification (i.e. *DepSyntRel* and *DepMorph*) returning rather different words. Secondly, there is quite a big degree of overlap between *DepMinimal* and *DepHeadChild*, but far less so with *DepSyntRel* and *DepMorph*. In other words, the lack of quantitative differences between the performance of the different models seems to mask the fact that they do in fact return quite different words in their nearest neighbors.

To further investigate the differences among the vector models, we examined the vectors of the nearest neighbors as compared to the ones of the target words, and identified which features have a high PPMI value in both vectors: these features would have a high influence on the cosine metric. More precisely, we selected a number of pairs of target words and nearest neighbors that were not synonymous or related

(to gain a deeper understanding on why these “erroneous” nearest neighbors words were retrieved). Next, we listed a number of features that were in the top 5% of highest PPMI values for both vectors. Table 9 summarizes these results, containing a (random) selection of these high-ranking features. For comparative purposes, we kept the target word constant.

Table 9: Features in top 5% of PPMI values for target words and their ‘erroneous’ nearest neighbors.

Model	Target word	Neighbor	Example features
BOW (Nouns)	σιωπή “silence”	δικαστής “judge”	καθέζομαι “sit down”, συκοφαντία “sycophancy”, ένθυμέομαι “desire”, φρίκη “shivering”, ήρωικός “heroic”, ακροάομαι “listen to”, μητριά “stepmother”, άτρεμέω “keep still”
BOW (Verbs)	όράω “see”	φεύγω “flee”	όσφραίνομαι “smell”, βδελύσσομαι “be loathsome”, περιβλέπω “look around”, προσπλέω “sail toward”, αίμάσσω “make bloody”, ένεργάζομαι “produce in”, ίππότης “horseman”, γλαυκός “gleaming”
DepMinimal (Nouns)	σιωπή “silence”	δήμος “populace”	καταδικάζω “convict”, καταψηφίζομαι “vote against”, εὐταξία “good order”, καρτερέω “be steadfast”, νεανίας “young man”, κλέω “celebrate”, στένω “groan”, θαύμα “wonder”
DepMinimal (Verbs)	όράω “see”	κάθηναι “sit”	έπιποθέω “desire”, πτήσσω “scare”, άσχημονέω “disgrace oneself”, όλιγάκις “seldom”, άποδειλιάω “be fearful”, προσελαύνω “drive to”, κρεμάννυμι “hang”
DepHead-Child (Nouns)	σιωπή “silence”	κίνδυνος “danger”	άσφαλής/head “safe”, ύποσημαίνω/head “indicate”, συνωθέω/head “force together”, έπιρριπτέω/head “throw oneself”, καρτερέω/head “be steadfast”, ύποπτέω/head “suspect”, γούν/child “at any rate”, πνίγω/head “choke”
DepHead-Child (Verbs)	όράω “see”	ΐστημι “stand”	πόρρωθεν/child “from far”, μακρόθεν/child “from far”, πρόσρημι/head “speak to”, άντα/child “over against”, έγγύθεν/child “from far”, διαταράσσω/head “confuse”, κάθηναι/child “sit”, όρχέομαι/child “dance”
DepSyntRel (Nouns)	σιωπή “silence”	χρόνος “time”	έξίστημι/head/adverbial “change”, καρός/head/coordinate “time”, κατέχω/head/adverbial “hold fast”, άγανακτέω/head/adverbial “be irritated”, παραδίδωμι/head/adverbial “hand over”, ύβριζώ/head/adverbial “maltreat”, έξεστι/head/adverbial “be possible”, δουλεύω/head/adverbial “serve”
DepSyntRel (Verbs)	όράω “see”	φημί “say”	άμελέω/child/object “neglect”, γελάω/child/object “laugh”, έπαίρω/child/object “raise”, τaráσσω/child/object “disturb”, ήσσάομαι/child/object “be inferior”, όρμάω/child/object “start”, κλαίω/child/object “weep”, διαλέγομαι/child/object “converse”
DepMorph (Nouns)	σιωπή “silence”	βία “violence”	παρέρχομαι/head/dative “pass by”, καταψηφίζομαι/head/accusative “vote against”, όχλος/child/genitive “crowd”, κατέχω/head/dative “hold fast”, παρήμι/head/dative “let go”, άποδέχομαι/head/genitive “accept”, ύπέικω/head/dative “withdraw”, συλλαμβάνω/head/dative “collect”
DepMorph (Verbs)	όράω “see”	εύρίσκω “find”	κάθηναι/child/participle_accusative “sit”, άναβαίνω/child/participle_accusative “go up”, χαλεπός/head/infinitive “difficult”, ΐστημι/child/participle_accusative “stand”, διάκειμαι/child/participle_accusative “be”, ρίπτω/child/participle_accusative “throw”, έρχομαι/child/participle_accusative “go”, προσέχω/child/participle_accusative “offer”

These data show that using a simple bag-of-words context model can lead to a large number of spurious associations. The association between δικαστής “judge” and μητριά “step-mother”, for instance, is based on the frequent use of the two words in a rhetorical speech without there being any direct link between the words (e.g. *άχθομαι μὲν οὖν, ὃ ἄνδρες δικασταί, ἐπὶ τῇ μητριᾷ χαλεπῶς ἐχούση* “I am in pain, **men of the jury**, because my **stepmother** is doing badly”). Similarly, the association between

γλαυκός “gleaming” and φεύγω “flee” is based on contexts in which the object of flight is described as γλαυκός, e.g. *γλαυκοῖο φουγῶν Τρίτωνος ἀπειλᾶς* “**fleeing** the threats of the **gleaming** Triton”. It is exactly these kinds of associations that the dependency-based models filter out.⁷

Examining the differences between the *DepMinimal* and *DepHeadChild* model, we can observe that in many cases it is quite obvious what the direction of the arc should be without knowing it in advance. For instance, a verb such as καταδικάζω “convict” would typically be the head of a noun such as σιωπή “silence” and δῆμος “people” and not its child, and an adverb such as ὀλιγάκις “seldom” would typically be the child of a verb such as ὁράω “see” and κάθημαι “sit” rather than its head, so adding the direction of the arc would be superfluous. In some cases adding the direction of the arc might even be detrimental. To give an example, nouns will typically be the head of relative clauses or attributive participles, while in a main clause they would be considered a child of the respective verb. Both ὁράω and θεάομαι “see”, for instance, have a feature κάλλος/head “beauty” with a high PPMI value from sentences such as *κάλλος οἷον οὐπω πρότερον ἐτεθέατο* “such a **beauty** as he **had never seen** before”, in which ἐτεθέατο (from θεάομαι) is considered to be the child of κάλλος, even though it also functions as the object of the relative clause. As a result, in such cases grouping these instances under a single feature “κάλλος” would be more effective.

Even in cases in which there is a clear hierarchical relationship, it is not obvious if this hierarchy is always relevant: in cases with adverbial clauses or participle groups, for instance, such as *ἀναβλέψας τοῖς ὀφθαλμοῖς εἶδεν αὐτὸν τὸν τόπον* “**looking up** with his eyes he **saw** this place” it is clear that the fact that the participle ἀναβλέψας (of ἀναβλέπω, “look up”) is in a dependency relationship with εἶδεν (of ὁράω, “see”) is relevant for the meaning of ὁράω, but it is less obvious that the fact that ἀναβλέψας is a child of εἶδεν is equally meaningful (a sentence such as *ἀνέβλεψε τοῖς ὀφθαλμοῖς καὶ εἶδεν αὐτὸν τὸν τόπον* “he **looked up** with his eyes and **saw** this place” would roughly convey the same meaning). This is not to say that the fact that ἀναβλέπω is in a subordinate relationship is entirely meaningless (otherwise the writer would obviously not have chosen to encode such a subordinate relationship explicitly by the use of the participle), but this might not be an aspect of meaning that is particularly useful to detect word similarity.

However, the direction of the arc is certainly not irrelevant in all cases. For instance, in the list of words that have a high PPMI value with both σιωπή “silence” and δῆμος “people” in the *DepMinimal* model, we can find nouns such as ὄχλος “crowd”, for which ὄχλος is usually the head (or in a coordinate relationship) in the case of δῆμος (e.g. *ὄχλοι παντοίων δῆμων*: “crowds of all sorts of people”), but in

⁷ Of course such less direct dependency links might sometimes be informative as well: in a sentence such as “fleeing the dangerous men”, for instance, the word “dangerous” does provide useful information about the meaning of “flee”. One possible way to include such contexts is to include indirect paths as well (such as *flee > man > dangerous*) and weigh the paths according to their length (as well as the type of syntactic relation), see Padó and Lapata (2007). Meanwhile, words which have no dependency path at all between them, such as *δικαστής* and *μητρική* in the example above, would still be excluded.

the case of *σιωπή* it usually is a child (e.g. *τῶν ὄχλων ἢ σιωπή*: “the silence of the crowds”) – “a crowd of silence” would be atypical. As there is little difference in performance between the two models, the advantages to explicitly code the dependency link on the feature seem to be as important as the drawbacks. Therefore a model that combines the strengths of both models would be preferable, i.e. only encode head/child information when it helps to make relevant semantic distinctions and not when it is e.g. simply related to specific conventions of the dependency-based format.

One way to further refine the dependency-based models is to add further syntactic and morphological labels to it, such as in the *DepSyntRel* and *DepMorph* models. However, a negative effect of such an approach would possibly be data sparsity, seeing that it further subdivides a given feature in several new features which each would be less frequently attested than the feature without label, and we are dealing with a relatively small corpus to start with. This would not be a problem if there was no connection between several syntactic uses of a word, if e.g. the “adverbial” use of word X would be entirely different in meaning from its “object” use: in such a case making this sub-distinction would only help to model meaning distinctions. However, this is clearly not always the case: looking at e.g. the top 5% of features with the highest PPMI values for both *σιωπή* and *σιγή* (both “silence”), we see several re-occurring features with a different syntactic label such as *κατέχω*/adverbial and *κατέχω*/subject, *ἀκούω*/adverbial and *ἀκούω*/object, and so on. One issue is that a specific semantic role can be encoded in different syntactic constructions, such as the patient, which would be encoded as the subject of an active verb but the object of a passive verb. Another issue is that the boundaries between labels such as “object” and “adverbial” are often rather fluid, which becomes increasingly problematic when dealing with an automatic parsing system. While this latter problem is not relevant for constructions that use morphology instead of syntactic relations, the problem of using different syntactic constructions to encode the same semantic role still arises there.

Finally, we can also see an important difference in the type of semantic information that is encoded in the *DepSyntRel* and *DepMorph* models as opposed to the other syntactic models. There does seem to be a greater emphasis on constructions that show a similar syntactic behavior: the nearest neighbors of *ὁράω* show a large number of verbs that are more broadly situated in the evidential domain rather than especially connected with acts of seeing such as *φημί* “claim”, *οἶδα* “know”, *μανθάνω* “learn”, *νομίζω* “think” and so on. Looking at the shared features with high PPMI values, almost all of them are verbal objects, denoting some kind of information that is manipulated, e.g. *ἰδοῦσα δὲ τὰς αἰγας τεταραγμένας* “seeing that the goats **had been disturbed**” and *τεταράχθαι μὲν αὐτήν [...] ἔφη μοι ἡ Θεοπάτρα* “Theopatra **said** to me that she **had been disturbed**”. Using morphology instead of syntactic labels further emphasizes the high co-occurrence of *ὁράω* with participial complementation, which is considered to be more objective than infinitival complementation: therefore verbs such as *νομίζω* “think” are pushed down from the 6th position in the list of nearest neighbors (with *DepSyntRel*) to the 41st (with *DepMorph*), while verbs such as *εὕρισκω* “find” appear in the top 10, from constructions such as *εὕρων*

παῖδα τὸν ἐμὸν **καθήμενον** “**finding** my child **sitting down**” which are quite comparable to something like τὸν Κροῖσον αὐτὸν **ὄρας** ἤδη ἐπὶ κλίνης χρυσοῦς **καθήμενον** “you already **see** Croesus himself **sitting** down on a golden throne”. In such constructions the meaning of ὄραω is in fact quite similar to εὐρίσκω, but the use of such syntactic and morphological features might overemphasize this specific aspect of the meaning of these verbs as opposed to other usages. Similarly, most features of σιωπή in *DepSyntRel* are related to its usage as an adverbial (specifically of manner). Since the label “adverbial” is used as a catch-all term for all sorts of adverbial relations, this can explain the high cosine similarity with χρόνος, which is similarly often used with an adverbial function, even though it is a quite different adverbial relation (of duration rather than manner). Using the morphological rather than the syntactic label further narrows it to usages with the dative case, which is common for manner adverbials (duration is typically expressed in the accusative), but the dative case is still quite broad and can be used to express all sorts of other semantic roles such as instrument (which would be the typical semantic role for βία “violence”). In other words, it is clear that the use of syntactic and morphological features does reveal aspects of meaning that are not present in other models, but it is less obvious that this information is also appropriate for tasks such as word similarity detection.

5.5 Performance with specific words

Next, we took a closer look at how well the models performed overall with specific words. Table 10 summarizes the average performance of some select noun classes across all five word models (the standard deviations per category are between brackets), see ‘Supplementary material’ for the full results. Starting with nouns, one category of nouns that performs particularly well are words in the natural domain: καρδία “heart”, ὀδούς “tooth”, ὄς “pig”, ἴρις “iris flower”, ἡίων “shore”, δελφίς “dolphin”, σκοπία “hill-top”, στρόχνον “winter cherry” and ὀποβάλαμον “balsam” return many synonyms or related words in their nearest neighbors, although this is the less the case with κῦμα “wave”, οὐλή “scar” and ψάμμος “sand”. As a general category, however, these words are clearly easier to model than other nouns, as can be seen in Table 10: the ratio related vs. unrelated words is clearly considerably higher than average (while they return less synonyms, this is probably because most of these words are so specific that they do not have a large number of synonyms to start with). Another group of nouns that seems to be modelled well are nouns referring to people, i.e. ἀδελφός “brother”, νεανίσκος “young man”, λοχαγός “commander” and πολεμιστής “soldier”. However, one of these words (πολεμιστής) performs somewhat worse than average, this category does not contain many words to start with, and the words in this category do have a higher token frequency than average. Concrete objects/structures also perform a little better than average (ἄγαλμα “statue”, πλοῖον “ship”, ταμειῖον “treasury”, ἵππόδρομος “chariot-road”, θήκη “case”, πέδιλον “sandal” and ἰμάσθλη “whip”), while qualities or emotions (ἀλήθεια “truth”, ἡδονή “pleasure”, ἔρις “strife”, ἄχος “distress”, οἶστρος “frenzy”, καλοκάγαθία “nobleness”) perform about average. Finally, the words that are clearly the most difficult to model refer to events or processes: μάχη “fight”, συμφορά “accident”, σιωπή “silence”, ἀνάμνησις

“remembrance”, προσευχή “prayer”, παραφυλακή “guard”, ἐστίασις “feasting”, πρόβασις “increase”, ἔλασις “driving away”, εὐπλοία “fair voyage” and εἰδωλατρία “idolatry”. This is slightly skewed by the outlier παραφυλακή (see also below), which returns on average 7.4 unrelated words, but most of them also have a lower than average ratio of related vs. unrelated words.

Table 10: Mean classification of 10 nearest neighbors per word class, with standard deviations between brackets.

	Synonym	Related	Distantly-related	Same domain	Unrelated
AVERAGE	1.0 (1.2)	3.0 (2.0)	3.3 (2.0)	1.8 (1.5)	0.8 (1.4)
Natural domain (N=12)	0.4 (0.6)	4.1 (2.4)	4.1 (2.3)	1.1 (1.3)	0.3 (0.5)
People (N=4)	0.7 (0.7)	4.2 (1.7)	3.2 (1.1)	1.7 (0.9)	0.3 (0.4)
Concrete objects (N=7)	2.0 (1.8)	2.3 (1.5)	3.6 (1.5)	1.7 (1.2)	0.4 (0.7)
Qualities/emotions (N=6)	0.9 (1.1)	4.5 (2.1)	3.0 (3.0)	0.8 (1.0)	0.8 (1.1)
Events/processes (N=10)	0.9 (1.0)	2.4 (1.4)	2.9 (1.4)	2.2 (1.4)	1.6 (2.1)

As for verbs, it is more difficult to exactly pinpoint a number of semantic classes that perform well, since the results seem more random there. There are some tendencies, however: many verbs that are easy to model refer to some concrete physical action such as οἴχομαι “go away”, ἀπελαύνω “drive away”, σκεπάζω “cover”, κρούω “knock” and ληίζομαι “plunder”. Verbs that belong to the mental domain also perform well (although they are all very frequent) such as δοκέω “seem”, μανθάνω “learn” and κρίνω “judge”. Other than that, there are no clear tendencies, although some bad-performing verbs are semantically quite vague or abstract, or have wide-ranging meanings, such as συμβαίνω (for which the LSJ dictionary lists meanings ranging from “stand with the feet together” to “come to an agreement”, “correspond with”, “to be an attribute of”, “happen” and so on), προαπαντάω (“go forth to meet”, “take steps in advance” or “to be interposed”) and ἀνασκευάζω (“pack up the baggage”, “remove”, “ravage”, “to be bankrupt”, “reverse a decision”, “build again”).

For verbs, these differences are probably best explained by their general semantic properties: it is not surprising that verbs that are semantically quite specific and concrete, e.g. physical contact verbs such as σκεπάζω “cover”, would have more useful context information than very ambiguous verbs such as συμβαίνω (see above), of which its meanings might be too disparate to model with a single vector. Animacy might also be a factor: verbs that have human objects might typically use pronouns or proper names to refer to these human referents, while these physical contact verbs typically have concrete non-animate objects, which might provide these models with more useful context information. This could also explain why verbs with typically verbal complements such as cognitive verbs are modelled well, since these complements are directly expressed as well. This is simply a hypothesis, however, that should be further explored in future research.

As for nouns, the same principles generally hold: nouns that are referentially more abstract such as nominalized processes tend to be modelled quite badly, while very concrete nouns perform well.

However, especially for nouns the influence of genre also seems to be an important factor. The most prominent example are nouns that typically belong to the scientific or natural domain, which were the easiest to model, as discussed above. We can give several reasons for this: first of all, there are many scientific texts in the Greek corpus. The works of four authors, i.e. Galen (medicine), Hippocrates (medicine), Aristotle (philosophy, including biology and physics) and Theophrastus (botany), together consist of 4.6 million tokens, or 1/8 of the total corpus. Secondly, such nouns tend to be well-demarcated, which makes them easier to model than more abstract concepts. Finally, these texts tend to be “definitional”, i.e. they precisely try to describe the concept under question, and as a result many useful context features are provided. See, for instance, some occurrences of the word ἴρις “iris” in Theophrastus’s *Enquiry into Plants*:

- (1) ἀνθεῖ δὲ καὶ ἡ **ἴρις** τοῦ θέρους καὶ τὸ στρούθιον καλούμενον· (...) ὁ μὲν ἀσφόδελος μακρὸν καὶ στενότερον καὶ ὑπόγλισχρον ἔχει τὸ φύλλον, (...), ἡ δὲ **ἴρις** καλαμωδέστερον· (...) ἔνια δὲ ἔχει, καθάπερ ἡ σκίλλα καὶ ὁ βολβὸς καὶ ἡ **ἴρις** καὶ τὸ ξίφιον· (Theophrastus, *Enquiry into Plants* 6.8.3)

The **iris** also blooms in summer, and the plant called soap-wort; (...) Asphodel has a long leaf, which is somewhat narrow and tough, (...), and **iris** one more like a reed. (...) some however have a stem, as squill purse-tassels **iris** and corn-flag (translation A. Hort).

The context features we find in those sentences are clearly suited to demarcate the meaning of ἴρις, e.g. ἀνθεῖ “blooms”, καλαμωδέστερον “more like reed”, and other flowery plants ἴρις coordinates with such as σκίλλα “squill”, βολβός “purse-tassels” and ξίφιον “corn-flag”.

Having more data for a given lemma obviously helps to model its meaning. However, this needs to be nuanced in two ways. First of all, there are situations in which having more data can be more detrimental, if the type of data is not really suited to model the meaning of the target word. This is, for instance, the case for παραφυλακή “guard”, which occurs in the majority of its usages in the papyri (124/149 times) in contexts such as the following:

- (2) παρὰ Αὐ]ρηλίου Παπνουθίου Πκυλίου μητρὸς [...]ιας ἀπὸ ἐπ[ο]ικείου Σεντοποῖο ὑπο [τὴν **παραφυλακὴν** τ[ῶ]ν ἀπὸ κόμης Πτι[μενκυρκ]εω[ς] Προμέν[ων] τοῦ Ἑρμοπολίτου[ο] νομοῦ] (BGU 6 1430)

“Of Aurelius son of Parnuthius son of Pkylius, his mother [...], from the hamlet Sentapouo under the **guard** of the Shepherds from the village Temencyrcis from the Hermopolites nome”

- (3) ἐν περιχώματι Τραισε ὑπὸ τὴν **παραφυλακὴν** τῶν ἀπὸ κόμης Ἄρεως τοῦ Ἑρμοπολίτου νομοῦ (SB 14 11373)

“(...) in the Traise dyke under the **guard** of the people from the Areos village of the Hermopolites nome”

- (4) συσταθεῖς ὑφ’ ὑμῶν εἰς **παραφυλακ(ήν)** [τῆς μητρο]πόλεως (P. Ryl. 2 88)
 “(...) being assigned by you for the guard of the metropolis”

While there are some context elements that may be useful to model the meaning of παραφυλακή, i.e. κώμη “village” and μητροπόλεως “metropolis”, in general these texts are quite formulaic, which has as a result that the same construction might be repeated several times, as in (2) and (3), and that these contexts might be quite generic (especially in texts such as contracts), e.g. “this person has done so and so in this place at this time”, as opposed to contexts such as (1). In other words, it is not only the quantity of the data that matters, but the quality as well: some types of data are clearly more suited to model lexical semantics than others.

