Glottometrics

54/2023

International Quantitative Linguistics Association

Glottometrics

Vol. 54, 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/.

International Quantitative Linguistics Association (IQLA)

Friedmanngasse 50
1160 Vienna
Austria

eISSN 2625-8226

Contents

History of quantitative linguistics

A comparison of two text specificity measures analyzing a heterogenous text corpus

Anton Oleinik[1*]

[1] Memorial University of Newfoundland, St. John's, Canada
[*] Corresponding author's email: aoleynik@mun.ca

DOI: https://doi.org/10.53482/2023_54_404

ABSTRACT

The article compares the performance of two term specificity measures, Cohen's d and Z-score, when analyzing political and media discourses on Russia's war in Ukraine in four languages and five countries. In addition to linguistic and stylistic heterogeneity, 3,347 texts included in the corpus have variable length. The two measures display convergent validity, as confirmed by various performance metrics. It is argued that the measures can be adapted to a broader range of tasks in information retrieval and digital humanities, in addition to their usefulness for text mining and content analysis.

Keywords: text specificity, term specificity, text mining, content analysis

1 Introduction

Although information retrieval and text mining refer to separate fields of knowledge and lie at the origin of different application systems (search engines as opposed to text analytics programs), they remain closely interconnected (Zhai, Massung, 2016). The present article highlights an aspect of their interconnectivity related to text specificity measures. On the one hand, such measures assist in content analysis of texts helping to identify terms (words, n-grams) that distinguish one document from the other. On the other hand, in information retrieval text specificity measures may allow to assess the extent to which searchable documents are about the same thing as a search query treated as a text, although short (Savoy, 2019). Text similarity also plays a role in several other NLP (Natural Language Processing) based tasks, such as automatic question answering, machine translation, dialogue systems and document matching (Wang, Dong, 2020) and digital humanities, broadly understood.

The idea of comparing text specificity measures emerged when content analyzing a corpus of political and media discourses on Russia's war in Ukraine in the framework of a larger ongoing research project. It was necessary to identify terms that would help distinguish discourses in function of their source (countries, political leaders, and media). No text specificity measure emerged as an obvious and uncontested choice, however. The performance of two measures, one based on Cohen's d and the other introduced by Savoy (2016), Z-score, is thus compared. The proposed analysis aims to identify convergent and divergent patterns in the outcomes obtained with their help. The research question can be formulated as to whether d and Z can be used interchangeably, or they have their own areas of application.

2 Related work

The need for having text representation and calculating distances between texts exists in information retrieval and text mining alike (Wang, Dong, 2020). Text representation involves viewing a text as a set of numerical features, for instance, as a bag-of-words. The order of words in the 'bag' is deemed to be irrelevant. Only their frequencies count. Since it was found that the usefulness of a term for content representation increases with the frequency of the term in the document but decreases with the number of documents, various weighting schemes, such as TF*IDF (term frequency by inverse document frequency), are commonly used (Salton, McGill, 1983; Evans et al., 2007; Jurafsky, Martin, 2008; Manning, Raghavan, Schütze, 2008; Savoy, 2016; Diermeier et al., 2011).

The transformation of texts into vectors paves the way to assessing their similarity and, ultimately, calculating distances between them. The relative position of texts included in a corpus is thus determined. Distance measures include Euclidean distance, Cosine distance, the Jaccard index, etc.

In corpus-based approaches text representation always has a relative, as opposed to absolute, character. If the same text is compared with the other set of documents, its representation changes. This caveat needs to be borne in mind when determining specific terms that characterize a text. Specificity refers to terms used to distinguish the text content (Salton, McGill, 1983). Specificity does not belong to a given document but terms that can be used to discriminate between two (or more) text categories (e.g., authors, text genres, etc.). For instance, what are the terms and expressions that best characterize each source of political and media discourses on Russia's war in Ukraine?

An approach to operationalizing specificity consists in building a dictionary. This process can be either theory- or data-driven (Simon, Xenos, 2004). In addition to content words serving to name things, express relations, perceptions, states or actions, the data-driven approach will extract many functional words (and, or, above, etc.). Functional words are normally excluded from the analysis. A data-driven dictionary includes m most frequent content word types or lemmas, with m varying from 50 to 1,000

(Burrows, 2002; Savoy, 2017; Savoy, 2019; Juola, Mikros, Vinsick, 2019). All words outside of the top m are also excluded from the analysis as uninformative. A theory-driven dictionary contains terms identified with the help of literature review. For instance, McClelland's motive imagery model borrowed from psychology was used to identify textual signals prefiguring military threats in political discourses on the Islamic Republic of Iran (Hogenraad, Garagozov, 2014).

It is at this stage that term specificity measures become necessary. They allow to quantitatively assess the specificity of terms included in the dictionary. The higher the value of a specificity measure for a term, the better it helps distinguish between texts included in a corpus. Several term specificity measures are known and used in application systems. A version of Cohen's d, a popular effect-size measure, is one (Shalak, 2004; Warner, 2013). It is implemented in several programs for content analysis, such as WordStat and VAAL. The other is Z-score, a version of the distance of a term score from the mean of a distribution expressed in unit-free terms (Savoy, 2016).

Both term specificity measures involve comparing the observed term frequency with its expected frequency calculated from corpus-level data. The calculation of the expected frequency is based on the assumption that the term is evenly distributed across all documents in the corpus. The idea of comparing the observed and the expected frequencies can be traced back to a generic chance-corrected measure (Goodman, Kruskal, 1954): $M_{cc} = \frac{M-E(M)}{M_{max}-E(M)}$, where M_{CC} is the chance-corrected measure, $M_{\max}$ is the maximal value M can reach, and $E(M)$ is the value expected for a null model. In the circumstances, the null model assumes that text category (e.g., authors, text genres, etc.) does not have an impact on the distribution of the term across documents.

The algorithms for calculating d and Z for the ith term denoted t_i differ, however.

(1) $$d\ (t_i) = \frac{\frac{tf_{i0}}{N_0} - \frac{tf_i}{N}}{\sqrt{\frac{\sum_i^m (\frac{tf_{i0}}{N_0} - \frac{tf_i}{N})^2}{m}}}$$

where m is the size of the dictionary (and also the query length in the context of information retrieval), tf_i – term i frequency, N_0 – the document word count, and N – the corpus word count.

There is a version of d adapted for the purposes of information retrieval, Standard Document Score, or SDS (Cummins, 2013):[1]

(2) $$SDS(m, D) = \frac{1}{\sqrt{|m|}} \sum_t^m \frac{(tf_{i0} - E[X^{t_i}])}{\sigma(X^{t_i})}$$

[1] The reviewer pointed out to some inconsistencies in this formula. However, it is kept in the version found in the source (Cummins, 2013, p. 114) since the possible errors do not affect the analysis in this article.

where $E[X^{t_i}]$ and $\sigma(X^{t_i})$ are the expected value and the standard deviation of the random variable X^{t_i} respectively, and |m| – the query length. In contrast to d, which is calculated at the term level, SDS is calculated at the document level. The SDS can be interpreted as the number of standard deviations a document is from the average document score for a specific query.

To calculate d, the difference between the observed and the expected *relative* frequencies is divided by the standard deviation. d has no lower or upper limit, whereas the cut-off values suggested in the literature are 0.8 for positive values and -0.8 for negative values (Warner, 2013).

(3)
$$Z\ (t_i) = \frac{tf_{i0} - N_0 \times \frac{tf_i}{N}}{\sqrt{N_0 \times \frac{tf_i}{N} \times (1 - \frac{tf_i}{N})}}$$

To calculate Z, the difference between the observed and the expected *absolute* frequencies is divided by the binomial variance. Z does not have a lower or upper limit either. The suggested cut-off values are 3 and -3: terms overused in a document have a Z score higher than 3 (and corresponding to 0.14% of the Gaussian distribution) whereas underused terms have a Z score lower than -3 (Savoy, 2016).

d is distinguishable from Z along the following lines:

1. Relative, as opposed to absolute, frequencies are used in numerator,
2. Standard deviation, as opposed to binomial variance, is used in denominator, and
3. The absolute value of d depends on m whereas that of Z does not depend on the size of the dictionary/query length.

Those differences suggest that Z may be more sensitive to small counts than d, whereas d appear to be suitable to analyze large corpora. To compare the behavior of d and Z when calculated for the same texts, their relative performance was assessed using standard performance metrics, precision, recall, accuracy, F_1-measure and Cohen’s Kappa (Evans et al., 2007; Jurafsky, Martin, 2008; Zhai, Massung, 2016; Vellino, Alberts, 2016; Khan, Qamar, Bashir, 2017; Dourado et al., 2019). Although human input is normally used as the gold standard in computer science (DiMaggio, 2015), in this case the other term specificity measure plays this role. In other words, an attempt to cross-validate d and Z was made. If their performance has convergent patterns, it is indicative of a consensus in the determination of term specificity (Mathet, Widlöcher, Métivier, 2015). Divergent patterns, on the other hand, would lead us to believe that d and Z cannot be used interchangeably, and it is up to the user to choose which one best serves her needs. Like many other problems in information retrieval and text mining, the choice of the term specificity measure is empirically defined. Which one works better cannot be answered by pure analytical reasoning or mathematical proofs (Zhai, Massung, 2016).

3 Political and media discourses on Russia's war in Ukraine

A highly heterogenous corpus was used for the purpose of comparing the performance of d and Z. It includes 3,347 texts in four languages of variable length, from 232 words to 145,237 words (55,599,283 words in total), discussing Russia's war in Ukraine. In addition to their variable length, texts also exemplify different genres: speeches of political leaders (political discourse) and news items (media discourse). The corpus covers five counties: two belligerents, the United States, and two European counties, the United Kingdom and France. In the case of Russia, Ukraine and the United States the coverage is more comprehensive since three media were monitored in each of those countries (*Kommersant*, *Izvestia*, and *Gazeta.ru*; *Ukrainska Pravda*, *RBC-Ukraina* and *Liga.net*; *New York Times*, *Washington Post* and *CNN* respectively). In the United Kingdom and France, one media was selected (*The Times* and *Le Monde* respectively). Transcripts of political leaders' speeches were retrieved from their official websites. Transcripts of speeches of members of legislative bodies were retrieved from the legislative bodies' official websites (the Russian Duma, the Ukrainian Rada, the US Congress, the British Parliament, and the French Assemblée Nationale). In total, the data came from 21 source.

Raw term frequencies (terms are only sequence of letters or short sequence of words, n-gram) were calculated using WordStat computer program. All subsequent calculations were performed with the help of custom algorithms.

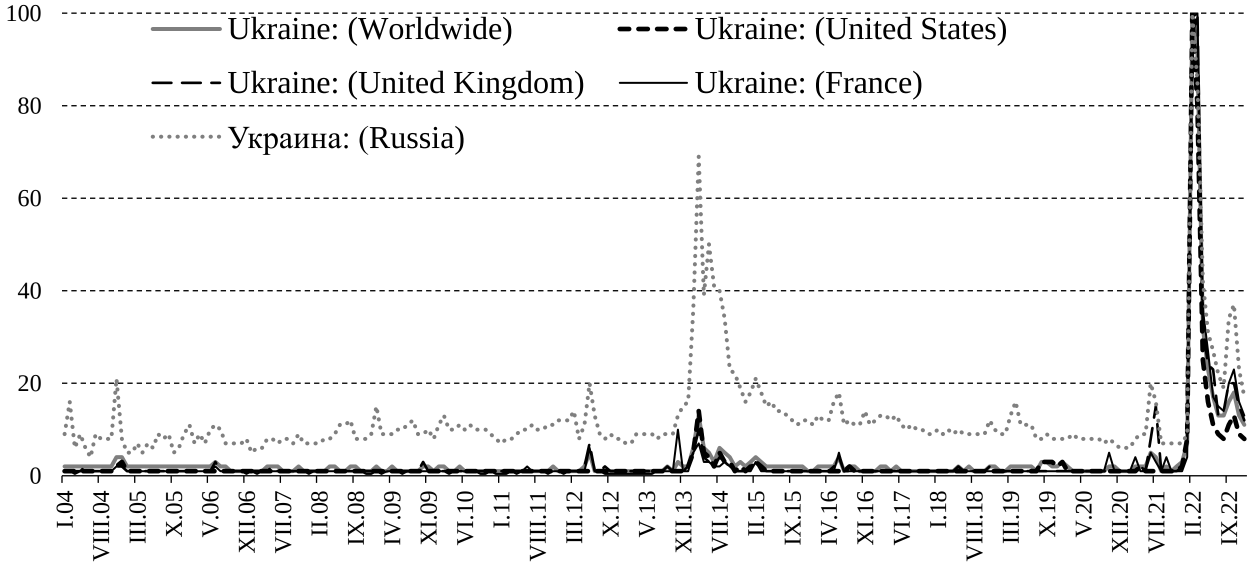

Figure 1: Relative popularity of 'Ukraine' as an internet search term in the US, the UK, France, Russia and worldwide, 2004-2022 (Source: Google Trends; a value of 100 is the peak popularity for the term).

Although Russia’s aggression against Ukraine started in 2014, it transformed into an all-out war eight years later, on February 24, 2022. The corpus covers first eight months of the full-fledged war which significantly increased the demand for information about Ukraine and the situation in this country at the international level (Figure 1). During this period, Ukraine managed to contain Russia’s initial attacks and started to progressively regain control over the territories occupied by Russia. Political and media discourses included in the corpus represent an informational dimension of the war. As Lasswell (1938) once observed, efforts to control opinions constitute one of the three chief implements of warfare, along with military pressure and economic pressure.

When studying informational warfare, one needs to identify the terms and expressions that can best characterize the discourse in each country directly or indirectly involved in the conflict. If such terms differ across the countries, then the hypothesis that war coverage does not allow establishing the truth finds some support (Lasswell, 1938; Knightley, 2003). One of the research questions addressed in the larger project is whether the differences in war coverage are territorially segregated according to national boundaries in the case of Russia’s war in Ukraine.

A dictionary was built for the purpose of content analyzing the corpus. When assessing the alternative research strategies, including topic analysis (DiMaggio, Nag, Blei, 2013; Zhai, Massung, 2016), the use of the dictionary was deemed to be a better option. Although Multilingual Probabilistic Topic Models allow discovering topics in corpora composed of texts written in different languages (Lind et al., 2022), they do not lessen the other requirement of homogeneity. In the circumstances, text lengths and genres (news items as opposed to speeches of political leaders) vary significantly. The quadrilingual dictionary includes 246 categories (about 400 words in each version – Ukrainian, Russian, English and French – since some categories include more than one word). The dictionary was compiled using a combination of theory- and data-driven approaches. Along with most frequent terms, it contains those commonly discussed in the extant literature on war coverage. For example, ‘war’ is often described in terms of ‘special military operation’ (Lukin, 2013), as in the Russian case. The mandated substitution of ‘special military operation’ for ‘war’ helps members of Russia’s power elite to create a perception of the aggression as limited in scope and not inherently violent. The 246 categories were weighted by TF*IDF whereas texts lengths were normalized. The TF*IDF values were calculated for the entire corpus using the formula $TF * IDF_i = tf_i \times \log\left(\frac{D}{df_i}\right)$, where D is the total number of documents in the corpus and df_i – the number of documents containing term i (Manning, Raghavan, Schütze, 2008; Jiang, Li, 2012).

4 Two measures of term specificity compared

The two lists generated for the group of leaders are discussed in detail for the purpose of illustration. As in all other subsamples, lengths of their speeches devoted to Russia’s war in Ukraine vary from 8,115 words (French President Macron) to 322,596 words (Ukraine’s President Zelensky). President

Zelensky delivered at least one address (sometimes up to four) to the national and the international audiences every day during the period under consideration. Lengths of war-related speeches of Russian President Putin's (49,169 words), US President Biden (31,175), and the then UK Prime Minister Johnson (25,242) lie in between.

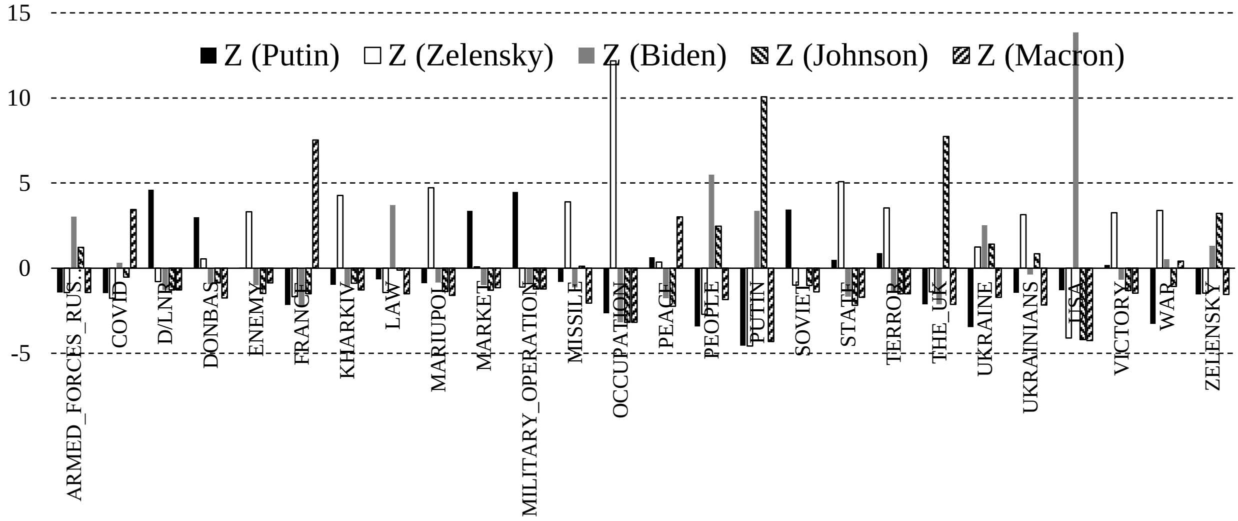

Figure 2: Z scores for the most specific terms of five political leaders.

Since in the cases of France and the UK one media only was monitored, the performance of d and Z was compared on the basis of four subsamples containing five sources of data each: the group of five leaders (Putin, Zelensky, Biden, Johnson and Macron), Russia, Ukraine, and the US. Most specific terms were identified for each subsample using d and Z, after which their lists so compiled were cross-checked. Six performance metrics, precision, recall, accuracy, F_1-measure, Cohen's Kappa and Pearson's r, informed the comparison of five pairs of the lists of most specific terms. The addition of Pearson's r allowed disregarding cut-off values that in the case of d may be arbitrary to some extent (they are set by convention): correlations were run between raw scores of d and Z.

The list of terms whose Z-scores exceed |3| includes 26 items (Figure 2). The list of terms whose d-values exceed |0.8| contains 39 items (Figure 3). Those lists overlaps to a significant extent. 25 terms are present on both lists: Covid, D/LNR (the acronyms for the Donetsk and the Luhansk People's Republics, two entities created in 2014 and supported by Russia in Eastern Donbas), Donbas, enemy, France, Kharkiv, law, Mariupol', market, military operation, missile, occupation, peace, people, Putin, Soviet, the State, terror, the UK, Ukraine, Ukrainians, USA, victory, war, and Zelensky.

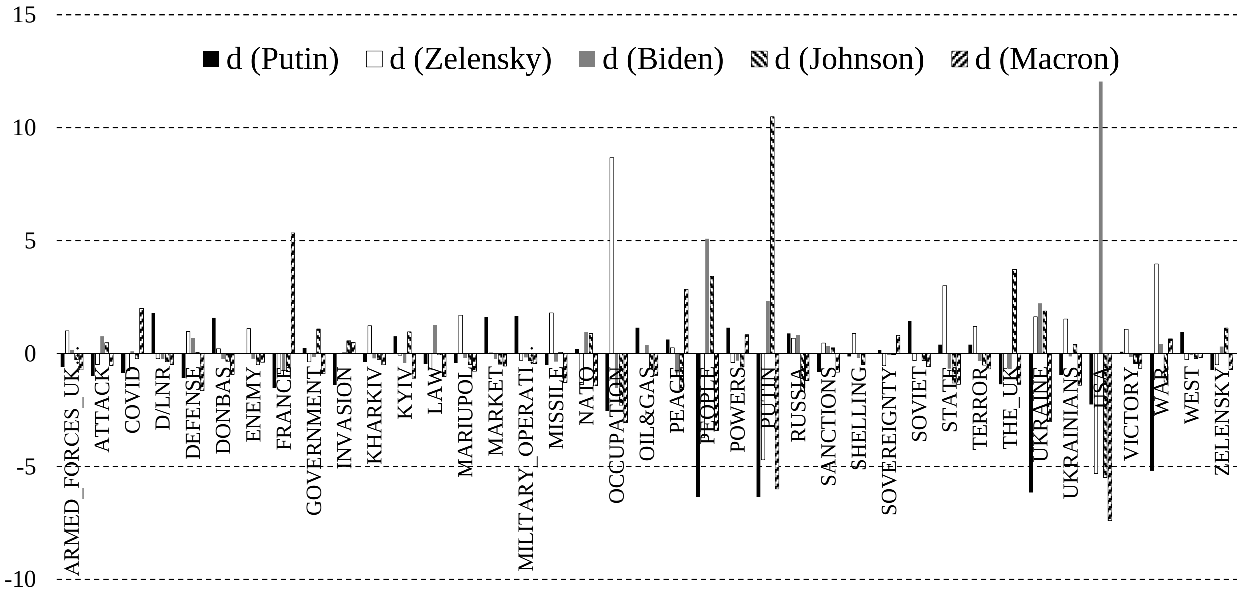

Figure 3: ‘d values for the most specific terms of five political leaders’.

The 25 terms constitute markers of the presidents’ discourses. When a president overuses a term, it becomes a positive marker of his discourse. When a president underuses a term, it is deemed to be a negative marker of his discourse. A negative character of a marker does not mean that a negative connotation is attached to it. The term is simply used by a president significantly less frequently than by his fellows. Vice versa, the term-positive marker has no value judgment attached to it. Sentiment analysis would be needed to discern value-judgments attached to the markers.

Positive markers for Putin are D/LNR, Donbas, market, military operation, Soviet; negative – people, Putin, Ukraine, and war. Waging a war against Ukraine, Putin nevertheless avoids naming his opponent and framing the aggression as a war. Positive markers for Zelensky are enemy, Kharkiv, Mariupol’, occupation, the State, terror, Ukrainians, victory, war; negative – Putin and the US. His discourse has more local detail than in other cases and is centered on the war’s impact on Ukraine instead of geopolitics. Positive markers for Biden are law, people, Putin, and the US; negative – occupation. Positive markers for Johnson are Putin, the UK, and Zelensky; negative – occupation and the US. Positive markers for Macron are Covid, France, and peace; negative – occupation, Putin, and the US.

Since the use of the term ‘military operation’ is mandated in Russia and the term ‘war’ – banned, the fact that the first is consistently (as per relevant d-value and Z-score) overused and the second – underused in this country comes as no surprise. The consistent overuse of the term ‘the State’ by Zelensky is more noteworthy (tf_0=1,974, tf=2,135, d=3.01, Z=5.08). Although the process of state-building started in Ukraine almost from scratch after the declaration of its independence in 1991 (Harasymiw, 2002), it appears that the ongoing war provided additional and powerful incentives to intensify this process: ‘36 days! 36! This is how long our State, our people have been able to stand against the army which was deemed to be among the best in the world' (Zelensky, March 31). 'I want to thank separately the inhabitants of our city of Energodar. Those brave Ukrainians who went down to the streets today to defend their city, to defend our State' (Zelensky, April 2).

There is one term with the Z-score exceeding the cut-off value yet with the d-value not reaching it, 'Russian Armed Forces.' It can relatively frequently be found in President Biden's speeches (tf_0=25, tf=37, d=3.04, Z=0.65): 'Thanks to the aid we've provided, Russian forces have been forced to retreat from Kyiv' (Biden, April 28).

The list of terms with d-values exceeding the cut-off value and Z-scores below the cut-off value contains 14 items: Ukrainian Armed Forces (Zelensky, tf_o=223, tf=239, Z=2.9, d=1.01), attack (Putin, tf_o=6, tf=176, Z=-1.63, d=-0.99), defense (Zelensky, tf_o=995, tf=1,195, Z=1.6, d=0.97), government (Johnson, tf_o=38, tf=250, Z=2.26, d=1.08), invasion (Putin, tf_o=3, tf=230, Z=-2, d=-1.4), Kyiv (Johnson, tf_o=37, tf=381, Z=1.71, d=0.97), NATO (Biden, tf_o=60, tf=265, Z=2.14, d=0.95), oil & gas (Putin, tf_o=60, tf=352, Z=1.5, d=1.14), powers (Putin, tf_o=93, tf=221, Z=2.21, d=1.15), Russia (Putin, tf_o=363, tf=3,136, Z=0.76, d=0.89), sanctions (Macron, tf_o=5, tf=591, Z=-1.1, d=-0.83), shelling (Zelensky, tf_o=236, tf=251, Z=2.72, d=0.89), sovereignty (Macron, tf_o=17, tf=174, Z=1.6, d=0.81), and the West (Putin, tf_o=51, tf=99, Z=2.4, d=0.95). For instance, references to sovereignty are common in the discourse of France's President Macron: 'France and Europe responded to this flagrant violation of the territorial integrity and the sovereignty of a European country with no delay and with resolution' (Macron, March 2).

Overall, the two term specificity measures show more convergency than divergency (Table 1). The average value of F_1, 0.64, is within the acceptable range. For instance, in a study of 11,089 front-page news articles using a dictionary of 20 categories, the reported F_1, 0.68, was similar (Burscher, Vliegenthart, De Vreese, 2015). The average value of Cohen's Kappa can be interpreted as substantial since it falls within the range from 0.61 to 0.8 (Warner, 2013). The average value of Pearson's r is also indicative of a substantial to strong relationship. One needs to bear in mind that the choice of the reference point, d or Z, affects only the values of precision and recall (precision becomes recall and vice versa), whereas the other performance metrics remain the same.

Table 1: Average values of recall, precision, accuracy, F_1, Cohen's Kappa and Pearson's r for four subsamples.

precision	0.7065
recall	0.7697
accuracy	0.9547
F_1	0.6412
Cohen's Kappa	0.6225
Pearson's r	0.8745

Since d and Z show convergent validity, they appear to measure the same thing, term specificity. At the same time, a look at the instances of misclassification in which one term specificity measure exceeds the cut-off point whereas the other does not suggests that Z tends to be more sensitive to small counts than d, whereas d – more suitable to analyze large texts. There are relatively more cases with small tf_0 when Z exceeds the cut-off value whereas d does not than when d exceeds the cut-off value whereas Z does not, although more tests are needed to confirm this pattern.

5 Conclusion

Two term specificity measures, d and Z, show convergent validity. They are not perfectly interchangeable, however. The assumption that Z is more sensitive than d when small wordcounts are imputed in their calculation needs further testing. It remains to be seen if text length should be taken into account when choosing between the two measures indeed.

The other promising direction for further research refers to the adaptation of d and/or Z to tasks in information retrieval. Although SDS suggests that it can be done, computational complexity constitutes an obstacle. The author of this measure was able to run tests at the price of its significant simplification as a result of imputing nominal-level data instead of ratio-level (Cummins, 2013). The underlying intuition is that the calculation of d and Z can be thought of as a method of outlier detection. The larger the values of d, Z and SDS, the further from an average score a term or a document deviates. Inversely, the smaller the values of specificity measures, the closer to an average score a term or a document is. The identification of centroids is important in information retrieval. For instance, by identifying documents with small SDS values for a given dictionary (query), it is possible to retrieve the closest matches. Under this scenario, instead of focusing on documents with largest values of specificity measures, principal attention is devoted to those with smallest values. They likely contain the information the author of a query is looking for. By typing a query, the user creates a dictionary with the help of which the aboutness of searchable documents, to use Cummins's term, is measured.

Acknowledgments

The author is grateful to one of the anonymous reviewers for useful suggestions and to the editorial office for the prompt handling of the submission. The remaining errors and inaccuracies are the author's.

References

Burscher, B., Vliegenthart, R., De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The Annals of the American Academy of Political and Social Science*, 659(1), pp. 122-131.

Cummins, R. (2013). A Standard Document Score for Information Retrieval. In: Kurland, O., Metzler, D., Lioma, C., Larsen, B., Ingwersen, P. (Eds.). *ICTIR'13: Proceedings of the 2013 Conference on the Theory of Information Retrieval*, Copenhagen Denmark 29 September 2013-2 October 2013, pp. 113-116. The Association for Computing Machinery.

Diermeier, D., Godbout, J.-F., Yu, B., Kaufman, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*. 42(1), pp. 31-55.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2).

