Current Issues in Linguistic Theory 356

Language and Text

Data, models, information and applications

EDITED BY
Adam Pawłowski
Jan Mačutek
Sheila Embleton
George Mikros

JOHN BENJAMINS PUBLISHING COMPANY

LANGUAGE AND TEXT

CURRENT ISSUES IN LINGUISTIC THEORY

AMSTERDAM STUDIES IN THE THEORY AND HISTORY
OF LINGUISTIC SCIENCE – Series IV
ISSN 0304-0763

Current Issues in Linguistic Theory (CILT) is a theory-oriented series which welcomes contributions from scholars who have significant proposals that advance our understanding of language, its structure, its function and especially its historical development. CILT offers an outlet for meaningful contributions to current linguistic debate.

A complete list of titles in this series can be found on *benjamins.com/catalog/cilt*

Volume 356

Adam Pawłowski, Jan Mačutek, Sheila Embleton and George Mikros (eds.)

Language and Text. Data, models, information and applications

LANGUAGE AND TEXT

DATA, MODELS, INFORMATION AND APPLICATIONS

Edited by

ADAM PAWŁOWSKI
University of Wrocław

JAN MAČUTEK
Mathematical Institute of Slovak Academy of Sciences & Constantine the Philosopher University in Nitra

SHEILA EMBLETON
York University

GEORGE MIKROS
Hamad Bin Khalifa University

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM & PHILADELPHIA

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/cilt.356

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2021041998 (PRINT) / 2021041999 (E-BOOK)

ISBN 978 90 272 1010 4 (HB)
ISBN 978 90 272 5838 0 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Part II. Empirical studies

Introduction

Adam Pawłowski[1], Sheila Embleton[2], Jan Mačutek[3,4] and George Mikros[5]
[1]University of Wrocław / [2]York University / [3]Mathematical Institute of Slovak Academy of Sciences / [4]Constantine the Philosopher University in Nitra / [5]Hamad Bin Khalifa University

The volume that we present here to the reader is unique for several reasons. It was created as the outcome of the 10th QUALICO 2018 conference of quantitative linguistics – organized for the first time in Poland (Wrocław), and at the same time the last before the outbreak of the pandemic, which has slowed down and complicated scientific life around the globe.

The theme of the conference "Information in language: Coding, extraction and applications" reflects the changes that have taken place in quantitative research on language in recent years. Although traditional areas, such as stylometry, quantitative dialectology, exploration of statistical laws of language, etc., are still very much present, new currents are increasingly developing. At the data level, this is evident in the domination of ever larger and more specialized text corpora, created and made available successively by repositories, libraries and media companies (e.g., repositories of scientific or media texts, corpora of large bibliographies). There is also a slow decline in quantitative research on fiction, which is being replaced by applied texts and those created by users in social media (Twitter, Facebook, various discussion lists, customers' opinions, etc.). The changes are also visible at the methodological level, where more and more complex data analysis algorithms are used to explore the semantic layer of language.

The reason for this slow but visible change is the evolution of the information landscape in the economy and culture of the entire world. There is a flood of electronic texts, unprecedented in history, which triggers the need for automatic analysis, covering not only the form, but above all semantics of language. Human beings are no longer able to follow the information flows – help in data extraction and analysis is brought by artificial intelligence systems, based on quantitative methods. But this change also has another aspect. People – regardless of culture and historical period – have always sought in their environment and universe elements

https://doi.org/10.1075/cilt.356.int

of order, arrangement, and sometimes even deeper meaning. However, the flood of information creates an impression of chaos, over which a human being loses control. The methods of quantitative text analysis and artificial intelligence develop so dynamically precisely because they allow language users to better organize mass information processes and restore a sense of order at the cognitive level.