Finally, even if we have a large amount of data with useful context features, the vectors we calculate might not always encode the desired semantic information. For instance, looking at the nearest neighbors of words such as πρόβασις “increase” and ἐπισυντίθημι “add successively”, we can see that most words are in the mathematical domain: e.g. διάμετρος “diameter”, ἀριθμός “number” and περίοδος “period, circumference” for πρόβασις and πολλαπλασιάζω “multiply”, διπλόω “double” and μερίζω “divide” for ἐπισυντίθημι. This is probably caused by the fact that the Greek corpus contains a large amount of mathematical material, with a specialized vocabulary (therefore these context features will receive high PPMI values), which pulls the vector toward the mathematical meaning of the word. However, these words have non-technical meanings as well, which might be subdued due to this factor – also note that in our evaluation we considered a word to be “synonymous” or “related” if this was true for at least one meaning, so the fact that some vectors might be “skewed” towards a particular meaning is not measured by the metrics we used above. There are multiple ways to resolve this issue: either by selecting or weighting parts of the corpus so that these non-technical meanings would also be represented, or by abandoning the use of one single vector to represent all meanings and either constructing vectors for specific genres or working with token-based models (see De Pascale 2019 for an application of both strategies in the context of dialectology). At any rate, it is necessary to take a closer look at the question of how the heterogeneity of the Greek corpus impacts the composition of our vector representation in the future.

5.6 Comparison with other benchmarks

As noted in Section 5.1, various other benchmarks for the evaluation of distributional semantic models for Ancient Greek exist, including Ancient Greek WordNet (Bizzoni et al. 2014), an automatically created WordNet for Ancient Greek based on bilingual English-Greek dictionaries, Justus Pollux’s *Onomasticon*, an ancient work from the second century AD describing semantically related words, Schmidt’s (1876-1886) *Synonymik der griechischen sprache* containing lists of Ancient Greek synonyms, and the AGREE benchmark (Stopponi et al. 2023), containing measures of word relatedness

scored by various independent researchers.⁸ All of these benchmarks consist of lists of related words, while the AGREE benchmark also contains a score from 0 to 100 how related these words were considered on average by the scholars.

To evaluate, for each pair that was considered semantically related in the benchmarks, we calculated how high each word of the pair appeared in the list of semantically related words (descending by cosine) of the other word. After that, we calculated the median of all these rankings as a metric of how well the models are able to detect closely semantically related words (we used median instead of mean since in many cases the ranking was very low, which would have a large effect on the mean).⁹ Additionally, for the AGREE benchmark, we calculated the correlation between the ratings of experts and the cosine similarity of the distributional models, using Spearman correlation. The results are presented in Tables 11-12 (since the benchmarks based on Schmidt and Pollux did not contain verbs, only results for nouns are presented there).

Table 11: Median rank of semantically similar word pairs according to the benchmarks among each other's neighbors returned by each word model (not SVD-scaled).

		BOW	DepMinimal	DepHC	DepSyntRel	DepMorph
WordNet	Nouns, N=11631	764	694	701	747	748
	Verbs, N=33015	1228	1188	1182	1177	1185
	Pollux (Nouns, N=2631)	527.5	463	490	547.5	585.5
	Schmidt (Nouns, N=2793)	238	209	209	248.5	278.5
AGREE	Nouns, N=129	23	22	22	25	33
	Verbs, N=67	31	18.5	23.5	27	22.5

Table 12: Spearman correlation of model ratings and expert ratings in the AGREE benchmark.

	BOW	DepMinimal	DepHC	DepSyntRel	DepMorph
Nouns, N=226	0.538	0.538	0.529	0.505	0.511
Verbs, N=234	0.370	0.414	0.420	0.441	0.441

⁸ The three first resources were used by Rodda et al. (2019), as noted in Section 2.2, and a digital (greatly abridged) version of the Onomasticon and Schmidt's lexicon were compiled by them (<https://github.com/alan-turing-institute/ancient-greek-semantic-space>). Ancient Greek WordNet can be found at <http://hdl.handle.net/20.500.11752/ILC-56>. The AGREE benchmark is found at <https://zenodo.org/record/7681749>.

⁹ Both Rodda et al. (2019) and Stopponi et al. (2023) evaluate similarity in terms of precision and recall, comparing the word pairs in the benchmarks to the k (5, 10, 15) nearest neighbors of these target words in distributional models. Recall represents how many of the related words in the benchmarks were included in the list of nearest neighbors, while precision represents how many of the nearest neighbors were included in the benchmarks. However, this seemed problematic to us as 1) the benchmarks are not generally exhaustive, complicating the calculation of precision, since the nearest neighbors might contain related words that are not included in the benchmarks, 2) k is an arbitrary choice, and some words might have much more related words than others and 3) some words might have less than k related words in the benchmarks, complicating the calculation of recall (as also noted by Stopponi et al. 2023).

Firstly, regarding the different benchmarks, the AGREE benchmark was the only resource explicitly created to evaluate distributional models, and it is clear that it is much more suited for this than the other benchmarks: the median word ranked much higher in the list of nearest neighbors than for any other benchmark we evaluated. The only benchmark that was somewhat close was Schmidt's *Synonymik*, and the median word included there was still very low in the list of nearest neighbors of its supposed semantically related words (at place 237 on average across all models) when compared to AGREE (at place 25 on average across all models for nouns). The median rank was in particular very low with Ancient Greek WordNet: Rodda et al. (2019) also noted that this resource did not match the results of their distributional models very well, which is likely an artifact of the substantial level of noise introduced by the automatic creation of this resource.

Regarding model performance, unlike the results discussed in the previous sections, in general there is not a large difference between bag-of-words models and syntactic models when evaluated against these benchmarks, with the bag-of-words model in several cases even performing best. This is true for both the comparison of the ranks of semantically related words as well as the correlation between experts' and models' ratings (although there seems to be a difference between nouns and verbs). Inspecting the data more closely, this is likely because several of these benchmarks contain words that are only related in a very topical way. For example, focusing on the differences between the *BOW* and *DepMinimal* model with the AGREE benchmark (the best performing benchmark), some word pairs that occur much lower on average in each other's list of nearest neighbors in *DepMinimal* vs. *BOW* are νόστος ('return home') vs. θάλασσα ('sea'), νόστος ('return home') vs. ὁδός ('way'), πατήρ ('father') vs. σέβας ('respect'), as well as words that are clearly very closely semantically related but will have a very different syntactic behavior, such as πόντος ('sea') vs. ἀλιεύς ('fisherman'), ῥῆσις ('speech') vs. ἀγορά ('marketplace', 'assembly'), and πρέσβυς ('old man') vs. ἡλικία ('age').¹⁰

6 Discussion

The aim of this study was to test the validity of distributional semantic models for Ancient Greek – and presumably, the results can be expanded to other highly inflectional and historical languages as well – in particular by focusing on the type of context features that are suited best to model lexical semantics. These context features involved an increasing level of analysis, ranging from (1) a simple 4 words window bag-of-words model, to all words that are in a dependency relationship, both excluding (2) and including (3) the direction of the dependency arc and the dependency relationship with a syntactic (4) and morphological (5) label (see Table 2).

¹⁰ Additionally, there were some words that are morphologically nouns but semantically adjectives that are typically combined with the other noun in the pair, such as ναῦς ('ship') vs. κορῶνις ('curved') and πόντος ('sea') vs. οἴνοψ ('wine-colored').

To evaluate the results of these different distributional models, we investigated how useful the (raw, PPMI-weighted) vectors are to detect word similarity, and what types of similarity they detected, by a (subjective) labeling of the nearest neighbors retrieved by each vector model. We found that dependency-based vectors are much better suited to return synonymous and/or taxonomically related words than a simple bag-of-words context model. This is especially striking since we used automatically parsed data, which still had a considerable error rate. The importance of using syntactic dependencies is likely caused by the free word order of Greek, since the relevant contextual information might not always be present in a small context window of preceding or following words.

Among the different dependency-based models, on the other hand, the differences are less pronounced. There are several reasons for this: (a) some technicalities of the dependency format (e.g. how coordination structures are encoded) create differences that are linguistically meaningless; (b) the direction of the arc might not always correspond to a meaningful relationship, at least not for the purpose of detecting word similarity (e.g. participles modifying other verbs); (c) some syntactic contrasts might in some cases be rather arbitrary (e.g. “adverbial” vs. “object”); (d) differences in syntactic structure do not always have a one-to-one correspondence to meaning differences (e.g. the object of an active construction and the subject of a passive construction both correspond to the patient or theme of the same verb); and (e) using syntactic and morphological features could introduce some high-level information about the syntactic usage of a word (e.g. the complementation patterns in which it typically takes part) which might not in all cases be optimal to detect word similarity. As a result, adding a too large amount of linguistic analysis could lead to data sparsity by dividing features in several sub-features of which the contrasts between them are not that significant. This is not to say that using a higher level of linguistic analysis is entirely detrimental: as there are no big quantitative differences between the different dependency models, it is rather the case that the benefits and the drawbacks of an increasing level of analysis outweigh each other. Therefore in the future it would be worthwhile to take a closer look at the different levels of granularity of specific labels and decide in which cases it would be beneficial for the detection of semantic similarity to make more fine-grained distinctions and in which cases it would not. Another, more automated way to reduce such “artificial” differences is to use a dimension reduction technique such as SVD, by including labels of various levels of granularity together in the PPMI matrix and letting the dimension reduction detect the most relevant distinctions.

Evaluating our results against independent benchmarks, we found that the difference between bag-of-words and syntactic models was less pronounced there, likely because these benchmarks contain several topically related words for which the syntactic models would reduce the strength of the association.

There are several ways to expand on this current work. First of all, we have shown that a wide mix of context features, i.e. bag-of-words context features, dependencies, syntactic relations and inflectional morphological features, all encode useful information for distributional semantic modelling. We could also add derivational morphological features to this list, which has already been noticed by Boschetti

(2010), but which we did not consider here due to a lack of derivational morphological annotation in the corpora we used. While we created a separate model for each of these categories of features, it would be useful to integrate the strengths of each of them in a single model, as detailed above.

Secondly, while this paper was specifically concerned with type-level distributional models, it would be useful to apply these insights to token-level models as well. Detecting word similarity on the type level ignores the fact that some words may be highly similar with respect to one meaning but highly dissimilar with respect to another meaning. Additionally, this study exclusively made use of a context-count architecture, which has been shown to perform inferiorly in comparison with context-predict architectures: therefore it will be useful to compare results with the latter models as well, both on the type level (e.g. *word2vec*, see also Stopponi et al. 2023) and on the token level (e.g. *RoBERTa*, see also Riemenschneider and Frank 2023).

Finally, we have shown that the lack of homogeneity of the Greek corpus with regard to genre is an important open problem – probably even more important than diachrony, seeing that many late literary writers wrote in a style similar to Classical Attic Greek. For many words the meaning is highly dependent on and/or predictable by the type of text in which they are used, and therefore their vectors can be skewed toward the meaning in some genres that are overrepresented in the corpus. In other words, this problem is highly related to the polysemy problem, and token-based models may therefore also be used to identify such genre-specific meanings. What is more, some text types provide more useful context features than others, e.g. highly descriptive scientific texts vs. formulaic texts such as contracts. As a result, even using more in-domain data might be detrimental if these data are less useful from a practical point of view (e.g. repetitive contexts). While this paper involved a very general task, in the future it will be necessary to take a closer look at the genre composition of the corpus from which the vectors are created, and filter out texts that are less suited for the task on hand or reduce their influence in some other way (e.g. by weighting them).

Acknowledgements

This paper is a thoroughly revised and updated version of a part of the PhD research of the first author (*A Computational Approach to the Greek Papyri*), supervised by the second author. We want to thank Toon Van Hal for his constructive feedback and his extensive work in annotating the dataset used in this study. We are also very grateful to the first reviewer for their constructive feedback, which has greatly improved the quality of this paper. Finally, we would like to thank the team of Martina Rodda, Philomen Probert and Barbara McGillivray, as well as Silvia Stopponi, for the highly valuable resources they created to evaluate distributional models which were also used in this study.

Supplementary material

All the material produced by this research can be found on <https://github.com/alekkeersmaekers/greek-count-vectors>

References

- Baroni, M., Dinu, G., Kruszewski, G.** (2014). Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247. Baltimore.
- Bizzoni, Y., Boschetti, F., Diakoff H., Del Gratta, R., Monachini, M., Crane, G. R.** (2014). The Making of Ancient Greek WordNet. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1140–1147. Reykjavik.
- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.** (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.
- Boschetti, F.** (2010). *A Corpus-Based Approach to Philological Issues*. University of Trento. (PhD thesis)
- Botha, J., Blunsom, P.** (2014). Compositional Morphology for Word Representations and Language Modelling. In: *International Conference on Machine Learning*, pp. 1899–1907. Beijing.
- Bullinaria, J. A., Levy, J. P.** (2007). Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39(3), pp. 510–526.
- Bullinaria, J. A., Levy, J. P.** (2012). Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD. *Behavior Research Methods*, 44 (3), pp. 890–907.
- Burns, P. J.** (2019). Building a Text Analysis Pipeline for Classical Languages. In: Berti, M. (Ed.). *Digital Classical Philology*, pp. 159–176. Berlin: De Gruyter Saur.
- Croft, W.** (2013). Radical Construction Grammar. In: Hoffmann, T., Trousdale, G. (Eds.). *The Oxford Handbook of Construction Grammar*, pp. 211–232. Oxford: Oxford University Press.
- De Pascale, S.** (2019). *Token-Based Vector Space Models as Semantic Control in Lexical Lectometry*. University of Leuven. (PhD thesis)
- Dik, H.** (1995). *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus*. Amsterdam: Gieben.
- Erk, K.** (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10), pp. 635–653.
- Firth, J. R.** (1957). A Synopsis of Linguistic Theory, 1930-1955. In: *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Blackwell.

- Heylen, K., Peirsman, Y., Geeraerts, D., Speelman, D.** (2008). Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pp. 3243–3249. Marrakech.
- Jones, H. S., Liddell, H.G., MacKenzie, R., Scott, R., Thompson, A. A.** (1996). *A Greek-English Lexicon*. Oxford: Clarendon.
- Keersmaekers, A.** (2020). The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In: *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pp. 39–50. Online.
- Kolb, P.** (2009). Experiments on the Difference between Semantic Similarity and Relatedness. In: *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pp. 81–88. Odense.
- Lapasa, G., Evert, S.** (2014). A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, 2, pp. 531–546.
- Lenci, A.** (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4, pp. 151–171.
- Levy, O., Goldberg, Y.** (2014). Dependency-Based Word Embeddings. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308. Baltimore.
- Lin, D.** (1998). Automatic Retrieval and Clustering of Similar Words. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pp. 768–774. Montreal.
- List, N.** (2022). How Can We Investigate Ancient Greek Categories Without the Influence of Our Own? Exploring Kinship Terminology Using Word2Vec. *International Journal of Lexicography* 35(2), pp. 137–152.
- Luong, T., Socher, R. Manning, C.** (2013). Better Word Representations with Recursive Neural Networks for Morphology. In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113. Sofia.
- McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., Vatri, A.** (2019). A Computational Approach to Lexical Polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4): pp. 893–907.
- Mercelis, W., Van Hal, T., Keersmaekers, A.** (Forthcoming). Tongue, Language or Noise? Word Sense Disambiguation in Ancient Greek with Corpus-Based Methods. In: *International Colloquium of Ancient Greek Linguistics*. Madrid.
- Padó, S., Lapata, M.** (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2): pp. 161–199.
- Peirsman, Y., Heylen, K., Geeraerts, D.** (2008). Size Matters: Tight and Loose Context Definitions in English Word Space Models. In: *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pp. 34–41. Hamburg.

- Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., McGillivray, B.** (2021). Lexical semantic change for Ancient Greek and Latin. In: Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., Hengchen, S. (Eds.). *Computational approaches to semantic change*, pp. 287–310. Berlin: Language Science Press.
- Riemenschneider, F., Frank, A.** (2023). Exploring Large Language Models for Classical Philology. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15181–15199. Toronto.
- Rodda, M. A., Senaldi, M. S. G., Lenci, A.** (2017). Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1): pp. 11–24.
- Rodda, M. A., Probert, P., McGillivray, B.** (2019). Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique des Langues*, 60p.(3), pp. 63-87.
- Schmidt, J. H. H.** (1876–1886). *Synonymik Der Griechischen Sprache*. Leipzig: Teubner.
- Singh, P., Rutten, G., Lefever, E.** (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 128–137. Punta Cana (online).
- Spanopoulos, A. I.** (2022). *Language Models for Ancient Greek*. National and Kapodistrian University of Athens. (BA thesis)
- Stopponi, S., Pedrazzini, N., Peels, S., McGillivray, B., Nissim, M.** (2023). Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work. In: *Proceedings of the 1st International Workshop on Ancient Language Processing (ALP) at RANLP*. Varna.
- Turney, P. D., Pantel, P.** (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37: pp. 141–188.
- Vatri, A., McGillivray, B.** (2018). The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 3(1): pp. 55–65.
- Yamshchikov, I., Tikhonov, A., Pantis, Y., Schubert, C., Jost, J.** (2022). BERT in Plutarch’s Shadows. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6071–6080. Abu Dhabi.

Active or descriptive: Textual activity and its dynamic changes of Ph.D. theses across disciplines

Shuyi Amelia Sun¹ , Wei Xiao^{2,3*} 

¹ School of Foreign Language Education, Jilin University, China

² Research Center for Language, Cognition and Language Application, Chongqing University, Chongqing, China

³ School of Foreign Languages and Cultures, Chongqing University, China

* Corresponding author's email: xiaoweiyx@126.com

DOI: https://doi.org/10.53482/2023_55_411

ABSTRACT

As an innovative and systematic genre in the academic community, Ph.D. theses have been heatedly researched in the field of English for Academic Purposes. Although research on the functional and formal features of Ph.D. theses has been abundant, their stylometric traits regarding textual activity have not been explored. Accordingly, this study explored the textual activity of Ph.D. theses and its dynamic changes across natural sciences, social sciences and humanities. A total of 150 Ph.D. theses (50 from each discipline) were analyzed, and the Q and χ^2 values were calculated to determine the textual activity of theses as well as its dynamic changes with the progression of texts. The results showed that, although the theses were found to be active in general, significant differences across disciplines do exist, in that the theses in natural sciences and humanities were more active while those in social sciences were more likely to lean towards the descriptive mode. This study has implications for widening the scope of cross-disciplinary academic genre analyses from an innovative quantitative linguistic perspective.

Keywords: textual activity, stylometrics, Ph.D. theses, disciplinary academic writing.

1 Introduction

With the rapid development of English for Academic Purposes (hereinafter EAP), a large body of research has looked into Ph.D. theses and their disciplinary linguistic features (Paltridge and Starfield 2020). As an innovative and systematic academic genre, Ph.D. theses reflect the frontiers and trends of an academic community (Xiao and Sun 2020). Considering academic writing is specific to the discipline and manifests variations among different academic communities (Xiao et al 2022, 2023a; Hyland 2012; Jiang 2022), cross-disciplinary research on Ph.D. theses can shed light on the textual variations across different disciplinary communities, explore how knowledge is rhetorically constructed and negotiated within each academic community, and provide more empirical evidence to support the pedagogy and practice of Ph.D. theses.

To date, previous studies on Ph.D. theses have mainly explored their functional and/or formal features by scrutinizing particular sections (e.g. the ‘introduction’ section, see Kawase 2018). The functional perspective concentrates primarily on ways to achieve certain communicative goals with appropriate linguistic resources, and the formal perspective is generally devoted to lexical/syntactic features drawing on manual coding or text-mining approaches. In light of the consensus that knowledge is constructed and negotiated within each discipline (Hyland 2012), the above-mentioned perspectives have been gradually filtering through to (cross-)disciplinary research. For example, Bunton (2002) explored generic moves in Ph.D. thesis introductions and found variations on specific steps across fields of science and technology, humanities, and social sciences. Hyland (2008) studied the forms, structures and functions of four-word clusters, providing evidence for the distinctive discipline-specific idiosyncracies of clusters in Ph.D. theses. Xiao and Sun (2020) investigated the lexical features of Ph.D. theses across disciplines, suggesting significant differences regarding lexical diversity and richness between natural sciences and humanities.

Despite the fruitful findings, little is yet known about the discipline-specific stylometric features of Ph.D. theses. Style generally refers to linguistic characteristics that people tend to express via spoken and/or written communication (Popescu et al. 2014). In the field of quantitative linguistics, style is taken as a quantifiable trait of language that can be detected using statistical techniques, and the statistical measurement of style is referred to as stylometrics (Schreibman et al. 2008). Among the stylometric features, textual activity is an important one that depicts activity-descriptivity (dis)equilibrium, i.e. whether texts tend to be active (plotted with substantial verbs) or descriptive (embellished with rich adjectives) (Jiang et al. 2020). To date, most studies on textual activity have focused on political and literary texts (Kubát and Čech 2016; Melka and Místecký 2019; Zörnig and Altmann 2016). For example, Kubát and Čech (2016) analyzed 50 US presidential inaugural speeches, and found that presidential speeches were influenced by speaker’s style and social affairs, such as wartime and financial crisis. Melka and Místecký (2019) explored the textual activity of Beam Piper’s novelette *Omnilingual*. Their findings suggested that most chapters of the novelette were highly active, which could be accounted for by the author’s stylistic preference, 20th-century fictions’ common features and the sub-genre conventions.

Previous studies on textual activity have been confined mostly to political and literary topics, whereas the embodied regularities are expected to be figured out by exploring more genres (Čech and Kubát 2016; Chen and Liu 2018), such as Ph.D. theses. An investigation into the textual activity of Ph.D. theses across disciplines can reveal their stylometric features and shed light on the construction and negotiation of disciplinary discourse. Besides, it should be noted that previous studies on Ph.D. theses tended to concentrate on only selected section(s), probably due to the compromise made between manual coding/annotation and the sheer size of Ph.D. theses (Thompson 2013). Although looking into separate sections can be more focused, it would lead to fragmented knowledge of how they are constructed

in the entirety (Kanoksilapatham 2015). Only a full-length analysis of Ph.D. theses can capture their global features (Xiao and Sun 2020). In addition, among the handful of existing text-mining research on Ph.D. theses, little attention has been paid to the dynamic changes of quantitative properties as texts progress, which fails to reveal how the text as a system regulates itself as it develops (Zörnig and Altmann 2016). Investigating the dynamic development of Ph.D. theses' textual activity can reveal the whole picture as to how Ph.D. theses manifest itself from a macro-perspective and how the disciplinary academic discourse stylometrically governs itself into the complex adapted system (Liu et al. 2017).

To address these issues, we would attempt to investigate the textual activity and its dynamic changes of Ph.D. theses across natural sciences, social sciences and humanities. The research questions are as follows:

- (1) What are the textual activity features of Ph.D. theses? Is there any variation across natural sciences, social sciences, and humanities?
- (2) How does the textual activity of Ph.D. theses change dynamically with the progression of texts? Is there any cross-disciplinary difference?