DiMaggio, P., Nag, M., Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), pp. 570-606.

Dourado, Í. C., Galante, R., Gonçalves, M. A., da Silva Torres, R. (2019). Bag of Textual Graphs (BoTG): A General Graph-Based Text Representation Model, *Journal of the Association for Information Science and Technology*, 70(8), pp. 817-829.

Evans, M., McIntosh, W., Lin, J., Cates, C. (2007). Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *Journal of Empirical Legal Studies*, 4(4), pp. 1007-1039.

Goodman, L. A., Kruskal, W. H. (1954). Measures of Association for Cross Classifications, *Journal of the American Statistical Association*, 49(268), pp. 732-764.

Harasymiw, B. (2002). *Post-Communist Ukraine*. Edmonton and Toronto: Canadian Institute of Ukrainian Studies Press.

Hogenraad, R. L., Garagozov, R. R. (2014). Textual fingerprints of risk of war. *Literary and Linguistic Computing*, 29(1), pp. 41-55.

Jiang, H., Li, W. (2012). Improved Algorithm Based on TFIDF in Text Classification. *Advanced Materials Research*, 403, pp. 1791-1794.

Juola, P., Mikros, G. K., Vinsick, S. (2019). A Comparative Assessment of the Difficulty of Authorship Attribution in Greek and in English. *Journal of the Association for Information Science and Technology*, 70(1), pp. 61-70.

Jurafsky, D., Martin, J. H. (2008). *Speech and Language Processing*. 2nd edition. Upper Saddle River, NJ: Pearson-Prentice Hall.

Khan, F. H., Qamar, U., Bashir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. *Knowledge and Information Systems*, 51(3), pp. 851-872.

Knightley, P. (2003). *The First Casualty: The War Correspondent as Hero, Propagandist, and Myth-maker from the Crimea to the Gulf War II*. London: André Deutsch.

Lasswell, H. D. (1938). *Propaganda technique in the World War*. New York: Peter Smith.

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., Boomgaarden, H. G. (2022). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures*, 16(2), pp. 96-114.

Lukin, A. (2013). The meanings of 'war': From lexis to context. *Journal of Language and Politics*, 12(3), pp. 424-444.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mathet, Y., Widlöcher, A., Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3), pp. 437-479.

Salton, G., McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Savoy, J. (2019). Text Categorization with Style. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (Eds.). *Advances in Information Retrieval: 41st European Conference on IR Research*, ECIR 2019 Cologne, Germany, April 14–18, 2019. Proceedings, Part II, pp. 408-409. Springer.

Savoy, J. (2017). Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. *Glottometrics*, 38, pp. 55-76.

Savoy, J. (2016). Text Representation Strategies: An Example With the State of the Union Addresses. *Journal of the Association for Information Science and Technology*, 67(8), pp. 1858-1870.

Shalak, V. I. (2004). *Kontent-analiz. Prilozhenija v oblasti politologii, psihologii, sociologii, kul'turologii, jekonomiki, reklamy* [Content analysis and its applications to political sciences, psychology, sociology, culturology, economic sciences and advertising]. Moscow: Omega-L.

Simon, A. F., Xenos, M. (2004). Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis, *Political Analysis*, 12(1), pp. 63-75.

Vellino, A., Alberts, I. (2016). Assisting the appraisal of e-mail records with automatic classification. *Records Management Journal*, 26(3), pp. 293-313.

Wang, J., Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information*, 11(9), p. 421.

Warner, R. M. (2013). *Applied Statistics: From Bivariate Through Multivariate Techniques*. 2nd edition. Thousand Oaks, CA: Sage.

Zhai, C., Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ACM Books.

The meaning distributions on different levels of granularity

Tsy Yih[1], Haitao Liu[1*]

Tsy Yih is the transliteration of the name of the first author in his mother tongue, Shanghai Wu Chinese. He is also known as ZI YE in Mandarin pinyin.

[1] Zhejiang University

[*] Corresponding author's email: lhtzju@yeah.net

DOI: https://doi.org/10.53482/2023_54_405

ABSTRACT

The meaning distributions of certain linguistic forms generally follow a Zipfian distribution. However, since the meanings can be observed and classified on different levels of granularity, it is thus interesting to ask whether their distributions on different levels can be fitted by the same model and whether the parameters are the same. In this study, we investigate three quasi-prepositions in Shanghainese, a dialect of Wu Chinese, and test whether the meaning distributions on two levels of granularity can be fitted by the same model and whether the parameters are close. The results first show that the three models proposed by modern quantitative linguists can both achieve a good fit for all cases, while both the exponential (EXP) model and the right-truncated negative binomial (RTBN) models behave better than the modified right-truncated Zipf-Alekseev distribution (MRTZA), in terms of the consistency of the goodness of fit, parameter change, rationality, and simplicity. Second, the parameters of the distributions on the two levels and the curves are not exactly the same or even close to each other. This has supported a weak view of the concept of 'scaling' in complex sciences. Finally, differences are found to lie between the distributions on the two levels. The fine-grained meaning distributions are more right-skewed and more non-linear. This is attributed to the openness of the categories of systems. The finer semantic differentiation behaves like systems with open set of categories, while the coarse-grained meaning distribution resembles those having a close set of few categories.

Keywords: semantic diversification, meaning distribution, scaling, self-similarity, level of granularity, the exponential distribution, right-truncated negative binomial

1 Introduction

Semantic diversification, or meaning distribution, is a phenomenon characterizing the differentiation of the meanings of words or other linguistic units1 (Altmann 1985, 2005, 2018; Köhler, 1991; Wang et al. 2021). If all the meanings of a word in texts are arranged in descending order in terms of their frequencies, meaning distributions generally follow a power-law-like curve (Altmann, 2018)2. This law is specifically called Beöthy's Law in memory of Beöthy and Altmann's three classical studies on the Hungarian prefixes (1984a, b; 1991). In quantitative linguistics, there has already been a large number of studies in this respect3.

However, meanings can be observed or classified on different levels of granularity. It is then natural to ask, whether the meaning distributions on different levels abide by the same law. Yet to our knowledge, no previous research in quantitative linguistics has been investigated in this way. In addition, if meaning distributions can be modelled by the same function, whether the parameters are the same. If not, whether they can be used to differentiate between the distributions on two levels.

Among all linguistic meanings or functions expressed by human language, semantic roles[4] have constituted a suitable topic of linguistic study since the idea to differentiate among different levels of granularity appeared. The most recent one we have found is the theory of three-level roles put forward by Van Valin (2005). To conclude the ideas, the traditionally perceived semantic roles such as agent, patient, instrument, recipient, beneficiary, etc., called meso-roles,

[1] Strictly speaking, the term 'semantic diversification' denotes a dynamic process prima facie, while in practice, it is generally used to describe the equilibrium state in that process. Therefore, it is synonymous with 'meaning distribution' or 'meaning diversification' (Fan & Altmann 2008, Fan et al. 2008) in some contexts. They will be used interchangeably in the present study.

[2] Two variable notations, N and V, respectively representing token size and type size are generally used in word frequency studies. In meaning distribution studies, N is kept while M, the counterpart of V, is used standing for the number of meanings of a linguistic form in texts, which will be employed in the remainder of this paper.

[3] For the collection of numerous case studies on semantic diversification, see Strauss & Altmann (2006) and Altmann (2018, Ch. 5).

[4] Semantics roles are alternatively known as theta roles, thematic roles, or participant roles in different traditions.

can be seen as the clustering of verb-specific or event-specific micro-roles. For instance, HITTER (the one who hits) and HITTEE (the one being hit) are micro-roles in the HIT event[5]. Following this model, Hartmann et al. (2014) have illustrated a semantic space where traditional meso-roles, such as agents and patients, can be seen as the clustering of micro-roles when zooming in from a coarse-grained level to a fine-grained level.

We, therefore, regard semantic roles as a suitable lens for semantic diversification on different levels. Semantic roles are generally encoded by case markers and adpositions[6] formally, on which there have already been abundant quantitative linguistic studies (Fuchs, 1991; Hennern, 1991; Roos, 1991; Rothe, 1991; Sanada & Altmann, 2009; Liu 2012; Kolenčíková & Altmann; 2020) due to their multifunctionality (Croft, 2003; Haspelmath, 2003). Thus, it is appropriate to proceed with the research in this line.

Specifically, in this paper, we intend to answer the following research questions:

1. Can meaning distributions on two different levels of granularity be fitted by the same model? Which model is the best?

2. Do the distributions on the two levels have the same parameters or look similar graphically?

3. What are the major differences between the distributions on the two levels and if there exist any potentially affected factors?

The present paper is structured as follows: Section 2 introduces the corpus and procedure. Section 3 shows the results, based on which we will answer the proposed questions in Section 4. Section 5 concludes the whole paper and points out some limitations.

[5] Note that Van Valin also propounds a third and most coarse-grained level of semantic roles, the macro-roles. There are only two macro-roles, actor and undergoer, similar to Dowty’s (1991) proto-agent and proto-patient, which serve as the poles lying on two ends of the continuum of actness. Yet binary classification hardly makes sense for a distribution. Therefore, in this research, macro-roles are not annotated.

[6] The adposition is a cover term for prepositions and postpositions. In languages like Chinese and English, adpositions are predominantly preposed, while postpositions are found in Japanese, Korean and the like. In the remainder of this study, prepositions will simply be used since the main language under investigation is a Wu Chinese, a Sinitic language.

2 Materials and methods

2.1 Materials

The target language under investigation is Wu Chinese, a language of the Sinitic family spoken in Eastern China. Geographically, Wu is distributed in the municipality of Shanghai and in parts of the provinces of Zhejiang, Jiangsu, and Jiangxi. To be more specific, Shanghainese or Shanghai dialect, which is a representative dialect of Wu, was selected. It is the dialect spoken in downtown Shanghai and is the mother tongue of the first author. Since Shanghainese is a dialect officially, it is rarely written down in spite that theoretically, the language experts claim that it can be written in Chinese characters. In recent years, there are folk groups who aim to revive the writing of this language and there have been attempts published in some newspapers such as 新民晚报 *Xinmin Wanbao* “Xinmin Evening News”, Wechat pushes, and even Wikipedia entries. Thus, thanks to these resources, we took a corpus-based approach in the present study. Being a dialect also means that Shanghainese lacks an official, authoritative dictionary. There are, nevertheless, two dictionaries of Shanghainese written by scholars, which are 上海话大词典 *Shànghǎihuà Dà-Cídiǎn* ‘The Grand Dictionary of Shanghainese’ (*SDC*) and 上海方言词典 *Shànghǎi Fāngyán Cídiǎn* ‘Shanghai Dialect Dictionary’ (*SFC*).

Specifically, we investigated three quasi-prepositions in this study. They are called ‘quasi-prepositions’ due to the characteristics of the Sinitic languages, where prepositions are generally grammaticalized from verbs or can grammaticalize into conjunctions. Therefore, there are many linguistic forms that stay at the middle stage and possess both the functions of prepositions and verbs or conjunctions. This is reflected in the terms, ‘coverbs’ and ‘prepositional conjunctions’ in some classic reference grammars of Chinese (Chao, 1968: 335, 791). The reason we do not exclude the verbal/conjunctional meaning is that there are obvious connections between different uses and these are not cases of homonymy. In other words, from a semasiological perspective, all meanings of the same word forms should be taken into consideration. Yet only the prepositional meanings, or meso-roles, will break down into micro-roles according to our definition of two levels of granularity.

Note that since quasi-prepositions are generally overlooked and less delved into in the traditions of Chinese dialectology, the abovementioned two dictionaries are both sketchy in this respect. A pilot study showed that the forced choice method according to the dictionaries gives poor results. In addition, a compiler of *SFC*, 陶寰 Tao Huan, told us that he deliberately omitted the meanings which are in common with the usage in Mandarin due to the limit of space since it is written in Mandarin and targeted at normal Chinese readers equipped with full lexical competence of Mandarin (p. c.). On the one hand, such background has left us a good chance to have a detailed look at the functions of its prepositions in this language. On the other hand, it calls for manual semantic annotation, which would be somehow subjective. However, due to the fact that on the micro-level, all the micro-roles were verb-specific in the framework adopted by us. We could rely on the verb forms to help discern the prepositional meanings, thereby reducing the degree of subjectivity. As for the meso-role level, we slightly modified the set of well accepted traditional roles according to each case as would be shown below.

The corpus employed was Shanghai Spoken Corpus (SSC) v2.0, compiled by University of Alberta (Han et al., 2017)[7]. In this corpus, all the data were transcribed in Chinese characters. We also transcribed them in Wuyu Pinyin 吴语拼音[8] for the sake of illustration in the remainder of the paper. The whole corpus consisted of six sub-corpora based on genre (conversation, interview, monologue, opera, TV script, song). While it was designed to be a balanced corpus, it was obviously biased towards spoken language. In addition, since in the genres of opera and song, texts usually did not conform to the grammatical pattern of everyday language, they were excluded from the study. For the rest four sub-corpora, the basic information is presented in Table 1.

Table 1: Sizes of sub-corpora in SSC v2.0.

Genre	Number of files	Number of words
conversation	2	28709
interview	5	31251
monologue	21	47663
TV script	23	20942
Sum	51	128565

[7] We appreciate Weifeng Han's help for providing the corpus.

[8] It is a kind of romanization of Wu language proposed by Wu Chinese Society (http://www.wu-chinese.com/).

The corpus querying software used in this study was Wordless v1.3.0 (Ye, 2019). We chose it over common software tools such as WordSmith and AntConc in that the user could choose the sentence rather than a small text within certain spans in all directions around the node word as context. However, truncated sentences were insufficient and confusing in semantics. Thus, a complete context was necessary for semantic annotation with the consideration of our research purpose. After the sentences containing the node words were extracted, they were imported to Microsoft Excel, where we did annotations and basic statistics.

After a simple pilot survey, we selected three representative quasi-prepositions in Shanghainese, 拿 *nau* (and its phonological variant *ne*), 把 *peh* (and its bisyllabic variant *pehla*), 搭 *tah* (and its phonological variants *teh*, *theh*). The basic statistics of the three quasi-prepositions are shown in Table 2. Those hits which were repetitive and unclear were eliminated.

Table 2: Frequencies of all three queries.

Form	Hits in the corpus	Effective hits
拿 nau	357	337
把 peh	386	377
搭 tah	71	63

In our study, we designed two sets of semantic annotations on the basis of dictionaries and the assumed theory of micro-roles. Meanings on two levels of granularity were then annotated manually assuming monosemy. In terms of coarse-grained meanings, we referred to the dictionaries and traditional meso-roles with modifications. As for the fine-grained semantics of prepositions, since micro-roles are verb-specific, the forms of verbs they collocate with are tangible and concrete criteria. Aspectual markers including but not limited to 过 *ku*, 脱 *theh*, 着 *zeh*, 辣海 *lahhe*, 好 *hau*, and directive complements such as 过去 *kuchi*, 进去 *cinchi* were omitted. The complete framework of meaning differentiations is presented in Table 3[9]:

[9] Bold represents the argument introduced by the preposition. ** indicates that the meaning is recorded in both *SDC* and *SFC*, while * in just *SDC*. For *peh*, the dictionary does not distinguish between the verbal usage ‘give’ and the prepositional usage of dative considering their translational counterparts in Mandarin share identical forms and close relationships on the grammaticalization path. Here we nevertheless make a distinction on both levels.

Table 3: Meaning differentiations of three quasi-prepositions in Shanghainese.

Words	Part-of-speech	Coarse-grained level	Fine-grained level
拿 nau	verb	‘take’**	‘take’
		‘hold’**	‘hold’
		desiderative	‘Give me/I want X.’
		‘use’**	‘use’
	preposition	instrument**	‘do sth. **with X**’
		patient	‘relieve **X**’
			‘process **X**’
			......
		theme	‘tell **X** to Y’
			‘conceive **X** as Y’
			……
		‘taking’	‘taking **X** as an example, VP’
把 peh	verb	permissive**	‘allow’
		causative*	‘cause’
		‘give’**	‘give’
	preposition	recipient**	‘give X **to Y**’
			‘tell X **to Y**’
			……
		patient	‘put **X** Y’
			……
		beneficiary	‘sing X **to Y**’
			‘buy X **for Y**’
			……
		agent (passive)**	‘V-ed **by X**’
		‘according to’	‘according to X, VP’
搭 tah	conjunction	NP conjunction	‘X **and** Y’
	preposition	companion	‘with X’
		recipient	‘tell **X** Y’
			……
		beneficiary	‘do X **for Y’**
			‘build X **for Y’**
			……
		‘same’	‘be the same **as X**’
			‘be different **from X**’
			…
		‘relation’	‘get along **with X**’
			……
		patient	‘meet **X**’
			……
		comparative	‘compared with X’

2.2 Methods

To address the research questions proposed above, we fit five models to the meaning distributions of each quasi-preposition on both coarse-grained and fine-grained levels.

The Zipfian or right-truncated zeta function in (1) is the most common candidate in the literature on rank-frequency distributions. Mandelbrot's formula or the Zipf-Mandelbrot distribution as in (2) introduced a displacement parameter (Mandelbrot 1965). These two fitting models have the advantage of simplicity and are widely used in other scientific disciplines.

(1) $P_x = Cx^{-a} \qquad x = 1, 2, \ldots, n$

(2) $P_x = C(x + b)^{-a} \qquad x = 1, 2, \ldots, n$

where C denotes an adjusting factor which helps make the sum of whole probabilities one.

Apart from these two, modern quantitative linguists have proposed several models to characterize semantic diversification. Among them, three rival models stand out. First, Altmann (1985) introduced the negative binomial distribution derived from a birth-and-death process. A second model is the Zipf-Alekseev distribution[10] (Hřebíček 1996). In practice, two variants called the right-truncated negative binomial distribution (RTNB) and the modified right-truncated Zipf-Alekseev distribution (MRTZA) are often used. The formula of RTNB and MRTZA are given respectively in (3) and (4):

(3) $P_x = \binom{k + x - 2}{x - 1} p^k (1 - p)^{x-1} \quad x = 1, 2, \ldots, n$

where $k > 0$, $0 < p < 1$.

(4) $P_x = \begin{cases} \alpha & x = 1 \\ \dfrac{(1 - \alpha)x^{-(a + b\ln x)}}{T} & x = 2, 3, \ldots, n \end{cases}$

where $T = \sum_{j=2}^{n} j^{-(a + b\ln j)}, a, b \in R, 0 < \alpha < 1.$

Another candidate is the exponential model, also called the stratificational approach, shown in (5), whose assumption is that the relative rate of change of ranked frequencies is constant. This

[10] It is also known as the Zipf-Dolinskij distribution.

distribution has been employed in Fan & Altmann (2008), Popescu et al. (2010), Altmann (2018) and a number of other studies.

(5) $y = 1 + ae^{-bx} \quad x = 1, 2, \ldots, n$

where a, b are parameters, and b stands for the rate of change.

In terms of the nature of models, the first four models are based on probability distributions, which is generally the case, while the last one is indeed a function[11]. The difference between the two cases lies in whether dependent variables add up to one. Popescu et al. (2010) attributed the peculiarity of the last model to the lack of fitting software of exponential distributions. For the rationale or motivation behind each model, interested readers are further referred to the original literature, or to several pieces of work in the handbooks or encyclopedias, such as Altmann (2005), Wimmer & Altmann (2005) and Strauss & Altmann (2006).

The fitting tool used included Altmann Fitter v3.1.0 (Altmann, 2000), which has been frequently employed in quantitative linguistics, and NLREG, which is employed to fit the exponential function. In the next section, we first compare the fitting results of five models and then discuss the research questions in turn.

3 Results

In this section, we present the results of model fitting and the graphical representations of distributions on the two levels. The original data of the observed frequencies can be found in the appendices of this paper.

[11] We thank anonymous reviewers for pointing this out.

Table 4: Parameters of models.

	EXP	**MRTZA**	**RTNB**	**RTZ**	**ZM**
nau coarse	a = 218.2461 b = 0.4731 R^2 = 0.8686	a = 0.0414 b = 1.0705 α = 0.3650 (n = 8) R^2 = 0.9304	k = 2.7438 p = 0.6856 (n = 8) R^2 = 0.9434	a = 1.2058 (n = 8) R^2 = 0.7301	a = 12.0000 b = 18.9835 (n = 8) R^2 = 0.8454
nau fine	a = 1674.6521 b = 2.6199 R^2 = 0.9825	a = 0.3414 b = 0.0482 α = 0.3650 (n = 126) R^2 = 0.9992	k = 0.2482 p = 0.0087 (n = 126) R^2 = 0.9548	NULL	a = 1.1625 b = 0.5499 (n = 126) R^2 = 0.6669
peh coarse	a = 195.6757 b = 0.3964 R^2 = 0.9538	a = 0.0342 b = 0.8091 α = 0.3263 (n = 8) R^2 = 0.9722	k = 2.6833 p = 0.6397 (n = 8) R^2 = 0.9865	a = 1.0761 (n = 8) R^2 = 0.8074	a = 11.9999 b = 23.1314 (n = 8) R^2 = 0.9294
peh fine	a = 295.5034 b = 0.8993 R^2 = 0.9853	a = 0.7630 b = 0.0446 α = 0.3263 (n = 108) R^2 = 0.9776	k = 0.2525 p = 0.0125 (n = 108) R^2 = 0.9842	a = 1.1425 (n = 108) R^2 = 0.9388	a = 1.2146 b = 0.3618 (n = 108) R^2 = 0.8978
tah coarse	a = 22.8774 b = 0.3146 R^2 = 0.8866	a = 0.5333 b = 0.3506 α = 0.2540 (n = 8) R^2 = 0.9125	k = 2.6936 p = 0.5612 (n = 8) R^2 = 0.9247	a = 0.8091 (n = 8) R^2 = 0.7357	a = 11.9998 b = 42.2678 (n = 8) R^2 = 0.8905
tah fine	a = 32.7177 b = 1.0265 R^2 = 0.9380	a = 0.2219 b = 0.0634 α = 0.2063 (n = 38) R^2 = 0.9813	k = 0.5621 p = 0.0405 (n = 38) R^2 = 0.8765	a = 0.7360 (n = 38) R^2 = 0.8641	a = 0.9193 b = 1.5354 (n = 38) R^2 = 0.7358

Table 4 demonstrates the parameters in the five models we have used. The parameters fall into three groups. The first group is put in parentheses and concerns the boundary conditions, including the maximal value of the domain (*n* in all related cases), and normalization constants (*C*s in RTZ and ZM though not shown in the table[12]). These parameters are case-specific but do not reflect inter-case universals. A second group contains the indicators of goodness-of-fit. In this case, we simply resort to the determination coefficient, R^2 ($R^2 > 0.90$, very good; $R^2 > 0.80$, good; $R^2 > 0.75$, acceptable; $R^2 < 0.75$, unacceptable). What is left constitutes the most important group. The parameters in this group are intrinsic to the models per se. When comparing models, the determination coefficient is a basic criterion.

On the basis of the data, we have the following findings. Assume that we move from a coarse-grained level to a fine-grained one. For MRTZA, in the cases of *nau* and *peh*, *a* increases while *b* decreases when while in the case of *tah*, on the contrary, *a* and *b* both decrease. The results are thus not consistent among the three quasi-prepositions for this model. In the model of RTNB, both *k* and *p* decrease significantly. The same is true for parameters *a* and *b* in the Zipf-Mandelbrot function. As in the exponential model, the parameters *a* and *b* increase significantly in all cases.

We could also note in some cells where the fitting results are bad. For instance, fitting RTZ to the data of *nau* on the fine-grained level fails, which makes the fitting results of this model not comparable for all quasi-prepositions. In addition, in several cases where we fit by means of RTZ and ZM, R^2 is less than 0.75, which indicates unacceptability.

Next, we present the graphical results for comparison between the two levels of meaning granularity.

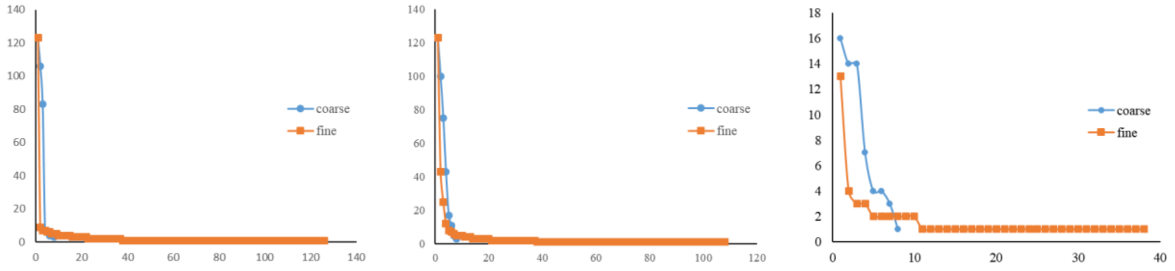

Figure 1. The rank-frequency distributions of meanings on two levels in linear coordinates (left: *nau*; middle: *peh*; right: *tah*).

[12] In fact, as shown in the formula (1–2), the models RTZ and ZM also have such a normalization constant *C*. Yet in the Altmann Fitter, they are regarded as probability distributions rather than functions, such as the exponential model fitted by means of NLREG. Hence, this parameter can be ignored given the additional constraint that the probabilities of all items add up to 1.

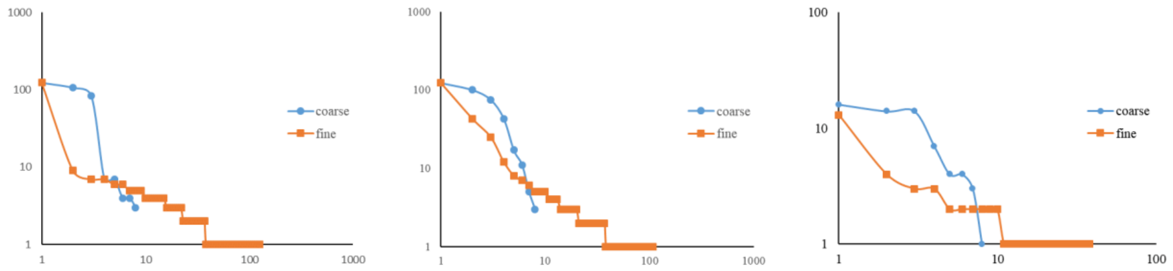

Figure 2. The rank-frequency distributions of meanings on two levels in log-log coordinates (left: *nau*; middle: *peh*; right: *tah*).

Figure 1 shows that the rank-frequency distributions of meanings on fine-grained ones are right-skewed compared with coarse-grained ones. In other words, fine meaning distributions have long tails. At first sight, the shapes of the two distributions are much different. In case there is information hidden by the linear coordinate, we also present them in log-log coordinates (Figure 2). It is shown that in no case are the distributions linear throughout the whole domain, or following a pure power law. Yet there could still be a ‘scaling range’ (Mandelbrot, 1997: 200). On the coarse-grained level, the curves first decrease slowly and then go down straight with a sudden change, while on the fine-grained level, there seem to be two stages. The first stage is linear and the second stage breaks down into steps.

4 Discussion

4.1 Which model is the best?

We have found that the three models of EXP, MRTZA and RTNB all give good results in terms of R^2. The determination coefficients of MRTZA are the largest in most cases. As for the rest two, sometimes the R^2 of the exponential models is larger than that of RTNB while other times the opposite happens. Prima facie, MRTZA is the best choice. However, R^2 is not the only criterion for comparing models. We argue that the exponential function and the RTNB model surpass MRTZA on several other aspects. First, we can see from our results that in terms of the change of parameters, both *a* and *b* in the exponential model, and *k* and *p* in RTNB change in a consistent way between two levels of granularity for each quasi-preposition. Specifically, the first group becomes larger as the semantic granularity goes finer, while the second group decreases. While for the case of MRTZA, the parameters *a* and *b* change in a different way for three quasi-prepositions. Second, the exponential model and RTNB have two intrinsic parameters, while the MRTZA has three. From the perspective of the Occam's Razor Principle, they perform both better than MRTZA. In fact, Altmann (2018: 4) also argued for the exponential function to be the main candidate of a unified model for the diversification phenomena, which is parallel to the status of Zipf-Alekseev function for length distribution. His primary motivation also pertains to simplicity, as the original differential equation and the rationale behind it are simpler than the other models. The exponential function simply follows the assumption that the relative rate of change of ordered frequencies is constant and negative, and the parameter *b* is that constant (Altmann, 2018: 3). On the other hand, before the advent of the exponential model, RTNB has always been among the best models characterizing the meaning diversification phenomena (see Beöthy & Altmann, 1984a, b and a number of papers in Rothe (ed.), 1991). Our findings again support the applicability of the model.