The arrangement of articles in the volume is dichotomous and, in some sense, reproduces the division mentioned above. The first, theoretical, part of the book consists of 8 chapters. For most of these contributions, the 'common denominator' is developing general (as opposed to language specific) mathematical models for language units and their properties, and testing their appropriateness (in terms of goodness of fit) on empirical data. The language units and properties include syllables, words, valency, grammar efficiency and complexity, as well as word order. All chapters in this section present new and very recent findings. They either generalize well-established concepts and ideas (e.g., Zipf's least effort principle, or the Menzerath-Altmann law, according to which larger wholes tend to consist of smaller parts) or apply them to completely new language material. The chapters from the theoretical section are not directly aimed at applications, but the models under consideration (and their parameters) still have, at least in a longer time horizon, the potential to contribute also to different applications in linguistics, such as text classification, authorship attribution, etc.

In the second, empirical, part of the book, there are 9 contributions. Although all of them focus on big data research at the lexical and partially phonetic level, they address several aspects of natural language analysis in the digital world. Epistemological aspects of big data research are the subject of one contribution ("The perils of Big Data"), while other texts focus on more detailed aspects, such as generating random texts using GAN (Generative Adversarial Networks) algorithm, automatic analysis of political language, comprehensive statistical analyses of large bibliographies (including a topic modelling method), automatic taxonomies of texts, automatic gender and genre recognition, authorial attribution, as well as linguistic data visualization.

Our first chapter, "On the impact of the initial phrase length on the position of enclitics in Old Czech" was written by Radek Čech, Pavel Kosek, Olga Navrátilová, and Ján Mačutek. In this paper, the authors present the results of an analysis of the most frequent pronominal enclitics in Old Czech, examining data from the oldest Czech Bible translation. They set up a hypothesis considering a relationship between the length of the initial syntactic phrase in a clause and the occurrence (or its absence) of the enclitic after this phrase. Data analysis revealed a negative correlation, i.e., the longer the phrase, the lower the proportion of enclitics in the post-initial position, although this tendency is not followed by all enclitics.

"Term distance, frequency, and collocations" is by Lars Johnsen. The chapter explores two different quantitative methods for extracting collocations from a corpus (one based on the frequency of terms and the other based on their distance). The two methods described have been tested on the text collection of digitized texts from the Norwegian National Library (nearly 440,000 books). All the collocations used were taken from a set of one thousand books that contain at least one occurrence of the target word. Both methods present a small computational cost since they don't need to calculate a reference statistic. Moreover, the distance-based methods seem theoretically appealing, and further research could give new insights about its exploitation in collocation analysis.

The analysis and comparison of linear sequences without any kind of prior knowledge is presented in "A method for comparison of general sequences via type-token ratio" by Vladimír Matlach, Diego Gabriel Krivochen, and Jiří Milička. The authors propose a general-purpose sequence analysis method that is generic enough to handle any type of digitized, linear, and discrete sequences and depends only on different single symbols with no need to utilize prior knowledge such as word boundaries or the existence of other units. Moreover, the proposed technique uses simple quantitative linguistics concepts (Type-to-Token Ratio) and produces results that can be used effectively for visualization and further processing using machine learning methods. The application of this method in real texts produced cohesive clustering solutions by grouping texts by language or type (random sequences, DNA, or source-codes).

A cross-linguistic study of the syllable is the focus of another interesting chapter, "Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian", by Biljana Rujević, Marija Kaplar, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović, and Ján Mačutek. The paper investigates some basic quantitative properties of the syllable in four Slavic languages: Croatian, Serbian, Russian, and Ukrainian. Both syllable frequencies and syllable length have been modeled using the Zipf-Mandelbrot distribution and the Dacey-Poisson distribution, respectively, following relevant word frequency and length modeling approaches. Moreover, the authors suggest a generalization of the Menzerath-Altmann law, which can also model the relation between word length and the mean syllable length in texts explaining the distribution of both linguistic levels through a unified quantitative expression.

The next chapter, entitled "*N*-grams of grammatical functions and their significant order in the Japanese clause", written by Haruko Sanada, investigates the statistically significant order of valency types (complements and adjuncts) in Japanese clauses by employing *n*-gram frequency data of valency types. One of the main results of this study is that the time and the place appear between the subject and

object with statistical significance. Moreover, the subject and object play the role of 'anchors' in the clause. Lastly, the occasion takes a position before the subject, between the subject and object, or after the object, giving the impression that Japanese is a free word order language.