2 Material

Ph.D. theses were collected using the ProQuest (Clarivate 2023) search engine¹. The selected Ph.D. theses satisfied the criteria that: (1) they were completed by doctoral candidates enrolled in the Ivy League universities in the U.S., (2) they were submitted to the universities within the recent ten years, (3) they were similar in length (30,000 words), and (4) they were organized in a typical 'Introduction-Literature Review-Methods-Results-Discussion-Conclusion' structure. The criteria were to ensure the validity and comparability of language material across disciplines. 50 theses were selected to represent natural sciences, social sciences, and humanities (Kagan 2009) respectively, and thus a total of 150 Ph.D. theses were enrolled.

These Ph.D. theses were first converted into plain texts using AntFileConverter (Anthony 2017) and then cleaned of the sections of abstract, acknowledgments, references and appendices. Details of the corpus are presented in Table 1. The one-way ANOVA test showed no significant difference in text length among the three disciplines ($p > .05$).

Table 1: Corpus information.

¹ ProQuest search engine for dissertations and theses can be accessed via the following link: <https://about.proquest.com/en/dissertations/>.

Discipline	Number of texts	Word count	Average text length
Natural sciences	50	1,552,615	31,052
Social sciences	50	1,641,342	32,827
Humanities	50	1,934,457	38,689
Total	150	5,128,414	34,189

3 Methodology

3.1 Indices and Formulas

The textual activity of Ph.D. theses was measured using Busemann's (1925) Q , rendered as:

$$(1) \quad Q = V/(V + A)$$

in which V and A are sums of verbs and adjectives respectively and Q stands for textual activity. The indicator draws on the assumption that texts are remarkably characterized by either action or description. As such, a more narrative text (e.g. short stories or fairy tales) is usually higher in the value of activity than a more descriptive one (e.g. rhetorically picturing a scenery in a travel book).

Based on Formula (1), textual activity can be roughly classified as active, neutral and descriptive (Zörnig et al. 2015). To be more precise, a chi-square test (see below) is suggested to be employed in combination (Melka and Místecký 2019).

$$(2) \quad \chi^2 = \frac{(V-A)^2}{A+V}$$

Based on the two indices, textual activity can be classified into five categories (cf. Table 2).

Table 2: Categories of textual activity.

Conditions	Textual activity
$Q > 0.55$ & $\chi^2 > 3.84$	significantly active (SA)
$Q > 0.55$ & $\chi^2 < 3.84$	active (AC)
$0.45 < Q < 0.55$	neutral (N)
$Q < 0.45$ & $\chi^2 < 3.84$	descriptive (DE)
$Q < 0.45$ & $\chi^2 > 3.84$	significantly descriptive (SD)

3.2 Data Analysis

We first calculated Q and χ^2 based on the full-length Ph.D. theses and accordingly identified the textual activity traits of full-length Ph.D. theses. Regarding dynamic changes, we calculated Q and χ^2 within each Ph.D. thesis upon accumulated text sizes that increase by 1000 words to figure out the textual activity of Ph.D. theses and the dynamic changes as texts progress. Then, we performed ANOVA tests

to examine whether the cross-disciplinary variations are statistically significant and further employed the TukeyHSD post-hoc analysis (Tukey 1949) to identify precisely where the significant difference lies².

4 Results

As to the full-length Ph.D. theses, the results show that the Q-value is relatively higher in natural sciences ($M=0.587$, $SD=0.049$) and humanities ($M=0.591$, $SD=0.056$), while it is lower in social sciences ($M=0.567$, $SD=0.048$).³ The ANOVA suggests a significant effect of discipline on the Q-value ($F(2, 147)=3.130$, $p<.05$). A post-hoc test of multiple comparisons further shows significant variation between social sciences and humanities ($p<.05$, 95% CI [-0.044, -0.004]). The χ^2 -value is higher in humanities ($M=400.988$, $SD=394.341$) compared with those in natural sciences ($M=260.628$, $SD=293.543$) and social sciences ($M=197.562$, $SD=266.164$). The ANOVA suggests a significant effect of discipline on χ^2 ($F(2, 147)=5.205$, $p<.01$). A post-hoc test of multiple comparisons shows noted variation between social sciences and humanities ($p<.01$, 95% CI [-0.044, -0.004]).

Based on the two indices, the majority of Ph.D. theses were found to be significantly active, and a minority were found to be neutral (cf. Figure 1). To be specific, the significantly active theses in natural sciences account for the largest proportion (82%), while neutral ones take up only 18%. In humanities, 74% of Ph.D. theses are significantly active, and 26% of them are neutral. Ph.D. theses of social sciences present a balanced distribution, where significantly active theses take up 56% and neutral ones account for 44%.

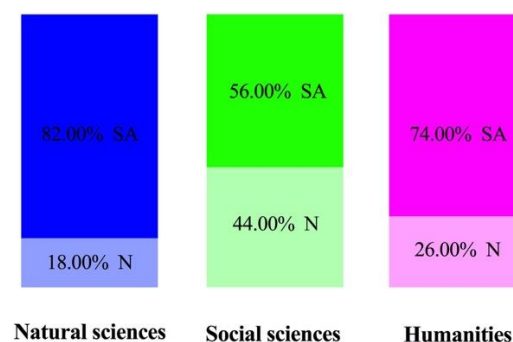


Figure 1: Textual activity of full-length Ph.D. theses. ‘SA’ stands for significantly active, and ‘N’ stands for neutral.

As to the dynamic changes, the mean Q-values alongside standard error of the mean (SEM, depicted as shadows) of each discipline are plotted in Figure 2, and the ANOVA and post-hoc results are shown in

² The procedure was adjusted by the Bonferroni correction.

³ M and SD represent ‘mean’ and ‘standard deviation’ respectively.

Table 3. The Q-values are low when texts are not lengthy and become higher as texts progress. As to disciplinary variations, the Q-values of humanities are significantly higher at the beginning (Chunks 1-2, $ps < .05$). After that, the curves of humanities and natural sciences gradually overlap, while that of social sciences tends to diverge, with Q-values significantly lower (Chunks 14-17, 23-24 and 26-28, $ps < .05$).

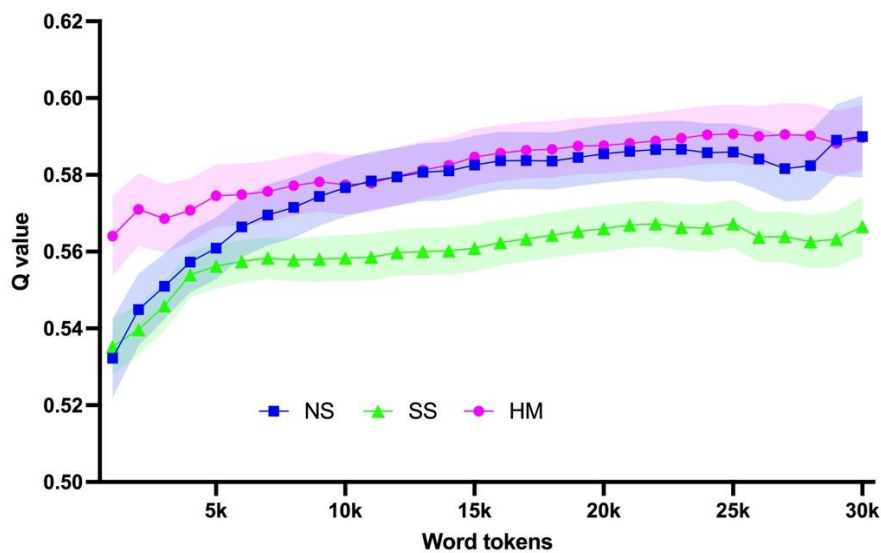


Figure 2: Q-value curves. ‘NS’, ‘SS’ and ‘HM’ represent natural sciences, social sciences and humanities respectively. The abbreviations have also been adopted in Figures 3 & 4 and Tables 3 & 4 below.

Table 3: Dynamic changes of Q-values across disciplines. F stands for the F -ratio. F -ratio would be close to 1 if the null hypothesis is true (i.e. no statistically significant variation lies across disciplines), and a larger F -ratio means that the variation among disciplinary groups is more than the possibility to see by chance (i.e. null hypothesis is rejected or statistically significant variation lies across disciplines). p stands for the p -value, which is to test the null hypothesis that data from all

disciplinary groups are drawn from populations with identical means. The two abbreviations are also adopted in Table 4.

Asterisks (*) are intended to flag levels of significance. If the p -value is less than 0.05, it is flagged with a star (*).

Chunk	Mean			F	p	Mean Difference					
	NS	SS	HM			NS-SS	NS-HM	SS-HM			
1	0.5322	0.5353	0.5641	3.432	*	0.035	-0.0031	-0.0318	*	-0.0288	
2	0.5449	0.5397	0.5710	3.809	*	0.024	0.0052	-0.0261		-0.0313	*
3	0.5510	0.5459	0.5686	2.208		0.114	0.0051	-0.0176		-0.0227	
4	0.5573	0.5540	0.5708	1.430		0.243	0.0033	-0.0135		-0.0168	
5	0.5609	0.5562	0.5746	1.654		0.195	0.0047	-0.0137		-0.0184	
6	0.5664	0.5575	0.5749	1.431		0.242	0.0089	-0.0085		-0.0174	
7	0.5696	0.5583	0.5757	1.494		0.228	0.0113	-0.0061		-0.0174	
8	0.5715	0.5578	0.5772	1.930		0.149	0.0137	-0.0057		-0.0194	
9	0.5744	0.5580	0.5782	2.254		0.109	0.0164	-0.0038		-0.0202	
10	0.5767	0.5583	0.5774	2.298		0.104	0.0184	-0.0007		-0.0191	
11	0.5784	0.5585	0.5778	2.526		0.083	0.0199	0.0006		-0.0193	
12	0.5795	0.5598	0.5796	2.619		0.076	0.0197	-0.001		-0.0198	
13	0.5807	0.5600	0.5813	2.925		0.057	0.0207	-0.0006		-0.0213	
14	0.5810	0.5602	0.5825	3.135	*	0.046	0.0208	-0.0015		-0.0223	
15	0.5825	0.5609	0.5846	3.467	*	0.034	0.0216	-0.0021		-0.0237	*
16	0.5836	0.5623	0.5857	3.401	*	0.036	0.0213	-0.0021		-0.0234	
17	0.5837	0.5633	0.5864	3.313	*	0.039	0.0204	-0.0027		-0.0231	
18	0.5836	0.5643	0.5866	3.048		0.050	0.0193	-0.003		-0.0223	
19	0.5845	0.5653	0.5875	3.006		0.053	0.0192	-0.003		-0.0222	
20	0.5855	0.5660	0.5876	2.902		0.058	0.0195	-0.0021		-0.0216	
21	0.5861	0.5669	0.5882	2.867		0.060	0.0192	-0.0021		-0.0213	
22	0.5866	0.5673	0.5889	2.920		0.057	0.0193	-0.0023		-0.0216	
23	0.5867	0.5664	0.5896	3.197	*	0.044	0.0203	-0.0029		-0.0232	
24	0.5858	0.5661	0.5904	3.290	*	0.040	0.0197	-0.0046		-0.0243	*
25	0.5860	0.5673	0.5907	2.987		0.054	0.0187	-0.0047		-0.0234	
26	0.5841	0.5638	0.5900	3.377	*	0.037	0.0203	-0.0059		-0.0262	*
27	0.5816	0.5640	0.5905	3.079		0.050	0.0176	-0.0089		-0.0265	*
28	0.5824	0.5626	0.5902	3.198	*	0.044	0.0198	-0.0078		-0.0276	*
29	0.5890	0.5632	0.5882	2.915		0.059	0.0258	0.0008		-0.0250	
30	0.5900	0.5666	0.5898	2.039		0.136	0.0234	0.0002		-0.0232	

The mean χ^2 -values of each discipline are plotted in Figure 3, and the χ^2 -values results are shown in Table 4. The χ^2 -values are low when texts are not lengthy and become higher as texts progress. The increase of the χ^2 -values is in fact due to the property of the indicator which generally increases as the sample size becomes larger (Mačutek and Wimmer 2013).

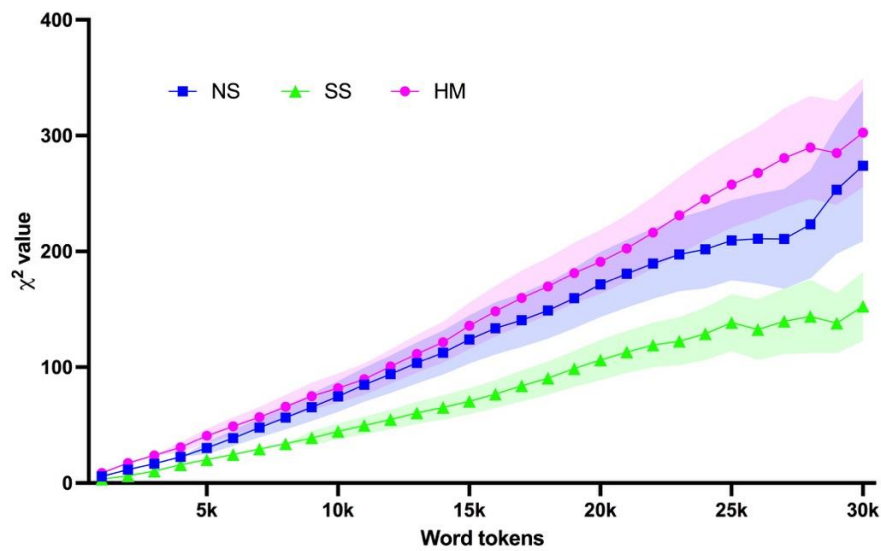


Figure 3: χ^2 -value curves.

Table 4: Dynamic changes of χ^2 -values.

Chunk	Mean		
	NS	SS	HM
1	5.702	3.387	8.716
2	11.516	6.362	17.253
3	16.663	10.297	23.785
4	22.528	15.718	30.841
5	30.299	20.277	40.764
6	38.896	24.657	49.149
7	47.934	29.266	56.993
8	56.630	33.838	65.913
9	65.509	39.046	75.027
10	74.745	44.768	82.014
11	84.961	49.786	89.674
12	94.000	54.998	100.612
13	103.675	60.604	111.571
14	112.682	65.416	121.538
15	123.908	70.553	135.859
16	133.675	76.901	148.336
17	140.601	83.964	159.753
18	148.956	90.584	169.661
19	159.648	98.706	181.283
20	171.447	106.343	191.071
21	180.746	113.133	202.493
22	189.570	119.266	216.336
23	197.489	122.594	231.119
24	201.755	128.856	245.169
25	209.485	138.685	257.769
26	210.953	132.572	267.733
27	210.696	139.751	280.554
28	223.476	143.855	289.773
29	253.300	137.983	284.917
30	274.061	152.669	302.592

Based on the joint conditions of Q and χ^2 , we can determine the dynamic changes of textual activity as texts progress. First, we assigned each chunk in each text a category of textual activity. We then calculated the percentages of texts in each category at each chunk. For example, at the first chunk in natural sciences, 28% of the texts are significantly active, 12% active, 50% neutral, 2% descriptive, and 8% significantly descriptive, and so forth (see Figure 4). It should be noted that the dynamic changes of textual activity in Figure 4 were not counted cumulatively in itself. Instead, as stated in Section 3.2, textual activity was determined by Q and χ^2 which were calculated within each Ph.D. thesis upon accumulated text sizes that increase by 1000 words.

As shown in Figure 4, at the beginning of theses, natural sciences are the least active and humanities are the most active, as is shown by the proportions of significantly active theses in each discipline. As texts progress, humanities remain active and natural sciences become even more active. Although social sciences drift towards the active mode, the change tendency is rather slow. Such tendencies last till the end of theses in that the significantly active theses in natural sciences (c.a. 80%) and humanities (c.a. 70%) far outnumber the neutral ones, suggesting an obvious active trend, whereas a considerable proportion (c.a. 40%) of theses in social sciences are neutral, suggesting a shift to the descriptive mode compared with the other two disciplines.

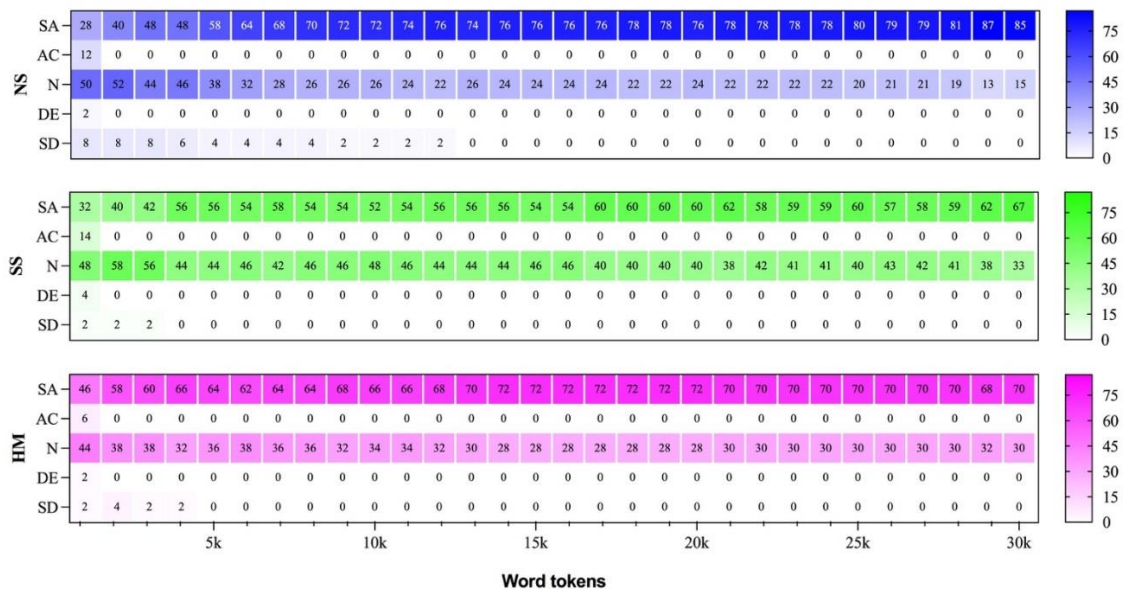


Figure 4: Dynamic changes of textual activity across disciplines.

5 Discussion

The present study aims to investigate the textual activity and its dynamic changes of Ph.D. theses across the natural sciences, social science and humanities. To this end, full-length texts were split into 1000-word chunks, and the Q and χ^2 values were calculated upon accumulated text sizes that increase by 1000 words to figure out the textual activity of Ph.D. theses as well as the dynamic changes with the progression of texts. According to our results, the Ph.D. theses were found to be active in general. However, disciplinary variations could still be witnessed, in the way that the theses in natural sciences and humanities were more active than those in social sciences. As for the dynamic changes, natural sciences are the least active and humanities are the most active at the beginning of theses. As texts progress, humanities remain active and natural sciences become even more active. Although social sciences drift towards the active mode, the change tendency is rather slow.

Our finding that Ph.D. theses were mostly found to be significantly active is in line with Xu and Jiang (2021) who also found that the academic genre is “generally active” (p. 118). Such a finding could be accounted for by the observation that verbs are central to the overall structure of sentences and play a pivotal role in sentences (Baker 2003). In the construction of sentences, verbs arguably carry the largest amount of syntactic and semantic information (Baker 2003; Goldberg 1995; Liu 2009), while adjectives are comparably dispensable in syntax and more likely to work just as modifiers (Jia and Liang, 2020; Zhou et al. 2022). In academic writing, although writers may adopt appraisal resources (e.g. *significant*, *satisfying*) to construct authorial stances and engage with readers (Hood 2006; Martin and White 2005), these adjectives usually occur alongside verbs (e.g. *it is significant to*), and writers would avoid an overuse of adjectives for it is the trustworthy contents rather than rhetorics that determine the quality of PhD theses (Sun and Crosthwaite 2022a, 2022b; Xiao et al. 2023b).

Regarding disciplinary variations, we found that the natural sciences and humanities, which use entirely different methodologies and discuss scientific evidence differently, are counter-intuitively close together in terms of textual activity. This closeness may be accounted for by their narrative nature. In natural sciences, knowledge is taken as a plain matter of facts and the procedures of uncovering knowledge depend on the accumulation of empirical inquiry (Kuteeva and Airey 2014). Theses in natural sciences would put more emphasis on the report of operating procedures, statistical/empirical results, strategies and activities. The language style, then, could be regarded as a typical narrative one that avoids rich adjectival embellishments (Jiang et al. 2020), giving rise to the rapid increase of activity in natural sciences. In humanities, knowledge is regarded as constructed interpretations due to the complicated nature of human beings (Kuteeva and Airey 2014). Thesis writers in humanities tend to resort to a wide range of multi-dimensional perspectives (Xiao et al. 2023a; Zhao et al. 2023; Coffin and Hewings 2003). For example, in English studies, students are generally required to interpret the message or themes of a literary text and support their interpretation by referring to the text as well as to literary

critics. In history, students are frequently expected to evaluate the plausibility of an interpretation of past events and to draw on documentary sources as evidence for their proposition (Coffin and Hewings 2003). The feature of multi-dimensionality requires the incorporation of a variety of “external facts” (Jiang et al. 2020, p. 10), which may result in the overtly narrative nature and the active style of Ph.D. theses in humanities.

In addition, we also found that Ph.D. theses in social sciences are more descriptive (less active) than the other two disciplines. The possible explanation may be that both natural sciences and humanities have a long tradition and are highly developed, while social sciences, as a combination of methods as in natural sciences and objects as in humanities, are lately emerging ones, and thus do not feature such a long tradition. From this perspective, the mid-way of social sciences can be regarded as in sharp contrast to natural sciences and humanities. The above-mentioned uniqueness of social sciences has been documented in some previous studies (Coffin and Hewings 2003; Flowerdew 2015; Paltridge and Starfield 2020). For example, Coffin and Hewings (2003) found that, as a result of empirical approaches and the compilation of social statistics, Ph.D. theses written by doctoral students from social sciences might feature quantitative data, which may appear in texts in the forms of tables, graphs and maps. Students have to organize the pictorial/numerical data, understand how to incorporate them convincingly, and eventually depict the complicated multimodal information in clear and logical words. Paltridge and Starfield (2020) also found that social sciences generally pay special attention to rhetorical issues, persuading the audience of the validity of authorial arguments. This argumentative trait requires writers to draw on substantial interpersonal resources (e.g. *clear, important*) to develop a convincing authorial voice (Martin and White 2005). Some scholars further argue that, in social sciences, writers’ abilities to use interpersonal strategies, introduce authorial voices, engage with alternative views and establish solidarity with disciplinary communities are generally perceived as key features of successful thesis writing (Flowerdew 2015). Therefore, the special trait of social sciences may tune the textual activity of Ph.D. theses in social sciences to the descriptive mode.