In sum, both the exponential model and RTNB have good rationales for being considered the best fitting models characterizing meaning distributions on both levels of granularity, and there seems to be no reason to argue for a winner between them based on the data provided in this paper. Moreover, the parameters can be used to differentiate between the two levels.

4.2 Parameters, same or different?

In this section, we aim to answer the question of whether the meaning distributions on two levels of granularity are similar. Both the parameters and graphical representations in Section 3 show that the distributions are very different between the two cases. On the one hand, the fitting results indicate that the parameters of the distributions change drastically, whereas on the other, the curves presented in either coordinate do not possess the same or similar shapes.

In what follows, we shall relate our finding to the concept of 'scaling' in complex sciences, i.e., the study of complex systems. 'Scaling' can be roughly understood as that systems observed on different scales manifest similar phenomena or follow the same rules (Kretzschmar, 2009). In several seminal works of Kretzschmar (2009, 2015, 2018, Kretzschmar et al. 2013), for instance, he investigated this issue from the perspective of the sociolinguistics of phonological systems. He found that the frequency distributions of phonemes on different scales in the acoustic space all manifest A-curves graphically[13], though he did not fit certain probability distributions to his data. Therefore, he deemed that he had proven the property of scaling at least in the field of sociophonetics.

Nevertheless, scaling may have two interpretations. The strong version of scaling sees it as the synonymy of 'self-similarity', which holds that meaning distributions on different levels of granularity follow the same fitting model and probably even have the same or similar parameters. This is a standpoint taken in Kretzschmar (2009). On the contrary, a weak or mild version of scaling says that distributions on different levels or scales are not strictly isomorphic. Rather, it is just that they all manifest A-curves in Kretzschmar's term, but do not necessarily have the same distribution functions, or other statistical parameters. This view was proposed in Kretzschmar et al. (2013). Our findings apparently support the weak version of scaling.

We shall next spend some space explaining why the strong version does not hold. In Kretzschmar (2009), he showed a strong favor of the idea that 'the part contains the information of the whole' which is a property of fractals based on his early non-quantitative study. For instance, he quoted the definition of Mandelbrot (1982) in Kretzschmar (p. 198). He also drew on the classical, well-known case of the length of the British coast studied by Mandelbrot (1967) (p. 179). However, a common misconception about the story is that a part of the coastline reflects the shape of the whole. In fact, Mandelbrot has already made it rather clear that it should be understood in a statistical sense. The related property is referred to as 'statistical self-similarity' rather than rigorous self-similarity in the sense of pure maths (as reflected by Koch snowflakes for instance). For real-life objects, a part is not the miniature of the whole generally. In other words, parts do not contain the information of the whole, and one could not deduce the total information about the whole from parts. As for the linguistic cases, it holds as well for

[13] He has named such distributions 'A-curves', mimicking 'S-curves' which are common in the field of language change. However, it seems inappropriate since there is no climbing-up part as in the graph of the letter 'A'. Rather, 'L-curve' appears to be more vivid.

the distributional patterns, and there is no such magic power that guarantees the isomorphism. Kretzschmar's prior understanding of 'scaling' falls into the Individualistic fallacy, the reverse of the Ecological fallacy, which is a classic statistical fallacy in science as pointed out by Horvath & Horvath (2003).

Later in Kretzschmar et al. (2013), there seems to be a change of idea. Kretzschmar has come to a milder conclusion with regard to the scaling property. That is, distributions on different scales are not strictly isomorphic. Rather, it is just that they all manifest in A-curves, but do not have the same distribution function, or the same statistical indicators. He also explicitly cited Horvaths' work and publicly support their standpoint. However, his attitude was still vacillating as reflected in his later monographs (Kretzschmar, 2015, 2018) which might be rather confusing to the reader. Therefore, it seems that Kretzschmar is not that certain about the interpretation of scaling, which is thus worth testing with real data. Based on our research, we agree with this moderate view of scaling, although this weak version itself seems to be a less significant claim than the strong version. Yet in other words, it also means that the parameters of models do have the ability to differentiate between levels of meaning granularity.

In the next section, we proceed to discuss the differences between the distributions and the primary factors.

4.3 Differences between the distributions on the two levels

Kretzschmar et al. (2013) claimed that a distribution with larger set of types tends to be more non-linear, and vice versa. This is supported by our results as shown in Figure 1. Therefore, we supplemented their conclusion that the fine-grained meaning distributions are more right-skewed.

In fact, this phenomenon can be explicated by the following proof. Remember that for this specific situation, we have constant N (number of tokens) and variant M (number of meaning types). Assume a distribution denoted as $\{f_r(x)\}$, $r = 1, 2, \ldots M$, where $\sum_{r=1}^{M} f_r(x) = N$. Since the total number of tokens N remains the same, once the group annotated as rank m is given a more fine-grained annotation, this class with frequency $f(m)$ will be broken down into several items with lower frequencies, thereby increasing the area of tail[14]. One extreme case is that if a person is able to identify different meanings in any different context, then $M = N$ and the absolute frequency of any item will become 1. Alternatively, if in all contexts is the word recognized as sharing the same meaning, then one meaning item takes all the frequencies.

In addition, we draw a key distinction between the two cases. Overall, the meaning distribution is similar to the case of rank-frequency distribution of various constituents. Yet a fine-grained distribution of meanings resembles that of words, whereas a coarse-grained one is alike that of letters or phonemes. The major difference lies in the openness of set of types. In the case of letters, phonemes and coarse-grained meanings, the set of all types M is closed, whereas for words or fine-grained meanings[15] here, it is an open set and grows with the number of tokens. It has been known in the literature that the distribution of words possesses a longer tail and has more hapaxes than that of letters or phonemes (Best & Rottmann, 2017, ch. 9), as well as being more non-linear. Thus in a similar vein, the same applies to the fine-grained meanings.

In sum, based on the graphical representations, we have pointed out the major difference between the two levels of meaning granularity, and attributed it to the openness of categories of the system.

4.4 Other general issues

In this final subsection, other factors that might influence our results are discussed.

First concern the genre of the corpus. Roos (1991) conducted a survey on the semantic diversification of Japanese *ni* and considered four homogenous texts and a mixed corpus. He found that the heterogeneity of the text does not play a crucial role. This has guaranteed the effectiveness of our research which also adopts a speech-biased corpus with several genres.

Second, we have only discussed the effect of the openness and numbers of the categories or

[14] A key condition here is that the set of fine-grained types must be the strict refinement of that of coarse-grained ones. Otherwise, this proof might not hold.

[15] Based on the approach taken in this study, fine-grained meanings are form-dependent, and thus form an open set. In other approaches, if one sets his own fine-grained level with the help of a dictionary or other sources, it will also be a closed set then. However, in real texts, it is common to find a meaning encoded by a word that is not predefined or recorded in the dictionary, a phenomenon caused by innovation in language use.

types, while Kretzschmar et al. (2013) mentioned that the shape of the distribution is also subject to the number of tokens. That is, a size effect might exist. Specifically, he deemed that only a sample with a large token size will manifest a non-linear distribution. In terms of our meaning distributions here, although the whole corpus is large enough with 130k tokens, the amount of the extracted form-meaning pairs might still be small, which is consequently expected to be expanded in the future.

One last facet concerns the identification of meaning-carrying units and the subjectivity of categorization. In all three cases, there are special items (*nau* in 拿……来讲 *nau ... lekaon* ‘taking … as an example’, *peh* in 把……讲起来 *peh ... kaonchile* ‘according to’ and *tah* in 搭……比起来 *tah ... pichile* ‘compared with’) for which the whole constructions rather than single words seem to be more appropriate meaning carriers. As far as we know, we have found no literature discussing the effect of unit identification on distributions so far. On the other hand, in Kretzschmar et al. (2013)’s study, speech as his scope of the study can be measured with accuracy, whereas in our case, we do not have a real semantic space as our foundation and the meaning annotation is more or less subject to subjectivity. The criterion of counting the number of meanings is inevitably vague (see Guiter, 1974 for a thorough discussion). In some studies, dictionaries were resorted to, which can serve as a golden standard. In most of the others, nevertheless, the methods were not clearly reported. However, even if one applies the dictionary approach, the actual use in texts might not be contained in the dictionaries, which leaves us only two remedies. The first is the forced choice method, which means to choose the closest meaning in the dictionary. The second is to go beyond the dictionary and add new meanings based on annotators’ intuitive judgment. It is a probable guess that there have long been such moves in that some researchers apparently annotate the word meanings subjectively. For example, in Rothe’s survey of the French word *et*, 72 different meanings are counted, which is usually too large a number of meanings for an entry in a dictionary to contain (Rothe, 1986, reported in Altmann, 2018: 41). Either way taken, this issue should hopefully have a better solution in the future studies.

5 Conclusions

In conclusion, this article attempts to investigate the features of semantic diversification on different levels of granularity. By way of extracting three quasi-prepositions from a corpus of the Shanghai dialect of Wu Chinese and annotating them semantically on two levels of granularity, we have answered three research questions.

First, several models are compared and those proposed by quantitative linguists show better performance than simple power functions. The exponential model and the right-truncated negative binomial model are found to be the best two considering the goodness of fit, consistency of parameter change, rationality, and simplicity. Second, our findings support the weak view of ‘scaling’ in complex sciences, that is, the meaning distributions on different levels of granularity all manifest the so-called A-curves by Kretzschmar in a rough sense. However, the parameters and shapes of models are different. In other words, the interpretation of scaling as self-similarity in a rigorous mathematical sense does not hold. Finally, there are several differences between the distributions on the two levels. The meaning distributions on a fine-grained level are found to be more right-skewed and more non-linear as compared with those on a coarse-grained one. This can also be proven mathematically given constant N (number of tokens) and variant M (number of types). The primary reason for the difference is attributed to the openness of the categories of systems.

The present study also adds to our understanding of the quantitative aspects of syntax-semantics interface or form-meaning mappings. Since the complex nature of ‘multiple-forms-to-multiple-meanings’ in natural language is widely acknowledged, in practice linguists start from the perspectives of synonymy (‘one-meaning-to-multiple-forms’) and polysemy (‘one-form-to-multiple-meanings’) in traditional terms, or onomasiological and semasiological approaches in usage-based, cognitive linguistic terms (Geeraerts, 2010). Köhler (1991) has made a similar distinction between two kinds of diversification from the perspective of quantitative linguistics. There has been such research into the former (Zhu & Liu, 2018) and we have contributed to the latter. On the macro level, the related distributional phenomena are attributed to the metaphorical language forces (Altmann, 1985; Altmann & Köhler, 1996), while on the micro level, they are reflections of several synergetic principles such as the minimization of efforts during language use or of inventories in language users' mind (Köhler, 2005, 2012).

Without doubt, this study also has its limitations. In the first place, it is still inevitable as we have pointed out that the differentiation of meanings is subjective. We have tried to minimize the degree of subjectivity such as resorting to dictionaries or basing the judgments on more concrete forms. Future studies might call for better measurements of meanings. Second, the size effect of the corpus is not tested in this survey, and we acknowledge that the size of hits may be criticized for being too small (up to a few hundreds). Third, we have only distinguished between two levels of granularity of meanings, while there is still a dearth of accurate measures of the hierarchical nature of meaning. Further investigations of these questions, along with a better distributional model or descriptive tool, are in need.

References

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica*, 19(1–2), pp. 177–200.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Eds.), *Quantitative Linguistics. An International Handbook*. Berlin: de Gruyter, pp. 646–658.

Altmann, G. (2018). *Unified Modeling of Diversification in Language*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler, R. (1996). "Language forces" and synergetic modelling of language phenomena. In: Schmidt, P. (Ed.), *Glottometrika 15*. Trier: WVT, pp. 62–76.

Beöthy, E., Altmann, G. (1984a). Semantic diversification of Hungarian verbal prefixes II. ki-. *Finnisch-ugrische Mitteilungen*, *8*, pp. 29–37.

Beöthy, E., Altmann, G. (1984b). Semantic diversification of Hungarian verbal prefixes III. "föl-", "el-", "be-". In: Rothe, U. (Ed.), *Glottometrika 7*, pp. 45–56. Bochum: Brockmeyer.

Beöthy, E., Altmann, G. (1991). The diversification of meaning of Hungarian verbal prefixes I. "meg-". In: Rothe, U. (Ed.). *Diversification Processes in Language: Grammar*. Hagen: Rottmann, pp. 60–66.

Best, K.-H., Rottmann, O. (2017). *Quantitative Linguistics, an Invitation*. Lüdenscheid: RAM-Verlag.

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Cong, J., Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews, 11*, pp. 598–618.

Croft, W. (2003). *Typology and Universals (2nd ed.).* Cambridge: Cambridge University Press.

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3), pp. 547–619.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics*, 17, pp. 66–78.

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics*, 17, pp. 79–96.

Fuchs, R. (1991). Diversifikation der Präposition *auf*. In: Rothe, U. (Ed.), *Diversification Process in Language: Grammar*, pp. 105–115. Hague: Rottmann.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Guiter, H. (1974). Les relations/Frequence-longueur-sens/Des mots (Langues Romanes et Anglais). In: Varvaro A. (Ed.). *XIV Congresso Internationale di Linguistica e Filologia Romanza: Napoli, 15–20 Aprile 1974. ATTI*, pp. 373–381. Amsterdam: John Benjamins.

Han, W., Arppe, A., Newman, J. (2017). Topic marking in a Shanghainese corpus: From observation to prediction. *Corpus Linguistics and Linguistic Theory,* 13(2), pp. 291–319. doi:10.1515/cllt-2013-0014

Hartmann, I., Haspelmath, M., Cysouw, M. (2014). Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language*, 38(3), pp. 463–484.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: Tomasello, M. (Ed.). *The New Psychology of Language: Cognitive and functional approaches to language structure* (Vol. 2). Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 211–242.

Hennern, A. (1991). Zur semantischen Diversifikation von *in* im Englischen. In Rothe, U. (Ed.), *Diversification Process in Language: Grammar*. Hague: Rottmann, pp. 116–126.

Horvath, B. M., Horvath, R. J. (2003). A closer look at the constraint hierarchy: Order, contrast, and geographical scale. *Language Variation and Change*, 15, pp. 143–170.

Hřebíček, L. (1996). Word associations and text. *Glottometrika*, 15, pp. 96–101.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (Ed.). *Diversification Process* *in Language: Grammar*, pp. 47–51. Hague: Rottmann.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (Eds.). *Quantitative Linguistics: An International Handbook*, pp. 760–774. Berlin/New York: de Gruyter.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin: de Gruyter.

Kolenčíková, N., Altmann, G. (2020). Analysis of Prepositions in Marína (Slovak Romantic Poem). *Glottometrics*, 48, pp. 88–107.

Kretzschmar, W. A. Jr. (2009). *The Linguistics of Speech*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr. (2015). *Language and Complex Systems*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr. (2018). *The Emergence and Development of English: An Introduction*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr., Kretzschmar, B. A., Brockman, I. M. (2013). Scaled measurement of geographic and social speech data. *Literary and Linguistic Computing*, 28, pp. 173–187.

Liu, H. (2012). Probability distribution of semantic roles in a Chinese treebank annotated with semantic roles. In: Naumann, S., Grzybek, P., Vulanović, R. Altmann, G. (Eds.). *Synergetic Linguistics. Text and Language as Dynamic Systems*, pp. 101–107. Vienna: Praesens.

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In: Wolman, B. B., Nagel, E. (Eds.). *Scientific Psychology*. Basic Books.

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, *156*, pp. 636–638.

Mandelbrot, B. (1982). *The Fractal Geometry of Nature*. New York: W. H. Freeman and Company.

Mandelbrot, B. (1997). *Fractals and Scaling in Finance: Discontinuity, concentration, risk*. New York: Springer.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law—another view. *Quality & Quantity*, *44*(4), pp. 713–731.

Roos, U. (1991). Diversifikation der japanischen Postposition "-ni". In: Rothe, U. (Ed.). *Diversification Process in Language: Grammar*, pp. 75–82. Hague: Rottmann.

Rothe, U. (1986). *Die Semantik des textuellen et*. Frankfurt: Peter Lang.

Rothe, U. (1991). The diversification of the case: genitive. In Rothe, U. (Ed.). *Diversification Process in Language: Grammar*, pp. 140–156. Hague: Rottmann.

Rothe, U. (Ed.). (1991). *Diversification Process in Language: Grammar*. Hague: Rottmann.

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics*, *19*, pp. 70–79.

Strauss, U., Altmann, G. (2006). *Diversification. In the Encyclopedia of Linguistic Laws and the Laws in Quantitative Linguistics.* Retrieved from http://lql.uni-trier.de/index.php/Diversification

Van Valin, R. D., Jr. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511610578

Wang, L., Guo, Y., Ren, C. (2021). A Quantitative Study on English Polyfunctional Words. *Glottometrics*, 50, pp. 42–56.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. (Eds.). *Quantitative Linguistics: An International Handbook*, pp. 791–807. Berlin/New York: Walter de Gruyter.

Zhu, J., Liu, H. (2018). The distribution of synonymous variants in Wenzhounese. *Glottometrics*, *41*, pp. 24–39.

Software

Altmann, G. (2000). *Altmann-Fitter 3.1.0* [Computer software]. Lüdenscheid: RAM-Verlag. Retrieved from http://www.ram-verlag.biz/altmann-fitter/

Ye, L. (2019). Wordless (Version 1.3.0) [Computer software]. Retrieved from https://github.com/BLKSerene/Wordless

Appendix I

The meaning distributions of three quasi-prepositions on the coarse level

Wordform	Meaning	x[i]	F[i]	NP[i][16]				
				EXP	MRTZA	RTNB	RTZ	ZM
nau	‘take’	1	123	137	123	124	145	140
	theme	2	106	86	116	102	63	81
	patient	3	83	54	52	59	39	48
	‘hold’	4	7	34	24	29	27	29
	instrument	5	7	21	12	13	21	18
	‘use’	6	4	14	6	6	17	11
	desiderative	7	4	9	3	2	14	7
	‘taking’	8	3	6	2	1	12	5
peh	‘give’	1	123	133	123	114	147	154
	recipient	2	100	90	113	110	70	88
	passive	3	75	61	62	73	45	53
	permissive	4	43	41	34	41	33	33
	beneficiary	5	17	28	20	21	26	21
	patient	6	11	19	12	10	21	14
	causative	7	5	13	8	5	18	9
	‘according to’	8	3	9	5	2	16	6
tah	recipient	1	16	18	16	14	20	17
	beneficiary	2	14	13	17	16	11	13
	companion	3	14	10	10	13	8	10
	‘relation’	4	7	8	7	9	6	7
	NP conjunction	5	4	6	5	6	5	6
	patient	6	4	4	4	3	5	4
	‘same’	7	3	4	3	2	4	4
	comparative	8	1	3	2	1	4	3

[16] The theoretical values are rounded here, as the frequencies are all integers.

Appendix II

The meaning distributions of *nau* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	123	43	1	85	1
2	9	44	1	86	1
3	7	45	1	87	1
4	7	46	1	88	1
5	6	47	1	89	1
6	6	48	1	90	1
7	5	49	1	91	1
8	5	50	1	92	1
9	5	51	1	93	1
10	4	52	1	94	1
11	4	53	1	95	1
12	4	54	1	96	1
13	4	55	1	97	1
14	4	56	1	98	1
15	4	57	1	99	1
16	3	58	1	100	1
17	3	59	1	101	1
18	3	60	1	102	1
19	3	61	1	103	1
20	3	62	1	104	1
21	3	63	1	105	1
22	3	64	1	106	1
23	2	65	1	107	1
24	2	66	1	108	1
25	2	67	1	109	1
26	2	68	1	110	1
27	2	69	1	111	1
28	2	70	1	112	1
29	2	71	1	113	1
30	2	72	1	114	1
31	2	73	1	115	1
32	2	74	1	116	1
33	2	75	1	117	1
34	2	76	1	118	1
35	2	77	1	119	1
36	2	78	1	120	1
37	2	79	1	121	1
38	1	80	1	122	1
39	1	81	1	123	1
40	1	82	1	124	1
41	1	83	1	125	1
42	1	84	1	126	1

Appendix III

The meaning distributions of *peh* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	123	37	2	73	1
2	43	38	1	74	1
3	25	39	1	75	1
4	12	40	1	76	1
5	8	41	1	77	1
6	7	42	1	78	1
7	6	43	1	79	1
8	5	44	1	80	1
9	5	45	1	81	1
10	5	46	1	82	1
11	4	47	1	83	1
12	4	48	1	84	1
13	4	49	1	85	1
14	3	50	1	86	1
15	3	51	1	87	1
16	3	52	1	88	1
17	3	53	1	89	1
18	3	54	1	90	1
19	3	55	1	91	1
20	3	56	1	92	1
21	2	57	1	93	1
22	2	58	1	94	1
23	2	59	1	95	1
24	2	60	1	96	1
25	2	61	1	97	1
26	2	62	1	98	1
27	2	63	1	99	1
28	2	64	1	100	1
29	2	65	1	101	1
30	2	66	1	102	1
31	2	67	1	103	1
32	2	68	1	104	1
33	2	69	1	105	1
34	2	70	1	106	1
35	2	71	1	107	1
36	2	72	1	108	1

Appendix IV

The meaning distributions of *tah* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	13	14	1	27	1
2	4	15	1	28	1
3	3	16	1	29	1
4	3	17	1	30	1
5	2	18	1	31	1
6	2	19	1	32	1
7	2	20	1	33	1
8	2	21	1	34	1
9	2	22	1	35	1
10	2	23	1	36	1
11	1	24	1	37	1
12	1	25	1	38	1
13	1	26	1		

Fellow or foe? A quantitative thematic exploration into Putin's and Trump's stylometric features

Yaqin Wang[1], Ting Zeng[2*]

[1] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies
[2] School of Foreign Languages, College of Arts and Sciences, Shanghai Polytechnic University
[*] Corresponding author's email: zengting@sspu.edu.cn

DOI: https://doi.org/10.53482/2023_54_406

ABSTRACT

Thematic concentration, a quantitative linguistic method, can reflect the speech style of a particular person. It may, to some degree, reflect the degree of a speaker's intention to communicate certain themes. There has been limited empirical research on the similarity between Trump and Putin with respect to their linguistic features. Thus, the present study aims to compare Putin's and Trump's stylometric features and political themes based on thematic concentration with a corpus of Putin's, Medvedev's, Trump's, and Obama's speeches. Results show that 1) Both Putin's and Trump's speeches' thematic concentration values are significantly or marginally significantly different from their precedents'. 2) Two leaders pay great attention to the concept of nationalism. 3) Thematic words of their speeches were slightly different across periods, reflecting the influence of external factors on the choice of language. The results of the present study may shed light on the stylometric studies of Putin and Trump.

Keywords: thematic concentration; stylometric features; Putin; Trump; authoritarianism

1 Introduction

Notoriously renowned for the slangy, vulgar, and violent political talk since the advent of the key phrase *Mochit' v sortire* 'to kill somebody in a toilet', the Russian politician Vladimir Putin, has attracted much scholarly attention from linguists and discourse analysts (Glukhova and Sorokina 2018; Sedykh 2016). Despite his offensive linguistic style, Putin was also commented on as a politician with a deliberate choice and strategy that serves political ends by legitimizing jargon or semi-jargon language in the official report (Glukhova and Sorokina 2018; Gorham 2014).

Sounds familiar? Commonly regarded as an object of comparison of Putin's political inclination (Hauser 2018), the former American president, Donald J. Trump, and his team also appear to have employed a deliberate or idiosyncratic campaigning style and rhetoric (Mercieca 2020; Reyes and Ross 2021). More importantly, Putin, on the one hand, was dubbed as an iron-fist father figure with an inevitably authoritarian inclination under the political scheme of Russia (Gorham 2005), with some political scientists defining his political strategies as "Putinism" (Fish 2017). The ideology of Trump (so-called Trumpology, Trumpism), on the other hand, has been regarded as a type of authoritarian leadership principle (Rivers and Ross 2020) as well. Van Dijk (2008) has pointed out that Putin is used to employing a positive self-presentation and a negative presentation of his opponents. Trump, in a similar vein, unsurprisingly resorted to similar construction of a 'self-versus opponent' image (Homolar and Scholz 2019; Ross and Caldwell 2020). Despite many similarities between Putin and Trump in terms of speech strategies, empirical linguistic research into the stylometric features has been much more limited. Exploration of this topic may help to clarify the relationship between Putin's and Trump's speech styles.

As one of the important measurements related to content analysis in quantitative linguistics, thematic concentration can indicate the speech style of a writer or speaker (Čech et al. 2015). As Čech (2016, p. 9, cited from Chen and Liu (2018, p. 68) and reformulated by authors) points out,

"the method of measuring thematic concentration can be classified among the types of textual analysis that are generally referred to as content analysis. In its nature, it is also close to quantitative analysis of the so-called 'keywords analysis'. However, as is evident from the title of this method, its primary aim is … to reveal the extent to which the author has addressed the topic(s) on the given theme or themes on the whole. From a more general perspective, it is a method for modeling a particular aspect of speech behavior."

This method has been used in investigating presidential inaugural speeches (Kubát and Čech 2016) and political debates (Savoy 2018). A number of studies have applied it to investigate linguistic features of official reports and political speeches (Čech 2014; Chen and Liu 2015, 2018; Wang and Liu 2018). Further, Čech (2014) reported significant differences in the levels of thematic concentration between Czechoslovak and Czech presidents from the totalitarian period and the period of democracy respectively. He suggested that the level of thematic concentration may, to some degree, indicate a tendency of ideology, be it a more totalitarian (a higher level of thematic concentration) or a more democratic one (a lower level). Wang and Liu (2018) reported a higher level of thematic concentration in Trump's campaign speeches, which is somehow consistent with the previous conclusion of his political inclination toward authoritarianism. These studies highlight the significance of thematic concentration in stylometric analyses.

Thus, the present study intends to compare Putin and Trump's speech style during their presidency based on the quantitative linguistic method, thematic concentration, by employing three indicators, viz., thematic concentration (TC), secondary thematic concentration (STC), and proportional thematic concentration (PTC). Since the value of thematic concentration is closely related to the indicator of h-point in scientometrics, which is rather sensitive to the language type (Popescu 2009), we compared theirs with those of their respective political predecessors, Medvedev and Obama[1]. Two sets of values, the Putin-Medvedev pair and the Trump-Obama pair, were collected. On top of that, thematic words reflecting the political themes of two political figures, namely, the Putin-Trump pair, were compared.

Research questions are as follows:

1. What is the relationship between Putin's and Medvedev's thematic concentration values?

2. What is the relationship between Trump's thematic concentration value compared with Obama's? Further, is Putin's position in the Putin-Medvedev pair different from Trump's in the Trump-Obama pair?

3. What are the thematic words of Putin and Trump, and what are the political themes they intend to emphasize?

The paper's layout is organized as follows: Section 1 introduces the general background information. Section 2 displays the details of the methods and materials employed in the study. Section 3 presents results and discussion, followed by conclusions and suggestions for further research in Section 4.

2 Methods and Materials

2.1 Materials

The organization of linguistic materials is shown in Table 1, 200 texts and 719,894 tokens in total. Putin's and Medvedev's materials were gleaned from the official website of the President of Russia,[2] and Trump's and Obama's were from the American Presidency Project.[3] Each political figure's speeches during their terms in office were chosen, including addresses to the Federal Assembly, or addresses before a joint session of the congress on the State of the Union, news conferences and remarks at special occasions. For each year, 6-14 texts were selected for each person. The composition of the corpus is displayed in Table 1 and specific information, i.e., date, place and theme, of each text is in Appendix A. It should be noted that the authorship of presidents' or political candidates' speeches is always disputable. President, however, is the one who delivers the speech. He is politically responsible for their speeches and thus can affect the text to some degree (Čech 2014).

[1] It would be more reliable to collect more former presidents' texts as the reference corpus. However, Putin has only one predecessor in the last two decades. Thus, we only chose speeches of Medvedev and Obama for comparison. In the future, texts of Russian politicians other than the president can be gleaned to further the research.
[2] http://www.kremlin.ru/
[3] https://www.presidency.ucsb.edu/

Table 1: The composition of the corpus.[4]

	Addresses to the Federal Assembly/ the State of the Union	News conference	Remarks at special occasions	Time range		Texts	Tokens
Putin	4	5	41	2017-2021	2017: 12 texts 2018: 11 texts 2019: 12 texts 2020-2021: 15 texts	50	157,051
Medvedev	4	4	42	2008-2012	2008: 9 texts 2009:12 texts 2010: 11 texts 2011: 11 texts 2012: 7 texts	50	126,514
Trump	3	4	43	2017-2021	2017: 13 texts 2018: 12 texts 2019: 11 texts 2020: 14 texts	50	209,225
Obama	7	6	37	2011-2016	2010-2011: 13 texts 2012: 7 texts 2013: 8 texts 2014: 9 texts 2015: 6 texts 2016:7 texts	50	227,104
Total	18	19	163	/		200	719,894

2.2 Methods

As an approach to measure the degree of the author's intention to communicate certain themes, thematic concentration (TC) was introduced by Popescu (2007) and further developed by a series of works (e.g., Popescu et al. 2009). The computation of TC is based on the concept of the h-point, which was conceived by Hirsch (2005) for scientometrics and then introduced into linguistics by Popescu (2007). If we rank word frequencies of a text in descending order, we can determine the value of the h-point when the rank of a particular word is equal to its occurrence. Figure 1 shows the position of an h-point in a rank-frequency distribution of a certain text.