The next chapter, "Linking the dependents: Quantitative-linguistic hypotheses on valency" by Petra Steiner, proposes a model that links the syntactic and semantic aspects of the case and valency to morphological case and other means of word-formation. Based on this model, the author forms the following two distinct research hypotheses: (a) the larger the number of arguments of a semantic predicate, the larger is the number of the syntactic arguments of the realized syntactic constructs, and (b) the larger the number of semantic arguments, the larger is the tendency for shortening on the syntactic level. Both hypotheses are confirmed empirically using the FrameNet database.

Relja Vulanović investigates the topic of the evaluation of grammar efficiency in his chapter "Grammar efficiency and the One-Meaning–One-Form Principle". The author proposes a new formula for evaluating grammar efficiency, which is more straightforward than previous formulas used for this task. Part of the new approach's simplicity is the inclusion of a new version of the measure of how much a linguistic system departs from the One-Meaning–One-Form Principle.

Makoto Yamazaki investigates the frequency distribution of the most common words across different texts in Japanese in his chapter "Distribution and characteristics of commonly used words across different texts in Japanese". Using the Balanced Corpus of Contemporary Written Japanese, he found that the distribution resembles Zipf's law with the differentiation that the curve always begins to increase shortly before the degree of commonality reaches its maximum. Moreover, the distribution trend is not affected either by the length or by the number of texts. Additionally, as the text length increases, the number of commonly used words also increases linearly.

Big Data in scientific research and, more specifically, in linguistic fieldwork research such as dialectometry is the focus of "The perils of Big Data", by Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler. The authors describe the advantages and the pitfalls of research in the era of Big Data. Their main conclusions are that researchers should care about the reliability and the validity of their methods and computational tools. Moreover, they should work on theoretically and empirically motivated sub-parts of their data pools. Whenever they use language resources developed by others, they should use them cautiously and understand both their theoretical and methodological framework.

The quantitative discrimination of natural and computer-generated random texts is the topic of "From distinguishability to informativity: A quantitative text

model for detecting random texts", by Maxim Konca, Alexander Mehler, Daniel Baumartz, and Wahed Hemati. The authors conduct several supervised classification experiments using features inspired by quantitative linguistics research on vocabulary distribution. The results show that the current random text models still generate easily distinguishable texts from non-random counterparts. Moreover, the discriminatory power of the classification models is based on a small set of relatively simple quantitative text characteristics.

The development of a novel method for calculating text readability is the focus of "A Modern Greek readability tool: Development of evaluation methods", by George Mikros and Rania Voskaki. The authors develop an automatic readability analysis tool that focuses on Modern Greek as a foreign language. The tool developed uses several stylometric indices inspired by work done in the field of quantitative linguistics that train Random Forests, a robust machine learning classification algorithm for predicting the reading level of texts. Its performance surpasses all previous readability tools for Modern Greek and creates a new state-of-the-art in readability detection in this language. Further analysis of the results with advanced visualization methods reveals the complex and fluid dynamics of the features used and their readability predictions.

The effects of the phonological structure of the Czech language on the online popularity that a blog post is getting is investigated by Jiří Milička and Alžběta Houzar Růžičková in their chapter on "Phonological properties as predictors of text success". The authors count several phonological natural classes and use their quantitative patterns as predictors of text success (measured as likes per view). The text success can be predicted satisfactorily using the beauty-in-averageness effect and the euphony principle as specific vowels and phoneme classes correlate with the popularity of online texts.

The stylometric analysis of the texts produced by candidates for the 2018 Czech presidential election is the focus of "Calculating the victory chances: A stylometric insight into the 2018 Czech presidential election" by Michal Místecký. The author used a wide range of stylometric features and keywords to profile the candidates' discourses quantitatively. The findings show that each candidate adopts a unique strategy to influence his electorate and that this strategy may be captured via stylometric methods.