6 Conclusion

This study investigated the textual activity of Ph.D. theses and dynamic changes across natural sciences, social sciences, and humanities from a stylometric perspective. The results show that in general, Ph.D. theses are significantly active, despite the fact that the theses in natural sciences and humanities are more active while those in social sciences are more likely to lean towards the descriptive mode. As to the dynamic changes, noted cross-disciplinary differences were also found. Similar trends of pro-activity were found in natural sciences and humanities, as opposed to the trend in social sciences that leans towards the descriptive mode. The findings could be accounted for by the different roles of verbs and adjectives in sentences (e.g. Baker 2003; Xu and Jiang 2021) as well as the features of academic/thesis

writing across disciplines (e.g. Xiao and Sun 2020; Sun and Crosthwaite 2022a, 2022b; Hyland 2012; Jiang 2022).

As an initial attempt, this study has methodological implications by showing the promising prospect of using textual activity as the stylometric method to unravel the stylistic features of Ph.D. theses, where traditional qualitative methods still prevail in the analyses of academic genres such as theses and research articles (Xiao and Sun 2020; Paltridge and Starfield 2020). The improved approach has increased the statistical soundness of results and may inspire EAP scholars to look into academic texts from an innovative quantitative linguistic perspective. In addition, from the theoretical perspective, our results confirm the active nature of the academic genre and complement previous disciplinary findings in a couple of ways. Such findings can be particularly vital to EAP and English for Research and Publication Purposes (ERRP) practitioners, who have to elaborate on such cross-disciplinary variations so as to equip green-hand students and novice academic writers with an awareness of the discipline-specific stylometric features in thesis writing.

Despite the meaningful findings, there remain some limitations. First, although a sample of 50 texts per disciplinary group has already exceeded the minimum requirement for the sample size (Roever and Phakiti 2017), the validity of the results could be improved with an enlarged sample. In addition, the scope of this study is but limited to textual activity of PhD theses. Future studies could measure more indicators (e.g. TTR, writer's view, Gini coefficient) to capture a wider picture of stylometric features of more academic genres. As stated in Section 1, previous research on textual activity has been confined mostly to political and literary topics, whereas the embodied regularities are expected to be figured out by exploring more genres (Čech and Kubát 2016; Chen and Liu 2018). It would be interesting to investigate textual activity of other academic genres such as research articles, which is also a key genre for knowledge creation and communication.

References

- Anthony, L.** (2017). AntFileConverter (Version 1.2.1) [Computer Software]. Tokyo: Waseda University. Retrieved from <https://www.laurenceanthony.net/software> (Accessed on Jan 1, 2022).
- Baker, M. C.** (2003). *Lexical categories: Verbs, nouns and adjectives*. Cambridge University Press.
- Bunton, D.** (2002). Generic moves in PhD theses introductions. In: Flowerdew, J. (Ed.). *Academic Discourse*, pp. 57-75. Harlow: Longman.
- Busemann, A.** (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik* [Youth's speech as an imprint of the rhythm of development]. Fischer.
- Čech, R., Kubát, M.** (2016). Text length and the thematic concentration of text. *Mathematical Linguistics*, 2(1), pp. 5-13.
- Chen, R., Liu, H.** (2018). Thematic concentration as a discriminating feature of text types. *Journal of Quantitative Linguistics*, 25(1), pp. 53-76. <https://doi.org/10.1080/09296174.2017.1339441>.
- Clarivate.** (2023). ProQuest [Database]. <https://about.proquest.com/en/>.
- Coffin, C., Hewings, A.** (2003). Writing for different disciplines. In: Coffin, C., Curry, M. J., Goodman, S., Hewings, A., Lillis, T., Swann, J. (Eds.). *Teaching Academic Writing: A Toolkit for Higher Education*, pp. 45-72. London: Routledge.
- Flowerdew, L.** (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, 20, pp. 58-68. <https://doi.org/10.1016/j.jeap.2015.06.001>.
- Goldberg, A.** (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Hood, S.** (2006). The persuasive power of prosodies: Radiating values in academic writing. *Journal of English for Academic Purposes*, 5(1), 37-49. <https://doi.org/10.1016/j.jeap.2005.11.001>.
- Hu, G., Cao, F.** (2015). Disciplinary and paradigmatic influences on interactional metadiscourse in research articles. *English for Specific Purposes*, 39, pp. 12-25. <http://dx.doi.org/10.1016/j.esp.2015.03.002>.
- Hyland, K.** (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), pp. 4-21. <https://doi.org/10.1016/j.esp.2007.06.001>.
- Hyland, K.** (2012). *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: Cambridge University Press.
- Jia, H., Liang, J.** (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua*, 239. <https://doi.org/10.1016/j.lingua.2020.102809>.
- Jiang, F.** (2022). *Metadiscursive Nouns: Interaction and Persuasion in Disciplinary Writing*. New York: Routledge.
- Jiang, F., Hyland, K.** (2022). "The datasets do not agree": Negation in research abstracts. *English for Specific Purposes*, 68, pp. 60-72. <https://doi.org/10.1016/j.esp.2022.06.003>.
- Jia, H., Liang, J.** (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua*, 239. <https://doi.org/10.1016/j.lingua.2020.102809>.



- Jiang, X., Jiang, Y., Hoi, C.** (2020). Is Queen's English drifting towards common people's English? — Quantifying diachronic changes of Queen's Christmas messages (1952–2018) with reference to BNC. *Journal of Quantitative Linguistics*, 29(1), pp. 1-36. <https://doi.org/10.1080/09296174.2020.1737483>.
- Kagan, J.** (2009). *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge: Cambridge University Press.
- Kanoksilapatham, B.** (2015). Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. *English for Specific Purposes*, 37, pp. 74-86. <https://doi.org/10.1016/j.esp.2014.06.008>.
- Kawase, T.** (2018). Rhetorical structure of the introductions of applied linguistics PhD theses. *Journal of English for Academic Purposes*, 31, pp. 18-27. <https://doi.org/10.1016/j.jeap.2017.12.005>.
- Kubát, M., Čech, R.** (2016). Quantitative analysis of US presidential inaugural addresses. *Glottometrics*, 34, pp. 14-27.
- Kuteeva, M., Airey, J.** (2014). Disciplinary differences in the use of English in higher education: Reflections on recent language policy developments. *Higher Education*, 67(5), pp. 533-549. <https://doi.org/10.1007/s10734-013-9660-6>.
- Liu, H.** (2009). *Dependency grammar: From theory to practice*. Science Press.
- Liu, H., Xu C., Liang, J.** (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, pp. 171-193. <https://doi.org/10.1016/j.plrev.2017.03.002>.
- Mačutek J., Wimmer, G.** (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3), pp. 227-240. <http://dx.doi.org/10.1080/09296174.2013.799912>.
- Martin, J. R., White, P. R. R.** (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.
- Melka, T., Místecký, M.** (2019). On stylistic features of H. Beam Piper's *Omnilingual*. *Journal of Quantitative Linguistics*, 27(3), pp. 1-40. <https://doi.org/10.1080/09296174.2018.1560698>.
- Paltridge, B., Starfield, S.** (2020). *Thesis and Dissertation Writing in a Second Language*. New York: Routledge.
- Popescu, I., Čech R., Altmann, G.** (2014). Descriptivity in special texts. *Glottometrics*, 29, pp. 70-80. Retrieved from <https://www.ram-verlag.eu/journals-e-journals/glottometrics/> (Accessed on Feb 12, 2022).
- Roever, C., Phakiti, A.** (2017). *Quantitative Methods for Second Language Research: A Problem-Solving Approach*. New York: Routledge.
- Schreibman, S., Siemens R., Unsworth, J.** (Eds.). (2008). *A Companion to Digital Humanities*. New Jersey: Wiley-Blackwell.
- Sun, S., Crosthwaite, P.** (2022a). "Establish a niche" via negation: A corpus-based study of negation within the Move 2 sections of PhD thesis introductions. *Open Linguistics*, 8(1), pp. 189-208. <https://doi.org/10.1515/opli-2022-0190>.
- Sun, S. A., Crosthwaite, P.** (2022b). "The findings might not be generalizable": Investigating negation in the limitations sections of PhD theses across disciplines. *Journal of English for Academic Purposes*, 59, pp. 101155. <https://doi.org/10.1016/j.jeap.2022.101155>.

- Thompson, P.** (2013). Thesis and dissertation writing. In: Paltridge, B., Starfield, S. (Eds.). *The Handbook of English for Specific Purposes*, pp. 283-299. Chichester: John Wiley & Sons.
- Tukey, J.** (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2), pp. 99-114. <https://doi.org/10.2307/3001913>.
- Xiao, W., Sun, S.** (2020). Dynamic lexical features of PhD theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics*, 27(2), pp. 114-133. <https://doi.org/10.1080/09296174.2018.1531618>.
- Xiao, W., Liu, J., Li, L.** (2022). How is information content distributed in RA introductions across disciplines? An entropy-based approach. *Research in Corpus Linguistics*, 10(1), 63-83. <https://doi.org/10.32714/ricl.10.01.04>.
- Xiao, W., Li, L., Liu, J.** (2023a). To move or not to move: An entropy-based approach to the informativeness of research article abstracts across disciplines. *Journal of Quantitative Linguistics*, 30(1), pp. 1-26. <https://doi.org/10.1080/09296174.2022.2037275>.
- Xiao, W., Guo, Y., Zhao, X.** (2023b). Towards Positivity: A Large-Scale Diachronic Sentiment Analysis of the Humanities and Social Sciences in China. *Fudan Journal of Humanities and Social Sciences*. <https://doi.org/10.1007/s40647-023-00380-2>.
- Xu, Z., Jiang, Y.** (2021). Activity of translational Chinese: A study based on three online corpora. *Foreign Language Teaching and Research*, 53(1), pp. 113-124. <https://doi.org/10.19923/j.cnki.fltr.2021.01.010>.
- Zhao, X., Li, L., Xiao, W.** (2023). The diachronic change of research article abstract difficulty across disciplines: a cognitive information-theoretic approach. *Humanities & Social Sciences Communications*, 10, pp. 194. <https://doi.org/10.1057/s41599-023-01710-1>.
- Zhou, H., Jiang, Y., Wang, L.** (2022). Are Daojing and Dejing stylistically independent of each other: A stylometric analysis with activity and descriptivity. *Digital Scholarship in the Humanities*. <https://doi.org/10.1093/llc/fqac042>.
- Zörnig, P., Altmann, G.** (2016). Activity in Italian presidential speeches. *Glottometrics*, 35, pp. 38-48. Retrieved from <https://www.ram-verlag.eu/wp-content/uploads/2018/08/glo35abstract.pdf> (Accessed on Feb 12, 2022).
- Zörnig, P., Stachowski, K., Popescu, I., Miyangah, T., Mohanty, P., Kelih, E., Chen, R., Altmann, G.** (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences*. Lüdenscheid: RAM-Verlag.

Funding statement

This work was supported by the Social Science Foundation of Chongqing [grant number 2019QNY51]; the Fund of the Interdisciplinary Supervisor Team for Graduates Programs of Chongqing Municipal Education Commission [grant number YDSTD1923]; the Fundamental Research Funds for the Central Universities [grant number 2021CDJSKZX07], and the Graduate Innovation Fund of Jilin University.

Swap distance minimization in SOV languages. Cognitive and mathematical foundations

Ramon Ferrer-i-Cancho^{1*}  (0000-0002-7820-923X), Savithry Namboodiripad²
 (0000-0002-7685-5895)

¹ Quantitative, Mathematical and Computational Linguistics Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.

² Linguistics Department, University of Michigan, Ann Arbor, Michigan, USA.

* Corresponding author's email: rferrericanch@cs.upc.edu.

DOI: https://doi.org/10.53482/2023_55_412

ABSTRACT

Distance minimization is a general principle of language. A special case of this principle in the domain of word order is swap distance minimization. This principle predicts that variations from a canonical order that are reached by fewer swaps of adjacent constituents are least costly and thus more likely. Here we investigate the principle in the context of the triple formed by subject (S), object (O) and verb (V). We introduce the concept of word order rotation as a cognitive underpinning of that prediction. When the canonical order of a language is SOV, the principle predicts $SOV < SVO$, $OSV < VSO$, $OVS < VOS$, in order of increasing cognitive cost. We test the prediction in three flexible order SOV languages: Korean (Koreanic), Malayalam (Dravidian), and Sinhalese (Indo-European). Evidence of swap distance minimization is found in all three languages, but it is weaker in Sinhalese. Swap distance minimization is stronger than a preference for the canonical order in Korean and especially Malayalam.

Keywords: word order preferences, canonical order, swap distance minimization

1 Introduction

Distance minimization pervades languages. In the domain of word order, there is massive evidence that the distance between words in a syntactic dependency representation of the sentence is minimized (Ferrer-i-Cancho et al., 2022; Futrell et al., 2015; Liu, 2008), a consequence of the syntactic dependency distance minimization principle (Ferrer-i-Cancho, 2004). A general principle of distance minimization in word order, which instantiates as syntactic dependency distance minimization, has been proposed (Ferrer-i-Cancho, 2014). Furthermore, the action of distance minimization in languages goes beyond the common notion of physical distance. Iconicity – which has also been argued to shape word order (Motamedi et al., 2022) – can be viewed as a response to a pressure to minimize the distance between a

linguistic form and meaning in production and interpretation (Dingemanse et al., 2015; Occhino et al., 2017; Perniss et al., 2010; Winter et al., 2022). Alignment in dialog (Garrod and Pickering, 2013; Pickering and Garrod, 2006) is the minimization of the distance between two or more speakers involved in a conversation. Because it operates across domains, distance minimization is likely to be one of the most general principles of language.

Distance minimization in word order (Ferrer-i-Cancho, 2014) presents itself as the syntactic dependency distance minimization principle (Ferrer-i-Cancho, 2004) and the swap distance minimization principle (Ferrer-i-Cancho, 2016). Critical characteristics of a compact but general theory of language are to specify (a) the cognitive origins of its principles (b) the cross linguistic support of its principles, and (c) the separation between principles and manifestations. Then compactness is achieved by uncovering the many distinct manifestations of the same principle (alone or interacting with other principles). Further, among the manifestations of a given principle, one has to distinguish direct from indirect manifestations.

1.1 Syntactic dependency distance minimization

Next we will revise the principle of syntactic dependency distance minimization from the standpoint of (a), (b) and (c) as a road map for research on swap distance minimization.

Concerning (a), syntactic dependency distance minimization is argued to result from counteracting interference and decay of activation in linguistic processes (Liu et al., 2017; Temperley and Gildea, 2018) and, accordingly, syntactic dependency distance in sentences is positively correlated with reading times (Niu and Liu, 2022).

Concerning (b), direct evidence of the principle of syntactic dependency distance minimization stems from the finding that syntactic dependency distances are smaller than expected by chance in samples of languages that have been growing in size and typological diversity (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho et al., 2022; Futrell et al., 2020; Futrell et al., 2015; Liu, 2008; Temperley, 2008).

Concerning (c), various manifestations of syntactic dependency distance minimization have been predicted. First, the acceptability of word orders and related word order preferences (Lin, 1996; Morrill, 2000). Second, formal properties of syntactic dependency structures such as the scarcity of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017) and the tendency to uncover the root (Ferrer-i-Cancho, 2008), thus predicting projectivity (continuous constituents) and planarity with high probability. Furthermore, syntactic dependency distance minimization predicts, in combination with projectivity, that the root of a sentence should be placed at the center (Alemany-Puig et al., 2022; Gildea and Temperley, 2007). An implication of the predictions is that verbs, which are typically the roots of a sentence, should be placed at the center, as in SVO orders or SVOI orders. For word orders in which the

verb appears first or last, syntactic dependency distance minimization predicts consistent branching for dependents of nominal heads (Ferrer-i-Cancho, 2015b), demonstrating the “unnecessity” of the headedness parameter of principles & parameters theory (Corbett, 1993; Ferrer-i-Cancho, 2015b).¹ The principle of swap distance minimization has received much less attention.

1.2 The order of S, V and O

Research on the order of S, V and O is biased towards SOV and SVO languages. SOV and SVO are the most attested dominant orders (76.5% according to Dryer (2013); 83.6% of languages and 69.6% families according to Hammarström (2016)). Accordingly, a large body of experimental research in the silent gesture paradigm has focused on factors that determine the choice between SOV and SVO (see Motamedi et al. (2022) and references therein). That bias neglects that there are languages that lack a dominant order (13.7% of languages according to Dryer (2013); 2.3% of languages and 6.1% of families according to Hammarström (2016)) or that exhibit two, rather than one, dominant orders (Dryer, 2013). Crucially, in many languages which do exhibit a dominant order, the other 5 non-dominant orders are produced. Though understanding such variation is vital, documentation and analyses of non-dominant orders receive relatively little attention (Levshina et al., 2023). This is reflected in psycholinguistic work, where the bulk of experimental research on the processing cost of word order focuses on just two orders, e.g. SVO versus OVS (Kaiser and Trueswell, 2004; Prabath and Ananda, 2017) or SVO versus VOS (Koizumi and Kim, 2016)². This challenge is the motivation of Namboodiripad’s research program on the cognitive cost of the six possible orders of S, V, and O in flexible order languages (Levshina et al., 2023; Namboodiripad et al., 2020; Namboodiripad, 2017, 2019). This is also why swap distance minimization is brought into play in this article.

1.3 Swap distance minimization

Swap distance minimization predicts pairs of primary alternating dominant orders (Ferrer-i-Cancho, 2016) and has been applied to shed light on the evolution of the dominant orders of S, V, and O from an ancestral SOV order (Ferrer-i-Cancho, 2015a, 2016). In general, the principle of swap distance minimization states that variations from a certain word order (canonical or not) that require fewer swaps of adjacent constituents are less costly (Ferrer-i-Cancho, 2015a, 2016). To illustrate how the principle works on triples, let us consider the case of the triple formed by subject (S), object (O) and verb (V). The so-called word order permutation ring is a graph where the vertices are all the six possible orderings

¹See Table 1 of Ferrer-i-Cancho and Gómez-Rodríguez (2021b) for further predictions.

²Note that practical challenges contribute to this. Comparing all six orders in an experiment requires more participants and different statistical tools as compared to simpler experimental designs; cf. Ohta et al. (2017).

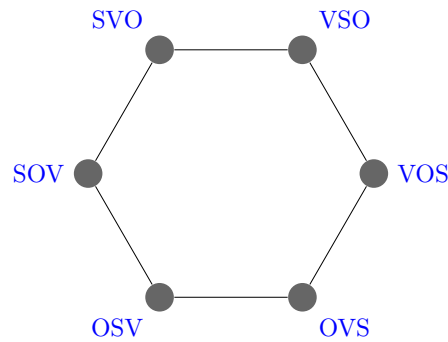


Figure 1: The word order permutation ring.

of the triple, and edges between two orders indicate that one order can be obtained from the other by swapping a pair of adjacent constituents (Figure 1). SOV and SVO are linked because swapping OV in SOV produces SVO, or equivalently, swapping VO in SVO produces SOV. For the case of triples, the permutation ring is an instance of a kind of graph which is called permutahedron in combinatorics (Ceballos et al., 2015). The swap distance between two orders is the distance (in edges) between two word orders in the permutahedron, namely, their distance is the minimum number of swaps of adjacent constituents that transforms one order into the other and vice versa.

A prediction of the swap distance minimization is that the cognitive cost of a word order will depend on its distance to the canonical order. When the canonical order of a language is SOV, SOV is at swap distance 0, SVO and OSV are at swap distance 1, VSO and OVS are at swap distance 2, and VOS is at swap distance 3 (Figure 1). Thus, the principle predicts (from easiest to most costly) the sequence ³

$$(1) \quad SOV < SVO, OSV < VSO, OVS < VOS.$$

For other canonical orders, the predictions that the permutahedron generates as a function of the canonical order are, in order of increasing processing cost (the canonical order appears first)

$$SVO < SOV, VSO < VOS, OSV < OVS$$

$$VSO < SVO, VOS < SVO, OVS < OSV$$

$$VOS < VSO, OVS < SVO, OSV < SOV$$

$$OVS < VOS, OSV < SOV, SVO < SVO$$

$$(2) \quad OSV < SOV, OVS < SVO, VOS < VSO.$$

³A sequence of this sort can be expressed with the following notation (Tamaoka et al., 2011)

$$SOV < SVO = OSV < VSO = OVS < VOS.$$

In our notation, = is replaced by a comma.

It is well-known that canonical orders are easier to process than non-canonical orders Menn, 2000; Meyer and Friederici, 2016 and thus canonical orders are processed faster than non-canonical orders (Hyönä and Hujanen, 1997; Kaiser and Trueswell, 2004; Tamaoka et al., 2011). The principle of swap distance minimization subsumes a preference for the canonical order but, crucially, it introduces a gradation for non-canonical orders, namely not all non-canonical orders are equally easy to process. The gradation is determined, by a precise definition of distance to the canonical order (Equation 1 and Equation 2). In contrast to Equation 1, just of preference of the canonical word order is expressed simply as

$$(3) \quad SOV < SVO, OSV, VSO, OVS, VOS.$$

1.4 The present article

Here we aim to contribute to research on swap distance minimization in the three directions above: (a), (b) and (c). We will increase the support for the principle both in terms of (a) and (b). As for (a), here we will introduce the concept of word order rotation as the analog of rotation in visual recognition experiments (Cooper and Shepard, 1973; Tarr and Pinker, 1989). In addition, we aim to validate the arguments using proxies of cognitive cost that are commonly used in cognitive science research such as reaction times and error rates (Cooper and Shepard, 1973; Tamaoka et al., 2011). As for (b), we will investigate the principle in languages from distinct linguistic families and quantify its effect with respect to other word order principles. As for (c), we will show that swap distance minimization predicts the acceptability of the order of subject, verb and object as syntactic dependency distance minimization predicts the acceptability of sentences (Lin, 1996; Morrill, 2000). Put differently, we will show that swap distance minimization manifests in the form of acceptability preferences.

We select three SOV languages which exhibit considerable word order flexibility, each from different language families: Sinhalese (Indo-European), Malayalam (Dravidian), and Korean (Koreanic). For each of these languages, all of the six possible orderings of S, V, and O are grammatical, attested, and have the same truth-conditional meaning (Namboodiripad, 2017, 2019; Tamaoka et al., 2011), though the degree of flexibility may vary depending on the context or measure of flexibility (Levshina et al., 2023; Yan and Liu, 2023). Sinhalese and Malayalam have been regarded as non-configurational (Mohan, 1983; Prabath and Ananda, 2017; Tamaoka et al., 2011). Interestingly, Malayalam exhibits more word order flexibility than Korean while, in turn, the flexibility of Korean is closer to that of English (Figure 8 of Levshina et al. (2023)).