[4] As shown in Appendix A, for Putin and Trump, only 2-3 texts were collected in 2021, thus we combine texts of 2020 and 2021 together. This also holds true for the group of 2010-2011 of Obama's texts (only one text was collected in 2010).

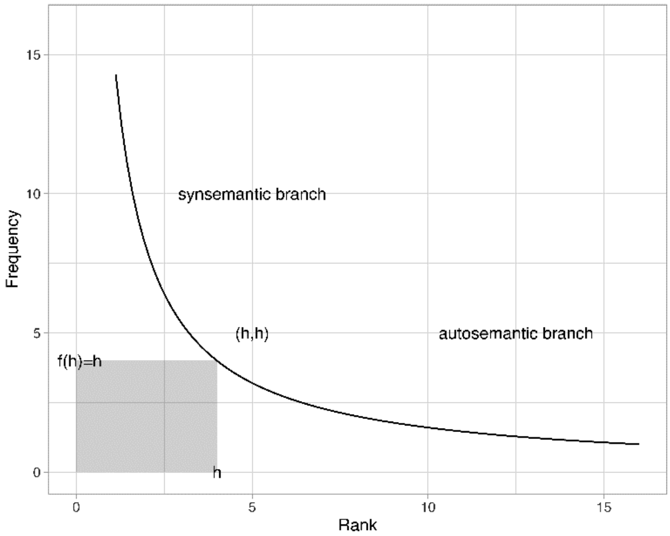

Figure 1: The position of the h-point in a rank-frequency distribution (cited from (Popescu et al. 2009, p. 17).

Popescu et al. (2009) demonstrated that the h-point fuzzily separates the frequent synsemantics (including prepositions, pronouns, particles, articles) from the autosemantics (including nouns, adjectives, and verbs), which build the major vocabulary of the text. Autosemantic words which occur before the h-point indicate that they are frequently used by the author. They represent the text themes (nouns) and descriptions and actions of certain central words (adjectives and verbs). This may signify that the author intends to communicate certain themes with others. The calculation of the h-point in the frequency distribution of lemmas[5] is shown below (for more details, see Popescu et al. 2009):

(1)
$$h = \begin{cases} \mathrm{r}_i, & \mathrm{r}_i = f(\mathrm{r}_i) \\ \frac{f(\mathrm{r}_i)\mathrm{r}_{i+1} - f(\mathrm{r}_{i+1})\mathrm{r}_i}{\mathrm{r}_{i+1} - \mathrm{r}_i + f(\mathrm{r}_i) - f(\mathrm{r}_{i+1})}, & \mathrm{r}_i \neq f(\mathrm{r}_i) \end{cases}$$

Based on the value of the h-point, the computation of thematic concentration can be defined as:

(2)
$$TC = 2\sum_{r'}^{T} \frac{(h-r')f(r')}{h(h-1)f(1)}$$

where *f(1)* is the frequency of the first rank, T is the number of autosemantics before the h-point, and r' is the average rank ($r' < h$).

[5] Čech (2014) and Čech at al. (2015) computed the h-point value based on the frequency distribution of lemmas (i.e., canonical forms of words). Hence the current study followed suit.

Table 2: Rank frequency distribution of Putin's speech of *Парад Победы на Красной площади* 'Victory Parade on Red Square' on May 9, 2021.

Rank	Average rank	Frequency	Lemma	English translation
1	1	60	*и*	and
2	2	20	*в*	at
3	3	17	*наш*	our
4	4	14	*мы*	we
5	5	14	*на*	on
6	6	10	*с*	with
7	**7.5**	**9**	***победа***	**victory**
8	7.5	9	*тот*	that
9	9.5	8	*который*	which
10	9.5	8	*кто*	who
11	11.5	7	*к*	to
12	**11.5**	**7**	***народ***	**people**
13	14	6	*быть*	be
14	**14**	**6**	***война***	**war**
15	14	6	*для*	for
16	18	5	*весь*	all
17	18	5	*год*	year
18	18	5	*за*	for
19	18	5	*по*	by
20	18	5	*сила*	power
21	21	4	*великий*	great

For example, the rank frequency distribution of Putin's speech of *Парад Победы на Красной площади* 'Victory Parade on Red Square' is displayed in Table 2. As Table 2 shows, there is no rank of a lemma that exactly equals its corresponding frequency, thus we calculate it by the second part of the Formula (1):

$$h_{sample\ text} = \frac{9 * 9 - 8 * 8}{9 - 8 + 9 - 8} = 8.5$$

Thus, there is one autosemantic word which lies in the pre-h domain, i.e., *победа* 'victory', as shown in Table 2. The TC value is calculated as follows according to Formula (2):

$$TC_{sample\ text} = 2 * \left(\frac{(8.5 - 7.5) * 9}{8.5 * (8.5 - 1) * 60}\right) = 0.0047$$

A problem occurs when the TC value of a certain text is 0, which poses a challenge for comparing thematic differences between texts. Therefore, Čech et al. (2015) proposed the indicator of secondary thematic concentration (STC) by doubling the h point.

(3) $$STC = \sum_{r'=1}^{2h} \frac{(2h-r')f(r')}{h(2h-1)f(1)}$$

The STC value of the sample text in Table 2 is displayed as well. 2h point of the text is 8.5*2=17, and there are three autosemantics before 2h point. STC value is:

$$STC = \frac{(17-7.5)*9}{8.5*(17-1)*60} + \frac{(17-11.5)*7}{8.5*(17-1)*60} + \frac{(17-14)*6}{8.5*(17-1)*60} = 0.0174$$

The third formula is called proportional thematic concentration (PTC). It is proposed to eliminate the circumstance where there is only one content word in the pre-h domain in a text (Čech et al. 2015). It is computed as:

(4) $$PTC = \frac{1}{N_h} \sum_{r'<h} f(r')$$

N_h refers to the frequency of all words $r_1, \ldots, r_h$, in the pre-h domain, the sum of *f (r')* is the frequency of all autosemantic words occurring before the h point. PTC value of the sample text is:

$$PTC = \frac{9}{153} = 0.0588$$

In sum, a higher level of TC, STC, and PTC signify the author'xss effort in communicating more intensive certain themes with others, while the lower one suggests the diversity of one's themes.

As argued by Čech (2016), TC and STC values are independent of text length of the range <200, 6500>. PTC values are said not to be a suitable tool for comparing texts with a length of N < 2000 words. In the present study, the lengths of most texts (171 texts) roughly fall into the interval of <200, 6500> and more than half of texts' (116 texts) lengths are greater than 2000. Thus, to investigate the influence of text size which may exert on indicators, we carried out three Pearson tests between TC, STC, and PTC and the text size. Results show that the correlation coefficient between TC, STC, and PTC and the text size is low (Pearson *r* = -0.13, -0.33, 0.04 respectively). Thus, we can compare indicators of texts with different sizes.

3 Results and Discussion

This section discusses quantitative results and possible factors for those phenomena. The comparisons of TC, STC, and PTC values in the Putin-Medvedev pair and the Trump-Obama pair are carried out, followed by analyses of the thematic words of the Putin-Trump pair. In both the comparison of thematic concentration and that of thematic words, diachronic comparisons or analyses of their speeches are shown after the general discussion.

3.1 Comparison of Putin's and Medvedev's thematic concentration

Table 3 displays descriptive statistics of three indicators from two Russian presidents during their terms of office.

Table 3: Descriptive statistics of three indicators of thematic concentration of Putin-Medvedev pair.

	TC	**STC**	**PTC**
Putin			
Min.	0	0.0054	0
First quartile	0.0098	0.0260	0.0717
Median	0.0234	0.0411	0.1118
Mean	**0.0321**	**0.0432**	**0.1215**
Third quartile	0.0403	0.0562	0.1613
Max.	0.1374	0.1081	0.3060
Standard Deviation	0.0318	0.0222	0.0731
Medvedev			
Min.	0	0.0028	0
First quartile	0.0018	0.0265	0.0340
Median	0.0203	0.0340	0.0878
Mean	**0.0201**	**0.0363**	**0.0835**
Third quartile	0.0324	0.0431	0.1247
Max.	0.0783	0.0937	0.2171
Standard Deviation	0.0190	0.0174	0.0612

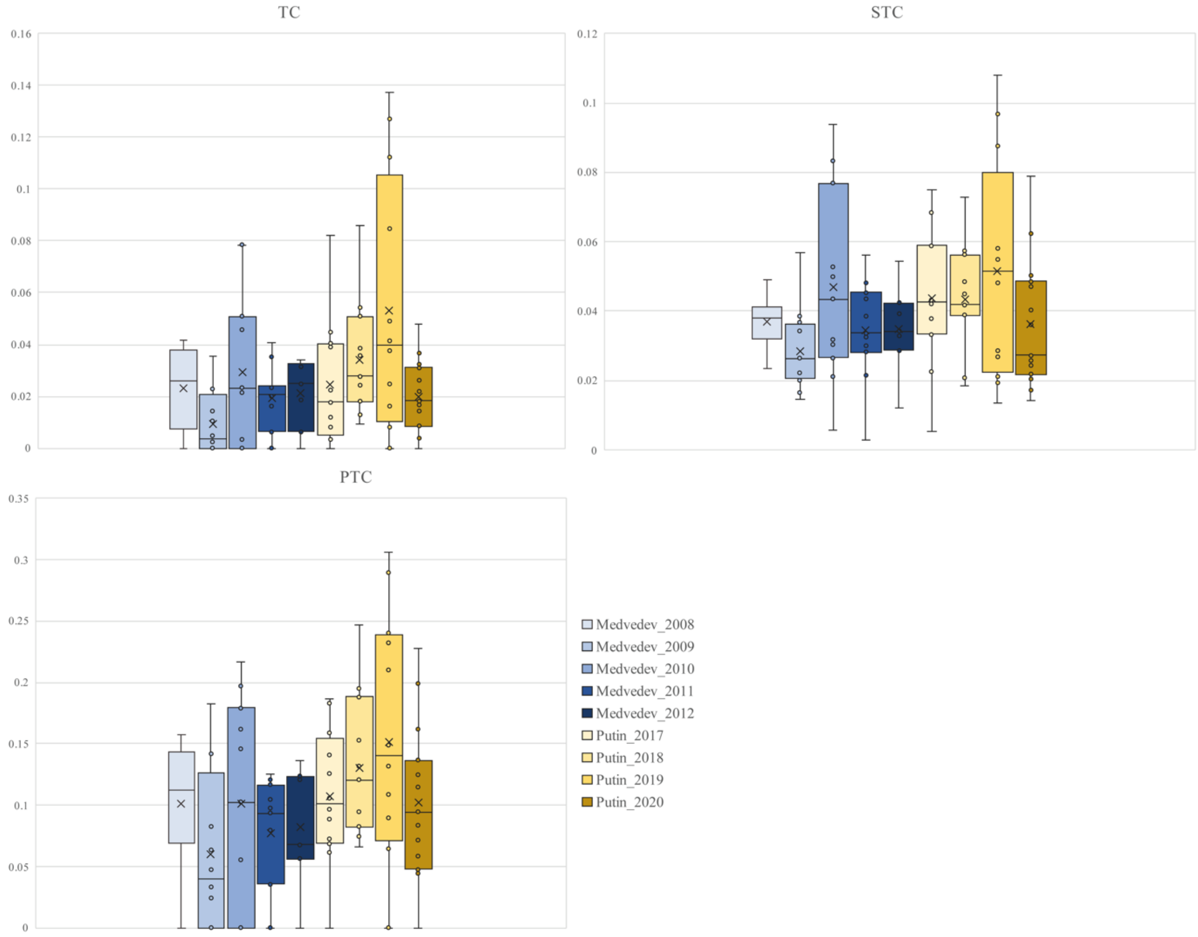

Figure 2: The distribution of three indicators of thematic concentration in the Putin-Medvedev pair. Boxes are the distribution of TC/STC/PTC values of two people each year as the legend displays. The blue series of boxes represents Medvedev's indicators and the yellow one is Putin's. The label "x" on each plot is the average value of each distribution.

As Table 3 and Figure 2 show, the average values of all indicators of Putin are greater than those of Medvedev. Then, a non-parametric test (Mann Whitney U) was carried out on TC values and two t-tests on STC and PTC values, respectively (since the set of TC values does not follow the normal distribution).

Results of the Mann Whitney U test show that values of Putin's TC (*Mdn* = 0.0272) is marginally significantly different from those of Medvedev's (*Mdn* = 0.0220, *U* = 966, *p* = .05 < .1). Regarding the values of the other two indicators, results of t-tests for two independent samples demonstrate that the difference between Putin's STC (*M* = 0.0432, *SD* = 0.0222) is marginally significant from Medvedev's (*M* = 0.0363, *SD* = 0.0174) values (*t* (98) = 1.738, *p* = .085 < .1). PTC values of Putin (*M* = 0.1215, *SD* = 0.0731) are significantly greater than those of Medvedev (*M* = 0.0835, *SD* = 0.0612, *t* (98) =2.823, *p* = .006 < .05). As Figure 2 shows, most of PTC values are higher in Putin's speeches.

These results indicate that regarding three indicators, Putin's thematic concentration is significantly (or marginally significantly) greater than Medvedev's. Three metrics, especially PTC values, can distinguish two people's degrees of thematic concentration. This implies that, to some degree, Putin's intention to communicate some topics is greater than that of Medvedev. In other words, his discursive practice contains relatively more central themes, while his predecessor's speeches reflect the diversity of themes.

As Čech et al. (2015) commented, texts with STC < TC can be regarded as extremely concentrated texts. We counted the number of texts and found that 9 texts of Putin's meet this requirement, while only 5 ones of Medvedev's do. This shows that, compared with Medvedev's speeches, more of Putin's speeches reach the extreme end of thematic concertation. This, additionally, reflects Putin's intense intention of communicative practice.

Diachronically, we compared Putin's speeches according to chronological order, i.e., based on four sets of speeches ranging from 2017 to 2021. Results of a One-way ANOVA test show no significant differences among speeches from 2017 to 2021 for STC and PTC values (p_{stc} = .364, p_{ptc} = .293). TC values show significant differences among different periods (*F* (3, 46) = 3.123, *p* = .035 < .05), however, the post-hoc test shows that only TC values of 2019 are significantly different from those of 2020-2021. It can be seen that in Figure 2, TC values for 2019 are greater than those for 2020-2021. This indicates that Putin did show differences across different periods diachronically in terms of the degree of concentration on certain themes.

3.2 Comparison of Trump's and Obama's thematic concentration

Table 4 displays descriptive statistics of three measurements from Trump and Obama during their terms of office.

Table 4: Descriptive statistics of three measurements of the thematic concentration of the Trump-Obama pair.

	TC	STC	PTC
Trump			
Min.	0	0.0115	0
First quartile	0.0035	0.0183	0.0335
Median	0.0095	0.0209	0.0560
Mean	**0.0114**	**0.0225**	**0.0619**
Third quartile	0.0165	0.0256	0.0892
Max.	0.0668	0.0529	0.1361
Standard Deviation	0.0114	0.0076	0.0376
Obama			
Min.	0	0.0025	0
First quartile	0.0021	0.0101	0.0206
Median	0.0045	0.0138	0.0361
Mean	**0.0054**	**0.0143**	**0.0363**
Third quartile	0.0080	0.0174	0.0538
Max.	0.0216	0.0323	0.0999
Standard Deviation	0.0046	0.0060	0.0251

As Table 4 shows, the mean values of three indicators of Trump are greater than those of Obama. Figure 3 displays that most of Trump's TC, STC, and PTC values are greater than those of Obama. Then, one non-parametric test (Mann Whitney U) was carried out on TC values and two t-tests on STC and PTC values, respectively (since only the set of TC values followed the normal distribution).

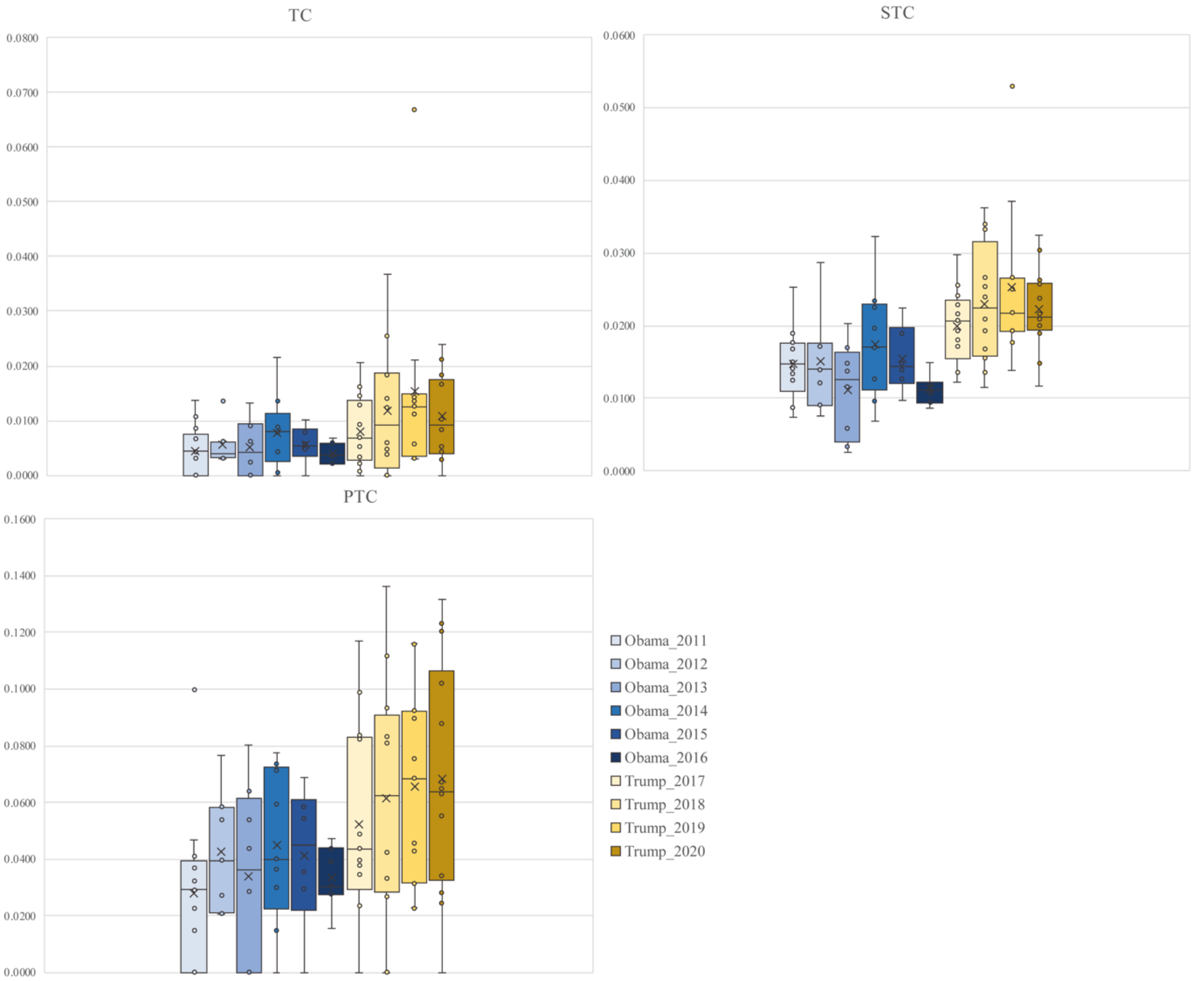

Figure 3: The distribution of three measurements of thematic concentration in the Trump-Obama pair. Boxes are the distribution of TC/STC/PTC values of two people each year as the legend displays. The blue series of boxes represents Obama's indicators and the yellow one represents Trump's. The label "x" on each plot is the average value of each distribution.

Results of the Mann Whitney U test show that values of Trump's TC (Mdn_{Trump} = 0.0095) is significantly different from those of Obama's (Mdn_{Obama} = 0.0045, U = 794, p = .002 < .01). Regarding the values of the other two indicators, results of t-tests for two independent samples demonstrate that Trump's STC (M = 0.0225, SD = 0.0076) is significantly greater than Obama's (M = 0.0143, SD = 0.0060) values (t (98) = -5.997, p < .0001). PTC values of Trump (M = 0.0619, SD = 0.0376) are significantly greater than those of Obama (M = 0.0363, SD = 0.0251) (t (98) = -4.013, p < .0001).

These results indicate that Trump's thematic concentration is significantly greater than Obama's regarding three indexes. This implies that, to some degree, Trump's intention to convey certain themes is greater than that of Obama; in other words, his speeches contain relatively more central themes while those of his predecessor reflect the diversity of themes. Here, levels of three indicators in Trump's speeches are significantly higher than those of Obama, suggesting his preference for an authoritarian leadership style. This result is consistent with Wang and Liu (2018)'s findings that the significantly greater TC levels of Trump's campaign speeches than those of Obama and Clinton.

Furthermore, in addition to the level of TC (in Wang and Liu's (2018) research), STC and PTC values applied in the current research also demonstrate a similar tendency of the distribution.[6] Results suggest that in addresses and remarks other than campaign speeches, Trump, as usual, demonstrates the tendency of concentrating on a handful of political themes. Moreover, 3 texts of Trump whose STC value is smaller than the TC value, while none of Obama's texts does so. As mentioned in Putin-Medvedev pair, 9 texts of Putin whose STC value is smaller than the TC value and 5 for Medvedev. This shows that generally speaking, Russian presidents' extremeness of TC is more evident than that of American presidents.

Likewise, we carried out a statistical test on values of three metrics across different periods. Results demonstrate no significant differences among speeches from 2017 to 2021 for Trump (p_{tc} = .480, $p_{s\text{tc}}$ = .402, $p_{p\text{tc}}$ = .718). This indicates that diachronically, Trump did not show obvious differences in terms of the degree of concentration on certain themes. Compared with Putin's results, Trump's intention to convey certain themes remains consistent no matter when the speech was delivered.

Together with what we have discussed so far, the values of TC, STC, and PTC in Putin's texts are significantly or marginally significantly greater than those of Medvedev's speeches; in a similar vein, those of Trump's are significantly greater than those of Obama's. Both Trump and Putin tend to concentrate on certain central themes compared with their predecessors. Čech (2014) suggested that the level of thematic concentration may, to some degree, indicate a tendency toward ideology. As noted by political scientists, e.g., Medvedev attempted to employ moderate reformism by promoting economic modernization and political liberalization (Noriega 2016). On the contrary, commonly reported as a strong leader with an iron fist, Putin is famous for his so-called father-figure leadership style. As for Trump, During the 2016 election, his authoritarian tendency has been one of the key factors in his winning the presidency (MacWilliams 2016; Homolar and Scholz 2019). The theme intensity of Putin and Trump may reflect their authoritarian leadership style to some extent. Future research, however, including more presidents and texts, is needed to explore this relationship.

[6] Since Čech (2014) and Wang and Liu (2018) only investigated the level of TC in speeches, this somehow suggests the applicability of STC and PTC values in the thematic concentration comparison.

More importantly, Trump's differences from his former president are more obvious than those of the Putin-Medvedev comparison. This reflects Trump's peculiarities again compared with traditional politicians. Davis (2020: 77) suggested, "neither Trump nor Putin made explicitly calls for authoritarianism…, despite evidence suggesting otherwise." By analyzing the political speeches of the two presidents, Davis then concluded that, though in their idiosyncratic ways, Trump and Putin constructed a kind of power centered around themselves, reflecting features of authoritarian leaders. This, to some extent, indicates that those two political leaders share a similar tendency from the aspect of thematic concentration.

Let us hence propose a question further, what are their thematic words and what kind of political themes do they want to emphasize?

3.3 Comparison of Putin's and Trump's thematic nouns

Due to limited space and the fact that nouns reflect political themes better, we only gleaned thematic nouns based on TC. The total frequencies of words (Frequency), the number of texts they occurred in (Occurrence), and the average value of ranks (Average rank) among all occurrences are shown in Table 5 and Table 6.

Table 5: The relevant information on Putin's thematic nouns.

	Thematic word	Translation	Frequency	Occurrence	Average rank
1	*год*	year	741	15	11
2	*человек*	man	544	13	20
3	*страна*	country	440	14	15
4	*Россия*	Russia	401	14	10
5	*вопрос*	question	216	3	35
6	*всё*	everything	177	3	43
7	*развитие*	development	166	6	19
8	*семья*	family	74	2	27
9	*работа*	work	71	3	19
10	*регион*	region	62	3	12
11	*процент*	percent	61	1	38
12	*система*	system	61	1	24
13	*решение*	solution	57	1	42
14	*восток*	East	44	1	15
15	*сотрудничество*	cooperation	44	3	9
16	*оружие*	weapon	42	1	33
17	*господин*	Sir	31	1	12
18	*гражданин*	citizen	31	1	29
19	*эпидемия*	epidemic	29	1	18
20	*отношение*	attitude	26	2	8

21	*коллега*	colleague	23	1	17
22	*проблема*	problem	22	1	15
23	*экономика*	economy	21	2	9
24	*интеллект*	intelligence	20	1	4
25	*Сербия*	Serbia	19	1	4
26	*Африка*	Africa	18	1	3
27	*бизнес*	business	18	1	8
28	*война*	war	18	2	6
29	*Монголия*	Mongolia	18	1	3
30	*государство*	state	17	1	10
31	*соотечественник*	compatriot	16	1	6
32	*ООН*	UN	15	1	8
33	*прокуратура*	Prosecutor's office	15	1	6
34	*спорт*	sports	15	1	4
35	*лауреат*	laureate	13	1	5
36	*право*	the right	13	1	8
37	*премия*	prize	12	1	9
38	*учитель*	teacher	12	1	3
39	*число*	number	12	1	10
40	*победа*	victory	9	1	7
41	*организация*	organization	8	1	7
42	*двадцатка*	G20	7	1	5

Table 6: The relevant information on Trump's thematic nouns.

	Thematic word	**Frequency**	**Occurrence**	**Average rank**
1	people	766	16	22
2	country	296	8	24
3	America	160	6	16
4	ballot	119	2	25
5	election	116	2	25
6	year	106	3	24
7	nation	92	3	15
8	vote	89	2	31
9	state	85	2	32
10	United	71	3	14
11	States	61	3	17
12	thing	58	1	38
13	Israel	43	1	28
14	Korea	38	1	12
15	tax	37	1	15
16	voter	34	1	30
17	drug	33	1	12

18	Coast	32	1	16
19	Guard	32	1	17
20	Lou	30	1	14
21	Afghanistan	25	1	13
22	price	25	1	18
23	vaccine	25	1	22
24	Dame	21	1	17
25	Notre	21	1	18
26	trade	21	1	17
27	God	19	1	13
28	Matt	18	1	14
29	Justice	17	1	15
30	virus	14	1	13

As shown in Table 5 and 6, it can be seen that in the past five years, there exist similarities and differences between two people's thematic words. For Putin, themes addressed most prominently mainly include the concept of nation and people (*человек* 'man', *Россия* 'Russia', *страна* 'country', *семья* 'family', *гражданин* 'citizen'), socio-economic development (*развитие* 'development', *экономика* 'economy', *бизнес* 'business'), other nations and foreign policy (*Сербия* 'Serbia', *Африка* 'Africa', *Монголия* 'Mongolia', *ООН* 'UN', *двадцатка* 'G20', *сотрудничество* 'cooperation'), security and wars (*оружие* 'weapon', *война* 'war'), epidemic (*эпидемия* 'epidemic'), etc. Putin focused on the idea of a strong, secure Russia (Davis, 2020), which is consistent with the most frequent thematic nouns (*человек* 'man', *страна* 'country', *Россия* 'Russia'). For Trump, he intensified topics related to nation and people (*people*, *country*, *America*, *nation*, *United States*) as well, election (*ballot*, *election*, *vote*, *voter*), economy (*tax, trade*, *price*), epidemic (*vaccine*, *virus*), social policy (*state*, *drug*), foreign policy (*Israel*, *Korea*, *Afghanistan*, *guard*), etc. The first two thematic words are consistent with the most frequent content words in his campaign corpus (Homolar and Scholz 2019).