The need to develop visualization methods for non-linear high-dimensional data abstracted from historical corpora is addressed in "Topological mapping for visualization of high-dimensional historical linguistic data", by Hermann Moisl. The author proposes topological mapping, a non-linear visualization method coupled with a Self-Organizing Map, a specific topological mapping technique to plot and analyze typological characteristics of a small historic text corpus.

Text classification research directed to the author's profiling and text genre is discussed in "Book genre and author's gender recognition based on titles: The example of the bibliographic corpus of microtexts" written by Adam Pawłowski, Elżbieta Herden, and Tomasz Walkowiak. The authors use word embedding features (word2vec, FastText) to develop supervised classification models applied successfully to a corpus of Polish microtexts (book titles derived from the Polish national bibliography).

The quantitative comparison of a corpus based on book titles and a general language corpus (both in Polish) is the aim of "Quantitative analysis of bibliographic corpora: Statistical features, semantic profiles, word spectra," by Adam Pawłowski, Krzysztof Topolski, and Elżbieta Herden. The authors compare a wide range of quantitative properties of the two corpora (word frequency distributions, parts of speech frequencies, word spectra, etc.) and find considerable differences between them.

Genre classification based on syntactic features (dependency types) is the focus of our last chapter, "Analysis of English text genre classification based on dependency types" by Yaqin Wang. This study explores whether the dependency types can be used as a distinctive text vector for classifying English genres. The author experimented with three different classification methods, namely principal component analysis, hierarchical clustering, and random forest. The results obtained show that the dependency type effectively distinguishes text genres, especially between spoken genre and written genre.

Generally speaking, a quantitative approach and automatic data processing techniques dominate in most of these studies. The analyses described are relevant to our contemporary digital reality, and some of them open new research fields. All chapters are innovative and/or bring new insights into a wide range of text-mining, digital humanities, and quantitative linguistics research.

PART I

Theory and models

On the impact of the initial phrase length on the position of enclitics in Old Czech

Radek Čech[1], Pavel Kosek[2], Olga Navrátilová[2] and Ján Mačutek[3,4]
[1]University of Ostrava / [2]Masaryk University / [3]Mathematical Institute of Slovak Academy of Sciences / [4]Constantine the Philosopher University in Nitra

This paper presents an analysis of the relationship between the length of the initial phrase and the positions of pronominal enclitics in a clause. The hypothesis predicting the negative correlation between the length of the phrase and the proportion of enclitics in the post-initial position was set up and tested. For testing the hypothesis, selected books – Genesis (Gn), Isaiah (Is), Job (Jb), Sirach (Sir), Gospel of St. Matthew (Mt), Gospel of St. Luke (Lk), Acts (Act), and Revelation (Rev) – from the first edition of the Old Czech Bible translation were used. The hypothesis was not rejected; however, some differences among particular pronouns were revealed.

Keywords: enclitics, word order, initial phrase, length, Old Czech

1. Introduction

Enclitics are language units with specific phonological as well as syntactic behaviour. They are defined as unstressed words that are joined prosodically with the preceding word. However, a syntactic relationship between the enclitic and that preceding word is not necessary. These facts have a crucial impact on their word order position. Further, several classes of enclitics are determined in linguistics (e.g., auxiliary, pronominal, clausal) and each type displays some specificity (cf. Uhlířová et al. 2017). Moreover, the behaviour of an enclitic depends also on a character of its host (the word to which it is attached). For instance, one can find different word order characteristics of Czech enclitics joined to infinitive verbs, on the one hand, and to finite verbs, on the other (Toman 2000). Consequently, enclitics are considered a heterogeneous set of units which, however, share some common qualities

https://doi.org/10.1075/cilt.356.01cec

(Zwicky 1994). One of the possibilities for a better understanding of the properties of enclitics is an analysis of their diachronic development. It allows us to observe forces and mechanisms which have an impact on them and, subsequently, it could bring new ways of how to explain their status in a synchronic language system.