In the context of Malayalam, the acceptability of a certain order has been argued to be determined by the position of the verb (Namboodiripad and Goodall, 2016). We will transform this specific proposal into

a general competing hypothesis, namely that the cost of a certain order (no matter how it is measured) is determined to some degree by the position of the verb, and link it with the theory of word order: a decrease in cost of processing of the verb as it is placed closer to the end is actually a prediction of the principle of minimization of the surprisal (maximization of the predictability) of the head (Ferrer-i-Cancho, 2017).⁴ In contrast to Equation 1, a preference for verb final would be expressed simply as

$$(4) \quad SOV, OSV < SVO, OVS < VSO, VOS.$$

The remainder of the article is organized as follows. Section 2 introduces the concept of word order rotation and a new mathematical framework. Section 3 justifies the choice of SOV languages and presents the data while Section 4 presents the statistical analysis methods. Section 5 shows evidence of swap distance minimization as predicted by Equation 1 in these three languages and compares it against two competing principles: a preference for the canonical order and a preference for the verb towards the end. Section 6 provides hawk-eye view of the results, speculates on their relation with the degree of word order flexibility of the languages, and proposes some issues for future research.

2 Theoretical foundations

2.1 Word order rotations

Here we present an argument on the cognitive support of the minimization of swap distance to the canonical order that is inspired by classic research on the cognitive effort of the visual recognition of objects (Cooper and Shepard, 1973; Tarr and Pinker, 1989). That research revealed that such cost depends on the rotation angle with respect to some canonical representation of the object. By analogy, the object is the triple formed by subject, object, and verb; we assume that its canonical representation is the order that language experts have identified as canonical; the rotation angle is the swap distance to the canonical order. However, the analogy with visual rotation can be made stronger by drawing the word order permutation ring on a circle as in Figure 1, placing a rotation axis at the center of the circle, and replacing the swap distance to the canonical order by the absolute value of the minimum angle of the rotation that is needed to put

- The word order of interest in the original position of the canonical order, or equivalently,

⁴A word of caution is necessary concerning the term competing hypothesis. It does not mean that maximization of predictability excludes swap distance minimization. Both forces can co-exist, and it is tempting to think that swap distance minimization implies the maximization of the predictability of the head for certain canonical orders, e.g., SOV or OSV. Indeed, we will show that swap distance and the position of the head (the verb) are significantly correlated.

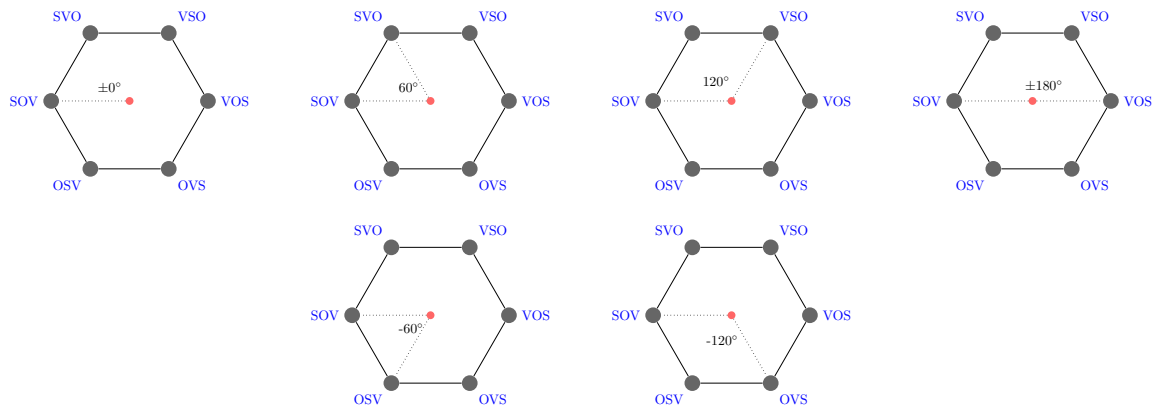


Figure 2: Rotations of word orders with respect to an axis at the center of the ring (marked in red). Recall that clockwise rotations have negative sign while anticlockwise rotations have positive sign. To become the canonical order SOV, (a) SOV needs a rotation of $\pm 0^\circ$, (b) SVO needs a rotation of 60° , (c) VSO needs a rotation of 120° , (d) VOS needs a rotation of $\pm 180^\circ$, (e) OSV needs a rotation of -60° , (f) OVS needs a rotation of -120° .

- The canonical order in the original position of the word order of interest.

The rotations that are needed to transform any order of S, V and O into SOV are shown in [Figure 2](#). Accordingly, the orders at distance 1 imply a rotation angle of $\pm 60^\circ$, orders at distance 2 imply a rotation of angle of $\pm 120^\circ$, and finally the order at distance 3 implies a rotation angle of $\pm 180^\circ$. In mathematical language, α , the angle of rotation (in degrees) that is required to transform a certain word order into the canonical word order, and d , the swap distance between an order and the canonical, obey

$$d = \frac{|\alpha|}{60}.$$

2.2 The correlation between a distance measure and cognitive cost

Here we present a new mathematical framework to measure the effect distinct word order principles by translating [Equation 1](#), [Equation 3](#), and [Equation 4](#) into Kendall τ correlations and also to understand how these principles interact.

We define s as the cognitive cost of a certain ordering of S, V, and O. Swap distance minimization predicts that s should increase following the ordering in [Equation 1](#). Accordingly, we test the swap distance minimization hypothesis by measuring $\tau(d, s)$, the Kendall τ correlation between the target score s and d , which is the swap distance between an order and the canonical order SOV. To test the hypothesis of the minimization of surprisal of the verb ([Equation 4](#)), we measure $\tau(p, s)$, namely the Kendall τ correlation between the target score s and p , the distance of the verb to the end (0 for verb-last, 1 for medial verb and 2 for verb first). Finally, as swap distance minimization subsumes a preference for

the canonical order (Equation 3), we also define a control hypothesis, namely that the effect is merely simply determined by the word order being canonical or not. That hypothesis is tested by means of $\tau(c, s)$, the Kendall correlation between the target score and c , a binary variable that is zero if the order is canonical and 1 otherwise. We refer to d , p and c as distance measures. c is a binary distance to the canonical order. The values of these distances in an SOV language are shown in Table 1.

Table 1: For each of the six possible orders, we show the swap distance to the canonical order SOV (d), the distance of the verb to the end of the triple (p), the binary distance to canonical order (c), the mean z -score acceptability according to the results of the experiments by Namboodiripad (2017, Table 2.7) and the corresponding rank transformation (the most acceptable has rank 1, the second most acceptable has rank 2 and so on).

Order	d	p	c	Acceptability	Rank transformation
SOV	0	0	0	1.05	1
OSV	1	0	1	0.80	2
SVO	1	1	1	0.36	3
OVS	2	1	1	0.30	4
VSO	2	2	1	-0.14	5
VOS	3	2	1	-0.36	6

Note: p takes the values 0 for verb final, 1 for verb medial, and 2 for verb initial. c takes a value of 0 if the order is canonical and 1 otherwise.

There are the three main variants of the Kendall τ correlation: τ_a , τ_b and τ_c (Kendall, 1970). The simplest definition is that of τ_a , that is defined, for a bivariate sample of size n , as

$$(5) \quad \tau_a = \frac{n_c - n_d}{\binom{n}{2}},$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs.

τ_a performs no adjustment for ties, while τ_b and τ_c do. In our study, adjustments for ties bother. As swap distance minimization subsumes the preference for the canonical order, we want to warrant that if $\tau(d, s)$ is sufficiently large then $\tau(d, s) > \tau(c, s)$ because swap distance minimization is a more precise hypothesis than a preference for the canonical order. In the Appendix, we show two very useful properties of τ_a : if τ_a is large enough, then one can be certain that swap distance minimization does not reduce to a preference for the canonical order or to a preference for verb-last. In the language of mathematics, if $\tau_a(d, s) > 0.3$ then $\tau_a(d, s) > \tau_a(c, s)$; if $\tau_a(d, s) > 0.8$ then $\tau_a(d, s) > \tau_a(p, s), \tau_a(c, s)$. We also want to ensure that the comparison between $\tau(d, s)$ and $\tau(p, s)$ is fair; notice that p has lower precision than d (d is on an integer scale between 0 and 3 while p is on an integer scale between 0 and 2). Adjustments for ties may cause the illusion of a weaker manifestation of swap distance minimization compared to other cognitive pressures.⁵ Hereafter τ means τ_a .

⁵Finally, another reason for not using τ_b is a further consequence of the adjustment for ties: τ_b is undefined when the variance of one of the variables is zero. With this respect, τ_a is robust across conditions and simplifies the coding as it does not require to deal with the special case of zero variance.

Finally, notice that distinct word order principles are related and thus the Kendall τ correlation between two distance measures are all positive (Table 2). Kendall τ correlation between d and p , $\tau(d, p)$ is significantly high while $\tau(d, c)$ and $\tau(p, c)$ are not (Table 2). Obviously, the fact that $\tau(d, c)$ is not significant is clearly due to a lack of statistical power. The arguments in the Appendix for the correlation between c and some other variable, allow one to conclude that $\tau(d, c)$ is maximum and its right p -value is minimum.

Table 2: Correlogram of Kendall τ correlation between each distance measure. We use right-sided exact tests of correlation with τ_a on the matrix in Table 1. Recall d is the swap distance to the canonical order, p is distance of the verb to the end of the triple and c is the binary canonical distance.

Variables	Kendall τ correlation	p -value
d and p	0.67	0.044
d and c	0.33	0.166
p and c	0.27	0.333

3 Material

3.1 Why SOV languages

The predictions in Equation 1 and 2 raise the question of the ideal conditions where swap distance minimization should be tested (point (b) in Section 1). One could naively argue that these predictions should hold for every language in any condition. The challenge is that swap distance minimization is just one of the various principles that shape word order in languages: word order is a multiconstraint satisfaction problem (Ferrer-i-Cancho, 2017; Xu et al., 2017). Thus, the observation of the action of a specific word order principle requires identifying the conditions where that principle will suffer from less interference from other word order principles. For instance, it has been predicted theoretically and demonstrated empirically that the action of surprisal minimization (predictability maximization) should be more visible in short sentences (Ferrer-i-Cancho and Gómez-Rodríguez, 2021a; Ferrer-i-Cancho et al., 2022). Interestingly, it has been shown that syntactic dependency distance minimization is weaker in Warlpiri, a non-configurational language (Ferrer-i-Cancho et al., 2022). Indeed, discontinuous constituents, one of the hallmarks of non-configurational languages (Austin and Bresnan, 1996; Hale, 1983) may indicate that dependency distance minimization is weaker, as it has been demonstrated that pressure to reduce the distance between syntactically related elements reduces the chance of discontinuity (Gómez-Rodríguez et al., 2022; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). Thus, interference from dependency distance minimization is expected to be weaker in non-configurational languages. Recall that dependency distance minimization alone would draw the verb, the root of the triple, towards the center of the triple (Alemany-Puig et al., 2022; Gildea and Temperley, 2007). In addition, we expect

that, in languages that exhibit word order flexibility, there is more room for capturing the manifestation of swap distance minimization. English, which is an SVO language, is an example of a non-ideal language to test this because of its word order rigidity (Figure 8 of Levshina et al. (2023)).

Given the considerations above, this article focuses on SOV languages. SOV languages are an ideal arena for testing this principle. In terms of representativity, SOV represents the most common dominant word order across languages (Dryer, 2013; Hammarström, 2016). Furthermore, SOV has been hypothesized to be an early stage in spoken languages (Gell-Mann and Ruhlen, 2011; Newmeyer, 2000), and it has been regarded as a default basic word order (Givón, 1979; Newmeyer, 2000). This view is supported by the fact that SOV is often the dominant order found in sign languages which are at the early stages of community-level conventionalisation (Meir et al., 2010; Sandler et al., 2005).

3.2 Data

Data is borrowed from existing publications but is available as a single file in the repository of the article.⁶ We borrow data from word order experiments in Malayalam (Namboodiripad, 2017), Korean (Namboodiripad et al., 2019), and Sinhalese (Tamaoka et al., 2011).⁷ In Korean and Malayalam, the target scores are average z -scored acceptability ratings from experiments in the spoken (listening) modality that are obtained from Namboodiripad (2017, Table 2.7 in Chapter 2) for Malayalam and Table 2 of Namboodiripad et al. (2019) for Korean. As is typical in acceptability judgment experiments, z -scores are used to control for individual variation in the use of the rating scale.

All participants in the Malayalam experiment ($N = 18$) grew up speaking Malayalam in Kerala, India, where it is the dominant language. For Korean, we consider three groups that are borrowed from Namboodiripad et al. (2019): bilingual speakers of Korean and English that are split into Korean-dominant ($N = 30$), English-dominant active (individuals who are fluent in comprehension and production of spoken Korean; $N = 13$), and English-dominant passive (individuals who are far more proficient in comprehension of spoken Korean than they are in production; $N = 14$).

For Sinhalese, the participants are described as native speakers. The target scores are mean reaction times and mean error rates in the spoken ($N = 42$) and written ($N = 36$) modality. Mean reaction times and mean error rates are borrowed from Table 1 and Table 2 of Tamaoka et al. (2011) for the written (reading) and spoken (listening) modality, respectively. Here, it is not clear how the authors controlled for individual variation (i.e., via z -scores or other statistical methods).

⁶In the *data* folder of <https://osf.io/b62ep/>.

⁷For each language, the target sentences have the same structure: animate subjects, inanimate objects, and active transitive verbs; sample stimuli can be found in each paper. Due to space limitations, we refer the reader to those original sources for further methodological details.

To validate findings in Malayalam as Namboodiripad (2017, 2019) did, we borrow frequencies of each of the six orders of S, V and O from an online corpus (Leela, 2016, Table 4) as an additional target score.⁸

By target score, we mean acceptability, reaction time, error, frequency, and the variants that result from pairwise contrasts. Every target score (other than frequency) yields a rank variant that results from comparing the scores of every pair of distinct orders by means of some statistical test. Here we adopt the convention that these ranks reflect cognitive cost: the least costly order has rank 1, the second least costly has rank 2 and so on. The pairwise contrasts for Malayalam give, in order of decreasing acceptability (Namboodiripad, 2017)

$$SOV, OSV > SVO, OVS > VSO, VOS.$$

Thus, SOV and OSV have acceptability rank 1, SVO and OVS have acceptability rank 2, and VSO and VOS have acceptability rank VSO and VOS. For Sinhalese, the pairwise contrasts for reaction time in spoken language give, in order of increasing reaction time (Tamaoka et al., 2011),

$$SOV < SVO, OVS < OSV, VSO, VOS$$

and thus SOV has reaction time rank 1, SVO and OVS have reaction time rank 2 and OSV, VSO and VOS have reaction time rank 3. For Korean, Namboodiripad et al. (2019) report in prose that the verb-medial orders and verb-initial orders group together, but the authors do not give more details. However, (Namboodiripad et al., 2020) report pairwise comparisons⁹ in a reanalysis of the same data. The ranking in order of decreasing acceptability is

$$SOV > OSV > SVO, OVS > VSO, VOS.$$

Thus, SOV has acceptability rank 1, OSV has acceptability rank 2, SVO and OVS have acceptability rank 3, and VSO and VOS have acceptability rank 4. All the pairwise contrasts for the languages investigated in this article are summarized in Table 3.

We define a condition as the combination of modality (spoken or written), the target score, and, optionally, a group.

The sign of certain scores that measure cognitive ease is inverted before the analyses to transform them into scores of cognitive cost. This is the case of acceptability ratings in Malayalam and Korean and word order frequencies in Malayalam. As we are using Kendall τ correlation, the transformation does

⁸The corpus comprises three types of discourse: interviews, discussions or debates, and conversations appearing in printed form in online media. The genres are relatively comparable with the experimental items because they come from more casual and conversational contexts. The whole corpus comprises 5598 monotransitive sentences but only 67.1% contain S, V and O according to Table 4 (Leela, 2016, Table 4). Thus we estimate that the frequencies of S, V and O are based on 3756 sentences. Further details be found at <http://hdl.handle.net/10803/399556> in Section 3.2.1 Methodology.

⁹Bonferroni corrected, with pooled SD.

Table 3: Summary of pairwise contrasts, in order of increasing cognitive cost for Korean (Namboodiripad et al., 2020), Malayalam (Namboodiripad, 2017) and (Tamaoka et al., 2011).

Language	Group	Score	Modality	Pairwise contrasts
Korean	Korean-dominant	acceptability	spoken	$SOV < OSV < SVO, OVS < VSO, VOS$
Korean	English-dominant active	acceptability	spoken	$SOV < OSV < SVO, OVS < VSO, VOS$
Korean	English-dominant passive	acceptability	spoken	$SOV < OSV < SVO, OVS < VSO, VOS$
Malayalam		acceptability	spoken	$SOV, OSV < SVO, OVS < VSO, VOS$
Sinhalese		reaction time	spoken	$SOV < SVO, OVS < OSV, VSO, VOS$
Sinhalese		reaction time	written	$SOV < SVO, OVS, OSV, VSO, VOS$
Sinhalese		error	spoken	$SOV < SVO, OVS, VSO < OSV, VOS$
Sinhalese		error	written	$SOV, SVO, VSO, VOS, OVS, OSV$

not alter the potential conclusions and has a clear advantage: all target scores can then be submitted to a right-sided Kendall correlation test. The resulting association between swap distance and acceptability rank is shown in Table 1.

4 Methodology

All the code used to produce the results is available in the repository of the article.¹⁰

4.1 Kendall τ correlation

We used R for the analyses. To compute Kendall τ correlation, we used neither the standard function to compute Kendall correlation, i.e. `cor` (that runs in $O(n^2)$ time, where n is the size of the sample), nor the faster implementation `cor.fk` (that runs in $O(n \log n)$ time) from the `pcaPP` library. The reason is that `cor` function computes Kendall τ_b instead of τ_a when there are ties¹¹. The documentation of `cor.fk` is not clear on this matter, but our experience suggests that it also implements τ_b : when we compute Kendall τ between the vector (1, 1, 2, 2, 3, 3) and itself, `cor` and `cor.fk` yield 1, the maximum value, as expected by the definition of τ_b . In contrast, our implementation of τ_a yields 0.8 because of the presence of ties. Therefore we computed τ_a using a naive implementation by us that runs in $O(n^2)$ time.

4.2 Kendall τ correlation test

The standard function for the Kendall correlation test, i.e. `cor.test`, fails to compute accurate enough p -values. To fix it, we implemented a function that computes, exactly, the right p -value of the Kendall correlation test by generating all permutations of the values of one of the variables and computing the Kendall τ correlation on each of those permutations. This exact test was also used for the differences $\tau(d, s) - \tau(p, s)$ and $\tau(d, s) - \tau(c, s)$.

¹⁰In the `code` folder of <https://osf.io/b62ep/>.

¹¹<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html>

4.3 Maximum correlation

We distinguish two reasons why a Kendall correlation is maximum:

- Maximum given a distance measure. Namely, given the sample as a matrix with two columns, one for the distance measure and the other for the score, there is no possible replacement of the values of the score that gives a higher correlation. See Property 3 for the maximum correlation and Property 5 for the minimum right p -value that is obtained when the correlation is maximum.
- Maximum given the sample. In this case, the correlation is the maximum given the bivariate sample used to compute the correlation. Namely, given the sample as a matrix with two columns, no permutation of a column of the sample matrix yields a higher correlation. This kind of maximum correlation is determined computationally from its definition.

It is easy to see that if a correlation is maximum given the distance measure, then it is also maximum given the sample. We also extend this notions to the differences $\tau(d, s) - \tau(p, s)$ and $\tau(d, s) - \tau(c, s)$.

4.4 A Monte Carlo global analysis

The Kendall τ correlation tests above suffer from lack of statistical power: the minimum p -value for the Kendall τ depends on the distance measure and ranges between $0.1\bar{6}$ for c and $0.00\bar{5}$ for d (Property 5). In the case of Sinhalese, none of the correlations across conditions and distance measures was statistically significant. To gain statistical power, we decided to perform a global statistical test for a given distance measure across all conditions. The statistic of that test is S , that is defined as the sum of all the Kendall correlations across all conditions for a given language and distance measure. The right p -value of the test was estimated by a Monte Carlo procedure as the proportion of $T = 10^6$ randomizations where S' , the value of S in a randomization, satisfied $S' \geq S$. Each randomization consists of producing a uniformly random permutation the values of one the target score that are assigned to the distance measure for each language and distance measure. Therefore, the smallest non-zero estimated p -value that this test can produce is $1/T = 10^{-6}$. The test was adapted to assess the significance of the difference between pairs of distance measures.

As an orientation for discussion, we assume a significance level of $\alpha = 0.05$ throughout this article. When we perform statistical tests over various individual conditions, we may suffer from multiple comparisons. When presenting results on individual conditions, we do not correct p -values for them because this problem is addressed by the Monte Carlo test, where we apply Holm correction in two contexts. When answering the question of when a distance measure yields significance, we adjust the p -values of $S(d)$, $S(p)$ and $S(c)$ for each language (9 comparisons). When answering the question of when the difference

between swap distance minimization and another principle yields significance, we adjust the p -values of $S(d) - S(c)$ and $S(d) - S(p)$ for each language (6 comparisons).

5 Results

5.1 Evidence of swap distance minimization

In Korean, the correlation between acceptability and swap distance to the canonical order, ($\tau(d, s)$) is statistically significant in all three groups: Korean-dominant, English-dominant active, and English-dominant passive (Table 4), suggesting that swap distance minimization is a robust effect. When acceptability ranks are used, the correlation turns out to be maximum given the sample. In the English-dominant active group, the correlation increases when mean acceptability is replaced by acceptability rank. In Malayalam, that correlation is statistically significant and maximum given the distance measure (Table 4). When raw mean acceptability scores are replaced by acceptability ranks resulting from pairwise contrasts, the correlation ($\tau(p, s)$) weakens (the opposite phenomenon with respect to group of English-dominant active in Korean) but it is still significant. That suggests that, in Malayalam, raw mean acceptability scores contain some information about swap distance minimization that is lost when using these ranks, likely due to lack of statistical power in the pairwise contrasts. The support for the swap distance minimization from the canonical order is confirmed when acceptability ratings are replaced by frequencies from Leela's corpus, which achieve a maximum correlation given the sample (Table 4). These findings suggest that swap distance minimization in Malayalam is a robust phenomenon because it is captured by independent measures.