Specifically, four words, namely, *год* 'year', *человек* 'man', *Россия* 'Russia', *страна* 'country', are the most frequent thematic words and occurred in more than 10 texts in Putin's speeches, while total frequencies of *people*, *country*, *America* rank the first three positions for Trump's texts, occurring in 5 or more texts. Both Trump and Putin emphasize the issues related to people and country. The concept of people is one of the basic concepts of political discourse (Yakoba 2017) and is often used as a tool of political manipulation. As stated by Yakoba (2017: 167), in several speeches delivered by Trump, no matter which topic he was talking about, "by emphasizing on the importance of the people, Trump...constructs a basis for creating an impression of concern for the nation." This works well in Putin's case, too, as he addressed the issue of people intensively.

As for the diachronic change in two presidents' thematic words, we calculated total frequencies of words and the number of texts they occurred in each year. As shown in Appendix B, the most frequent thematic nouns in Putin's texts from 2017 to 2019 are *год* 'year' while the most frequent one is *страна* 'country' in the year 2020 and 2021. The theme of nation and people (*Россия* 'Russia', *страна* 'country', *человек* 'man') ranks in the first several positions for five years, which again highlights Putin's intention on emphasizing the concept of country and people when addressing to his audience.

In 2017, development and security (*развитие* 'development', *оружие* 'weapon') were given enough attention, in 2019, the topics on global issues and foreign policy (*восток* 'East', *регион* 'region', *Сербия* 'Serbia', *Африка* 'Africa', *Монголия* 'Mongolia', *сотрудничество* 'cooperation') were repeatedly mentioned by Putin. When in 2020 and 2021, the period of COVID 19, the theme related to the pandemic and socio-economic development (*эпидемия* 'epidemic', *проблема* 'problem', *экономика* 'economy') was mentioned for many times.

For Trump, as shown in Appendix C, the most frequent thematic noun is always *people* from 2017 to 2021. The concept of nation and people (*country*, *America*, *United States*), in addition, is highlighted in 2017 and 2018. Apart from that, global and economic issues (*Korea*, *Afghanistan*, *trade*, *tax*) were emphasized by Trump in 2017 and 2018. In 2019, the concept of people remained to be concentrated by him while the intensity of the concept of nation and country decreased to some degree. In fact, during the 2020 and 2021, i.e., the 2020 US presidential election, Trump turned to topics serving his own political ends, which are essential for promoting himself, viz., the election (*ballot*, *election*, *vote*, *voter*). In contrast, the issue of pandemics (*vaccine*, *virus*) seems to be given less attention. His intensity revolved around the election, or more specifically, legal vs. illegal ballots, the issue he valued much more than the epidemic.

4 Conclusion

In sum, the present study explored the intensity of thematic concentration of Russian and American presidents using quantitative linguistics methods and qualitative analysis. Values of thematic concentration, secondary thematic concentration, and proportional thematic concentration of Putin's speeches are significantly or marginally significantly different from those of Medvedev's texts. All of Trump's three indicators are significantly greater than those of Obama. Diachronically, Putin's speeches contain more central themes in 2019 than in 2020-2021. By contrast, Trump remains a consistent tendency toward conveying a small number of themes in his communicative practice.

The quantitative-linguistic method, thematic concentration, employed in the current study may gain insight into the relationship between Trump and Putin and their predecessors, Obama and Medvedev, respectively, in terms of their choice of language. This also reflects the feasibility of combining the quantitative linguistic metric, thematic concentration in discourse analysis and stylistic studies. Further research on thematic words can be conducted, such as words, synonyms, and their references to a greater set (or list), usually called *hreb,* proposed by Ziegler and Altmann (2002).

5 Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and remarks. This study is supported by the Social Science Foundation of the Education Ministry of China (Grant No. 21YJC740060), the Social Science Foundation of the Education Ministry of China (Grant No. 19YJC740003), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (22JJD740018), the Philosophy and Social Science Planning Project of Guangdong Province (Grant No. GD23YWY02).

6 References

Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics* 22(3), pp. 215–232. https://doi.org/10.1080/09296174.2015.1037157.

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity* 48(2), pp.899–910. https://doi.org/10.1007/s11135-012-9811-3.

Čech, R. (2016). *Tematická koncentrace textu v češtině* [The thematic concentration of the text in Czech language]. Praha: ÚFAL.

Chen, R., Liu, H. (2015). Ideologies of supreme court justices: Quantitative thematic analysis of multiple opinions of "Bush V. Gore 2000." *Glottotheory* 6(2), pp. 299–322.

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.

Davis, C. M. (2020). *Presidential Authoritarianism in the United States and Russia During the Metamodern Era*. Fort Lauderdale, FL: Nova Southeastern University dissertation.

Fish, M. S. (2017). The Kremlin Emboldened: What Is Putinism? *Journal of Democracy* 28(4), pp. 61–75.

Glukhova, I., Sorokina, O. (2018). Linguistic implementation of Persuasive Strategies and Tactics in Political Public Communication (a case study of V. Putin's election campaign speeches). In: Uslu, F., Güçlü, T., Özdemir, M., Altan, K., Aslan, S. (Eds.). *Proceedings of INTCESS2018-5th International Conference on Education and Social Sciences, 5-7 February 2018, Istanbul, Turkey*.

Gorham, M. S. (2005). Putin's Language. *Ab Imperio* 2005(4), pp. 381–401.

Gorham, M. S. (2014). *After Newspeak: Language Culture and Politics in Russia from Gorbachev to Putin*. Ithaca, NY: Cornell University Press.

Hauser, M. (2018). Metapopulism in-between democracy and populism: tranformations of Laclau's concept of populism with Trump and Putin. *Distinktion: Journal of Social Theory* 19(1), pp. 68–87. https://doi.org/10.1080/1600910X.2018.1455599.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46), pp. 16569–16572. https://doi.org/10.1073/pnas.0507655102.

Homolar, A., Scholz, R. (2019). The power of Trump-speak: Populist crisis narratives and ontological security. *Cambridge Review of International Affairs* 32(3), pp. 344–364. https://doi.org/10.1080/09557571.2019.1575796.

Kubát, M., Čech, R. (2016). Quantitative Analysis of US Presidential Inaugural Addresses. *Glottometrics* 34, pp. 14–27.

MacWilliams, M. C. (2016). Who decides when the party doesn't? Authoritarian voters and the rise of Donald Trump. *PS: Political Science & Politics* 49(4), pp. 716–721. https://doi.org/10.1017/S1049096516001463.

Mercieca, J. (2020). *Demagogue for president: the rhetorical genius of Donald Trump*. College Station, TX: Texas A&M University Press.

Noriega, A. C. (2016). The Putin System: Russian Authoritarianism Today. *Revista Mexicana de Análisis Político y Administración Pública* 5(1), pp. 75–92.

Popescu, I. I., Altmann, G., Grzybek, P., Jayaram B. D., Köhler, R., Krupa, V., Macutek, J., Pustet, R., Uhlirova, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin & New York: De Gruyter Mouton.

Popescu, I. I. (2007). Text ranking by the weight of highly frequent words. In Peter Grzybek & Reinhard Köhler (Eds.). *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, pp. 555–566. Berlin & Boston: De Gruyter Mouton.

Popescu, I. I. (2009). Thematic concentration of the text. In: Popescu, I., Altmann, G., Grzybek, P., Jayaram B. D., Köhler, R., Krupa, V., Macutek, J., Pustet, R., Uhlirova, L., Vidya, M. N. (Eds.). *Word Frequency Studies*, pp. 95–100. Berlin & New York: De Gruyter Mouton.

Reyes, A., Ross, A. (2021). From the White House with anger: Conversational features in President Trump's official communication. *Language & Communication* 77, pp. 46–55. https://doi.org/10.1016/j.langcom.2020.12.003.

Rivers, D. J., Ross, A. S. (2020). Authority (de) legitimation in the border wall Twitter discourse of President Trump. *Journal of Language and Politics* 19(5), pp. 831–856. https://doi.org/10.1075/jlp.19105.riv.

Ross, A. S., Caldwell, D. (2020). 'Going negative': An APPRAISAL analysis of the rhetoric of Donald Trump on Twitter. *Language & Communication* 70, pp. 13–27. https://doi.org/10.1016/j.langcom.2019.09.003.

Savoy, J. (2018). Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities* 33(1), pp. 143–159. https://doi.org/10.1093/llc/fqx007.

Sedykh, A. P. (2016). К вопросу об идиополитическом дискурсе В.В. Путина [On the problem of the idiopolitical discourse of V.V. Putin in the Russian language]. *Political linguistics* 1(55), pp. 35–41.

Van Dijk, T. A. (2008). *Discourse and power*. New York: Macmillan International Higher Education.

Wang, Y., Liu, H. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse & Society* 29(3), pp. 299–323. https://doi.org/10.1177/0957926517734659.

Yakoba, I. A. (2017). Деконструкция дискурса дональда Трампа (на примере его предвыборных выступлений 2016 г.) [Deconstruction of Donald Trump discourse (on example of his pre-election speeches in 2016) in the Russian language]. *Discourse-Pi* 14(1), pp. 164–170.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse: ein textlinguistisches Arbeitsbuch*. Vienna: Praesens.

7 Appendix

Due to limited space, appendix can be found at https://osf.io/xf38v?view_only=68b2f6e335aa4dd8b661209e2e29a889

Direct and indirect evidence of compression of word lengths. Zipf's law of abbreviation revisited

Sonia Petrini[1] (0000-0002-0514-6223), Antoni Casas-i-Muñoz[2] (0000-0001-5690-316X), Jordi Cluet-i-Martinell[2] (0000-0003-4188-6728), Mengxue Wang[2] (0000-0002-8262-9333), Christian Bentz[3] (0000-0001-6570-9326), Ramon Ferrer-i-Cancho[1*] (0000-0002-7820-923X)

[1] Quantitative, Mathematical and Computational Linguistics Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.
[2] Universitat Politècnica de Catalunya (UPC), Barcelona School of Informatics, Barcelona, Catalonia, Spain.
[3] Department of Linguistics, University of Tübingen, Tübingen, Germany.
[*] Corresponding author's email: rferrericancho@cs.upc.edu

DOI: https://doi.org/10.53482/2023_54_407

Abstract

Zipf's law of abbreviation, the tendency of more frequent words to be shorter, is one of the most solid candidates for a linguistic universal, in the sense that it has the potential for being exceptionless or with a number of exceptions that is vanishingly small compared to the number of languages on Earth. Since Zipf's pioneering research, this law has been viewed as a manifestation of a universal principle of communication, i.e. the minimization of word lengths, to reduce the effort of communication. Here we revisit the concordance of written language with the law of abbreviation. Crucially, we provide wider evidence that the law holds also in speech (when word length is measured in time), in particular in 46 languages from 14 linguistic families. Agreement with the law of abbreviation provides indirect evidence of compression of languages via the theoretical argument that the law of abbreviation is a prediction of optimal coding. Motivated by the need of direct evidence of compression, we derive a simple formula for a random baseline indicating that word lengths are systematically below chance, across linguistic families and writing systems, and independently of the unit of measurement (length in characters or duration in time). Our work paves the way to measure and compare the degree of optimality of word lengths in languages.

Keywords: word length, compression, law of abbreviation

1 Introduction

It has been argued that linguistic universals are a myth (Evans and Levinson, 2009), but this neglects the statistical regularities that the quantitative linguistic community has been investigating for many decades. A salient case is Zipf's law of abbreviation, the tendency of more frequent words to be shorter

(Zipf, 1949). It holds across language families (Bentz and Ferrer-i-Cancho, 2016; Koplenig et al., 2022; Levshina, 2022; Meylan and Griffiths, 2021; Piantadosi et al., 2011), writing systems (Sanada, 2008; Wang and Chen, 2015) and modalities (Börstell et al., 2016; Hernández-Fernández and Torre, 2022; Torre et al., 2019), and also when word length in characters is replaced by word duration in time (Hernández-Fernández et al., 2019). Furthermore, the number of species where a parallel of this law has been confirmed in animal communication is growing over time (Semple et al., 2022).[1] In language sciences, research on the law of abbreviation in languages measures word length in discrete units (e.g., characters) whereas, in biology, research on the law in other species typically uses duration in time. Here, we aim to reduce the gulf that separates these two traditions by promoting research on the law of abbreviation on word durations.

G. K. Zipf believed that the law of abbreviation constituted *indirect* evidence of the minimization of the cost of using words (Zipf, 1949). At present, Zipf's view is supported by standard information theory and its extensions: the main argument is that the minimization of L, the mean word length, that is indeed a simplification of Zipf's cost function,[2] leads to the law of abbreviation (Ferrer-i-Cancho et al., 2019; Ferrer-i-Cancho et al., 2013). Using the terminology of information theory, the minimization of mean word length is known as compression. Using the terminology of quantitative linguistics, L is the average length of tokens from a repertoire of n types, that is defined as

$$L = \sum_{i=1}^{n} p_i l_i, \tag{1}$$

where p_i and l_i are, respectively, the probability and the length of the i-th type. In practical applications, L is calculated replacing p_i by the relative frequency of a type, that is

$$p_i = f_i/T,$$

where f_i is the absolute frequency of a type and T is the total number of tokens, i.e.

$$T = \sum_{i=1}^{n} f_i.$$

This leads to a definition of L that is

$$L = \frac{1}{T}\sum_{i=1}^{n} f_i l_i.$$

At present, the mathematical link between the law of abbreviation and compression has been established under the assumption that words are coded optimaly so as to minimize L. If words are coded optimaly, the correlation between the frequency of a word and its duration cannot be positive (Ferrer-i-Cancho

[1] The interested reader can check the latest discoveries on this law in "Bibliography on laws of language outside human language" at https://cqllab.upc.edu/biblio/laws/.

[2] He referred to the cost function as "minimum equation" (Zipf, 1949).

et al., 2019). Thus, a lack of correlation between the frequency of a word and its duration does not imply absence of compression. Furthermore, it is not a warranted assumption that languages code words optimaly. Therefore, an approach to find *direct* evidence of compression getting rid of the assumption of optimal coding is required.

As a first approach, one could compare the value of L of a language against L_{max}, the maximum value that L could achieve in this language. The larger the gap between L and L_{max}, the higher the level of compression in the language. However, the problem is that L_{max} can be infinite *a priori*. To fix that problem, one could restrict L_{max} to be finite but then this raises the question of what should be the finite value of L_{max} and why. For these reasons, here we resort to the notion of random baseline, that here is defined assuming some random mapping of word types into strings. In previous research, the random baseline was defined by the average word length resulting from a shuffling of the current length/duration of types so as to check if L was smaller than expected by chance in that random mapping (Ferrer-i-Cancho et al., 2013; Heesen et al., 2019). Critically, an exact method to compute the random baseline, namely the expected word length in these shufflings, is missing.

The remainder of the article is organized as follows. In Section 2, we introduce the definition of L_r, the random baseline, that we will use to explore direct evidence of compression. In particular, we derive a simple formula for L_r that will simplify future research on compression in natural communication systems. In Section 3 and Section 4, we present, respectively, the materials and methods that will be used to provide further evidence of compression and the law of abbreviation in real languages with emphasis on word durations. In Section 4, we present a new unsupervised method to exclude words with foreign characters in line with good practices for research on linguistic laws and communicative efficiency (Meylan and Griffiths, 2021). In Section 5, we show that the law of abbreviation holds without exceptions in a wide sample of languages, independently of the unit of measurement of word length, namely characters or duration in time, providing further indirect evidence of compression in languages. In addition, the random baseline indicates that word lengths are systematically below chance, across linguistic families and writing systems, independently of the unit of measurement (length in characters or duration in time), providing direct evidence of compression. Finally, in Section 6, we discuss the findings in relation to the potential universality of the law of abbreviation and the universality of compression in languages. We also make proposals for future research.

2 A random baseline revisited

In our statistical setting, the null hypothesis states that compression (minimization of word lengths) has no effect on word lengths. The alternative hypothesis states that compression has an effect on word lengths as Zipf hypothesized. If the null hypothesis is rejected then word lengths are shorter than expected by chance.

Table 1: Matrix indicating the frequency and length of three types. The mean type length is $L = \frac{235}{125} = 1.88$.

i	f_i	l_i
1	100	2
2	20	1
3	5	3

Consider a matrix with two columns, f_i and l_i, that are used to compute the average word length L. The matrix in Table 1 gives $L = \frac{235}{125} = 1.88$. We consider the null hypothesis of a random mapping of probabilities into lengths, namely that the ordering of the f_i's or the l_i's in Table 1 is arbitrary and results from a random shuffling of one of these variables or both. We use f_i', l_i' and p_i' for the new values of f_i, l_i and p_i that result from one of these shufflings.

This null hypothesis was introduced in research on compression in human language and animal communication to test if L is significantly small using a permutation test (Ferrer-i-Cancho et al., 2013; Heesen et al., 2019). Later, it was used to estimate the degree of optimality of word lengths (Moreno Fernández, 2021; Pimentel et al., 2021). Our new contribution here is a precise mathematical characterization of the null hypothesis and the derivation of a simple formula the expected word length.

In the context of computing average word length, the matrix in Table 1 is equivalent to a matrix where the column f_i is replaced by a column with p_i thanks to

$$p_i' = \frac{f_i'}{T}.$$

Indeed, the null hypothesis has three variants

1. Single column shuffling. Only the column of f_i or p_i is shuffled.

2. Single column shuffling. Only the column of l_i is shuffled.

3. Dual column shuffling. The column of f_i or p_i and the column of l_i are both shuffled.

In each of the variants, all random shufflings of a specific column are equally likely. In case of dual shuffling, the shuffling of one column is independent of the shuffling of the other column. The outcome of a dual shuffling on Table 1 is shown in Table 2.

The random baseline, L_r, is the expected value of L under the null hypothesis.[3] L_r can be defined in more detail in two main equivalent ways:

[3] Notice that L is indeed the expected value of the length of a token but under a distinct setting (a distinct null hypothesis), where one picks a token uniformly at random over all tokens of a text and looks at its length.

Table 2: Matrix indicating the frequency and length of three types. The mean type length is $L = \frac{345}{125} = 2.76$.

i	f'_i	l'_i
1	20	2
2	100	3
3	5	1

1. The value of L that is expected if L is recomputed after pairing the f_i's and the l_i's at random and recomputing L. The new value of L depends on the variant of the null hypothesis. When shuffling the column for f_i in the matrix (Table 1), the new L is

$$L' = \frac{1}{T}\sum_{i=1}^{n} f'_i l_i.$$

When shuffling the column for l_i and recomputing L, the new L is

$$L' = \frac{1}{T}\sum_{i=1}^{n} f_i l'_i.$$

When shuffling both columns, the new L is

$$L' = \frac{1}{T}\sum_{i=1}^{n} f'_i l'_i.$$

2. The average value of L that is expected over all possible shufflings in one of the variants of the null hypothesis. In the example in Table 3, on shuffling only the l_i column,

$$L_r = \frac{\frac{155}{125} + \frac{170}{125} + \frac{235}{125} + \frac{265}{125} + \frac{330}{125} + \frac{345}{125}}{6} = \frac{155 + 170 + 235 + 265 + 330 + 345}{125 \cdot 6} = 2.$$

We use $\mathbb{E}[X]$ to refer to the expected value of a random variable X under some variant of the null hypothesis above. Then

$$L_r = \mathbb{E}[L'],$$

where L' is the value of L resulting from some shuffling.

In quantitative linguistics, the mean length of tokens (L) is also known as dynamic word length (Chen et al., 2015) and corresponds to the mean length of the words in a text. The mean length of types (M), defined as

$$M = \frac{1}{n}\sum_{i=1}^{n} l_i,$$

is also known as the static word length and corresponds to average length of the headwords in a dictionary (Chen et al., 2015). Interestingly, the following property states that L_r turns out to be M independently of the variant of the null hypothesis under consideration.

Property 2.1. *The expected value of L' under any variant of the null hypothesis is $L_r = M$.*

Proof. We analyze $\mathbb{E}[L']$ under each of the variants of the null hypothesis.

Dual shuffling. Applying the linearity of expectation and independence between the shuffling of the p_i column of the that of the l_i column, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p'_i l'_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i l'_i] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i]\,\mathbb{E}[l'_i].\end{aligned}$$

Noting that

$$\begin{aligned}\mathbb{E}[p'_i] &= \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n} \\ \mathbb{E}[l'_i] &= \frac{1}{n}\sum_{i=1}^{m} l_i = M,\end{aligned}$$

we finally obtain

(2)
$$\mathbb{E}[L'] = \sum_{i=1}^{n} \frac{M}{n} = M.$$

Single shuffling of the l_i column. Applying the linearity of expectation and the fact that the column of p_i remains constant, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p_i l'_i\right] \\ &= \sum_{i=1}^{n} p_i\,\mathbb{E}[l'_i].\end{aligned}$$

Recalling $\mathbb{E}[l'_i] = M$, we finally obtain

(3)
$$\mathbb{E}[L'] = M\sum_{i=1}^{n} p_i = M.$$

Single shuffling of the p_i column. Applying the linearity of expectation and the fact that the column of l_i remains constant, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p'_i l_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i] l_i.\end{aligned}$$

Recalling $\mathbb{E}[p'_i] = \frac{1}{n}$, we finally obtain

(4)
$$\mathbb{E}[L'] = \frac{1}{n}\sum_{i=1}^{n} l_i = M.$$

□

Table 3: All the $3! = 6$ permutations of the column l_i in Table 1 that can be produced. Each permutation is indicated with letters from A to F. L', the mean length of types in a shuffling, is shown at the bottom for each permutation.

		A	B	C	D	E	F
i	f_i	l'_i	l'_i	l'_i	l'_i	l'_i	l'_i
1	100	1	1	2	2	3	3
2	20	2	3	1	3	1	2
3	5	3	2	3	1	2	1
	L'	$\frac{155}{125} = 1.24$	$\frac{170}{125} = 1.36$	$\frac{235}{125} = 1.88$	$\frac{265}{125} = 2.12$	$\frac{330}{125} = 2.64$	$\frac{345}{125} = 2.76$

The previous finding indicates that the random baseline for L is equivalent to assuming that all word types are equally likely, namely, replacing each p_i by $1/n$.

3 Material

3.1 General information about corpora and languages

We investigate the relationship between the frequency of a word and its length in languages from two collections: Common Voice Forced Alignments (Section 3.2.1), hereafter CV, and Parallel Universal Dependencies (Section 3.2.2), hereafter PUD.

All the preprocessed files used to produce the results from the original collections are available in the repository of the article.[4]

PUD comprises 20 distinct languages from 7 linguistic families and 8 scripts (Table 4). CV comprises 46 languages from 14 linguistic families (we include 'Conlang', i.e. 'constructed languages', as a family for Esperanto and Interlingua) and 10 scripts (Table 5). Both PUD and CV are biased towards the Indo-European family and the Latin script. The typological information (language family) is obtained from Glottolog 4.6[5]. The writing systems are determined according to ISO-15924 codes[6]. In Table 4 and Table 5, we show the scripts using their standard English names. For example, most languages from the Indo-European family are written in Latin scripts. We also categorize Chinese Pinyin and Japanese Romaji as Latin scripts.

[4] In the *data* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.

[5] https://glottolog.org/

[6] https://unicode.org/iso15924/iso15924-codes.html

Table 4: Summary of the main characteristics of the languages in the PUD collection. For each language, we show the linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters: A, the observed alphabet size (number of distinct characters), n, the number of word types, and T, the number of word tokens.

Language	Family	Script	A	n	T
Arabic	Afro-Asiatic	Arabic	20	3309	11667
Indonesian	Austronesian	Latin	23	4501	16702
Russian	Indo-European	Cyrillic	23	4666	11749
Hindi	Indo-European	Devanagari	44	4343	20071
Czech	Indo-European	Latin	33	7073	15331
English	Indo-European	Latin	25	5001	18028
French	Indo-European	Latin	26	5214	20407
German	Indo-European	Latin	28	6116	18331
Icelandic	Indo-European	Latin	32	6035	16209
Italian	Indo-European	Latin	24	5606	21266
Polish	Indo-European	Latin	31	7188	15191
Portuguese	Indo-European	Latin	38	5661	21855
Spanish	Indo-European	Latin	32	5750	21067
Swedish	Indo-European	Latin	25	5624	16378
Japanese	Japonic	Japanese	1549	4852	24737
Japanese-strokes	Japonic	Japanese	1549	4852	24737
Japanese-romaji	Japonic	Latin	24	4849	24734
Korean	Koreanic	Hangul	379	6218	12307
Thai	Kra-Dai	Thai	50	3573	20860
Chinese	Sino-Tibetan	Han (Traditional variant)	2038	4970	17845
Chinese-strokes	Sino-Tibetan	Han (Traditional variant)	2038	4970	17845
Chinese-pinyin	Sino-Tibetan	Latin	50	4970	17845
Turkish	Turkic	Latin	28	6587	13799
Finnish	Uralic	Latin	24	6938	12701

Table 5: Summary of the main characteristics of the languages in the CV collection. For every language we show its linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters: A, the observed alphabet size (number of distinct characters), n, the number of word types, and, T, the number of word tokens. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense as Conlang languages are not related in the common linguistic family sense.

Language	Family	Script	A	n	T
Arabic	Afro-Asiatic	Arabic	31	6397	45825
Maltese	Afro-Asiatic	Latin	31	8058	44112
Vietnamese	Austroasiatic	Latin	41	370	938
Indonesian	Austronesian	Latin	22	3768	44210
Esperanto	Conlang	Latin	27	27759	406261
Interlingua	Conlang	Latin	20	5126	30504
Tamil	Dravidian	Tamil	29	1210	6439
Persian	Indo-European	Arabic	38	13115	1662508
Assamese	Indo-European	Assamese	43	971	1813
Russian	Indo-European	Cyrillic	32	31827	637686
Ukrainian	Indo-European	Cyrillic	34	14337	120760
Panjabi	Indo-European	Devanagari	37	84	98
Modern Greek	Indo-European	Greek	33	5813	37880
Breton	Indo-European	Latin	28	4228	38237
Catalan	Indo-European	Latin	39	79112	3294206
Czech	Indo-European	Latin	33	15518	147582
Dutch	Indo-European	Latin	23	10225	316498
English	Indo-European	Latin	28	173023	9828713
French	Indo-European	Latin	49	160243	3729370
German	Indo-European	Latin	30	148436	4230565
Irish	Indo-European	Latin	23	2251	22593
Italian	Indo-European	Latin	34	54996	811783
Latvian	Indo-European	Latin	27	7251	29456
Polish	Indo-European	Latin	32	25340	595411
Portuguese	Indo-European	Latin	27	11509	283048
Romanian	Indo-European	Latin	29	6423	33341
Romansh	Indo-European	Latin	26	9614	43792
Slovenian	Indo-European	Latin	24	5937	26304
Spanish	Indo-European	Latin	33	75010	1842474
Swedish	Indo-European	Latin	25	4371	62951
Welsh	Indo-European	Latin	22	11143	539621
Western Frisian	Indo-European	Latin	30	8383	63073
Oriya	Indo-European	Odia	41	764	1700
Dhivehi	Indo-European	Thaana	27	111	1284
Georgian	Kartvelian	Georgian	25	6505	12958
Basque	Language isolate	Latin	21	24748	458071
Mongolian	Mongolic	Mongolian	31	14608	70217
Kinyarwanda	Niger-Congo	Latin	26	133815	1939810
Abkhazian	Northwest Caucasian	Cyrillic	28	119	156
Hakha Chin	Sino-Tibetan	Latin	23	2499	17776
Chuvash	Turkic	Cyrillic	22	4311	13583
Kirghiz	Turkic	Cyrillic	30	10130	61844
Tatar	Turkic	Cyrillic	34	21823	144356
Yakut	Turkic	Cyrillic	28	7904	22577
Turkish	Turkic	Latin	31	8926	107686
Estonian	Uralic	Latin	23	28691	121549

3.2 The datasets

We measure word length in two main ways: *duration in time* and *length in characters*. Concerning Chinese and Japanese, we additionally consider the number of strokes and the number of characters of their romanization (i.e. Pinyin for Chinese and Romaji for Japanese).

Given these datasets, word durations are obtained only from CV. Word lengths in characters are obtained from both CV as well as from PUD. Word lengths in strokes, and word lengths in characters after romanization, are obtained only from PUD.

3.2.1 Common Voice Forced Alignments

The Common Voice Corpus[7] is an open source dataset of recorded voices uttering sentences in many different languages. The amount of data, as well as the source and topic of each sentence, depends considerably on the language and the corpus version. Specifically, the Common Voice Corpus 5.1 contains information on 54 languages and dialects.