In this paper, we introduce the results of an analysis of the most frequent pronominal enclitics in Old Czech, namely, the enclitics *sě* (accusative reflexive), *mi* ("to me"), and *tě* ("you"). In the investigated texts, the reflexive forms *sě* and *tě* preserve some remnants of stressed (orthotonic) forms. This situation can be viewed as a symptom of the historical change from an unstable enclitic into a stable enclitic / enclitic tantum (Uhlířová et al. 2017). This process probably accelerated during the Old Czech period (Trávníček 1956: 147; Šlosar 1967: 252). As we have shown earlier, in the overwhelming majority of examples the investigated forms are characterized by the properties of an enclitic word (Kosek et al. 2018a). For illustration, we give examples of typical usage of *sě* (accusative reflexive), *mi* ("to me"), and *tě* ("you") in Old Czech in (1), (2), and (3), respectively (the enclitic is underlined).

(1) *Kde jest ten*, [*jenž <u>sě</u> jest narodil*,]
Who.LOC.F.SG REFL.ACC be.AUX.PRS.3SG born.PTCP.PST.M.SG
král židovský?
king.NOM.M.SG *Jewish*.NOM.M.SG
"Where is the one who has been born king of the Jews?"
The Olomouc Bible, Mt 2,2

(2) *nepodal-s <u>mi</u> vody*
Neg-give.PTCP.PRET.M.SG-be.AUX.PRS.2.SG me.DAT.SG water.GEN.F.SG
mým nohám
my.DAT.F.PL foot.DAT.F.PL
"You gave me no water for my feet" The Olomouc Bible, Lk 7,44

(3) *A v prosbě za Izmahele sem <u>tě</u>*
and in prayer.LOC.F.SG for Ishmael.ACC.M.SG be.AUX.PRS.1.SG you.ACC.SG
uslyšal
hear.PTCP.PRET.M.SG
"As for Ishmael, I have heard you" The Olomouc Bible, Gen 17,20

In this study, we set up a hypothesis considering a relationship between the length of the initial syntactic phrase in a clause and the occurrence (or its absence) of the enclitic after this phrase. The theoretical background of the hypothesis – namely, the longer the first phrase of the clause, the lower the probability of the occurrence of the enclitic after this phrase – puts together both prosodic and syntactic properties of observed phenomena. Specifically, the enclitic cannot appear after a pause. The longer the initial phrase, the higher the probability that a pause will occur after it, and consequently, the probability that the enclitic occurs after the initial phrase

decreases with increasing phrase length. This mechanism is valid only for the initial phrase which does not contain the governor of the enclitic. This phenomenon was described, among others, by Ertl (1924), Radanović-Kocić (1996: 435) who called it a "heavy constituent constraint", or by Ćavar and Wilder 1999: 443–444), who considered it a result of the "clitic third" rule.

For the analysis, data from the oldest Czech Bible translation are used. These data were chosen because this study is one of the first results of a larger project (Kosek et al. 2019) which aims at both describing and explaining word order characteristics of pronominal enclitics and their development from Old Czech to the present. In other words, we start with the oldest language material available in Czech because it represents a 'starting point' for our long-term research interest.

2. Classification of the word ordering of enclitics in Old Czech

A detailed descriptive analysis of pronominal enclitics in finite verb phrases reveals several tendencies of its positions in the Old Czech word order (Kosek et al. 2018a). Kosek et al. determine several word order positions of pronominal enclitics and, further, they found significant differences among frequencies of these forms in particular positions. For instance, a pronominal enclitic occurs most frequently after an initial syntactic phrase / the first word of the clause and, on the contrary, it never occurs as the first word of the clause. Two positions are dominant: (1) the post-initial position and (2) the contact position in the middle or at the end of a clause (which we call 'non-post-initial positions' for the sake of simplicity). These positions cover about 95% of all occurrences. As for the former, it is defined as the position after the initial phrase, which can be comprised of one or more words, as is illustrated in examples (4) and (5) (square brackets mark the initial phrase, the enclitic is underlined).