In Sinhalese, we find no support for swap distance minimization on individual conditions except for reaction times in the written modality, where the correlation between reaction time and swap distance to the canonical order yields a borderline p -value (p -value=0.061). When the raw mean reaction times in that modality are replaced by ranks obtained from pairwise contrasts, the correlation $\tau(d, s)$ decreases ($\tau(d, s)$ drops from 0.6 to $\tau(p, s) = 0.3$), suggesting that raw reaction times may contain some information about swap distance minimization that is lost during the pairwise contrasts. Interestingly, the correlation with these ranks is maximum given the sample (Table 4). In contrast, the rank transformation resulting from pairwise contrasts has the opposite effect for reaction time and error in the spoken modality: $\tau(d, s)$ increases after applying that transformation. That suggests that mean reaction time and mean error rate are noisy in the spoken modality.

Although statistical support for swap distance minimization is missing on individual conditions in Sinhalese, the Monte Carlo global analysis (Table 5) indicates that the sum of Kendall τ correlations over all conditions is significantly high ($S(d) = 2.4$, p -value = $1.5 \cdot 10^{-3}$), suggesting that swap distance

Table 4: The outcome of three correlation tests. First, the Kendall τ correlation test between s , the target score, and d is its swap distance to the canonical order SOV. Second, the Kendall τ correlation test between s and p , the distance of the verb to the end. Second, the Kendall τ correlation test between s and c , a binary variable that indicates if the order is canonical or not. For each correlation test, red indicates that the correlation is maximum (and the p -value is minimum) given the distance measure; orange indicates that the correlation is maximum (and p -value is minimum) given the sample.

Language	Group	Score	Modality	$\tau(d, s)$	p -value	$\tau(p, s)$	p -value	$\tau(c, s)$	p -value
Korean	Korean-d	acceptability	spoken	0.733	0.022	0.8	0.011	0.333	0.167
Korean	Korean-d	acceptability rank	spoken	0.733	0.022	0.8	0.011	0.333	0.167
Korean	English-d a	acceptability	spoken	0.667	0.033	0.8	0.011	0.333	0.167
Korean	English-d a	acceptability rank	spoken	0.733	0.022	0.8	0.011	0.333	0.167
Korean	English-d p	acceptability	spoken	0.733	0.022	0.8	0.011	0.333	0.167
Korean	English-d p	acceptability rank	spoken	0.733	0.022	0.8	0.011	0.333	0.167
Malayalam	-	acceptability	spoken	0.867	0.006	0.8	0.011	0.333	0.167
Malayalam	-	acceptability rank	spoken	0.667	0.044	0.8	0.011	0.267	0.333
Malayalam	-	frequency	-	0.8	0.011	0.8	0.011	0.333	0.167
Sinhalese	-	reaction time	spoken	0.333	0.228	0.267	0.289	0.333	0.167
Sinhalese	-	reaction time rank	spoken	0.467	0.117	0.4	0.133	0.333	0.167
Sinhalese	-	reaction time	written	0.6	0.061	0.4	0.167	0.333	0.167
Sinhalese	-	reaction time rank	written	0.333	0.167	0.267	0.333	0.333	0.167
Sinhalese	-	error	spoken	0.267	0.239	0.133	0.422	0.333	0.167
Sinhalese	-	error rank	spoken	0.4	0.15	0.2	0.333	0.333	0.167
Sinhalese	-	error	written	0	0.6	-0.133	0.733	0.2	0.5
Sinhalese	-	error rank	written	0	1	0	1	0	1

Note: c is 0 if the order is canonical and 1 otherwise. p is 0 for verb-last, 1 for verb-medial and 2 for verb first. In Korean, the groups are *Korean-d* (Korean-dominant), *English-d a* (English-dominant active) and *English-d p* (English-dominant passive).

minimization is present but weak in Sinhalese. In Korean and Malayalam, the Monte Carlo global analysis just confirms the findings on individual languages (Table 5; p -value $< 10^{-5}$ in both languages).

5.2 Evidence of maximization of the predictability of the verb

The correlation between the distance from the verb to the end of the sentence and each of the scores ($\tau(p, s)$) was statistically significant for Korean and Malayalam over all conditions, and it was indeed maximum given the distance measure (Table 4). In both languages and across all conditions, $\tau(p, s)$ was maximum given the distance measure. However, the global analysis (Table 5) revealed that the sum of Kendall τ correlations over all conditions is borderline significant in Sinhalese ($S(p) = 1.53$, p -value = 0.066), suggesting that the maximization of the predictability of the verb has some global effect on that language. In Korean and Malayalam, the Monte Carlo global analysis based on $S(p)$ just confirms the findings on individual languages (Table 5; p -value $< 10^{-5}$ in both languages).

5.3 Evidence of a preference for the canonical order

The correlation between the binary distance to the canonical order and each of the scores ($\tau(p, s)$) was never statistically significant across languages and conditions (Table 4), but this is due to the lack of the

Table 5: Summary of the outcome of the Monte Carlo global analysis over all conditions for each language S is the sum of the Kendall τ correlation over all conditions for a certain distance measure. d is swap distance to the canonical order, p is distance of the verb to the end of the triple, and c is binary canonical distance. p -values have been adjusted with Holm correction (as explained in Section 4).

Language	$S(d)$	p -value	$S(p)$	p -value	$S(c)$	p -value	$S(d) - S(c)$	p -value	$S(d) - S(p)$	p -value
Korean	4.33	$< 10^{-6}$	4.8	$< 10^{-6}$	2	$2.4 \cdot 10^{-5}$	2.33	$9 \cdot 10^{-4}$	-0.47	1
Malayalam	2.33	$1.8 \cdot 10^{-5}$	2.4	$1.4 \cdot 10^{-5}$	0.93	$9.3 \cdot 10^{-3}$	1.4	$2.1 \cdot 10^{-3}$	-0.07	1
Sinhalese	2.4	$4.6 \cdot 10^{-3}$	1.53	0.065	2.2	$1.6 \cdot 10^{-5}$	0.2	1	0.87	0.11

statistical power of the test (the minimum p -value is $0.1\bar{6}$ as explained in the Appendix). Indeed, the Monte Carlo global analysis based on $S(c)$ shows that a preference for the canonical order has a significant effect in all languages but much more strongly in Korean and Sinhalese (Table 5; p -value $< 10^{-2}$ in all languages). The latter could be due to the larger amount of conditions in Sinhalese and Korean, which may amplify the statistical effect.

Table 6: The outcome of two Kendall correlation difference tests. The first test is on $\tau(d, s) - \tau(c, s)$. The second test is on $\tau(d, s) - \tau(p, s)$. In each correlation test, orange indicates that the correlation is maximum (and then the p -value is minimum) given the sample.

Language	Group	Score	Modality	$\tau(d, s) - \tau(c, s)$	p -value	$\tau(d, s) - \tau(p, s)$	p -value
Korean	Korean-d	acceptability	spoken	0.4	0.1	-0.067	0.753
Korean	Korean-d	acceptability rank	spoken	0.4	0.078	-0.067	0.728
Korean	English-d a	acceptability	spoken	0.333	0.133	-0.133	0.778
Korean	English-d a	acceptability rank	spoken	0.4	0.078	-0.067	0.728
Korean	English-d p	acceptability	spoken	0.4	0.1	-0.067	0.753
Korean	English-d p	acceptability rank	spoken	0.4	0.078	-0.067	0.728
Malayalam	-	acceptability	spoken	0.533	0.006	0.067	0.5
Malayalam	-	acceptability rank	spoken	0.4	0.078	-0.133	0.833
Malayalam	-	frequency	-	0.467	0.022	0	0.558
Sinhalese	-	reaction time	spoken	0	0.6	0.067	0.5
Sinhalese	-	reaction time rank	spoken	0.133	0.35	0.067	0.433
Sinhalese	-	reaction time	written	0.267	0.233	0.2	0.247
Sinhalese	-	reaction time rank	written	0	0.5	0.067	0.5
Sinhalese	-	error	spoken	-0.067	0.611	0.133	0.256
Sinhalese	-	error rank	spoken	0.067	0.383	0.2	0.167
Sinhalese	-	error	written	-0.2	0.883	0.133	0.267
Sinhalese	-	error rank	written	0	1	0	1

Note: $\tau(d, s)$ is the correlation between a score and swap distance. $\tau(c, s)$ is the correlation between a score and the binary distance to canonical order. $\tau(p, s)$ is the correlation between a score and the distance to end of the verb. In Korean, the groups are *Korean-d* (Korean-dominant), *English-d a* (English-dominant active) and *English-d p* (English-dominant passive).

5.4 Can the results be reduced to simply a preference for the canonical order?

It could be argued the finding of swap distance minimization effects is a mere consequence of a rather obvious expectation: canonical orders are easier to process than non-canonical orders. Indeed, swap

distance minimization also predicts a preference for canonical orders but adds a gradation on non-canonical orders. However, we find that the correlation between a target score and swap distance to canonical order ($\tau(d, s)$) as well as the correlation between a target score and distance of the verb to the end ($\tau(p, s)$) are always greater than the correlation between the target score and being canonical or not ($\tau(c, s)$) in both Korean and Malayalam; this is also the case in Sinhalese with two exceptions: error in the spoken and written modality (Table 4 and Table 6). In Korean, the difference $\tau(d, s) - \tau(c, s)$ is always positive but never significant. However, the difference is borderline significant in all groups when acceptability ranks are used (p -value = 0.078). In Malayalam, the analysis of $\tau(d, s) - \tau(c, s)$ (Table 6) indicates that swap distance minimization has a significantly stronger effect than a preference for a canonical order across conditions (although the p -value of acceptability ranks, i.e. 0.078 is borderline). Furthermore, concerning mean acceptability, the difference is maximum given the sample. The Monte Carlo global analysis shows that indeed $S(d) - S(c)$ is significantly large in both Korean and Malayalam (p -value $< 10^{-4}$), indicating that swap distance minimization is significantly stronger than a preference for a canonical order (Table 5).

In Sinhalese, the difference $\tau(d, s) - \tau(c, s)$ is never statistically significant across conditions and that is confirmed by the Monte Carlo global analysis (p -value = 0.369). (Table 5).

5.5 Swap distance minimization versus maximization of the predictability of the verb

In Korean, the effect of swap distance minimization is weaker than the force that drags the verb towards the end. In particular, the correlation between acceptability and swap distance to the canonical order ($\tau(d, s)$) is always smaller than the correlation between mean acceptability and verb position ($\tau(p, s)$). In Table 4 and Table 6, we can check that $\tau(d, s) < \tau(p, s)$ in all conditions. The p -value of $\tau(d, s)$ are greater than those of $\tau(p, s)$ (Table 4). Unsurprisingly, we find that the $\tau(d, s) - \tau(p, s)$ is never significant – neither on individual conditions (Table 6), nor on the global analysis (see $S(d) - S(p)$ in Table 5).

In Malayalam results are mixed: the sign of $\tau(d, s) - \tau(p, s)$ depends on the condition but $\tau(d, s)$ beats $\tau(p, s)$ in the condition where both $\tau(d, s)$ and $\tau(p, s)$ are maximum given the distance measure ($\tau(d, s) = 0.867 > \tau(p, s) = 0.8$ in Table 4). Thus, in that condition, swap distance minimization has an effect in Malayalam that cannot be reduced to preference for verb-last. The lack of verb initial orders with two overt arguments in Leela's corpus, in spite of being grammatically possible, suggests that undersampling may be limiting the observation of a stronger swap distance minimization effect when frequencies are used as a proxy for cognitive cost. As it happened with Korean, we find that the $\tau(d, s) - \tau(p, s)$ is never significant neither on individual conditions (Table 6) nor on the global analysis (see $S(d) - S(p)$ in Table 5).

In Sinhalese we find the opposite phenomenon with respect to Korean: the effect of swap distance minimization is stronger: given a score and a condition, $\tau(d, s) > \tau(p, s)$ in all cases. Interestingly, we find that the $\tau(d, s) - \tau(p, s)$ is never significant on individual conditions (Table 6) and this is confirmed in the global analysis (see $S(d) - S(p)$ in Table 5).

6 Discussion

We have seen that an effect consistent with swap distance minimization is found in all three languages (Table 4). However, we have seen that in Sinhalese, the effect is weak and requires a global analysis over all conditions for it to become statistically significant (Table 5).

We have demonstrated that swap distance minimization is significantly stronger than a preference for the canonical order in Korean and Malayalam by means of a global analysis across conditions (Table 5). In Malayalam, swap distance minimization is so strong that its superiority with respect to a preference for the canonical order manifests also on individual conditions (Table 6). Notice that the acceptability ranks in Table 1 coincide with a labelling of the vertices of the permutahedron following a traversal of the permutahedron from SOV (Figure 3), which is known as breadth first traversal in computer science (Cormen et al., 1990). There are $5! = 120$ possible traversals starting at SOV, but only 4 four of them are breadth first traversals; the acceptability rank (that results from transforming mean acceptability scores into ranks) has hit one of them. In Sinhalese, swap distance minimization is neither significantly stronger than a preference for the canonical order nor significantly stronger than the preference for verb-last (Table 5) that is believed to explain acceptability in Malayalam (Namboodiripad and Goodall, 2016; Namboodiripad, 2017).

We have provided evidence that swap distance minimization is cognitively relevant in capturing human behavior: it is significantly stronger than the principle it subsumes, i.e. the preference for the canonical order, in Korean and in Malayalam. In Sinhalese, we failed to find that swap distance minimization is acting significantly stronger than a preference for the canonical order. It is possible that swap distance minimization is acting beyond a preference for the canonical order, but its additional contribution with respect to other word order principles may remain statistically invisible. First, recall that swap distance minimization subsumes the preference for the canonical word order. Second, swap distance minimization and preference for verb-last are strongly correlated. Recall that the Kendall τ correlation between d and p , $\tau(d, p)$ is significantly high while $\tau(d, c)$ and $\tau(p, c)$ are not (Table 2). This is in line with the view that word order is a multiconstraint satisfaction principle, and word orders can compete or collaborate (Ferrer-i-Cancho, 2017). Third, our analyses on Sinhalese are based on data which is averaged across

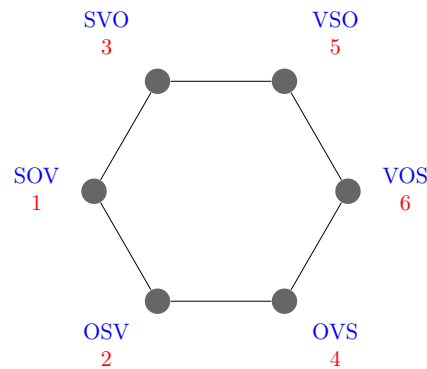


Figure 3: The word order permutation ring with the acceptability rank of every word order marked in red below each word order. The word order with the highest mean acceptability has rank 1, the word order with the 2nd highest mean acceptability has rank 2 and so on.

participants. Because we could not control for individual variation in that language as in Namboodiripad’s dataset (Section 3), the effects of swap distance minimization could indeed be stronger than what our analysis has revealed. Thus, controlling for individual variation in Sinhalese should be the subject of future research. Finally, the behavioral measures are not uniform across languages, as we currently do not have acceptability scores for Sinhalese, which could contribute to apparent differences across languages.

In neurolinguistics, it has been found that activity in certain brain regions (e.g., the left inferior frontal gyrus) is higher for non-canonical orders than for canonical orders (Meyer and Friederici, 2016). We suggest an interpretation of this finding as a consequence of a mental “rotation” operation to retrieve the canonical order (Figure 2) and propose a new research line: the use of swap distance as a more fine grained predictor of brain activity with respect to the traditional binary contrast of canonical versus non-canonical order (Meyer and Friederici, 2016, Table 48.1).

The strength of the swap distance minimization compared to the effect of other principles depends on the language. In Korean, the manifestation of swap distance minimization is weaker than that of the maximization of the predictability of the verb but stronger than a preference for the canonical order (Table 6). In Malayalam, swap distance minimization exhibits the strongest effect (Table 4). In Sinhalese, swap distance minimization is the second strongest, as in Korean, but the preference for a canonical order exhibits the strongest effect (Table 4).

We speculate that the major findings summarized above are consistent with the following scenario. First, recall that there is evidence that Korean exhibits a word order flexibility close to that of English and that Korean is more rigid than Malayalam (Levshina et al., 2023). The proposals of Sinhalese and Malayalam as non-configurational languages (Mohanani, 1983; Prabath and Ananda, 2017; Tamaoka et al., 2011)

suggest these two languages exhibit more word order freedom than Korean. ¹²

Second, consider the following arguments. As we discussed in [Section 1](#), strong evidence of swap distance minimization requires that interference from other word order principles is reduced. The fact that Korean is the only language where the maximization of the predictability of the verb has the strongest effect, provides additional support for the rigidity of Korean and the possible interference of that principle with swap distance minimization. As one moves from more rigid word orders to more flexible word orders, one expects that the manifestation of swap distance minimization becomes clearer. Accordingly, Malayalam exhibits the strongest manifestation of swap distance minimization but a weaker effect of the maximization of the predictability of the verb. However, an excess of word order flexibility may shadow the manifestation of swap distance minimization. If we assume that Sinhalese has the highest degree of word order flexibility, it is not surprising that none of the principles has a significant effect on individual conditions ([Table 4](#)) and that swap distance minimization does not show a significantly stronger effect than other word order preferences after a global analysis over conditions ([Table 5](#)).

A weakness of the arguments above is that, for Sinhalese, we are not measuring word order flexibility in the same way as for Korean and Malayalam. We are just assuming it should be very flexible according to the non-configurational hypothesis (Prabath and Ananda, 2017; Tamaoka et al., 2011), and, as argued in (Levshina et al., 2023), going from categorical to gradient characterizations of constituent order typology is critical to building explanatory models in this domain (see also Yan and Liu (2023) for research on categorical versus gradient characterizations). Thus, an urgent task is to investigate word order flexibility in Sinhalese in a cross-linguistically comparable way, perhaps with the same methodology as in Namboodiripad's research program (Namboodiripad, 2017, 2019; Namboodiripad et al., 2019). The complementary is also another important question for future research, namely, investigating reaction times and error rates in Malayalam and Korean with the methodology of (Tamaoka et al., 2011). We hope this research stimulates researchers also to investigate languages with canonical orders other than SOV (cf. Garrido Rodriguez et al., 2023). The predictions of swap distance minimization on non-SOV languages are already available in [Equation 2](#).

Finally, an implication of swap distance minimization for word order evolution is a tendency to preserve the canonical order, as variants that deviate from it will be more costly (contra misinterpretations of efficiency-based explanations which might lead one to predict that SOV languages should eventually

¹²Non-configurationality can be seen from a strong *a priori* theoretical assumption, namely that non-configurationality is an adjustable parameter in a language as opposed to an emergent property which becomes apparent via the interaction of a constellation of other factors Ferrer-i-Cancho, 2017. We take the position of Levshina et al., 2023, that languages are not separable into configurational or non-configurational, but rather that they vary along a cline in degree of flexibility. However, we do currently mention a role for non-configurationality on Page 19.

change to SVO). That tendency would be reinforced by other principles that determine the optimality of the canonical word order, e.g., in verb final languages, the placement of the verb is optimal with respect to maximization of the predictability of the verb (Ferrer-i-Cancho, 2017), and we have shown that a preference for verb-last and swap distance minimization are strongly correlated (Table 2). Therefore, it is not surprising that grammars are robustly transmitted even during instances of rapid discontinuities in language change, such as the emergence of creole languages; the dominant word order in creoles is overwhelmingly that of the lexifiers (Blasi et al., 2017). As such, swap distance minimization provides one potential answer for why languages vary when it comes to how much they minimize dependencies. Moreover, the findings here exemplify cases where general efficiency-based explanations do not lead to the same outcomes for every language, even when those languages on the surface seem to be very similar. Additional typological features, such as degree of flexibility, interact with swap distance minimization and dependency length minimization, leading us to predict structured variation across languages in how these very general principles are applied and manifest.

Acknowledgments

We are very grateful to L. Alemany-Puig for a careful revision of the manuscript and to L. Meyer for helpful comments. We also thank V. Franco-Sánchez and A. Martí-Llobet for helpful discussions on swap distance minimization. We became aware of the concept of permutahedron in combinatorics thanks to V. Franco-Sánchez. RFC is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya) and the grants AGRUPS-2022 and AGRUPS-2023 from Universitat Politècnica de Catalunya.