Common Voice Forced Alignments (CVFA)[8] were created by Josh Meyer using the Montreal Forced Aligner[9] on top of the Common Voice Corpus 5.1. Kabyle, Upper Sorbian and Votic were left out of the alignments for an undocumented reason. Therefore, CVFA contains information on 51 languages.

In our analyses, Japanese and the three Chinese dialects were excluded as the forced aligner failed to correctly extract words from sentences. In addition, both Romansh dialects were fused into a single Romansh language. Indeed, given the nature of this corpus, all languages are likely to be represented by more than one dialect.

Notice that Abkhazian, Panjabi, and Vietnamese have a critically low number of tokens ($T < 1000$ in Table 5). However, we decided to include them in the analyses so as to understand their limitations related to corpus size.

3.2.2 Parallel Universal Dependencies

The Universal Dependencies (UD)[10] collection is an open source dataset of annotated sentences, in which the amount of data depends on each language. The Parallel Universal Dependencies (PUD) collection is a parallel subset of 20 languages from the UD collection, consisting of 1000 sentences. It allows for a cross-language comparison, controlling for content and annotation style.

In Table 4, we show the characteristics of the languages in PUD. For traditional Chinese and Japanese,

[7] https://commonvoice.mozilla.org/en/datasets

[8] https://github.com/JRMeyer/common-voice-forced-alignments

[9] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[10] https://universaldependencies.org/

we also include word lengths in romanizations (Pinyin and Romaji respectively), as well as word lengths measured in strokes, resulting in a total of 24 language files. Notice that three Japanese words that are hapax legomena could not be romanized and thus the number of tokens and types varies slightly with respect to the original Japanese characters (Table 4).

4 Methodology

All the code used to produce the results is available in the repository of the article.[11]

4.1 The units of length

4.1.1 Duration

The duration of a word for a given language is estimated by computing the median duration in seconds across all its occurrences in utterances in the CV corpus. All words with equal orthographic form are assumed to be the same type. The median is preferred over the mean as it is less sensitive to outliers (that may be produced by forced alignment errors) and better suited to deal with heavy-tailed distributions (Hernández-Fernández et al., 2019). Given the oral nature of the data, we do expect to observe some variation in the duration of words, due to differences between individuals, and variation within a single individual. This is more generally in line with speakers acting as complex dynamical systems (Kello et al., 2010). For these reasons, median duration is preferred for research on the law of abbreviation in acoustic units (Torre et al., 2019; Watson et al., 2020).

4.1.2 Length in characters

Word length in characters is measured by counting every Unicode UTF-8 character present in a word. Special characters such as "=" were removed. Characters with stress accents are considered as different from their non-stressed counterpart (e.g. "a" and "à" are considered separate characters). Following best practices from (Meylan and Griffiths, 2021), characters were always kept in UTF-8.

4.1.3 Length in strokes

Japanese Kanji and Chinese Hanzi were turned into strokes using the *cihai* Python library.[12] In Japanese characters other than Kanji, namely Japanese Kana, the number of strokes in printed versus hand-written modality can differ (Chinese Hanzi and Japanese Kanji have the same number of strokes in printed version or hand-written version). Here we counted the number of strokes in printed form. Japanese Kana were converted into printed strokes by using a hand-crafted correspondence table, since Kana is not part

[11] In the *code* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.

[12] https://github.com/cihai/cihai

of the CJK unified character system. This table was created by us and checked by a native linguist (S. Komori from Chubu University, Japan). It is available in the repository of the article.[13]

In case of discrepancies on the number of strokes for a given character, the most typical printed version was chosen.

4.1.4 Length in Pinyin and Romaji

Chinese Pinyin was obtained using the *cihai* package as above, while the Japanese Romaji was obtained with the *cutlet* Python library.[14] The latter uses Kunrei-shiki romanization (since it is the one used officially by the government of Japan) and the spelling of foreign words is obtained in its native reading (e.g. "カレー" is romanized as "karee" instead of "curry"). There are some particularities with the romanization of Kanji characters by *cutlet*. For example, in the case of the word "year" (年), it chose the reading of "Nen" instead of "Tosi", which would be the expected one.

A more systematic issue with Japanese romanization is that it does not provide means to indicate pitch accents, which are implicitly present in Kanji. For example, "日本" "Ni↑hon" ("Japan") is romanized as simply "Nihon". Therefore, the alphabet size of romanized Japanese is smaller than it should be, compared to other languages where, as stated before, stress accents are counted as distinctive features of characters.

4.2 Tokenization

Tokenization is already given in each dataset and we borrow it for our analyses. Thus tokenization methods are not uniform for CV and PUD and are not guaranteed to be uniform among languages even within each of these datasets.

4.3 Filtering of tokens

Examining our datasets, we noticed that in some text files there was a considerable number of unusual character strings, as well as foreign words (written in different scripts). These need to be filtered out in order to obtain a "clean" set of word types. To this end we filter out tokens following a two step procedure:

1. *Mandatory elementary filtering*. This filter consists of:

    - *Common filtering*. In essence, it consists of the original tokenization and the removal of tokens containing digits. In each collection, the original tokenizer yields tokens that may contain certain punctuation marks. Due to the nature of the CV dataset, the bulk of punctuation was

[13] In the *data/other* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.

[14] https://github.com/polm/cutlet

already removed via the Montreal Forced Aligner with some exceptions. For instance, single quotation (in particular “’”) is a punctuation sign that is kept within a word token in CV, as it is necessary for the formation of clitics in multiple languages, such as in English or French. In PUD, as a part of UD, contractions are split into two word types. “can’t” is split into “ca” “n’t” (in CV “can’t” would remain as just one token). In both collections, words containing ASCII digits are removed because they do not reflect phonemic length and can be seen as another writing system.

- *Specific filtering*. In case of the PUD collection, we excluded all tokens with Part-of-Speech (POS) tag ‘PUNCT’. In case of the CV collection, we removed tokens tagged as <unk> or null tokens, namely tokens that either could not be read or that represent pauses.
- *Lowercasing*. Every character is lowercased. In the case of CV, this is already given by the Montreal Forced Aligner, while in the case of PUD, tokens are lowercased by means of the *spaCy* Python package.[15]

2. *Optional filtering*. This is a new method that is applied after the previous filter and described in Section 4.4.

4.4 A new method to filter out unusual characters

It has been pointed out that “chunk” words and loanwords can distort the results of quantitative analyses of word lengths (Meylan and Griffiths, 2021). Indeed, especially the files of the Common Voice Corpus feature a considerable number of word tokens which do not consist of characters belonging to the primary alphabet of the respective writing system. Meylan and Griffiths (2021) proposed to use dictionaries to exclude such anomalous words. However, this is not feasible for our multilingual datasets, as loanword dictionaries are not available for this large number of diverse languages (Table 4 and Table 5). The Intercontinental Dictionary Series,[16] for example, contains only around half of the languages in our analysis, so it is not applicable to many of them. Hence, this approach would lead to a non-uniform treatment of different languages and texts. Selecting a matched set of semantic concepts across languages using a lexical database is also infeasible due to similar reasons.

Against this backdrop, we decided to develop an unsupervised method to filter out words which contain highly unusual characters. For a given language, the method starts by assuming that the strings (after the mandatory filtering illustrated above) contain characters of two types: characters of the working/primary alphabet as well as other characters. We hypothesize that the latter are much less frequent than the former.

[15] https://spacy.io/

[16] https://ids.clld.org/

Following this rationale, we apply the k-means algorithm of the *Ckmeans R* package[17] to split the set of characters into the two groups based on the logarithm of the frequency of the characters.[18] To maximize the power of the clustering method, we use the exact method with $k = 2$ for one dimension instead of the customary approximate method. We then keep the high frequency cluster as the real working alphabet and filter out the word tokens that contain characters not belonging to this high frequency cluster.

We illustrate the power of the method by showing working alphabets that are obtained on CV, that is the noisiest one of the collections.

In English, the working alphabet is defined by the 26 English letters and quotation marks (“’”, “’”). These quotation marks are used often in clitics, and as such are correctly identified as part of the encoding, since, for example, “can’t” and “cant” are different words in meaning, with “can’t” meaning “can not”, while “cant” is a statement on a religious or moral subject that is not believed by the person making the statement, with the differentiating feature being the “’”. Therefore, the working alphabet becomes 5 vowels (“a”, “e”, “i”, “o”, “u”), 21 consonants (“b”, “c”, “d”, “f”, “g”, “h”, “j”, “k”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “v”, “w”, “x”, “y”, “z”) and 2 kinds of quotation marks (“’”, “’”).

In Russian, the working alphabet comprises 9 vowels (“а”, “о”, “у”, “ы”, “э”, “я”, “ю”, “и”, “е”), a semivowel / consonant “й”, 20 consonants (“б”, “в”, “г”, “д”, “ж”, “з”, “к”, “л”, “м”, “н”, “п”, “р”, “с”, “т”, “ф”, “х”, “ц”, “ч”, “ш”, “щ”) and 2 modifier letters (“ъ”, “ь”).

In Italian, it comprises 5 vowels (“a”, “e”, “i”, “o”, “u”), 21 consonants (“b”, “c”, “d”, “f”, “g”, “h”, “j”, “k”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “v”, “w”, “x” , “y”, “z”) and 6 instances of the 5 vowels containing a diacritic mark (“à”, “è”, “é”, “ì”, “ò”, “ù”).

The unsupervised filter method filter is not applied to Chinese, Japanese and Korean as, given their nature, this would exclude letters that actually belong to the real alphabet. In Section B.1 we analyze the impact of the optional filter and provide arguments for not applying the unsupervised filter to these languages. As a compensation, strings that contain non-CJK characters are filtered out in Chinese and Japanese as a part of the optional filter. In Korean, only a few characters are not proper Hangul and thus such a complementary filtering is not necessary.

[17] https://cran.r-project.org/web/packages/Ckmeans.1d.dp/index.html

[18] The motivation for taking logarithms of frequencies is three-fold: First, this brings observations closer together. Note that the k-means algorithm prefers high-density areas. Second, this transforms the frequencies into a measure of surprisal, following standard information theory (Shannon, 1948). Third, manual inspection suggests that the logarithmic transformation is required to produce an accurate split.

4.5 Immediate constituents in writing systems

When measuring word length in written languages, we are using *immediate constituents* of written words. In Romance languages, the immediate constituents are letters of the alphabet, which are a proxy for phonemes. For syllabic writing systems (as Chinese in our dataset), these are characters that correspond to syllables. In addition, for Chinese and Japanese, we are considering two other possible units for word length, which are not immediate constituents, but alternative ways of measuring word lengths which could provide useful insights: strokes and letters in Latin script romanizations. That means that for each of these languages words are unfolded into three systems, one for each unit of encoding (original characters, strokes, romanized letters/characters). In the hierarchy from words to other units, only the original characters are immediate constituents.

4.6 Statistical testing

4.6.1 Correlation

When measuring the association between two variables, we use both Pearson correlation and Kendall correlation (Conover, 1999). Note that the traditional view of Pearson correlation as a measure of linear association and thus not suitable for non-linear association has been challenged (van den Heuvel and Zhan, 2022).

4.6.2 How to test for the law of abbreviation

We used a left-sided correlation test to verify the presence of the law of abbreviation. In a purely exploratory or atheoretic exploration, one should use a two-sided test. In an exploration guided by theory, namely regarding the law of abbreviation as a manifestation of compression, the test should be left-sided as theory predicts that $\tau(p, l)$ cannot be positive in case of optimal coding (Ferrer-i-Cancho et al., 2019).

4.6.3 How to test for compression

In the context of the null hypothesis of a random mapping of type probabilities into type lengths, testing that compression (minimization of L) has some effect on actual word lengths is easy because L is a linear function of r, the Pearson correlation between word length and word probability (Appendix A). In particular,

$$L = ar + L_r,$$

where $a = (n-1)s_p s_l$, being n the number of types and s_p and s_l, respectively,the standard deviation of type probabilities and type lengths. In such random mappings, L_r, s_p and s_l remain constant and then testing if r is significantly small is equivalent to testing if L is significantly small (notice $a \geq 0$).

4.6.4 Controlling for multiple testing

When performing multiple correlation tests at the same time, it becomes easier to reject the null hypothesis simply by chance. To address this problem we used a Holm-Bonferroni correction to p-values.[19] We applied the correction when checking the law of abbreviation in the languages of a collection, so as to exclude the possibility that the law of abbreviation is found many times simply because we are testing it in many languages.

5 Results

In Section 1, we highlighted the importance of distinguishing between direct and indirect evidence of compression. Against this theoretical backdrop, here we first investigate the presence of Zipf's law of abbreviation in languages. Then we investigate direct evidence of compression with the help of the new random baseline.

5.1 The law of abbreviation revisited

We investigate the presence of the law of abbreviation by means of left-sided correlation tests for the association between frequency and length. We use both Kendall correlation, as suggested by theory on the origins of the law (Ferrer-i-Cancho et al., 2019), and Pearson's. For each language, we show the significance level of the relationship, color-coded by the value of the correlation coefficient. Figure 1 (a,b) indicates that the law holds in all languages – regardless of the definition of word length – when Kendall τ correlation is used. In both collections, we find Kendall τ correlation coefficients significant at the 99% confidence level, except for Dhivehi in the CV collection when length is measured in characters, and Abkhazian, Dhivehi, Panjabi and Vietnamese when length is measured in duration. However, note that these are all still significant at the 95% confidence level. When Pearson correlation is used instead, Figure 1 (c) shows that the picture remains the same in PUD. The main findings are the same also in CV (Figure 1 (d)), but when length is measured in duration Panjabi ceases to be significant at the 95% confidence level. Overall, we only fail to find the law of abbreviation in Panjabi given word durations, and using Pearson correlation. This is most probably related to undersampling, as this particular language only features 98 tokens (Table 5).

[19] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html

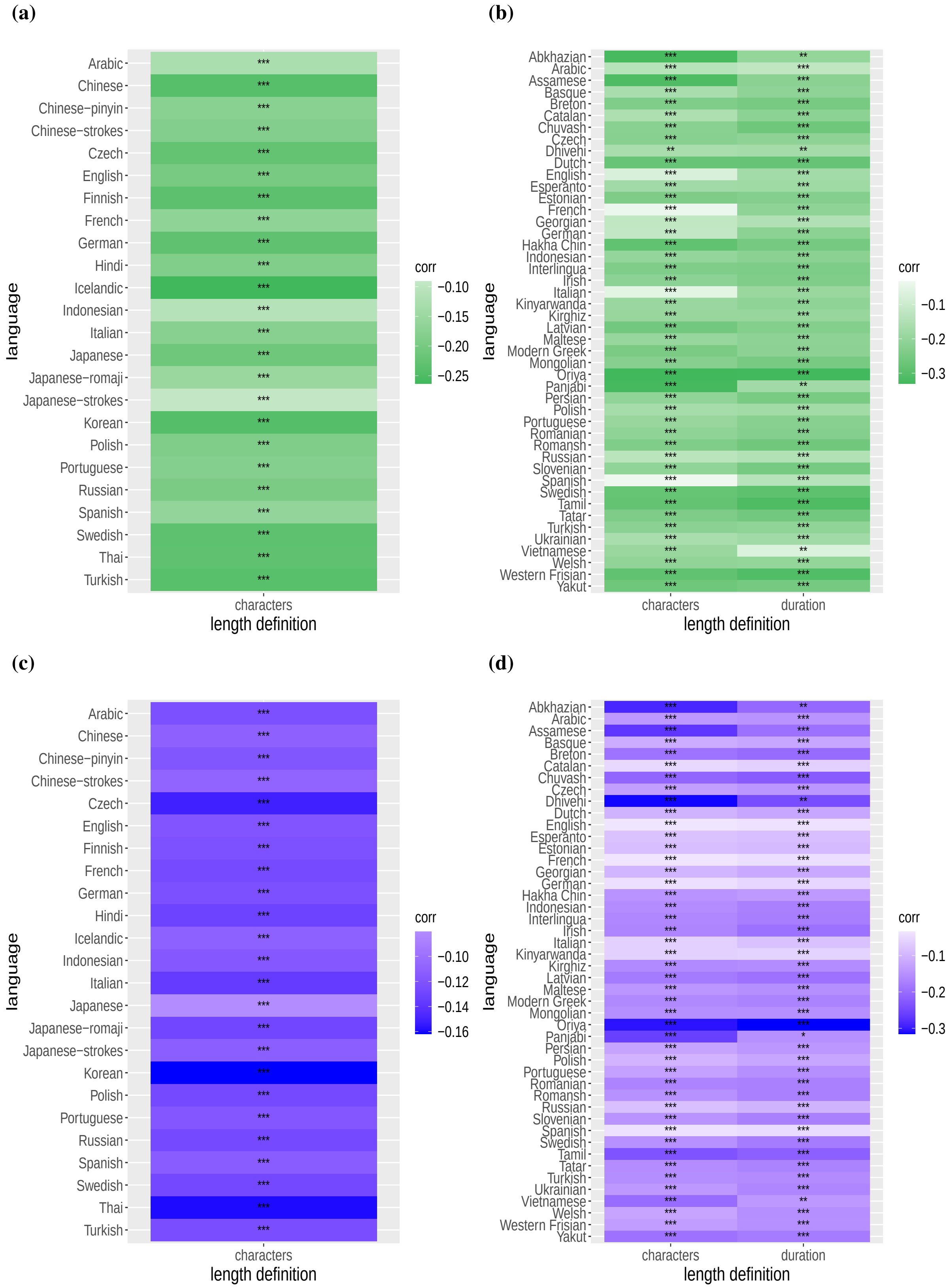

Figure 1: The correlation between frequency and length across languages. '***' indicates a Holm-Bonferroni corrected p-value lower than or equal to 0.01, '**' indicates lower than or equal to 0.05 but smaller than 0.1 and '*' indicates lower than or equal to 0.1. Here '*' symbols are not used to indicate significance but p-value ranges. (a) Kendall τ correlation in PUD (word length in characters). (b) Kendall τ correlation in CV (left: word length in characters; right: word length in duration). (c) Same as (a) with Pearson r correlation. (d) Same as (b) with Pearson r correlation.

5.2 Real word lengths versus the random baseline

We investigate the relationship between the actual mean word length (L) and the random baseline (L_r). We find that $L < L_r$ for all languages in every collection (Figure 2 and Tables B3, B4, B5). Interestingly, there is a large gap between L and L_r in the majority of languages, which is more compelling in CV with word durations (Figure 2). Exceptions to the large gap – as in the case of Panjabi and Abkhazian when length is measured in duration – mainly concern languages with reduced sample sizes. The result holds even when alternative units of measurement are considered for Chinese and Japanese.

Figure 2 is reminiscent of Figure 4 of Pimentel et al. (2021) but our setting is much simpler (it only involves L and L_r).

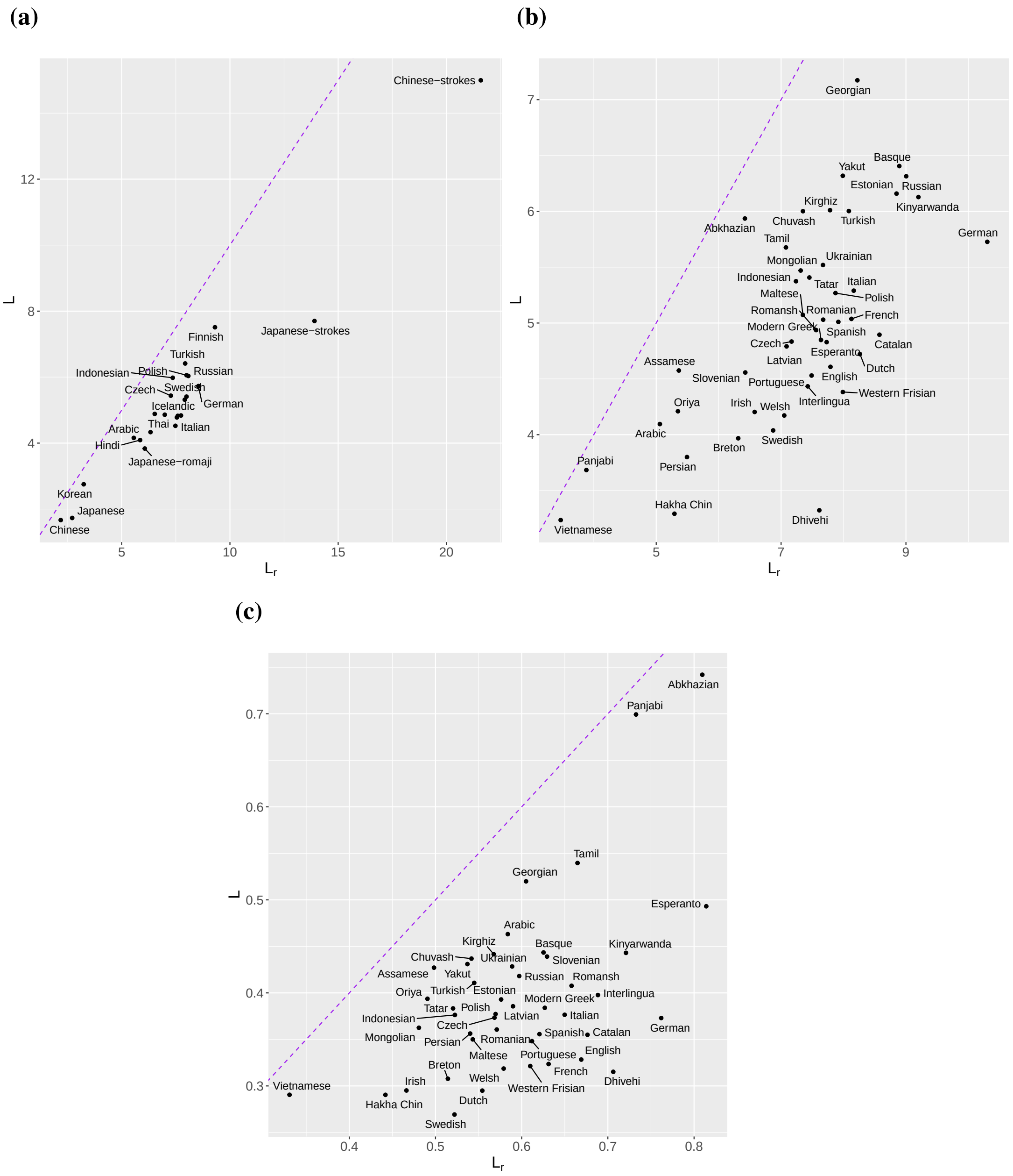

Figure 2: Mean word length (L) as a function of the random baseline (L_r) in languages. Every point stands for a language. The diagonal (long dashed line) indicates the line $L = L_r$. Languages with $L < L_r$ are located below the diagonal. (a) Languages in PUD with word length measured in characters (or strokes for Chinese and Japanese). (b) Languages in CV with word length measured in characters. (c) Languages in CV with word length measured in duration (seconds).

5.3 Impact of disabling the filter of words that contain "foreign" characters

All results presented in this section have been obtained after applying the new method to filter out highly unusual characters and words described in Section 4.4. If the filter is disabled, we obtain some slight changes in the values, but the qualitative results summarized above remain the same.

6 Discussion

6.1 The universality of Zipf's law

The first step of our analysis consisted in checking the universality of the law of abbreviation in the languages of our samples through a Kendall τ correlation test. Here, we introduced two methodological improvements with respect to previous research: using the Bonferroni-Holm correction for p-values, as well as word length in time given spoken utterances, rather than just characters in written form (Bentz and Ferrer-i-Cancho, 2016). We also computed Pearson correlations for two reasons: (a) to verify the robustness of the conclusions and (b) to check the significance of the gap between L and L_r (the case of (b) is addressed in the next subsection). We find that the law of abbreviation holds in nearly all languages in our sample at a 95% confidence level, independently from how word length is measured, and even after controlling for multiple testing. The only exception is Panjabi in CV, but only when length is measured in duration and Pearson r correlation is used. Panjabi is also the language suffering most from under-sampling (only 98 tokens). Therefore, Panjabi cannot be considered a true exception to the law of abbreviation.

Given the rather scarce evidence of the law of abbreviation in word durations in human language (Torre et al., 2019), we have taken step forward by providing evidence of it in 46 languages from 14 linguistic families. The massive agreement of the law of abbreviation even when orthographic word lengths are replaced by word durations in human languages provides stronger support for the law of abbreviation as a potentially universal pattern of human languages with respect to previous research relying on word length in characters (Bentz and Ferrer-i-Cancho, 2016) and often on a small number of linguistic families (Koplenig et al., 2022; Levshina, 2022; Meylan and Griffiths, 2021; Piantadosi et al., 2011).

6.2 Direct evidence of compression

We have found that word lengths are shorter than expected by chance ($L < L_r$) in all languages in every collection (Figure 2). Such a systematic finding is unlikely to be accidental and strongly indicates that compression is acting in all languages in our sample. Crucially, the finding holds independently of how word length is measured. The ample evidence of compression even when orthographic word lengths are replaced by word durations in human languages provides stronger support for compression as a universal principle of the organization of languages with respect to previous research relying on word length in characters (Ferrer-i-Cancho et al., 2013).

It could be argued that these findings constitute evidence of compression in ensembles of language but not in individual languages. The reason is that $L < L_r$ does not imply that the difference between the actual word length and the random baseline is statistically significant for a single language. However, we have shown that the Pearson correlation is indeed a linear function of L and L_r (Appendix A) and thus L is significantly small in every language where the law of abbreviation has been confirmed using a Pearson correlation test.

Finally, the direct correspondence we have established between the average length of types (M) and the random baseline sheds new light on previous research. For instance, it has been shown that $M < L$ in Chinese characters in six time periods spanning two millennia (Chen et al., 2015, Fig. 4), which now can be reinterpreted as a sign of compression of word lengths in Chinese in light of our theoretical findings.

Future research

In this article, we have introduced a new random baseline and unveiled a systematic gap between that random baseline and real mean word lengths that we have interpreted as direct evidence of compression. Figure 2 suggests that the gap is wider when word lengths are measured in duration rather than in characters. However, we have not quantified the magnitude of that gap and we have neither taken into consideration the gap between actual mean words lengths and the minimum baseline, that would be defined as the minimum word length that could be achieved under certain constraints (Cover and Thomas, 2006; Ferrer-i-Cancho et al., 2019; Pimentel et al., 2021). Future research should quantify the first gap in relation to the minimum baseline. As the random baseline is crucial to asses the degree of optimality of word lengths, we have paved the way for exploring the degree of optimality of word lengths in characters or duration in languages.

Authors’ contributions

SP: Conceptualization, Formal Analysis, Investigation, Software, Supervision, Validation, Visualization, Writing-original draft, Writing-review & editing; ACM: Data curation, Resources, Software; JCM: Writing-review & editing, Resources; MW: Writing-review & editing; CB: Conceptualization, Writing-review & editing; RFC: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project Administration, Supervision, Writing-original draft, Writing-review & editing.

Acknowledgments

This article is one of the products of the research project of the 1st edition of the master degree course “Introduction to Quantitative Linguistics” at Universitat Politècnica de Catalunya. We are specially grateful to two students of that course: L. Alemany-Puig for helpful discussions and computational support, and M. Michaux for comments on early versions of the manuscript. We thank M. Farrús and A. Hernández-Fernández for advice on voice datasets. We also thank S. Komori for advice on Japanese, Y. M. Oh for advice on Korean and S. Semple for helping us to improve English. Last but not least, we are very grateful to an anonymous reviewer for very helpful comments.

SP is funded by the grant ’Thesis abroad 2021/2022’ from the University of Milan. CB was partly funded by the *Deutsche Forschungsgemeinschaft* (FOR 2237: Words, Bones, Genes, Tools - Tracking Linguistic, Cultural and Biological Trajectories of the Human Past), and the *Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung* (Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora, 176305). RFC is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya).

References

Balasubrahmanyan, V. K., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*(3), 177–228. https://doi.org/10.1080/09296179608599629

Bentz, C., Ferrer-i-Cancho, R. (2016). Zipf’s law of abbreviation as a language universal. In C. Bentz, G. Jäger, I. Yanovich (Eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen. http://hdl.handle.net/10900/68639

Börstell, C., Hörberg, T., Östling, R. (2016). Distribution and duration of signs and parts of speech in Swedish Sign Language. *Sign Language & Linguistics*, *19*, 143–196. https://doi.org/10.1075/sll.19.2.01bor

Chen, H., Liang, J., Liu, H. (2015). How does word length evolve in written Chinese? *PLOS ONE*, *10*(9), 1–12. https://doi.org/10.1371/journal.pone.0138567

Conover, W. J. (1999). *Practical nonparametric statistics* [3rd edition]. Wiley.