(4) [*Tehda*] *sě* *otevřě* *chrám* *boží*
then REFL.ACC open.FUT.3SG temple.NOM.M.SG god.ADJ.POS.NOM.M.SG
v nebi
in heaven.LOC.N.SG
"Then the temple of God in heaven was opened"
The Olomouc Bible, Rev 11,19

(5) [*Co*] *sě* *tobě* *vidí*, *Šimone?*
what.NOM REFL.ACC you.DAT see.PRAES.3.SG Simon.VOC.SG.M
"What do you think, Simon?" The Olomouc Bible, Mt 17,24

As for the non-post-initial position, it is defined as the position in which the enclitic does not occur in the second position and is immediately adjacent to its syntactically/morphologically superordinate item, see example (6), (7), and (8) (squared brackets mark the initial phrase, the superordinate item of the enclitic is bolded, and the enclitic is underlined).

(6) [*Ale mládenečky*] ***hnětla*** *sě* *v*
but children.INSTR.M.PL struggle.PART.PRET.ACT.F.SG REFL.ACC in
životě *jsúce*
womb.LOC.M.SG being.PART.PRAES.ACT.NOM.PL
"But the children struggled in her womb" The Olomouc Bible, Gen 25,22

(7) [*Volánie* *Sodomských* *a* *Gomorrejských*]
outcry.NOM.N.SG sodom.ADJ.GEN.M.PL and gomorrha.ADJ.GEN.M.PL
rozmnožilo *sě* ***jest***
multiply.PTCP.PRET.ACT.N.SG REFL.ACC be.AUX.PRET.3SG
"The cry of Sodom and Gomorrha is multiplied"
The Olomouc Bible, Gen 18,20

(8) [*Narozenie*] *pak* *našeho* *Jezukrista* *takto*
birth.NOM.N.SG then our.GEN.M.SG Jesus.GEN.M.SG Christ.GEN.M.SG so
sě ***jest*** ***stalo***
REFL.ACC be.AUX.PRET.3SG happen.PTCP.PRET.ACT.N.SG
"This is how the birth of Jesus the Messiah came about"
The Olomouc Bible, Mt 1,18

Because the purpose of the study is to analyze the strongest tendencies of word ordering of the enclitics as well as to try to find mechanisms that can explain them, we do not follow the rather fine-grained classification of Kosek et al. (2018a). For the testing of the hypothesis, we determine only two positions: (1) post-initial and (2) non-post-initial (for more details, see § 4).

3. Language material

We used the first edition of the Old Czech Bible translations. We chose the Bible text for the following reasons: (a) it is one of the oldest Old Czech prose texts (older texts from the first half of the 14th century are poetic and they cannot be used to observe word order characteristics typical for the Czech language of that time), (b) the results can be compared with the ones based on later Czech Bible translations which enables us to observe the historical development of the phenomena under study. According to Kyas (1997: 43) and Vintr (2008: 1883a), the complete Old Czech translation of the Bible probably dates from the 1350s and it is considered

to be a work of around ten anonymous translators. However, no autograph of the translation is known. The oldest version of the Old Czech Bible has survived in later copies: the Dresden Bible (*Bible drážďanská*, 1360s), the Litoměřice-Třeboň Bible (*Bible litoměřicko-třeboňská*, 1411–1414), and the Olomouc Bible (*Bible olomoucká*, 1417 – Kyas 1997: 57; Vintr 2008: 1883b). Unfortunately, none of these copies are entirely identical to the original version of the text. Specifically, the original text was slightly revised in later copies and, moreover, some parts of the original have been replaced by more recent translations – e.g., the Litoměřice-Třeboň and Olomouc Bibles include a different translation of the Gospel of Matthew (known as the Gospel of Matthew with homilies), and the Olomouc Bible incorporates some epistles from the Acts of the Apostles, which were taken from the second edition of the Old Czech Bible translation (Kyas 1997: 42, 61–62; Vintr 2008: 1883b). Further, the copies have not survived in their entirety. For instance, the Dresden Bible, which represents the oldest version, was unfortunately entirely destroyed during the First World War, and only part of the original text has survived in the form of photocopies and copies.