References

- Alemany-Puig, L., Esteban, J. L., Ferrer-i-Cancho, R.** (2022). Minimum projective linearizations of trees in linear time. *Information Processing Letters*, 174, 106204. <https://doi.org/10.1016/j.ipl.2021.106204>
- Austin, P., Bresnan, J.** (1996). Non-configurationality in Australian aboriginal languages. *Natural Language and Linguistic Theory*, 14(2), 215–268. <https://doi.org/10.1007/bf00133684>
- Blasi, D. E., Michaelis, S. M., Haspelmath, M.** (2017). Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour*, 1(10), 723–729. <https://doi.org/10.1038/s41562-017-0192-4>
- Ceballos, C., Manneville, T., Pilaud, V., Pournin, L.** (2015). Diameters and geodesic properties of generalizations of the associahedron. *Discrete Mathematics & Theoretical Computer Science, DMTCS Proceedings, 27th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2015)*. <https://doi.org/10.46298/dmtcs.2540>

- Cooper, L. A., Shepard, R. N.** (1973). Chronometric studies of the rotation of mental images. In W. G. CHASE (Ed.), *Visual information processing* (pp. 75–176). Academic Press. <https://doi.org/10.1016/B978-0-12-170150-5.50009-3>
- Corbett, G.** (1993). The head of Russian numeral expressions. In *Heads in grammatical theory* (pp. 11–35). Cambridge University Press Cambridge. <https://doi.org/https://doi.org/10.1017/CBO9780511659454>
- Cormen, T. H., Leiserson, C. E., Rivest, R. L.** (1990). *Introduction to algorithms*. The MIT Press.
- Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., Monaghan, P.** (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10), 603–615. <https://doi.org/https://doi.org/10.1016/j.tics.2015.07.013>
- Dryer, M. S.** (2013). Order of subject, object and verb. In M. S. Dryer M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/81>
- Ferrer-i-Cancho, R.** (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70, 056135. <https://doi.org/10.1103/PhysRevE.70.056135>
- Ferrer-i-Cancho, R.** (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, 11(3), 393–414. <https://doi.org/10.1142/S0219525908001702>
- Ferrer-i-Cancho, R.** (2014). Towards a theory of word order. Comment on “Dependency distance: A new perspective on syntactic patterns in natural language” by Haitao Liu et al. *Physics of Life Reviews*, 21, 218–220. <https://doi.org/10.1016/j.plrev.2017.06.019>
- Ferrer-i-Cancho, R.** (2015a). The placement of the head that minimizes online memory. A complex systems approach. *Language Dynamics and Change*, 5(1), 114–137. <https://doi.org/10.1163/22105832-00501007>
- Ferrer-i-Cancho, R.** (2015b). Reply to the commentary “Be careful when assuming the obvious”, by P. Alday. *Language Dynamics and Change*, 5(1), 147–155. <https://doi.org/10.1163/22105832-00501009>
- Ferrer-i-Cancho, R.** (2016). Kauffman’s adjacent possible in word order evolution. *The evolution of language: Proceedings of the 11th International Conference (EVOLANG11)*.
- Ferrer-i-Cancho, R.** (2017). The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics*, 39, 38–71.
- Ferrer-i-Cancho, R., Gómez-Rodríguez, C.** (2021a). Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1), 50–76. <https://doi.org/10.1080/09296174.2019.1645547>
- Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., Alemany-Puig, L.** (2022). Optimality of syntactic dependency distances. *Physical Review E*, 105(1), 014308. <https://doi.org/10.1103/PhysRevE.105.014308>
- Ferrer-i-Cancho, R., Gómez-Rodríguez, C.** (2021b). Dependency distance minimization predicts compression. *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, 45–57. <https://aclanthology.org/2021.quasy-1.4/>

- Futrell, R., Levy, R. P., Gibson, E.** (2020). Dependency locality as an explanatory principle for word order. *Language*, 96(2), 371–412. <https://doi.org/10.1353/lan.2020.0024>
- Futrell, R., Mahowald, K., Gibson, E.** (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences USA*, 112(33), 10336–10341. <https://doi.org/https://doi.org/10.1073/pnas.1502134112>
- Garrido Rodriguez, G., Norcliffe, E., Brown, P., Huettig, F., Levinson, S. C.** (2023). Anticipatory processing in a verb-initial Mayan language: Eye-tracking evidence during sentence comprehension in Tseltal. *Cognitive Science*, 47(1), e13292. <https://doi.org/https://doi.org/10.1111/cogs.13219>
- Garrod, S., Pickering, M. J.** (2013). Dialogue: Interactive alignment and its implications for language learning and language change. In *The language phenomenon* (pp. 47–64). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36086-2_3
- Gell-Mann, M., Ruhlen, M.** (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences USA*, 108(42), 17290–17295. <https://doi.org/10.1073/pnas.1113716108>
- Gildea, D., Temperley, D.** (2007). Optimizing grammars for minimum dependency length. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 184–191. <https://www.aclweb.org/anthology/P07-1024>
- Givón, T.** (1979). *On understanding grammar*. Academic.
- Gómez-Rodríguez, C., Christiansen, M., Ferrer-i-Cancho, R.** (2022). Memory limitations are hidden in grammar. *Glottometrics*, 52, 39–64. https://doi.org/10.53482/2022_52_397
- Gómez-Rodríguez, C., Ferrer-i-Cancho, R.** (2017). Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96, 062304. <https://doi.org/10.1103/PhysRevE.96.062304>
- Hale, K.** (1983). Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, 1(1). <https://doi.org/10.1007/bf00210374>
- Hammarström, H.** (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1), 19–29. <https://doi.org/10.1093/jole/lzw002>
- Hyönä, J., Hujanen, H.** (1997). Effects of case marking and word order on sentence parsing in Finnish: An eye fixation analysis. *Quarterly Journal of Experimental Psychology*, 50, 841–858. <https://doi.org/10.1080/713755738>
- Kaiser, E., Trueswell, J. C.** (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2), 113–147. <https://doi.org/10.1016/j.cognition.2004.01.002>
- Kendall, M. G.** (1970). *Rank correlation methods* (4th). Griffin.
- Koizumi, M., Kim, J.** (2016). Greater left inferior frontal activation for SVO than VOS during sentence comprehension in kaqchikel. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01541>

- Leela, M.** (2016). Early acquisition of word order: Evidence from Hindi, Urdu and Malayalam (Doctoral dissertation). Universitat Autònoma de Barcelona. <http://hdl.handle.net/10803/399556>
- Levshina, N., Namboodiripad, S., Allasonnière-Tang, M., Kramer, M., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T. M., Kidd, E., Liu, Z., Naccarato, C., Nordlinger, R., Panova, A., Stoyanova, N.** (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4), 825–883. <https://doi.org/10.1515/ling-2021-0098>
- Lin, D.** (1996). On the structural complexity of natural language sentences. *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. <https://aclanthology.org/C96-2123>
- Liu, H.** (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9, 159–191. <https://doi.org/10.17791/jcs.2008.9.2.159>
- Liu, H., Xu, C., Liang, J.** (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, 171–193. <https://doi.org/10.1016/j.plrev.2017.03.002>
- Meir, I., Sandler, W., Padden, C., Aronoff.** (2010). Emerging sign languages. In M. Marschark P. E. Spencer (Eds.), *Oxford handbook of deaf studies, language, and education* (pp. 267–280). Oxford University Press Oxford. <https://doi.org/10.1093/oxfordhb/9780195390032.013.0018>
- Menn, L.** (2000). It's time to face a simple question: Why is canonical form simple? *Brain and Language*, 71(1), 157–159. <https://doi.org/10.1006/brln.1999.2239>
- Meyer, L., Friederici, A. D.** (2016). Chapter 48 - neural systems underlying the processing of complex sentences. In G. Hickok S. L. Small (Eds.), *Neurobiology of language* (pp. 597–606). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-407794-2.00048-1>
- Mohanan, K.** (1983). Lexical and configurational structures. *The Linguistics Review*, 3, 113–139. <https://doi.org/10.1515/tlir.1983.3.2.113>
- Morrill, G.** (2000). Incremental processing and acceptability. *Computational Linguistics*, 25(3), 319–338. <https://aclanthology.org/J00-3002>
- Motamedi, Y., Wolters, L., Schouwstra, M., Kirby, S.** (2022). The effects of iconicity and conventionalization on word order preferences. *Cognitive Science*, 46(10). <https://doi.org/10.1111/cogs.13203>
- Namboodiripad, S., Garcia-Amaya, L., Kramer, M., Tobin, S., Sedarous, Y., Henriksen, N., Boland, J., Coetzee, A.** (2020). Verb position and flexible constituent order processing: Comparing verb-final and verb-medial languages. *Poster at 33rd CUNY Conference on Human Sentence Processing. Amherst, Massachusetts*. <https://osf.io/d9wq8/>
- Namboodiripad, S., Goodall, G.** (2016). Verb position predicts acceptability in a flexible SOV language. *Poster at 29th CUNY Conference on Human Sentence Processing. Gainesville, Florida*.
- Namboodiripad, S.** (2017). An Experimental Approach to Variation and Variability in Constituent Order (PhD Thesis). UC San Diego. <https://escholarship.org/uc/item/2sv6z8bz>

- Namboodiripad, S.** (2019). A gradient approach to flexible constituent order. <https://doi.org/10.31234/osf.io/rvjn5>
- Namboodiripad, S., Kim, D., Kim, G.** (2019). English dominant and Korean speakers show reduced flexibility in constituent order. *Proceedings of Chicago Linguistics Society 53*. <http://savi.ling.lsa.umich.edu/publications/CLSmanuscript.pdf>
- Newmeyer, F. J.** (2000). On the reconstruction of 'proto-world' word order. In C. K. *et al.* (Ed.), *The evolutionary emergence of language* (pp. 372–388). Cambridge University Press.
- Niu, R., Liu, H.** (2022). Effects of syntactic distance and word order on language processing: An investigation based on a psycholinguistic treebank of English. *Journal of Psycholinguistic Research*, 51(5), 1043–1062. <https://doi.org/10.1007/s10936-022-09878-4>
- Occhino, C., Anible, B., Wilkinson, E., Morford, J. P.** (2017). Iconicity is in the eye of the beholder: How language experience affects perceived iconicity. *Gesture*, 16(1), 100–126.
- Ohta, S., Koizumi, M., Sakai, K. L.** (2017). Dissociating effects of scrambling and topicalization within the left frontal and temporal language areas: An fMRI study in Kaqchikel Maya. *Frontiers in Psychology*, 8, 748.
- Perniss, P., Thompson, R., Vigliocco, G.** (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1. <https://doi.org/10.3389/fpsyg.2010.00227>
- Pickering, M. J., Garrod, S.** (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2-3), 203–228. <https://doi.org/10.1007/s11168-006-9004-0>
- Prabath, K., Ananda, M. L.** (2017). Configurationality and mental grammars: Sentences in Sinhala with re-duplicated expressions. *International Journal of Multidisciplinary Studies*, 3(2), 25. <https://doi.org/10.4038/ijms.v3i2.4>
- Sandler, W., Meir, I., Padden, C., Aronoff, M.** (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences USA*, 102, 2661–2665. <https://doi.org/10.1073/pnas.0405448102>
- Tamaoka, K., Kanduboda, P., Sakai, H.** (2011). Effects of word order alternation on the sentence processing of Sinhalese written and spoken forms. *Open Journal of Modern Linguistics*, 1, 24–32. <https://doi.org/10.4236/ojml.2011.12004>
- Tarr, M. J., Pinker, S.** (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21, 233–282. [https://doi.org/10.1016/0010-0285\(89\)90009-1](https://doi.org/10.1016/0010-0285(89)90009-1)
- Temperley, D.** (2008). Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3), 256–282. <https://doi.org/10.1080/09296170802159512>
- Temperley, D., Gildea, D.** (2018). Minimizing syntactic dependency lengths: Typological/Cognitive universal? *Annual Review of Linguistics*, 4(1), 67–80. <https://doi.org/10.1146/annurev-linguistics-011817-045617>
- Winter, B., Sóskuthy, M., Perlman, M., Dingemanse, M.** (2022). Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-021-04311-7>

Xu, C., Liang, J., Liu, H. (2017). DDM at work. *Physics of Life Reviews*, 21, 233–240. <https://doi.org/10.1016/j.plrev.2017.07.001>

Yan, J., Liu, H. (2023). Basic word order typology revisited: A crosslinguistic quantitative study based on UD and WALs. *Linguistics Vanguard*. <https://doi.org/10.1515/lingvan-2021-0001>

Appendix

The maximum Kendall correlation

Recall the definition of τ in Equation 5. Let n_0 be the number of pairs that are neither concordant nor discordant.

Property 1.

$$(6) \quad \frac{n_0}{\binom{n}{2}} - 1 \leq \tau \leq 1 - \frac{n_0}{\binom{n}{2}}.$$

Proof. By definition,

$$n_c + n_d + n_0 = \binom{n}{2}.$$

The substitution

$$n_c = \binom{n}{2} - n_d - n_0$$

transforms Equation 5 into

$$\tau = 1 - \frac{2n_d + n_0}{\binom{n}{2}}.$$

The latter and the fact that $n_d \geq 0$ by definition leads to

$$\tau \leq 1 - \frac{n_0}{\binom{n}{2}}.$$

By symmetry, the substitution

$$n_d = \binom{n}{2} - n_c - n_0$$

transforms Equation 5 into

$$\tau = \frac{2n_c + n_0}{\binom{n}{2}} - 1.$$

The latter and the fact that $n_c \geq 0$ by definition leads to

$$\tau \geq \frac{n_0}{\binom{n}{2}} - 1.$$

Hence we conclude Equation 6. □

Consider the Kendall τ correlation between x and y . Let N_x be the number of distinct values of x and N_y be the number of distinct values of y . Let us group the values of x in a tie and define t_i the number of tied values in the i -th group. Let us group the values of y in a tie and define u_i the number of tied values in the i -th group. Then

Property 2.

$$(7) \quad n_0 \geq \max \left(\sum_{i=1}^{N_x} \binom{t_i}{2}, \sum_{i=1}^{N_y} \binom{u_i}{2} \right).$$

Proof. Notice that pairs formed with values in a tie cannot be neither concordant nor discordant. Then the i -th tie group of x contributes with $\binom{t_i}{2}$ pairs of points that are not concordant nor discordant. Then, the overall contribution to pairs of this sort by x is

$$\sum_{i=1}^{N_x} \binom{t_i}{2}.$$

Similarly, the contribution by y to pairs of points that are neither concordant nor discordant is

$$\sum_{i=1}^{N_y} \binom{u_i}{2}.$$

Combining the contributions of x and y one retrieves [Equation 7](#). The reader with some statistical background may have already realized that the summations over the number of distinct pairs in a group above are the ingredients of the adjustment for ties in the denominator in the definition of τ_b (Kendall, 1970). □

The next property presents the range of variation of τ for each distance measure

Property 3. Consider the Kendall correlation, i.e $\tau(x, y)$ where x is some distance measure and y can be any (for instance, y can be some score s). We have that

$$\begin{aligned} -\frac{13}{15} = -0.8\bar{6} &\leq \tau(d, y) \leq \frac{13}{15} = 0.8\bar{6} \\ -\frac{4}{5} = -0.8 &\leq \tau(p, y) \leq \frac{4}{5} = 0.8. \\ -\frac{1}{3} = -0.\bar{3} &\leq \tau(c, y) \leq \frac{1}{3} = 0.\bar{3}. \end{aligned}$$

Proof. Now we will derive the range of variation of τ for each distance measure by applying an implication of [Equation 7](#), namely

$$n_0 \geq \sum_{i=1}^{N_x} \binom{t_i}{2}.$$

Notice that

$$n_0 = \sum_{i=1}^{N_x} \binom{t_i}{2}$$

This happens when all the values of y are different. This is a typical situation when using continuous scores, as repeated values are unlikely except in case of lack of numerical precision.

Consider the matrix in [Table 1](#). In case of $\tau(d, s)$, there are four groups with $t_1 = t_4 = 1$ (for $d = 1$ and $d = 3$) and $t_2 = t_3 = 2$ (for $d = 1$ and $d = 2$), that yield

$$n_0 = \sum_{i=1}^{N_x} \binom{t_i}{2} = 2 \binom{2}{2} = 2$$

and then [Equation 6](#) gives

$$\tau(d, s) \leq 1 - \frac{2}{15} = \frac{13}{15}.$$

In case of $\tau(p, s)$, there are three groups with $t_1 = t_2 = t_3 = 2$ (two points in a tie for $p = 0$, $p = 1$ and also $p = 2$), that yield

$$n_0 = 3 \binom{2}{2} = 3$$

and then [Equation 6](#) gives

$$\tau(p, s) \leq 1 - \frac{3}{15} = \frac{4}{5}.$$

Finally, in case of $\tau(c, s)$, there are only two groups with $t_1 = 1$ and $t_2 = 5$ (5 points in a tie for $c = 1$), that yield

$$n_0 = \binom{5}{2} = 10$$

and then [Equation 6](#) gives

$$\tau(c, s) \leq 1 - \frac{10}{15} = \frac{1}{3}.$$

The lower bounds are obtained just by inverting the sign thanks to [Equation 6](#). □

The following corollary indicates that if $\tau(d, y)$ is sufficiently large then no other distance measure can give a higher correlation and also the symmetric, namely, if $\tau(d, y)$ is sufficiently small then no other distance measure can give a smaller correlation.

Corollary 1. *If $\tau(d, y) > 1/3$ then $\tau(d, y) > \tau(c, y)$. If $\tau(d, y) > 4/5$ then $\tau(d, y) > \tau(p, y), \tau(c, y)$.*

If $\tau(d, y) < -1/3$ then $\tau(d, y) < \tau(c, y)$. If $\tau(d, y) < -4/5$ then $\tau(d, y) < \tau(p, y), \tau(c, y)$.

Proof. A trivial consequence of [Proposition 3](#). □

The minimum p -value of the Kendall correlation test

As we explain in [Section 4](#), the p -value of the Kendall τ correlation test is computed exactly by enumerating all the $6! = 720$ permutations. In general,

$$p\text{-value} \geq \frac{m}{n!},$$

where m is the number of permutation with the same τ as the actual one. Notice that $m \geq 1$ because the permutation that coincides with the current ordering yields the same τ . As the test is one-sided and $m \geq 1$, one obtains

$$p\text{-value} \geq 1/6! = \frac{1}{720} = 0.0013\bar{8}.$$

However, a more accurate lower bound of m is given by

Property 4.

$$(8) \quad m \geq \max \left(\prod_{i=1}^{N_x} t_i!, \prod_{i=1}^{N_y} u_i! \right).$$

Proof. Every permutation of values in the same tie group does not produce a different sequence. For the i -th group of x , there are $t_i!$ permutations of values in the same group that do not produce a different sequence. Integrating all the groups, one obtains that there are

$$\prod_{i=1}^{N_x} t_i!$$

permutations of the x column of the matrix that produce the same sequence. By symmetry, there are

$$\prod_{i=1}^{N_y} u_i!$$

permutations of the y column of the matrix that produce the same sequence. Combining the contributions of x and y , we obtain [Equation 8](#). □

[Equation 8](#) leads to more accurate lower bounds of the p -value of τ that are presented in the following property.

Property 5. Consider the p -value of the exact right sided correlation test of $\tau(x, y)$ where x is some distance and y can be any (for instance, y can be some score s). The p -value of $\tau(d, y)$ satisfies

$$p\text{-value} \geq \frac{1}{180} = 0.00\bar{5}.$$

The p -value of $\tau(p, y)$ satisfies

$$p\text{-value} \geq \frac{1}{90} = 0.0\bar{1}.$$

The p -value of $\tau(c, y)$ satisfies

$$p\text{-value} \geq \frac{1}{6} = 0.1\bar{6}.$$

Proof. Now we will derive a lower bound of the p -value for each distance measure neglecting any information of about the distribution of the values of y , namely applying an implication of [Equation 8](#), that is

$$m \geq \prod_{i=1}^{N_x} t_i!.$$

Notice that

$$m = \prod_{i=1}^{N_x} t_i!$$

holds when all the values of y are different. This is a typical situation when using continuous scores, as we have explained above.

For $\tau(d, s)$, the four groups with $t_1 = t_4 = 1$ (for $d = 1$ and $d = 3$) and $t_2 = t_3 = 2$ (for $d = 1$ and $d = 2$) give

$$p\text{-value} \geq \frac{4}{6!} = \frac{1}{180}.$$

For $\tau(p, s)$, the three groups with $t_1 = t_2 = t_3 = 2$ (two points in a tie for $p = 0$, $p = 1$ and also $p = 2$) give

$$p\text{-value} \geq \frac{8}{6!} = \frac{1}{90}.$$

Finally, for $\tau(c, s)$, the only two groups with $t_1 = 1$ and $t_2 = 5$ (5 points in a tie for $c = 1$) give

$$p\text{-value} \geq \frac{5!}{6!} = \frac{1}{6} = 0.1\bar{6}.$$

□

Modality of verbs as stylometric feature in Czech genres

Miroslav Kubát¹ , Xinying Chen^{1*} 

¹ University of Ostrava

* Corresponding author's email: cici13306@gmail.com

DOI: https://doi.org/10.53482/2023_55_413

ABSTRACT

The goal of the study is to analyze modality of verbs from the perspective of its usage in different types of Czech texts. The analysis is based on data from the Czech National Corpus, specifically the balanced corpus of contemporary written Czech SYN2020, which contains 100 million words. The proportion of modal verbs to all verbs is used to measure modality. Furthermore, different types of modality are considered: necessity and possibility. The findings reveal distinct patterns in the use of modal verbs in different genres. Thus, index of modality seems to be a promising stylometric feature. Non-fiction literature, especially administrative texts, exhibits the highest modality. In contrast fiction texts, namely poetry, has the lowest modality.

Keywords: syntax, modality, stylometry, Czech.

1 Introduction

Stylometry is a branch of linguistics that focuses on the quantitative analysis of various styles and the identification of distinctive features within texts. Stylometry has numerous applications, including authorship attribution, genre classification, and text clustering, and has become an increasingly popular tool for literary scholars, linguists, and forensic investigators (cf. Holmes 1998; Juola 2007; Savoy 2020). By providing insights into the stylistic characteristics of texts, stylometry can offer a deeper understanding of language use.

There are two main branches of stylometry. The first one is based on simple quantitative indicators such as word frequencies, mean sentence length, or text features like lexical diversity. This approach is more traditional and allows straightforward interpretation of the obtained data, which is important when one wants to understand different styles of writing. In the second approach, the methods usually belong to machine learning algorithms, most recently neural networks (see e.g. Matthews and Merriam 2020; Savoy 2020). Thus, this approach belongs to the black box method category, in which linguistic interpretation is rather difficult or even impossible at this stage.

Our study belongs to the traditional stylometric approach based on simple and straightforward indices. The study deals particularly with one feature of verbs - modality. The methodology is inspired by similar stylometric indicators such as subjectivity, objectivity, descriptivity, activity or nominality (cf. Kubát et al. 2021, Zörnig et al. 2014). These indicators are based on simple ratios expressing text features. Despite their simplicity, they have shown to be useful for distinguishing different genres or authors (see e.g. Chen & Kubát 2022; Kubát et al. 2021; Místecký 2018; Zhou et al. 2022).

In this study, we propose a new index of modality. Modality of text is defined as the ratio of the number modal verbs to the number of all verbs in a text. Furthermore, we also focus on a ratio between modal verbs expressing possibility and necessity. Our goal is to discover how modality varies across different styles and genres in a big balanced corpus of contemporary written Czech SYN2020.

Modal verbs in Czech, as in many languages, exhibit a high degree of variability in their usage patterns. This variability is not random but is closely tied to genre-specific conventions and the communicative purposes of texts. For example, academic writing might favor certain modal verbs to express certainty or probability, while fiction may use them differently to depict character intentions or hypothetical scenarios. Analyzing the frequency and context of these modal verbs can thus provide valuable insights into the stylistic fingerprints of different text types. (cf. Chong et al. 2023; Huschová 2015)

2 Material

The language material comes from a large balanced corpus of contemporary written Czech SYN2020 (Křen et al. 2020) belonging to the series of synchronous corpora developed by Czech National Corpus. SYN2020 covers texts mainly from 2015–2019. The size of the corpus is 100 million words. SYN2020 is divided into three equally sized parts: FIC: fiction, NFC: non-fiction, NMG: newspapers and magazines. These three text-type groups are then divided into subcategories such as novel, poetry, humanities, etc. The text-type structure of SYN2020 can be seen in Table 1. A detailed description can be found on the website of SYN2020 <https://wiki.korpus.cz/doku.php/cnk:syn2020>.

Table 1: Text-type structure of SYN2020.