Cover, T. M., Thomas, J. A. (2006). *Elements of information theory* [2nd edition]. Wiley.

Deng, W., Allahverdyan, A. E., Li, B., Wang, Q. A. (2014). Rank–frequency relation for Chinese characters. *The European Physical Journal B*, *87*(2), 47. https://doi.org/10.1140/epjb/e2014-40805-2

Evans, N., Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429–492. https://doi.org/10.1017/s0140525x0999094x

Ferrer-i-Cancho, R., Bentz, C., Seguin, C. (2019). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, *29*(2), 165–194. https://doi.org/10.1080/09296174.2020.1778387

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578. https://doi.org/10.1111/cogs.12061

Heesen, R., Hobaiter, C., Ferrer-i-Cancho, R., Semple, S. (2019). Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20182900. https://doi.org/10.12775/3991-1.039

Hernández-Fernández, A., G. Torre, I., Garrido, J.-M., Lacasa, L. (2019). Linguistic laws in speech: The case of Catalan and Spanish. *Entropy*, *21*(12). https://doi.org/10.3390/e21121153

Hernández-Fernández, A., Torre, I. G. (2022). Compression principle and Zipf's Law of brevity in infochemical communication. *Biology Letters*, *18*, 20220162. https://doi.org/10.1098/rsbl.2022.0162

Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., Orden, G. C. V. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005

Koplenig, A., Kupietz, M., Wolfer, S. (2022). Testing the relationship between word length, frequency, and predictability based on the German reference corpus. *Cognitive Science*, *46*(6), e13090. https://doi.org/10.1111/cogs.13090

Levshina, N. (2022). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, *24*(2). https://doi.org/10.3390/e24020280

Meylan, S. C., Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, *45*(6), e12983. https://doi.org/10.1111/cogs.12983

Moreno Fernández, J. (2021). The optimality of word lengths (Master's thesis). Barcelona School of Informatics. Barcelona. http://hdl.handle.net/2117/361054

Naranan, S., Balasubrahmanyan, V. K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, *52*, 728–738.

Piantadosi, S. T., Tily, H., Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., Blasi, D. (2021). How (non-)optimal is the lexicon? *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.naacl-main.350

Sanada, H. (2008). *Investigations in japanese historical lexicology*. Peust & Gutschmidt Verlag.

Semple, S., Ferrer-i-Cancho, R., Gustison, M. (2022). Linguistic laws in biology. *Trends in Ecology and Evolution*, *37*(1), 53–66. https://doi.org/10.1016/j.tree.2021.08.012

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379–423, 623–656.

Torre, I. G., Luque, B., Lacasa, L., Kello, C. T., Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, *6*(8), 191023. https://doi.org/10.1098/rsos.191023

van den Heuvel, E., Zhan, Z. (2022). Myths about linear and monotonic associations: Pearson's r, Spearman's ρ, and Kendall's τ. *The American Statistician*, *76*(1), 44–52. https://doi.org/10.1080/00031305.2021.2004922

Wang, Y., Chen, X. (2015). Structural complexity of simplified Chinese characters. In A. Tuzzi J. M. M. Benesová (Eds.), *Recent contributions to quantitative linguistics* (pp. 229–239). De Gruyter. https://doi.org/10.1515/9783110420296-019

Watson, S. K., Heesen, R., Hedwig, D., Robbins, M. M., Townsend, S. W. (2020). An exploration of Menzerath's law in wild mountain gorilla vocal sequences. *Biology Letters*, *16*, 20200380. https://doi.org/10.1098/rsbl.2020.0380

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley.

Appendices

Appendix A Theory

Here we review the relationship between L, L_r and Pearson correlation

Given two random variables x and y and a sample of n points, $\{(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)\}$, the sample covariance is defined as

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right),$$

where $\bar{x}$ is the sample mean of x and $\bar{y}$ is the sample mean for y, i.e.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Now consider than the random variables are p (the probability of a type) and l (the length/duration of a type) instead of x and y. Then our sample of n points is $\{(p_1, l_1), ..., (p_i, l_i), ..., (p_n, l_n)\}$, one point per type. Accordingly, the covariance between p and l in a sample of points is

$$s_{pl} = \frac{1}{n-1}\left(\sum_{i=1}^{n} p_i l_i - n\bar{p}\bar{l}\right).$$

Recalling the definition of L (Equation 1) and noting that $\bar{p} = \frac{1}{n}$ and $\bar{l} = M = L_r$ (recall Property 2.1), we finally obtain

$$s_{pl} = \frac{1}{n-1}(L - L_r).$$

The sample Pearson correlation is

$$r = \frac{s_{xy}}{s_x s_y},$$

where s_x and s_y are the sample standard deviation of x and y, i.e.

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Proceeding as we did for the covariance, we find that the Pearson correlation between p and l is

$$r = \frac{L - L_r}{(n-1)s_p s_l}.$$

Then it is easy to see that L is a linear function of the Pearson correlation r or s_{pl}. For instance,

$$L = ar + b,$$

where

$$a = (n-1)s_p s_l$$
$$b = L_r.$$

Other linear relationships can be shown similarly.

Appendix B Analysis

We here present complementary analyses, tables and plots.

B.1 The impact of the unsupervised filter

Table B1 and Table B2 show the impact of the unsupervised filter in the optional filter. PUD is a controlled setting for the impact of the filter because it is a collection where tokens are of high quality compared to CV. Thus we expect that the impact of the optional filter is low in PUD. Unexpectedly, the number of tokens reduces substantially (a reduction of the order of thousands) in Chinese, Japanese and Korean. An additional drastic reduction in the observed alphabet size in these languages strongly suggests that the optional filter is not adequate for them. For these reasons, we believe we should not apply the unsupervised filter to these languages because their writing system is essentially a syllabary. We suspect that the actual need for the exclusion could be a combination of sampling problems relating to a large alphabet size (compared to the Latin script) and a heavy- tailed rank distribution that breaks the optional filter. It is well-known that the rank distribution of Chinese characters is long-tailed, spanning two orders of magnitude (Deng et al., 2014), while that of phonemes (the counterpart of letters in many languages using the Latin script) is exponential-like (Balasubrahmanyan and Naranan, 1996; Naranan and Balasubrahmanyan, 1993). However, that issue should be the subject of future research.

In CV, we find that the optional filter has a similar impact in languages concerning the reduction in the number of tokens but higher impacts concerning the reduction of the alphabet sizes, suggesting that presence of strings with strange characters. The three languages with the most marked reduction in alphabet size are French, Spanish, German and Italian, with an alphabet size greater then 100.

Table B1: The impact of the unsupervised filter in the PUD collection. For every language, we show its linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters after applying the mandatory filter but before applying the unsupervised filter, that are A, the observed alphabet size (number of distinct characters), n, the number of types, and, T, the number of tokens. A', n' and T' are the respective values of A, n and T after applying the unsupervised filter.

Language	Script	Family	A	A'	n	n'	T	T'
Arabic	Arabic	Afro-Asiatic	47	39	6600	6596	18214	18201
Indonesian	Latin	Austronesian	39	23	4596	4501	16819	16702
Russian	Cyrillic	Indo-European	61	31	7358	7113	15870	15588
Hindi	Devanagari	Indo-European	84	50	4920	4716	21184	20796
Czech	Latin	Indo-European	49	33	7360	7073	15700	15331
English	Latin	Indo-European	39	25	5082	5001	18135	18028
French	Latin	Indo-European	48	26	5593	5214	21084	20407
German	Latin	Indo-European	39	28	6215	6116	18446	18331
Icelandic	Latin	Indo-European	43	32	6175	6035	16385	16209
Italian	Latin	Indo-European	42	24	5944	5606	21815	21266
Polish	Latin	Indo-European	47	31	7329	7188	15386	15191
Portuguese	Latin	Indo-European	47	38	5678	5661	21873	21855
Spanish	Latin	Indo-European	39	32	5765	5750	21083	21067
Swedish	Latin	Indo-European	39	25	5842	5624	16653	16378
Japanese	Japanese	Japonic	1549	609	4990	3345	24899	22538
Japanese-strokes	Japanese	Japonic	1549	609	4852	3345	24737	22538
Japanese-romaji	Latin	Japonic	23	19	4984	4860	24892	24743
Korean	Hangul	Koreanic	1002	401	8031	6424	14475	12540
Thai	Thai	Kra-Dai	89	52	3818	3599	21642	21121
Chinese	Han (Traditional variant)	Sino-Tibetan	2038	814	5224	3154	18129	15436
Chinese-strokes	Han (Traditional variant)	Sino-Tibetan	2038	814	4970	3154	17845	15436
Chinese-pinyin	Latin	Sino-Tibetan	49	44	5224	5038	18129	17885
Turkish	Latin	Turkic	42	28	6793	6587	14092	13799
Finnish	Latin	Uralic	39	24	7076	6938	12853	12701

Table B2: The impact of the unsupervised filter in the CV collection. The content is the same as in Table B1. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense as Conlang languages are not related in the common linguistic family sense.

Language	Script	Family	A	A'	n	n'	T	T'
Arabic	Arabic	Afro-Asiatic	44	31	7497	6397	49448	45825
Maltese	Latin	Afro-Asiatic	40	31	8148	8058	44272	44112
Vietnamese	Latin	Austroasiatic	86	41	574	370	1300	938
Indonesian	Latin	Austronesian	28	22	3817	3768	44336	44210
Esperanto	Latin	Conlang	38	27	27932	27759	406725	406261
Interlingua	Latin	Conlang	27	20	5552	5126	31428	30504
Tamil	Tamil	Dravidian	44	29	1525	1210	7580	6439
Persian	Arabic	Indo-European	105	38	13240	13115	1665428	1662508
Assamese	Assamese	Indo-European	60	43	1115	971	2000	1813
Russian	Cyrillic	Indo-European	54	32	31921	31827	638782	637686
Ukrainian	Cyrillic	Indo-European	44	34	14399	14337	120984	120760
Panjabi	Devanagari	Indo-European	48	37	95	84	110	98
Modern Greek	Greek	Indo-European	46	33	5834	5813	37926	37880
Breton	Latin	Indo-European	41	28	4322	4228	38493	38237
Catalan	Latin	Indo-European	67	39	79213	79112	3294506	3294206
Czech	Latin	Indo-European	44	33	16032	15518	150312	147582
Dutch	Latin	Indo-European	41	23	10666	10225	320992	316498
English	Latin	Indo-European	97	28	173522	173023	9829660	9828713
French	Latin	Indo-European	244	49	162740	160243	3732822	3729370
German	Latin	Indo-European	152	30	150362	148436	4235094	4230565
Irish	Latin	Indo-European	31	23	2311	2251	22751	22593
Italian	Latin	Indo-European	110	34	55480	54996	812604	811783
Latvian	Latin	Indo-European	35	27	7792	7251	30358	29456
Polish	Latin	Indo-European	38	32	25365	25340	595613	595411
Portuguese	Latin	Indo-European	41	27	13049	11509	295042	283048
Romanian	Latin	Indo-European	36	29	6449	6423	33370	33341
Romansh	Latin	Indo-European	40	26	9801	9614	44192	43792
Slovenian	Latin	Indo-European	28	24	5994	5937	26402	26304
Spanish	Latin	Indo-European	186	33	75617	75010	1843646	1842474
Swedish	Latin	Indo-European	30	25	4454	4371	63282	62951
Welsh	Latin	Indo-European	43	22	11488	11143	547345	539621
Western Frisian	Latin	Indo-European	42	30	8419	8383	63127	63073
Oriya	Odia	Indo-European	59	41	921	764	1929	1700
Dhivehi	Thaana	Indo-European	40	27	155	111	1388	1284
Georgian	Georgian	Kartvelian	34	25	7945	6505	15481	12958
Basque	Latin	Language isolate	28	21	24998	24748	460188	458071
Mongolian	Mongolian	Mongolic	36	31	14844	14608	70638	70217
Kinyarwanda	Latin	Niger-Congo	96	26	135328	133815	1945038	1939810
Abkhazian	Cyrillic	Northwest Caucasian	37	28	150	119	189	156
Hakha Chin	Latin	Sino-Tibetan	28	23	2515	2499	17806	17776
Chuvash	Cyrillic	Turkic	36	22	5565	4311	16270	13583
Kirghiz	Cyrillic	Turkic	38	30	10497	10130	62687	61844
Tatar	Cyrillic	Turkic	47	34	22313	21823	145458	144356
Yakut	Cyrillic	Turkic	42	28	8041	7904	22795	22577
Turkish	Latin	Turkic	37	31	8957	8926	107910	107686
Estonian	Latin	Uralic	34	23	30135	28691	123895	121549

B.2 Mean word length and the law of abbreviation

In Table B3, Table B4 and Table B5, we show the mean word length (L) and the random baseline (L_r) as well as the outcome of the correlation test between length and frequency for PUD and for CV when length is measured in characters and also in duration, respectively.

Table B3: Mean word length and the correlation between frequency and length in PUD. Word length is measured in number of characters. Mean word length (L) is followed by the random baseline (L_r). Each correlation statistic (Kendall τ or Pearson r) is followed by p-values after applying Holm-Bonferroni correction (rather than being the direct output of the correlation test).

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Arabic	Afro-Asiatic	Arabic	4.03	5.54	-0.13	8.32×10^{-32}	-0.13	1.12×10^{-20}
Czech	Indo-European	Latin	5.44	7.27	-0.22	1.20×10^{-113}	-0.15	2.47×10^{-36}
English	Indo-European	Latin	4.87	7.00	-0.20	2.52×10^{-66}	-0.12	6.98×10^{-17}
French	Indo-European	Latin	4.81	7.47	-0.16	2.44×10^{-49}	-0.12	4.24×10^{-19}
German	Indo-European	Latin	5.74	8.56	-0.23	1.25×10^{-108}	-0.12	3.85×10^{-21}
Indonesian	Austronesian	Latin	5.96	7.35	-0.11	6.37×10^{-21}	-0.12	6.53×10^{-15}
Italian	Indo-European	Latin	4.85	7.64	-0.16	4.09×10^{-54}	-0.13	8.45×10^{-23}
Polish	Indo-European	Latin	6.07	8.00	-0.19	1.12×10^{-80}	-0.13	2.78×10^{-26}
Portuguese	Indo-European	Latin	4.35	7.47	-0.20	9.96×10^{-67}	-0.12	1.12×10^{-17}
Russian	Indo-European	Cyrillic	6.04	8.08	-0.19	4.58×10^{-88}	-0.13	4.85×10^{-26}
Spanish	Indo-European	Latin	4.83	7.59	-0.16	4.10×10^{-51}	-0.11	1.89×10^{-17}
Swedish	Indo-European	Latin	5.41	7.99	-0.23	3.99×10^{-101}	-0.13	6.28×10^{-21}
Turkish	Turkic	Latin	6.43	7.94	-0.24	4.26×10^{-124}	-0.12	4.20×10^{-23}

Table B4: Mean word length and the correlation between frequency and length in CV. Word length is measured in number of characters. Content is the same as in B3. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense, and Conlang languages are not related in the common family sense.

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Abkhazian	Northwest Caucasian	Cyrillic	5.94	6.42	-0.32	4.48×10^{-5}	-0.29	1.43×10^{-3}
Arabic	Afro-Asiatic	Arabic	4.10	5.06	-0.14	5.32×10^{-43}	-0.14	2.04×10^{-28}
Assamese	Indo-European	Assamese	4.57	5.36	-0.31	4.73×10^{-31}	-0.27	3.09×10^{-17}
Basque	Language isolate	Latin	6.41	8.89	-0.16	2.68×10^{-262}	-0.11	6.95×10^{-69}
Breton	Indo-European	Latin	3.97	6.31	-0.24	4.93×10^{-86}	-0.19	4.09×10^{-35}
Catalan	Indo-European	Latin	4.90	8.58	-0.15	0.00	-0.05	9.53×10^{-51}
Chuvash	Turkic	Cyrillic	6.00	7.35	-0.22	5.49×10^{-74}	-0.21	3.80×10^{-43}
Czech	Indo-European	Latin	4.83	7.17	-0.22	1.75×10^{-295}	-0.13	1.69×10^{-58}
Dhivehi	Indo-European	Thaana	3.32	7.61	-0.16	1.65×10^{-2}	-0.31	1.24×10^{-3}
Dutch	Indo-European	Latin	4.72	8.26	-0.28	0.00	-0.10	1.35×10^{-24}
English	Indo-European	Latin	4.61	7.79	-0.07	0.00	-0.03	3.45×10^{-45}
Esperanto	Conlang	Latin	4.83	7.73	-0.18	0.00	-0.08	1.12×10^{-41}
Estonian	Uralic	Latin	6.16	8.85	-0.24	0.00	-0.09	2.55×10^{-48}
French	Indo-European	Latin	5.04	8.13	-0.04	3.57×10^{-85}	-0.04	8.56×10^{-46}
Georgian	Kartvelian	Georgian	7.17	8.22	-0.12	3.67×10^{-31}	-0.10	3.47×10^{-15}
German	Indo-European	Latin	5.73	10.30	-0.12	0.00	-0.04	4.21×10^{-59}
Hakha Chin	Sino-Tibetan	Latin	3.29	5.29	-0.29	4.31×10^{-72}	-0.15	3.88×10^{-13}
Indonesian	Austronesian	Latin	5.37	7.24	-0.20	1.13×10^{-59}	-0.16	2.73×10^{-21}
Interlingua	Conlang	Latin	4.43	7.43	-0.24	8.95×10^{-101}	-0.16	7.39×10^{-31}
Irish	Indo-European	Latin	4.20	6.58	-0.21	2.38×10^{-41}	-0.17	5.18×10^{-15}
Italian	Indo-European	Latin	5.29	8.16	-0.06	2.24×10^{-67}	-0.06	4.19×10^{-49}
Kinyarwanda	Niger-Congo	Latin	6.13	9.20	-0.19	0.00	-0.06	3.32×10^{-117}
Kirghiz	Turkic	Cyrillic	6.01	7.78	-0.19	1.45×10^{-141}	-0.16	6.13×10^{-57}
Latvian	Indo-European	Latin	4.79	7.09	-0.26	5.81×10^{-160}	-0.18	1.36×10^{-53}
Maltese	Afro-Asiatic	Latin	5.07	7.35	-0.20	2.32×10^{-107}	-0.14	1.58×10^{-36}
Modern Greek	Indo-European	Greek	4.85	7.64	-0.24	3.73×10^{-124}	-0.16	1.77×10^{-34}
Mongolian	Mongolic	Mongolian	5.47	7.31	-0.23	1.73×10^{-263}	-0.15	2.31×10^{-76}
Oriya	Indo-European	Odia	4.21	5.35	-0.33	2.00×10^{-28}	-0.31	2.94×10^{-17}
Panjabi	Indo-European	Devanagari	3.68	3.88	-0.32	8.69×10^{-4}	-0.26	8.60×10^{-3}
Persian	Indo-European	Arabic	3.80	5.49	-0.21	2.38×10^{-229}	-0.12	2.06×10^{-45}
Polish	Indo-European	Latin	5.27	7.87	-0.17	8.03×10^{-292}	-0.10	2.47×10^{-58}
Portuguese	Indo-European	Latin	4.53	7.49	-0.19	1.09×10^{-168}	-0.13	1.21×10^{-41}
Romanian	Indo-European	Latin	5.03	7.67	-0.21	3.27×10^{-97}	-0.17	6.46×10^{-41}
Romansh	Indo-European	Latin	4.94	7.56	-0.24	5.91×10^{-184}	-0.15	5.42×10^{-48}
Russian	Indo-European	Cyrillic	6.31	9.00	-0.13	7.75×10^{-225}	-0.09	3.03×10^{-52}
Slovenian	Indo-European	Latin	4.56	6.43	-0.21	1.47×10^{-88}	-0.15	4.71×10^{-29}
Spanish	Indo-European	Latin	5.01	7.92	-0.03	5.95×10^{-32}	-0.04	3.48×10^{-29}
Swedish	Indo-European	Latin	4.04	6.87	-0.28	6.91×10^{-129}	-0.15	1.62×10^{-22}
Tamil	Dravidian	Tamil	5.68	7.08	-0.28	1.01×10^{-35}	-0.23	5.21×10^{-16}
Tatar	Turkic	Cyrillic	5.41	7.45	-0.24	0.00	-0.16	3.15×10^{-118}
Turkish	Turkic	Latin	6.00	8.09	-0.22	1.32×10^{-158}	-0.16	2.51×10^{-48}
Ukrainian	Indo-European	Cyrillic	5.52	7.67	-0.16	3.01×10^{-136}	-0.14	1.74×10^{-61}
Vietnamese	Austroasiatic	Latin	3.24	3.47	-0.19	2.98×10^{-5}	-0.20	1.96×10^{-4}
Welsh	Indo-European	Latin	4.17	7.05	-0.21	2.40×10^{-185}	-0.12	4.39×10^{-38}
Western Frisian	Indo-European	Latin	4.38	7.99	-0.29	1.19×10^{-244}	-0.13	2.62×10^{-33}
Yakut	Turkic	Cyrillic	6.32	7.99	-0.26	5.48×10^{-185}	-0.19	2.12×10^{-65}

Table B5: Mean word length and the correlation between frequency and length in CV. Word length is measured in duration. Content is the same as in B4.

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Abkhazian	Northwest Caucasian	Cyrillic	0.74	0.81	-0.20	1.23×10^{-2}	-0.21	2.52×10^{-2}
Arabic	Afro-Asiatic	Arabic	0.46	0.58	-0.12	1.75×10^{-40}	-0.15	2.00×10^{-31}
Assamese	Indo-European	Assamese	0.43	0.50	-0.22	1.25×10^{-17}	-0.19	3.14×10^{-9}
Basque	Language isolate	Latin	0.44	0.63	-0.21	0.00	-0.12	1.29×10^{-78}
Breton	Indo-European	Latin	0.31	0.51	-0.25	1.92×10^{-107}	-0.20	4.94×10^{-39}
Catalan	Indo-European	Latin	0.35	0.68	-0.21	0.00	-0.06	8.70×10^{-69}
Chuvash	Turkic	Cyrillic	0.44	0.54	-0.26	1.18×10^{-116}	-0.22	6.89×10^{-49}
Czech	Indo-European	Latin	0.37	0.57	-0.21	6.40×10^{-295}	-0.14	5.07×10^{-70}
Dhivehi	Indo-European	Thaana	0.32	0.71	-0.17	2.40×10^{-2}	-0.24	1.51×10^{-2}
Dutch	Indo-European	Latin	0.29	0.55	-0.28	0.00	-0.12	1.47×10^{-33}
English	Indo-European	Latin	0.33	0.67	-0.17	0.00	-0.04	4.83×10^{-62}
Esperanto	Conlang	Latin	0.49	0.81	-0.18	0.00	-0.09	1.25×10^{-47}
Estonian	Uralic	Latin	0.39	0.58	-0.23	0.00	-0.09	4.65×10^{-55}
French	Indo-European	Latin	0.32	0.63	-0.21	0.00	-0.04	7.25×10^{-71}
Georgian	Kartvelian	Georgian	0.52	0.61	-0.15	6.51×10^{-51}	-0.12	9.17×10^{-21}
German	Indo-European	Latin	0.37	0.76	-0.22	0.00	-0.05	1.57×10^{-96}
Hakha Chin	Sino-Tibetan	Latin	0.29	0.44	-0.25	9.56×10^{-64}	-0.14	1.10×10^{-11}
Indonesian	Austronesian	Latin	0.38	0.52	-0.22	1.29×10^{-76}	-0.17	8.41×10^{-26}
Interlingua	Conlang	Latin	0.40	0.69	-0.24	9.77×10^{-114}	-0.18	3.80×10^{-36}
Irish	Indo-European	Latin	0.30	0.47	-0.24	1.42×10^{-55}	-0.19	1.02×10^{-19}
Italian	Indo-European	Latin	0.38	0.65	-0.19	0.00	-0.08	4.19×10^{-87}
Kinyarwanda	Niger-Congo	Latin	0.44	0.72	-0.21	0.00	-0.06	7.54×10^{-101}
Kirghiz	Turkic	Cyrillic	0.44	0.57	-0.20	1.38×10^{-159}	-0.16	1.63×10^{-55}
Latvian	Indo-European	Latin	0.39	0.59	-0.23	1.70×10^{-141}	-0.19	4.58×10^{-60}
Maltese	Afro-Asiatic	Latin	0.35	0.54	-0.21	9.96×10^{-140}	-0.15	3.89×10^{-42}
Modern Greek	Indo-European	Greek	0.38	0.63	-0.21	3.20×10^{-105}	-0.17	1.46×10^{-37}
Mongolian	Mongolic	Mongolian	0.36	0.48	-0.25	0.00	-0.15	3.12×10^{-73}
Oriya	Indo-European	Odia	0.39	0.49	-0.33	2.21×10^{-31}	-0.32	1.59×10^{-18}
Panjabi	Indo-European	Devanagari	0.70	0.73	-0.18	4.63×10^{-2}	-0.15	8.44×10^{-2}
Persian	Indo-European	Arabic	0.36	0.54	-0.25	0.00	-0.14	4.66×10^{-58}
Polish	Indo-European	Latin	0.38	0.57	-0.17	0.00	-0.12	2.88×10^{-82}
Portuguese	Indo-European	Latin	0.35	0.61	-0.22	1.15×10^{-243}	-0.15	4.38×10^{-59}
Romanian	Indo-European	Latin	0.36	0.57	-0.23	2.49×10^{-127}	-0.17	2.82×10^{-43}
Romansh	Indo-European	Latin	0.41	0.66	-0.26	7.70×10^{-248}	-0.17	2.06×10^{-64}
Russian	Indo-European	Cyrillic	0.42	0.60	-0.15	2.13×10^{-299}	-0.10	5.30×10^{-75}
Slovenian	Indo-European	Latin	0.44	0.63	-0.25	3.37×10^{-146}	-0.17	3.04×10^{-40}
Spanish	Indo-European	Latin	0.36	0.62	-0.14	0.00	-0.05	2.05×10^{-41}
Swedish	Indo-European	Latin	0.27	0.52	-0.29	1.03×10^{-156}	-0.18	4.76×10^{-32}
Tamil	Dravidian	Tamil	0.54	0.66	-0.31	2.06×10^{-48}	-0.22	5.35×10^{-14}
Tatar	Turkic	Cyrillic	0.38	0.52	-0.26	0.00	-0.17	8.68×10^{-141}
Turkish	Turkic	Latin	0.41	0.54	-0.21	7.11×10^{-158}	-0.16	9.43×10^{-50}
Ukrainian	Indo-European	Cyrillic	0.43	0.59	-0.18	3.01×10^{-176}	-0.16	3.53×10^{-86}
Vietnamese	Austroasiatic	Latin	0.29	0.33	-0.07	4.63×10^{-2}	-0.14	1.40×10^{-2}
Welsh	Indo-European	Latin	0.32	0.58	-0.20	6.25×10^{-197}	-0.15	3.21×10^{-58}
Western Frisian	Indo-European	Latin	0.32	0.61	-0.31	0.00	-0.15	8.59×10^{-43}
Yakut	Turkic	Cyrillic	0.43	0.54	-0.25	2.41×10^{-186}	-0.18	9.50×10^{-58}

The journal SMIL – Statistical Methods in Linguistics (1962–1976) – some notes about the history of quantitative linguistics in Scandinavia and beyond

Emmerich Kelih[1]

[1] Department of Slavonic Studies, University of Vienna

DOI: https://doi.org/10.53482/2023_54_408

ABSTRACT

This article deals with the history of quantitative linguistics. The focus of this paper is the journal *SMIL – Statistical Methods in Linguistics*, which was published by Hans Karlgren in Stockholm from 1962 to 1976 (with a short interruption between 1966 and 1969). SMIL is a representative example of the process of differentiation in quantitative linguistics during the seventies and can be seen as one early major “Scandinavian” contribution to statistical and quantitative linguistics.

Key words: history of quantitative linguistics, statistical linguistics, *SMIL*, Hans Karlgren (1933–1996)

A scientific discipline needs not only a group of researchers but also a corresponding institutional organisation. In particular, possibilities to disseminate current research results and to promote the exchange of information and scientific knowledge are required. While there are currently several journals explicitly devoted to questions of quantitative linguistics (*Journal of Quantitative Linguistics*, *Glottometrics*, *Glottotheory*, and many others), a look at the more recent history of science shows that the establishment of journals with such a focus is quite laborious and that, overall a corresponding infrastructure in this field has only developed slowly.