Faced with this state of affairs, we chose the Olomouc Bible as the main text for the investigation for the following reasons. It is the oldest known complete text of the Old Czech Bible and it is the text which forms the basis of a critical edition of the Old Czech Bible as conceived by Kyas (Kyas ed. 1981, 1985, 1988; Kyas et al. eds. 1996; Pečírková et al. eds. 2009). Because all observed phenomena had to be annotated manually, we used only a sample of the Bible. We attempted to select books which (1) as far as possible, differ in their text structure and style, and (2) were the work of different translators, in the view of Kyas (1997: 43) who identifies two distinct groups of translators on the basis of the Czech equivalents they used for specific Latin words such as *adorare, benedicere, benedictus*, etc. In this way, we attempted to at least partially compensate for the limitations inherent in analyzing the specific language of Bible texts. As a result, four books from the Old Testament and four books from the New Testament were chosen: Genesis (Gn), Isaiah (Is), Job (Jb), Sirach (Sir), Gospel of St. Matthew (Mt), Gospel of St. Luke (Lk), Acts (Act), and Revelation (Rev). Since the Acts of the Apostles were taken from the second edition of the Old Czech Bible translation in the Olomouc Bible, this text is taken from the Litoměřice-Třeboň Bible (meaning that the translation is from the first edition of the Old Czech Bible) into our sample.

4. Methodology

To test the hypothesis predicting a relationship between the length of the initial phrase and word ordering of the enclitic, first, we manually determined both the initial phrase and the position of the enclitic in each clause. Next, the length of the initial phrase was measured by the number of (a) letters and (b) words. Finally, for each value of length of the initial phrase, a proportion of the post-initial positions of the enclitic was calculated. However, for some values of length (especially for longer ones) of the initial phrase, we had too few occurrences of enclitics. Therefore, we pooled the data to get at least 5 occurrences of the enclitic in each group and, consequently, we computed the average length of the phrase in such a group. To take the frequency of the occurrence of the enclitic in a particular position into account, we computed average weighted lengths (ALi), where weights are represented by particular frequencies.[1]

5. Results

According to the hypothesis, the longer the initial phrase, the smaller the proportion of enclitics in the post-initial position. The length of the initial phrase was measured by the number of letters, see Table 1 and Figure 1–3. For all these enclitics, we can see a general tendency that is in accordance with the hypothesis. However, closer observation also reveals other important findings. First, there are obvious differences between the characteristics of *sě*, on the one hand, and *mi* and *tě*, on the other. Specifically, in the case of the shortest initial phrases, the proportion of the post-initial position of the enclitic *sě* is strikingly smaller than the proportions of *mi* and *tě*. Further, *sě* displays a different shape of distributions of post-initial positions (hereinafter pp) with regard to the ALi in comparison to *mi* and *tě*. In the case of *sě*, a more or less gradual decreasing tendency appears (except for ALi = 1); however, in the case of *mi* and *tě*, one can see no clear tendency up to ALi ≈ 10, then the proportions of post-initial enclitics decrease. To compute the relationship between ALi and pp, Kendall's rank correlation coefficient was used with results as follows. For *sě*, we get $\tau = -0.75$ (p-value < 0.001), for *mi* and *sě,* $\tau = -0.37$ (p-value < 0.115) and $\tau = -0.45$ (p-value < 0.032).

1. The original data are available here: http://cechradek.cz/data/Cech_etal_On_the_Impact_of_Initial_Phrase_Length_Data.pdf

Table 1. Proportions of post-initial positions (pp) of enclitics *sě*, *mi*, and *tě* in the analyzed books in the Olomouc Bible. The length of the initial phrase is measured by the number of letters. ALi means the average weighted length of the post-initial phrase, fp the frequency of the particular enclitic in post-initial position, fn the frequency of the particular enclitic in non-post-initial position