Text-Type Group	Text-Type
FIC: fiction	NOV: novels
	COL: short stories
	VER: poetry
	SCR: drama, screenplays
NFC: non-fiction	SCI: scientific literature
	PRO: professional literature
	POP: popular literature
	MEM: memoirs and autobiographies
	ADM: administrative
NMG: newspapers and magazines	NEW: news
	LEI: leisure magazines

SYN2020 is a syntactically annotated corpus, using a parser from the NeuroNLP2 toolkit trained on data from the Prague Dependency Treebank (Bejček et al. 2012) and the FicTree corpus (Jelínek 2017). It marks dependency relations between words and assigns syntactic functions. The corpus achieves high accuracy rates of 92.39% for UAS (unlabeled attachment score) and 88.73% for LAS (labeled attachment score). While errors are more common in less frequent syntactic functions, the most frequent functions have an error rate of less than 5% (<https://wiki.korpus.cz/doku.php/cnk:syn2020>). Despite some errors, SYN2020 is an outstanding syntactically annotated corpus. This is especially due to its size and balanced structure of various types of text.

3 Methodology

Verb modality expresses the speaker's attitude towards the action or state described by a verb. There are three verbs that are considered to be primary modal verbs in Czech language: *muset* [must], *moci* [can], *smět* [may]. These verbs express necessity or possibility (see Table 2). These primary modal verbs are also defined by several syntactic characteristics (see. Karlík & Šimík 2017)¹. The verb *mít* [to have] can also be used as modal verb in Czech. However, *mít* is mainly used as non-modal verb.² That is why *mít* is not counted as basic modal verb in this study.

¹ Karlík and Šimík (2017) define following several syntactic characteristics traditionally attributed to modal verbs:

- (a) They only go with infinitives, not with subordinate clauses.
- (b) They cannot be expanded with a noun phrase.
- (c) They do not form imperatives (commands).
- (d) They do not form passive voice.
- (e) They do not have aspectual counterparts (different forms showing the completion or duration of the action).
- (f) They do not form action nouns or verbal nouns.
- (g) It is possible to separately expand the modal verb and the full verb in a sentence.
- (h) Both the modal verb and the infinitive can be negated separately.

² Based on our small analysis of 300 random occurrences in SYN2020, 67 had a modal meaning. Since *mít* it is mainly used as a non-modal verb and the existing SYN2020 annotation framework does not support automatical distinction between the two, we decided to exclude the verb *mít* from our observation.

Although other verbs expressing desire, intention, or ability (e.g., *chtít* [to want], *potřebovat* [to need], *doufat* [to hope]) can be also considered as modal verbs in a broader sense, we work only with three basic aforementioned modal verbs in this research. The list of analyzed modal verbs can be seen in Table 2.

Table 2: Analyzed modal verbs.

Possibility	Necessity
<i>moci</i> ‘can’	<i>muset</i> ‘must’
<i>smět</i> ‘may’	<i>nemoci</i> ‘cannot’
<i>nemuset</i> ‘need not’	<i>nesmět</i> ‘must not’

We measure the level of modality by a ratio of modal verbs to all verbs³:

$$\text{modality} = \frac{\text{number of modal verbs}}{\text{number of all verbs}}$$

4 Results

The results in Figure 1 and Figure 2 show that non-fiction literature (NFC) has a higher modality than fiction (FIC). The newspapers and magazines (NMG) are then positioned in the middle. As for text-types (see Figure 2), administrative texts (ADM) have the highest modality (0.054) among all the analyzed text-types. Poetry (VER) has the lowest value (0.020). Although memoirs and autobiographies (MEM) are in SYN2020 structure assigned to non-fiction literature, these texts have a rather fiction-like modality. That can be explained by the fact that the writing style of memoirs and autobiographies is very close to fiction literature (cf. Soukupová 2015). This genre is therefore naturally somewhere between fiction and non-fiction literature.

This observed phenomenon can be attributed to the intrinsic purposes and contexts inherent in each genre. Non-fiction texts, particularly administrative documents, are largely concerned with providing directives, guidelines, and regulations. These texts need to articulate rules and expectations clearly, often dictating what actions are required, permitted, or prohibited in specific situations. See following examples from the corpus:

- “Souhrn finančních potřeb se vždy musí rovnat souhrnu finančních zdrojů.” [The sum of financial needs must always equal the sum of financial resources.]

³ In SYN2020, for searching all verbs, we use CQL query [tag="V.*"]; for searching modal verbs [lemma = "moci"], [lemma = "muset"], [lemma = "smět"].

- “Hygienické zařízení vyčleněné pro personál smí využívat jen personál.” [Sanitary facilities reserved for staff may only be used by staff.]
- “Žadatel nesmí požádat v průběhu roku, ve kterém mu byla poskytnuta dotace z Programu rozvoje venkova na stejný předmět dotace.” [The applicant may not apply during a year in which a subsidy from the Rural Development Programme was granted for the same object of subsidy.]

Conversely, fiction texts, which primarily encompass narratives, tend to focus on storytelling rather than instructing or informing. The narrative style of fiction is more about weaving events into a storyline, and less about prescribing behaviors or outcomes. Thus, the relative scarcity of modal verbs in fiction is indicative of a stylistic choice that aligns with the genre's focus on creative expression and character development.

Journalistic texts exhibit a modality that strikes a balance between the definitive nature of non-fiction and the narrative freedom of fiction. This intermediate modality reflects journalism's objective to inform with factual precision while also crafting compelling stories. Modal verbs in journalism navigate between asserting facts and suggesting possibilities, providing a versatile approach to engaging readers with news and narratives.

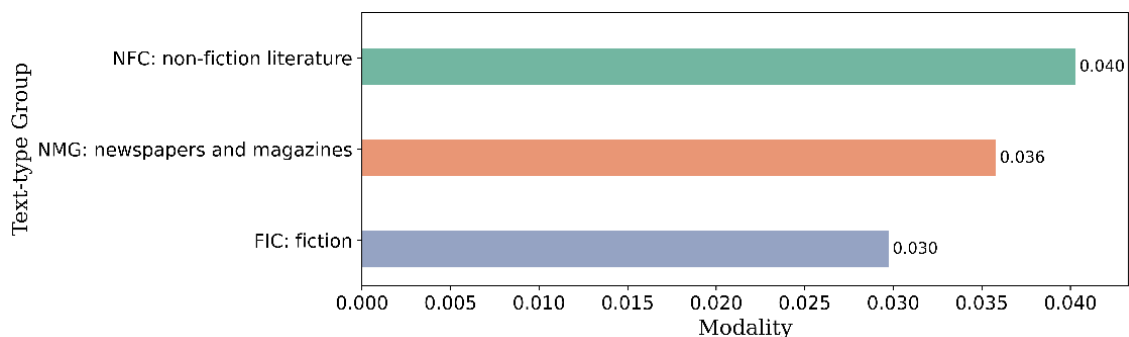


Figure 1: Modality in text-type groups.

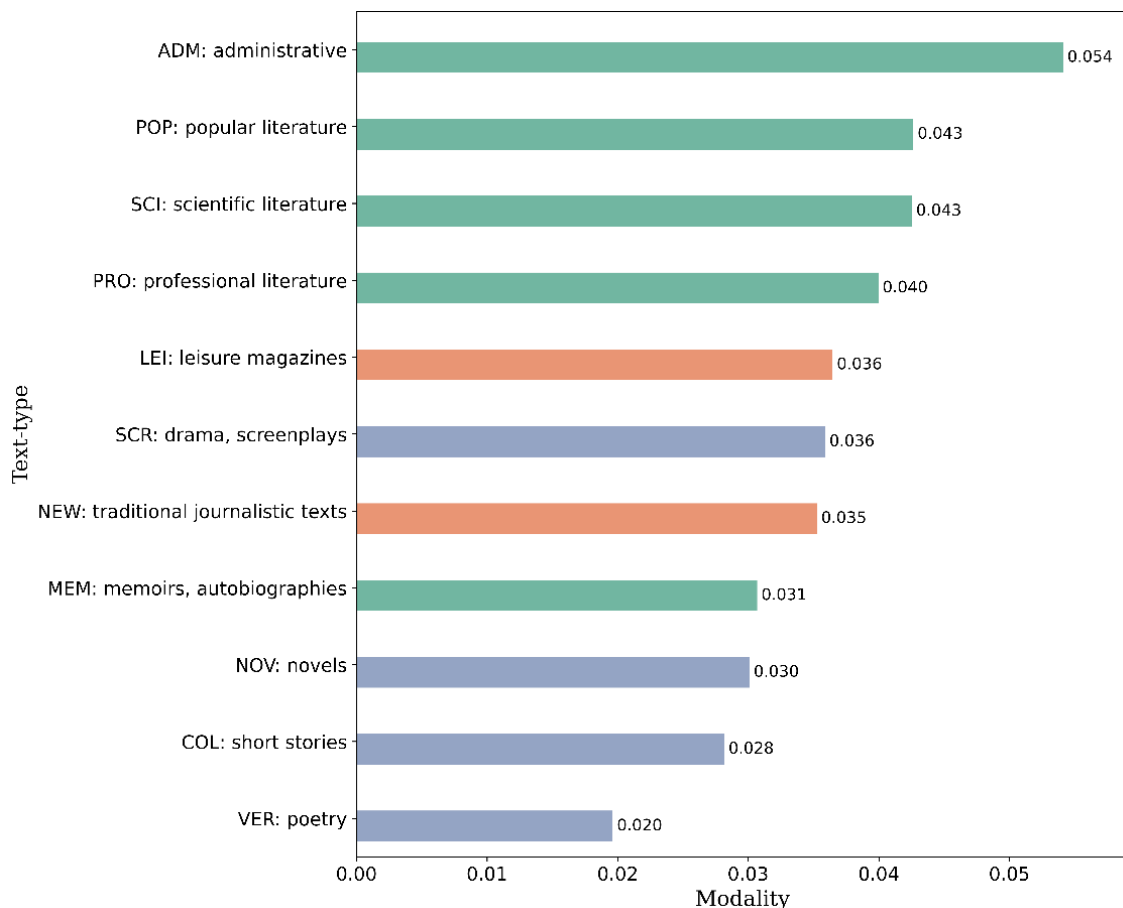


Figure 2: Modality in text-types.

Since the analyzed modal verbs belong to two groups of modality (possibility and necessity), we also focus on their more detailed usage. A percentage representation of modal verbs of possibility and modal verbs of necessity is calculated. In Figure 3, we can see that newspapers and magazines (NMG) and non-fiction texts (NFC) tend to use modal verbs that express possibility rather than necessity. Conversely, fiction (FIC) has more modal verbs of necessity. As shown in Figure 4, the ratio between possibility and necessity in particular text-types is fairly consistent without any clear outliers inside text-type groups. The only exceptions are memoirs and autobiographies (MEM) which again tend towards fiction (FIC) rather than non-fiction literature (NFC). It is interesting to note how consistent the values of fiction (FIC) are. Even so different genres such as fiction, drama, and poetry use modal verbs in similar proportions.

The results suggest that in fiction, there is often a stronger emphasis on situations where characters are compelled to act or are restricted from doing so, reflecting the conflicts and constraints that drive narrative tension. On the other hand, non-fiction texts, particularly scientific literature, show a preference for possibility modal verbs. This could be because scientific writing frequently explores hypotheses, suggests potential explanations, and discusses findings that are not absolute but rather indicative of a

probability. The high use of possibility modal verbs aligns with the tentative and exploratory nature of scientific inquiry.

It is also interesting to note that contemporary poetry, which is more intimate and personal form of fiction, demonstrates the lowest possibility and the highest necessity. This could be due to its focus on the human condition and personal experiences, which are often framed by necessity and constraints.

In general, the preference for necessity modal verbs in fiction can be seen as a reflection of the genre's focus on dramatizing human experiences, while the prevalence of possibility modal verbs in non-fiction, especially scientific literature, underscores a stylistic approach that accommodates the uncertainty and openness inherent in scientific exploration. Tables 3 and 4 present exact frequencies of specific modal verbs, shedding light on their individual roles in expressing possibility and necessity.

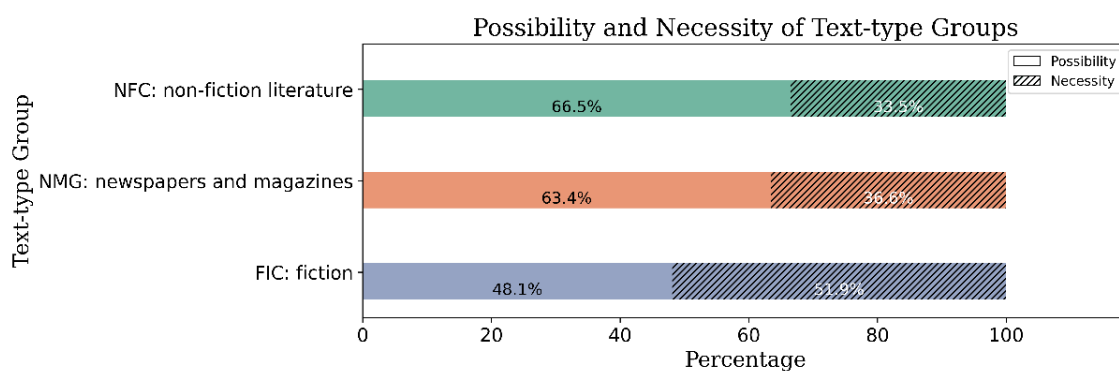


Figure 3: Percent representation of possibility and necessity modal verbs in text-type groups.

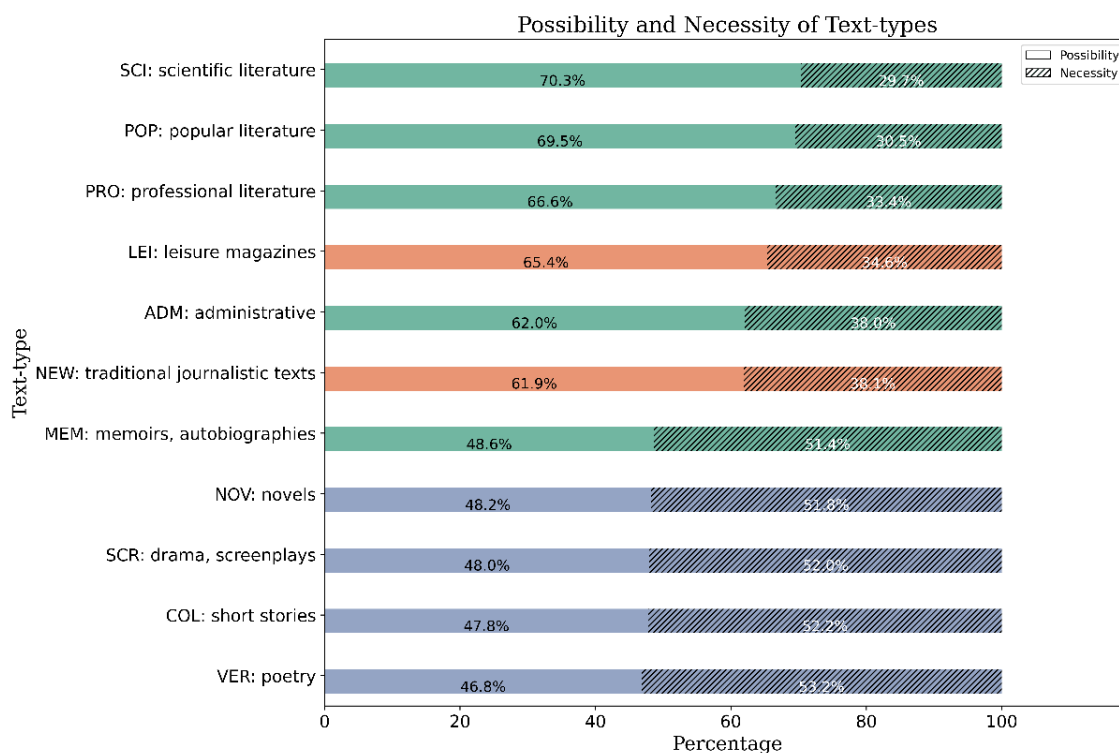


Figure 4: Percent representation of possibility and necessity modal verbs in text-types.

Table 3: Relative frequencies per milion (ipm) of modal verbs in text-type groups.

Text-type group	Possibility			Necessity		
	<i>moci</i>	<i>smět</i>	<i>nemuset</i>	<i>muset</i>	<i>nemoci</i>	<i>nesmět</i>
FIC: fiction	2886.335	53.65681	262.3288	2125.354	1155.582	174.0779
NFC: non-fiction	3703.843	27.41155	246.4694	1362.105	519.4416	123.6891
NMG: newspapers and magazines	3400.447	20.67973	281.7208	1533.359	482.8113	125.3138

From Table 3, it is evident that journalistic texts use *moci* less than non-fiction but more than fiction, suggesting a balance between expressing capabilities and acknowledging limitations in practical discourse. Fiction, while less frequently employing *moci*, shows a greater use of *muset*, reflecting a narrative emphasis on necessity and inevitability. Journalistic texts demonstrate a moderate usage of *nesmět*, possibly indicating a focus on the boundaries of societal norms and regulations within reported events. These variations in modal verb frequencies underscore the different linguistic strategies employed across genres, offering insightful data for stylistic studies.

Table 4: Relative frequencies per milion (ipm) of modal verbs in text-types.

Text-type	Possibility			Necessity		
	<i>moci</i>	<i>smět</i>	<i>nemuset</i>	<i>muset</i>	<i>nemoci</i>	<i>nesmět</i>
NOV: novels	2976.97	50.38	265.08	2193.69	1175.91	170.53
COL: short stories	2598.65	43.17	238.09	1896.15	1072.73	175.11
VER: poetry	1424.51	156.62	172.58	1118.26	663.38	214.47
SCR: drama, screenplays	3527.30	91.72	409.76	2579.18	1561.26	220.33
SCI: scientific literature	3644.69	22.51	181.80	1076.71	447.94	102.86
PRO: professional literature	3273.35	14.12	230.84	1353.94	300.13	112.35
POP: popular literature	4254.46	28.37	307.56	1355.69	543.72	119.27
MEM: memoirs, autobiographies	2767.64	52.35	227.00	2029.83	1003.50	184.08
ADM: administrative	3550.96	100.31	154.76	1653.67	303.79	378.31
NEW: traditional journalistic texts	3208.11	21.09	231.30	1546.24	477.51	110.04
LEI: leisure magazines	3688.83	20.07	357.32	1514.04	490.76	148.22

From Table 4, we can see that in general, *moci* largely shapes the concept of possibility, while *muset* and *nemoci* are key in defining the necessity. This distinction in modal verb usage between non-fiction and fiction genres suggests different stylistic approaches to expressing potentiality and obligation, which is a valuable observation for stylistic analysis. Furthermore, the results indicate that a) fiction texts use *muset* and *nemoci* more heavily, underscoring the themes of obligation and constraint in storytelling. b) Non-fiction texts, particularly scientific literature, rely more on *moci*, highlighting the prevalence of capability and theoretical possibility in academic discourse. c) Administrative texts use *nesmět* extensively, which is aligned with the genre's regulatory nature.

5 Conclusion

The study shows that non-fiction literature, especially administrative texts, reach the highest modality. In contrast, fiction, especially poetry, shows the lowest possibility. Journalistic texts are between them. In terms of usage modal verbs expressing possibility and necessity, fiction tends to use more modal verbs of necessity compared to non-fiction and journalism.

In conclusion, the observed consistency of modality values within text-type groups suggests that modality may offer a reliable and insightful tool for understanding different styles and genres. Index of modality seems to be therefore a useful measure for analyzing a wide variety of texts. It could thus be used in stylometry alongside other similar measures, such as subjectivity, attributivity, nominality, descriptivity, etc.

It is important to note that this is only an initial attempt to employ modality index in stylometry. Our preliminary conclusions need to be validated by further research. Furthermore, this technique may also be used for authorship attribution to determine whether modality is indicative of distinct writing styles among various authors. Since this study is limited to the Czech language, it would be interesting to examine this feature in other languages as well.

Acknowledgements

The research was supported by the Czech Science Foundation (GAČR), project No. 22-20632S.

References

- Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z.** (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, pp. 231-246.
- Chen, X., Kubát, M.** (2022). Rural versus urban fiction in contemporary Chinese literature – Quantitative approach case study. *Digital Scholarship in the Humanities*, 37(3), pp. 681-692.
- Chong, S. T., Ng, Y. J., Karthikeyan, J., Lee, S. Y.** (2023). The use of modals in academic discourse: A comparative analysis. In *AIP Conference Proceedings* (Vol. 2685, No. 1). AIP Publishing.
- Holmes, D. I.** (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), pp. 111-117.
- Huschová, P.** (2015). Exploring modal verbs conveying possibility in academic discourse. *Discourse and Interaction*, 8(2), pp. 35-47.
- Jelínek, T.** (2017). FicTree: a Manually Annotated Treebank of Czech Fiction. In: Hlaváčová, J. (Ed.). *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, pp. 181-185. Accessible at <http://ceur-ws.org/Vol-1885/181.pdf>.

- Juola, P.** (2007). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3), pp. 233-334.
- Karlík, P., Šimík, R.** (2017). Modální sloveso. In: Karlík, P., Nekula, M., Pleskalová, J. (Eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Accessible at <https://www.czechency.org/slovník/MOD%C3%81LN%C3%8D%20SLOVESO>.
- Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kovářiková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., Škrabal, M.** (2020). *SYN2020: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2020. Accessible at: <http://www.korpus.cz>.
- Kubát, M., Čech, R., Chen, X.** (2021). Attributivity and Subjectivity in Contemporary Written Czech. In: Chen, X., Čech, R. (Eds.): *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pp 58-64. Sofia: Association for Computational Linguistics.
- Kubát, M., Mačutek, J., Čech, R.** (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1), pp. 138-152.
- Matthews, R. A., Merriam, T. V.** (2020). Distinguishing literary styles using neural networks. In *Handbook of neural computation* (pp. G8-1). CRC Press.
- Místecký, M.** (2018). Counting Stylometric Properties of Sonnets: A Case Study of Machar's Letní sonety. *Glottometrics*, 41, pp. 1-12.
- Savoy, J.** (2020). *Machine learning methods for stylometry. Authorship Attribution and Author Profiling*. Cham: Springer.
- Soukupová, K.** (2015). Autobiografie: žánr a jeho hranice. *Česká literatura*, 63(1), pp. 49-72.
- Zhou, H., Jiang, Y., Wang, L.** (2022). Are Daojing and Dejing stylistically independent of each other: A stylometric analysis with activity and descriptivity. *Digital Scholarship in the Humanities*, 38(1), pp. 434-450.
- Zörnig, P.** (2014). *Descriptiveness, activity and nominality in formalized text sequences*. Lüdenscheid: RAM-Verlag.