Some years ago, the editor of *Glottometrics* called for contributions on the history of quantitative linguistics and/or to introduce individual researchers who have worked in this field. We are happy to comply with this request in this article and would like to take a closer look at a small cornerstone in the history of modern quantitative linguistics in Scandinavia. It is about the journal *SMIL – Statistical Methods in Linguistics*, which was published under this title from 1962 to 1976; the successor project *SMIL Quarterly: Journal of Linguistic Calculus* was then to appear until 1981, but with a clear and explicit focus on computational linguistics only. *SMIL* was founded in 1962 by the Swedish linguist Hans Karlgren (1933–1996) and published by the privately financed publishing house *Skriptor* (the full Swedish name is *Stockholm Språkförlaget Skriptor*). Karlgren himself was interested in the application of statistical methods in linguistics, but was later to make a name for himself in computational linguistics, and is for instance regarded as the initiator of the well-known COLING conferences. Hammarström (2012: 84-87) describes Hans Karlgren in his memoirs as "[...] the most friendly, generous, intelligent and original person one could meet". In particular, he also refers to Karlgren's good organisational skills. Despite his relatively young age (he was "only" thirty years old when he founded *SMIL*) he was able to acquire papers from linguists who were quite well-known at that time. For further details about this see below.

The founding of *SMIL* coincided with the information-theoretical or so-called cybernetic "revolution" in the sciences, which also left its mark on linguistics. This information-theoretical enthusiasm also lead to a boost in the application of statistical methods in language and text analysis in general. This, in turn, led to the question of "how to name the child", and this is precisely the period when terms such as *statistical linguistics*, *mathematical linguistics* and *quantitative linguistics* were coined and then, after long-lasting discussions, were differentiated. As one reads in the preface to *SMIL* 1 (cf. Karlgren 1962a), the founding of *SMIL* was preceded by the First Scandinavian Symposium on Statistical Linguistics in 1960 at Stundyblom Castle, with over 40 speakers and 20 papers presented. At the same time, as the interest in statistics in linguistics grew, it also became apparent that there was a lack of a platform where current contributions could be published, but also where relevant bibliographical information on research literature in this field could be provided. Thematically, Karlgren (1963a: 69) described the focus of the journal in *SMIL* 2, 1963 as follows:

> Statistical linguistics calls for both advanced mathematical analyses of the models employed and for experiments hugging the linguistic ground where the methods are put to test on concrete material. *SMIL* will provide papers of both kinds, averaging, we hope, adequate proportions.

This reflects an incipient process of differentiation, where the application of statistical methods becomes the field of work of so-called statistical linguistics or later then quantitative linguistics, while so-called mathematical linguistics preferred primarily formal mathematical methods. However, *SMIL* was on its way to have two hearts beating in its chest from the beginning on. This is especially evident from the fact that from *SMIL* 3, 1964, onwards, the well-known Hungarian mathematical linguist Ferenc Kiefer (1931–2020) became the co-editor of *SMIL*. Before going into more detail about some of *SMIL*'s thematic focuses, the language policy pursued should also be mentioned. The imprint of *SMIL* explicitly states that "Contributions will be printed in English, German or French. Writers may feel free to submit manuscripts in any reasonable language." This means that a multilingual language policy was followed (and also realised in practice, at least partly). In addition – and one has to remember the political tensions in East–West relations at that time – scientific exchange with the Soviet Union was proactively promoted by translating the titles of selected interesting articles into Russian and in the other direction by referring to important publications in the field of language statistics or automatic language processing from the Soviet Union (later issues of *SMIL* would feature articles by prominent Soviet authors). All in all, *SMIL* developed over time from an initially Scandinavian-focused journal into a truly international publication, which can also be seen in the successive expansion of the authorship and the readership.

In the following, an attempt will be made to present some main thematic focuses, with our interest directed exclusively towards contributions that correspond to the focus of today's quantitative linguistics.

It is particularly noticeable that in the initial phase, among the so-called statistical works, there was definitely a focus on phonetic-phonological issues. Sigurd (1963) should be mentioned, who dealt intensively with the modelling of the frequency of phoneme inventories. This topic was later addressed again by Altmann/Lehfeldt (1980), among others, and the statistical modelling has not found a satisfactory solution to this day. The contribution by Ladefoged (1970), which deals with the quantitative measurement of phonetic similarity, raised an important research question which would also stimulate research in quantitative linguistics. Other topics included the measurement of entropy as an information-theoretical measure based on phoneme frequencies (cf. Piotrowsky 1969). The functional load of phonemes (cf. Rischel 1962) was discussed from a quantitative point of view. Weiss (1962) reported on phoneme frequencies in Swedish, an investigation mainly motivated by applied aspects (speech therapy). Other "applied" works dealt with experimental-phonetic questions on the speed of speech in syllables and words spoken in speeches given in the Hungarian Parliament (cf. Nosz 1964), where the motivation for investigating this was stenographical[1] issues.

Beyond that phonetic-phonological focus, however, no particularly strong focus of content can actually be detected based on the statistical contributions published in *SMIL*. Among other things, "classical" questions of stylometry (cf. Anttila 1963, who refers to the different distribution of indigenous and borrowed lexemes in Early Modern English), of automatic speech recognition (with the help of multivariate procedures, see Mustonen 1965) and those of "language mixture" were presented. The latter aspect was dealt with by a well-known representative of quantitative linguistics, namely Gustav Herdan (1897–1968), who attempted to investigate (cf. Herdan 1963) the degree to which texts are influenced by other languages with the help of the frequency of initial letters of a word occurring in particular text samples. One of the few contributions to syntactic analysis with the help of statistical methods was made by Uhlířová (1969), an important representative of Czech quantitative linguistics, who continued to deal intensively with this question.

The first years of *SMIL* (1962–1965) provided good insight into the statistical and quantitative linguistics of the 1960s. Although the organisational centre of the journal was in Sweden and in the hands of Hans Karlgren, *SMIL* succeeded in providing a platform for international authors who played an important role later on in statistical linguistics. From 1964 onwards, statistical linguistic research has been

[1] Interestingly enough, Hans Karlgren was also occupied as a stenographer in the Swedish Parliament in the beginning of his career. Maybe his interest in the statistical analysis of language was triggered there.

institutionalised in Stockholm in the form of the Research Group for Quantitative Linguistics (*KVAL – Forskningsgruppen för Kvantitativ Lingvistik*). In *SMIL* 4, 1965, the imprint refers to the fact that contacts were also established with the Mathematical Society of Japan and that Mizutani Sizuo became one of the co-editors of *SMIL*. With regard to *SMIL* and KVAL, Hammarström (2012: 84) reports that Hans Karlgren chose the respective titles or abbreviations in Swedish with care and supposedly also with some ironic purpose; while *SMIL* can be interpreted as "smile" in English, KVAL – as in the German *Qual* – is to be read as "pain". When one considers how time-consuming and financially expensive any statistical evaluation with the help of computers was at that time, then humour, irony, and perseverance were certainly good companions of statistical linguistic research. Another point worth mentioning is that *SMIL* and its editors (Ferenc Kiefer may have played a major role as co-editor in this respect) sought, as already pointed out above, close contact with Eastern European and Soviet colleagues from the very beginning. This is evident not only from the bibliographical references to works from this field, but also from the fact that, for example, several synoptic works by Soviet colleagues, especially from the field of formal mathematical linguistics (e.g. Šrejder 1971, Rozentsveig 1971), appeared in *SMIL* 7, 1971.

At this point it should be mentioned that, contrary to a prior announcement, *SMIL* did not appear from 1965 to 1969 at all. The editorial from 1969 (cf. Karlgren 1969a) notes that there were financial reasons for this, but also problems with the acquisition of explicitly "statistical" works. However, Karlgren (1969a: 2) nevertheless "decided to make another attempt to fulfil the promises to regularly bring out a publication on statistical methods in linguistics" and he also specified that "statistical methods must not be understood as opposed to mathematical methods but as a subset of these", and that generally there is no strict demarcation line between these two disciplines.

This is reflected in the works published in *SMIL* from 1970 onwards. The high number of reviews by the editors themselves (H. Karlgren and F. Kiefer) published from the 1970s onwards is striking. Furthermore, the editorial of 1972 states that "[...] some of the papers treat problems which are not conspicuously statistical in nature. We do not regret this; in fact, it is one of our major points that there is no sharp demarcation between statistical and other mathematical linguistics" (Karlgren 1972a: 3). In fact, this shows that successive statistical-quantitative works are indeed more and more in the background of *SMIL*. Nevertheless, Karlgren (1972b) presented current works that deal with Markov models in linguistics, or he refers to works that deal with the incipient automated creation of concordances (cf. Karlgren 1972c). What is noticeable when reading the papers that appeared in the 1970s in *SMIL* is an increasing perspective towards application, although the contribution by Szanser (1973), for example, provides interesting quantitative insights into the quantitative structure of paragraphs, where the problem of a theoretical frequency distribution of the length of sentences in paragraphs is addressed.

SMIL 11, 1975, contains one of the few explicitly theoretical contributions on the question of the epistemological orientation of quantitative linguistics in general. This is the contribution by Hans Karlgren entitled *Quantitative Models – of What?* (Karlgren 1975b), which deals explicitly with the question of what status quantitative methods can have in contemporary linguistics. What is noticeable – at least one can read this between the lines – is a certain disillusionment[2] or ambivalence about the significance of quantitative methods. In general, however, the right questions are asked and the proper keywords for a certain kind of quantitative linguistics are provided. It is said that quantitative approaches promote "thinking in hypotheses" (examples are given that deal with the similarity of languages or authorship determination) and that an independent description of quantitative phenomena does indeed produce exciting results per se. As an example, the study of lexical frequencies, including the length of units, or word abbreviations in the context of language change are named, which need to be, according to the author's point of view, interpreted in terms of modern information theory. At the same time, however, the fear is expressed that an explanation of these phenomena based on communication theory falls short and that other[3] explanations should therefore also be considered. In any case, quantitative methods are seen as having great potential for achieving generalizations, but at the same time it is emphasized that quantification is a reduction of phenomena. Furthermore, it is generally stated rather pessimistically that a "shift of interests" has taken place in linguistics and that quantitative linguistics "treats problems which are no longer in vogue" (Karlgren 1975: 29). In the same issue from 1975, however, one finds two – methodologically seen – publications of quite high quality. Powers (1975) analyses active/passive constructions in English with the help of Bayesian statistics. Lee/Ross (1975) present an interesting contribution on the order of monosyllabic and polysyllabic words in texts, which they contrast with a theoretical card-shuffling model for the word length distribution.

With *SMIL* 12, 1976, it became clear that the journal was giving up its focus on statistical-quantitative linguistics (which had never actually really taken place), stating that "Our journal, once dedicated to statistical methods in linguistics, has successfully widened its scope and from next year the scope will be officially defined as computational linguistics in general" (Karlgren 1976a: 3), making statistical methods a sub-discipline and auxiliary discipline of computational linguistics. In order to meet the new requirements, *SMIL – Statistical Methods in Linguistics* was renamed *SMIL Quarterly: Journal of Linguistic Calculus* in 1976. From 1977 onwards, the journal appeared quarterly, albeit with a clear focus on computational linguistics, and in 1981 it finally ceased publication.

As an irony of history, it should be mentioned that the last "genuinely" statistical contribution in *SMIL* 12 is Muller (1976), who gives a good overview of the state of quantitative lexical studies, not only

[2] Interestingly enough in this paper Karlgren (1962) is not mentioned, where already a „modern" outline of statistical linguistics as autonomous discipline has been proposed. There also some ideas about the relevance of (text) statistics for diachronic problems are given.

[3] The contributions of *SMIL* certainly deserve the predicate international, but there is no evidence that, for example, the text by Altmann (1972) on the status and aims of quantitative linguistics, which is important from today's point of view, would have been received at the same time. There, the advantages of measurement at different scale levels were discussed, and it was said that statistical procedures should not only be used inductively, but that deductive hypotheses are also of interest (an aspect that G. Altmann later developed more precisely). In particular in Altmann's paper the idea of uncovering latent dependencies, which was later to lead to the formulation of a synergetic view, is also given.

with regard to French, but he also refers in general to mean frequency, repetition rate, to statistical modelling of vocabulary richness etc., and thus summarises well the theoretical state of the art in quantitative lexicology of the 1970s. In addition, Muller reflects on many future tasks of statistical linguistics, which he sees primarily in depicting linguistic facts at the model level or in pointing out discrepancies between model and text.

In summary, the following can be said. The journal *SMIL* (1962–1976) brought together a rather heterogeneous linguistic spectrum in the 12 years of publication examined here. This oscillates between application, utilisation of statistical procedures, the emerging formal mathematical linguistics and also application-oriented computational linguistics. The contributions to statistical/quantitative linguistics form a relatively small part of the total number of published articles, but are nevertheless to some extent representative of the state of affairs in the 1960s and early 1970s. This period is accompanied by a permanent search for the "essence" of quantitative linguistics, which is ultimately understood by the editor Hans Karlgren in the sense of an integrative approach, i.e. in fact a reduction to the application of statistical methods as a subfield of computational linguistics. This approach has, to a certain extent, become entrenched, since statistics are nowadays used in many sub-fields of linguistics. At the same time, from a contemporary perspective, the contributions show the beginning of an independent quantitative linguistics as we know it today. In any case, *SMIL* is one of the first publication forums that actually had a designated focus on statistical/quantitative methods only. Thus, at least in terms of intention, the journal succeeded in overcoming the particularism in quantitative linguistic research of that time. Moreover, it is clear that Hans Karlgren succeeded in acting beyond the Scandinavian area and skilfully stimulated international exchange in the field of statistical and quantitative linguistics.

References

Altmann, G. (1972). Status und Ziele der quantitativen Sprachwissenschaft. In: Jäger, S. (ed.). *Linguistik und Statistik*, pp. 1-9. Braunschweig: Vieweg. (= Schriften zur Linguistik, 6)

Altmann, G.; Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie.* Bochum: Brockmeyer. (= Quantitative Linguistics, 7).

Hammarström, G. (2012). *Memories of a linguist 1940-2010.* Muenchen: Lincom Europa. (= Linguistics edition, 85)

Karlgren, H. (1962). Die Tragweite lexikalischer Statistik. *Språkvetenskapliga Sällskapets i Uppsala färhandlingar* 27, pp. 77-108.

Bibliography of published articles and reviews (1962-1976)

SMIL 1, 1962

Karlgren, H. (1962a). Editor's note. *Statistical Methods in Linguistics*, 1, p. 5.

Karlgren, H. (1962b). Microfiche section. *Statistical Methods in Linguistics*, 1, pp. 6-7.

Karlgren, H. (1962c). Bibliographical service. *Statistical Methods in Linguistics*, 1, p. 8.

Karlgren, H. (1962d). The Scandinavian Symposium on Statistical Linguistics. *Statistical Methods in Linguistics*, 1, pp. 9-12.

Rischel, J. (1962). On functional load in phonemes. *Statistical Methods in Linguistics,* 1, pp. 13-23.

Ulvestad, B. (1962). On the use of transitional probability estimates in programming for mechanical translation. *Statistical Methods in Linguistics*, 1, pp. 24-40.

Weiss, M. (1962). Über die relative Häufigkeit der Phoneme des Schwedischen. *Statistical Methods in Linguistics*, 1, pp. 41-55.

SMIL 2, 1963

Karlgren, H. (1963a). Editor's note. *Statistical Methods in Linguistics,* 2, p. 69.

Karlgren, H. (1963b). Microfiche section. *Statistical Methods in Linguistics*, 2, pp. 70-71.

Karlgren, H. (1963c). Bibliographical service. *Statistical Methods in Linguistics,* 2, p. 72.

Anttila, R. (1963). Loanwords and statistical measures of style in the Towneley Plays. *Statistical Methods in Linguistics*, 2, pp. 73-93.

Sigurd, B. (1963). A note on the number of phonemes. *Statistical Methods in Linguistics*, 2, pp. 94-99.

Gyldén, Y. (1963). Rational construction of trade marks. *Statistical Methods in Linguistics*, 2, pp. 100-109.

Herdan, G. (1963). A method for the quantitative analysis of language mixture. *Statistical Methods in Linguistics*, 2, pp. 110-123.

Karlgren, H. (1963d). News on Scandinavian research. *Statistical Methods in Linguistics*, 2, pp. 124-125.

SMIL 3, 1964

Karlgren, H. (1964a). Editor's note. *Statistical Methods in Linguistics*, 3, p. 5.

Karlgren, H. (1964b). Bibliographical service. *Statistical Methods in Linguistics*, 3, pp. 6-7.

Kiefer, F. (1964). Some aspects of mathematical models in linguistics. *Statistical Methods in Linguistics*, 3, pp. 8-26.

Nosz, G. (1964). Redegeschwindigkeit in Silben und Worten pro Zeiteinheit. *Statistical Methods in Linguistics*, 3, pp. 27-42.

Kanger, St. (1964). The notion of a phoneme. *Statistical Methods in Linguistics,* 3, pp. 43-48.

Brodda, B., Karlgren, H. (1964). Relative positions of elements in linguistic strings. *Statistical Methods in Linguistics*, 3, pp. 49-101.

Karlgren, H. (1964c). News on recent research: The Kval-group conference. *Statistical Methods in Linguistics*, 3, p. 102.

SMIL 4, 1965

Abraham, S., Kiefer, F. (1965). An algorithmic definition of the morpheme. *Statistical Methods in Linguistics*, 4, pp. 4-9.

Holstein, A. P. (1965). A statistical analysis of schizophrenic language. Preliminaries to a study. *Statistical Methods in Linguistics*, 4, pp. 10-14.

Mackay, A. (1965): On the type-fount of the Phaistos disc. *Statistical Methods in Linguistics,* 4, pp. 15-25.

Marcus, S. (1965). Sur un ouvrage de Stig Kanger, concernant le phonème. *Statistical Methods in Linguistics*, 4, pp. 27-36.

Mustonen, S. (1965). Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics*, 4, pp. 37-44.

Siméonoff, E. (1965). On the distributions of the "costs" of combinations of k letters in a written language. *Statistical Methods in Linguistics*, 4, pp. 45-50.

Stene, J. (1965). [Rev.] Alvar Ellegård, A statistical method for determining authorship. The Junius Letters, 1769-1772. 115 pp. + XLVII. Gothenburg Studies in English. No. 13. Gothenburg, 1862. *Statistical Methods in Linguistics*, 4, pp. 51-53.

Karlgren, H. (1965). SMIL. Meždunarodnyj žurnal po statističeskim metodam i lingvistike. *Statistical Methods in Linguistics*, 4, pp. 54-56.

SMIL 5, 1969

Karlgren, H. (1969a). Editorial. *Statistical Methods in Linguistics*, 5, p. 2.

Brodda, B., Karlgren, H. (1969). Synonyms and synonyms of synonyms. *Statistical Methods in Linguistics,* 5, pp. 3-17.

Uhlířová, L. (1969). On statistical experimenting in syntax. *Statistical Methods in Linguistics*, 5, pp. 18-33.

Piotrowsky, R. (1969). Entropy and redundancy in four European languages. *Statistical Methods in Linguistics*, 5, pp. 34-35.

Sharf, D. J., Baehr, T. J. (1969). Quantitative analysis of articulation correspondences. *Statistical Methods in Linguistics,* 5, pp. 36-43.

Anttila, R. (1969). Sound preference in alliteration. *Statistical Methods in Linguistics*, 5, pp. 44-48.

Sgall, P. (1969). A multilevel generative description of language. *Statistical Methods in Linguistics*, 5, pp. 49-58.

Karlgren, H. (1969b). Announcement. *Statistical Methods in Linguistics*, 5, pp. 59-60.

SMIL, 6, 1970:

Karlgren, H. (1970a). Editorial. *Statistical Methods in Linguistics*, 6, p. 2.

Brodda, B. (1970). Document retrieval - A "topological" problem. *Statistical Methods in Linguistics*, 6, pp. 3-14.

Dolby, J. L. (1970). On word decomposition and semantic relatedness. *Statistical Methods in Linguistics*, 6, pp. 15-22.

Ladefoged, P. (1970). The measurement of phonetic similarity. *Statistical Methods in Linguistics*, 6, pp. 23-32.

Schank, R. C., Tesler, L. (1970). A conceptual dependency parser for natural language. *Statistical Methods in Linguistics,* 6, pp. 33-51.

Szanser, A. J. (1970). Automatic error-correction in natural languages. *Statistical Methods in Linguistics*, 6, pp. 52-59.

Verburg, P. A. (1970). Hobbes' calculus of words. *Statistical Methods in Linguistics*, 6, pp. 60-65.

SMIL 1971, 7

Karlgren, H. (1971a). Editorial. *Statistical Methods in Linguistics*, 7, p. i.

Karlgren, H. (1971b). Similarity between trade mark devices. *Statistical Methods in Linguistics*, 7, p. 1-20.

Stanley, P. (1972). The use of context-sensitive rules in immediate constituent analysis. *Statistical Methods in Linguistics,* 8, pp. 21-31.

Rozentsveig, V. Y. (1971). Models in Soviet linguistics. *Statistical Methods in Linguistics*, 7, pp. 32-49.

Šrejder, J. A. (1971). Basic trends in the fields of semantics. *Statistical Methods in Linguistics*, 7, pp. 50-59.

Kiefer, F. (1971). Rev. Zellig S. Harris: Papers in Structural and Transformational Linguistics, Formal Linguistics Series, Vol. 1., D. Reidel Publishing Company/Dordrecht, Holland, 1970. *Statistical Methods in Linguistics*, 7, pp. 60-62.

Karlgren, H. (1971c). Rev. E. Zierer, The Theory of Graphs in Linguistics, Janua Linguarum, Series Minor 94, Mouton, The Hague & Paris, 1970. *Statistical Methods in Linguistics*, 7, pp. 63-68.

Karlgren, H. (1971d). Rev. A. Juilland and H. H. Lieb, „Klasse“ und Klassifikation in der Sprachwissenschaft, Janua Linguarum, Series Minor No. 74, The Hague & Paris, 1968. *Statistical Methods in Linguistics*, 7, pp. 69-76.

Karlgren, H. (1971e). Rev. Mongé, Alf and Landsverk, O. G., Norse medieval cryptography in Runic Carvings, The Norseman Press, Glendale, 1967. Pp. 224, illustrations. *Statistical Methods in Linguistics*, 7, pp. 77-83.

SMIL 8, 1972

Lindh, E.-C. (1972). To our subscribers. *Statistical Methods in Linguistics*, 8, p. i.

Karlgren, H. (1972a). Editorial. *Statistical Methods in Linguistics*, 8, pp. 3.

Dostert, B. H., Thompson, F. B. (1972). Syntactic analysis in REL English. *Statistical Methods in Linguistics*, 8, pp. 5-39.

Engström, G. (1972). Automatic phonemization in practice. *Statistical Methods in Linguistics*, 8, p. 39-55.

Ishiwata, T. (1972). Méthode pour résoudre l’ambiguité dans le traitement automatique. *Statistical Methods in Linguistics*, 8, pp. 56-63.

Klein, Sh., Oakley, J. D., Suurballe, D. J., Ziesemer, R. A. (1972). A program for generating reports on the status and history of stochastically modifiable semantic models of arbitrary universes. *Statistical Methods in Linguistics*, 8, pp. 64-93.

Kiefer, F. (1972). Rev. Brainerd, Barron, Introduction to the mathematics of language study, mathematical linguistics and automatic language processing, American Elsevier Publishing Co., New York, 1971, pp. 313. *Statistical Methods in Linguistics*, 8, pp. 94-95.

Karlgren, H. (1972b). Rev. Damerau, Frederick J., Markov models and linguistic theory, Mouton, The Hague, 1971, pp. 196. *Statistical Methods in Linguistics*, 8, pp. 96-101.

Karlgren, H. (1972c). Rev. Finkenstaedt, T., Leisi, E., Wolff, D., A chronological English dictionary. Carl Winter Universitätsverlag, Heidelberg, 1970, xvi, pp. 1395. *Statistical Methods in Linguistics*, 8, pp. 102-106.

Karlgren, H. (1972d). Rev. Rolf Gavare, Graph description of linguistic structures, Almqvist & Wiksell, Stockholm 1972, pp. 96. *Statistical Methods in Linguistics*, 8, pp. 107-108.

Karlgren, H. (1972e). Rev. Meninger, Karl: Number words and number symbols. A cultural history of numbers: M.I.T. Press, 2nd printing 1970 (trans. of revised German edition 1958, pp. pp. 480. *Statistical Methods in Linguistics*, 8, pp. 109-110.

Kiefer, F. (1972). Smaby, Richard M., Paraphrase Grammars, Formal linguistic series, Vol. 2. D. Reidel Publishing Co., Dordrecht-Holland, 1971, viii + 145 pp. *Statistical Methods in Linguistics*, 8, pp. 111-114.

Karlgren, H. (1972f). Rev. Smith, Raymond G., Speech Communication: Theory and Models, Harper & Row, New York et al., 1970, pp. 230. *Statistical Methods in Linguistics*, 8, pp. 115-117.

SMIL 9, 1973

Karlgren, H. (1973a). Editorial. *Statistical Methods in Linguistics*, 9 p. i.

Brodda, B. (1973). Some classes of solvable categorial expressions. *Statistical Methods in Linguistics*, 9, pp. 5-41.

Kiefer, F. (1973). A propos derivational morphology. *Statistical Methods in Linguistics*, 9, pp. 42-59.

Ureland, St. (1973). Nominalized complements occurring after Swedish “Höra” (‘hear’). *Statistical Methods in Linguistics*, 9, pp. 60-78.

Szanser, A. J. (1973). A study of the paragraph structure. *Statistical Methods in Linguistics*, 9, pp. 79-90.

Karlgren, H. (1973b). Rev. Wunderli, Peter, Ferdinand de Saussure und die Anagramme, Linguistik und Literatur, Max Niemeyer Verlag, Tübingen 1972. *Statistical Methods in Linguistics*, 9, pp. 91-100.

SMIL 10, 1973

Karlgren, H. (1974a). Editor's Note. *Statistical Methods in Linguistics*, 10, p. i.

Karlgren, H. (1974b). Categorial grammar calculus. *Statistical Methods in Linguistics*, 10, pp. 1-128.

Kiefer, F. (1974c). Rev. Benny Brodda. Koverta kasus i svenska, Papers from the Institute of Linguistics, University of Stockholm 1973. *Statistical Methods in Linguistics*, 10, pp. 129-133.

Kiefer, F. (1974d). Rev. Sture Ureland, Verb complementation in Swedish and other Germanic languages, Skriptor, 1973. *Statistical Methods in Linguistics*, 10, p. 134-139.

SMIL, 11, 1975

Karlgren, H. (1975a). Editorial. *Statistical Methods in Linguistics*, 11, p. i.

van Meerten, R. J. (1975). On the printing direction of the Phaistos disc. *Statistical Methods in Linguistics*, 11, pp. 5-24.

Karlgren, H. (1975b). Quantitative Models – of What? *Statistical Methods in Linguistics*, 11, pp. 25-31.

Powers, J. E. (1975). A Bayesian analysis of linguistic data. *Statistical Methods in Linguistics*, 11, pp. 32-50.

Lee, K. S., Ross, D. (1975). A card-shuffling model for the distribution of monosyllabic and polysyllabic English words. *Statistical Methods in Linguistics*, 11, pp. 51-63.

Karlgren, H. (1975c). Rev. Index Thomisticus. *Statistical Methods in Linguistics*, 11, pp. 64-67.

SMIL 12, 1976

Karlgren, H. (1976a). Editorial. Computational linguistics. *Statistical Methods in Linguistics*, 12, pp. 3-4.

Fillmore, Ch. J. (1976). The need for a frame semantics within linguistics. *Statistical Methods in Linguistics*, 12, pp. 5-29.

Hajičová, E. (1976). Question and answer in linguistics and man-machine communication. *Statistical Methods in Linguistics*, 12, pp. 30-46.

Joshi, A. K., Rosenschein, St., J. (1976). Some problems of inferencing: relation of inferencing to decomposition of predicates. *Statistical Methods in Linguistics*, 12, pp. 47-70.

Sheil, B. A. (1976). Observations on context free parsing. *Statistical Methods in Linguistics*, 12, pp. 71-109.

Thompson, H. (1976). Towards a model of language production: Linguistic and computational foundations. *Statistical Methods in Linguistics.* 12, pp. 110-126.

Vauquois, B. (1976). Automatic translation – A survey of different approaches. *Statistical Methods in Linguistics*, 12, pp. 127-135.

Muller, Ch. (1976). Some recent contributions to statistical linguistics. *Statistical Methods in Linguistics*, 12, pp. 136-147.

Karlgren, H. (1976b). SMIL Quarterly Journal of Linguistic Calculus. *Statistical Methods in Linguistics*, 12, p. 149.