sě				*mi*				*tě*			
ALi	**fp**	**fn**	**pp**	**ALi**	**fp**	**fn**	**pp**	**ALi**	**fp**	**fn**	**pp**
2	16	13	0.55	2	25	1	0.96	1.88	14	2	0.88
3	53	27	0.66	3	62	2	0.97	3	25	5	0.83
4	61	32	0.66	4	33	4	0.89	4	24	1	0.96
5	74	38	0.66	5	48	0	1	5	16	4	0.8
6	56	33	0.63	6	26	5	0.84	6	22	4	0.85
7	44	31	0.59	7	21	0	1	7	18	1	0.95
8	27	25	0.52	8	9	1	0.9	8	17	1	0.94
9	27	24	0.53	9	6	2	0.8	9	6	3	0.67
10	17	18	0.49	10	5	0	1	10	6	0	1
11	5	20	0.2	12.17	3	3	0.5	11.83	3	3	0.5
12	3	12	0.2	19.29	1	6	0.14	13.88	2	6	0.25
13	3	18	0.14					16	2	4	0.33
14	0	18	0					20.4	1	3	0.25
15	0	20	0								
16	0	11	0								
17	0	16	0								
18	1	9	0.1								
19	0	5	0								
21.57	0	7	0								
24.4	0	5	0								
39.14	1	6	0.14								

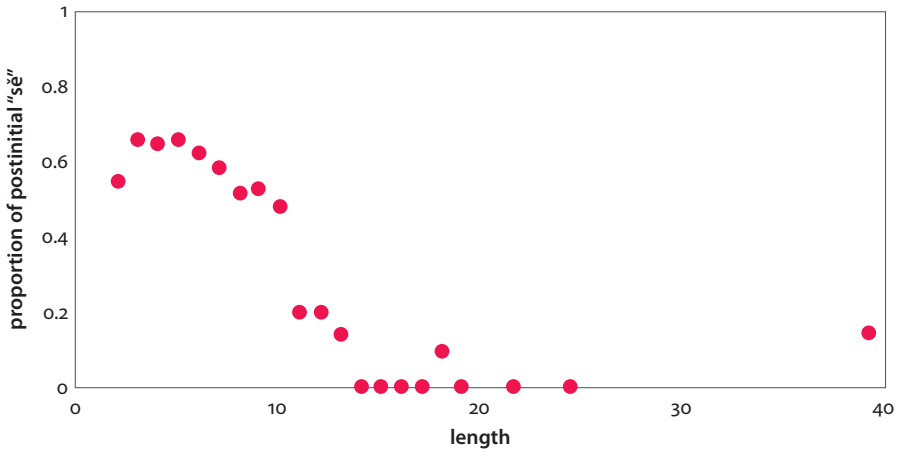


Figure 1. Proportions of post-initial positions of enclitic *sě*. The length of the initial phrase is measured by the number of letters

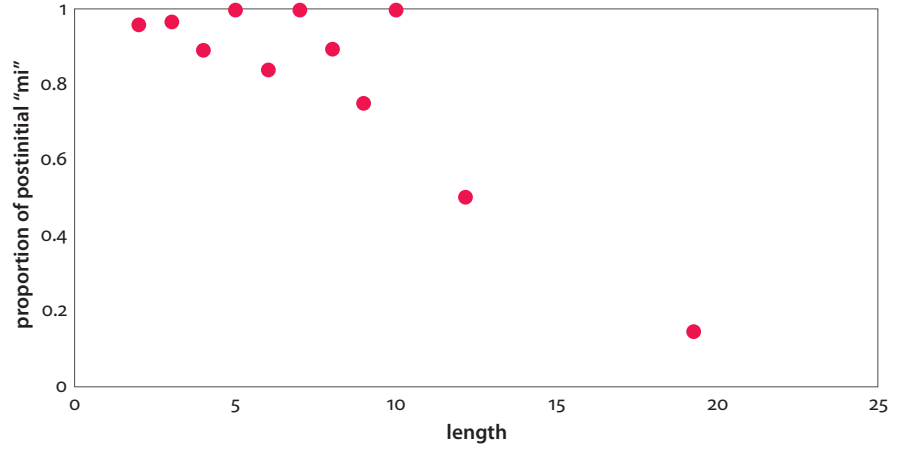


Figure 2. Proportions of post-initial positions of enclitic *mi*. The length of the initial phrase is measured by the number of letters

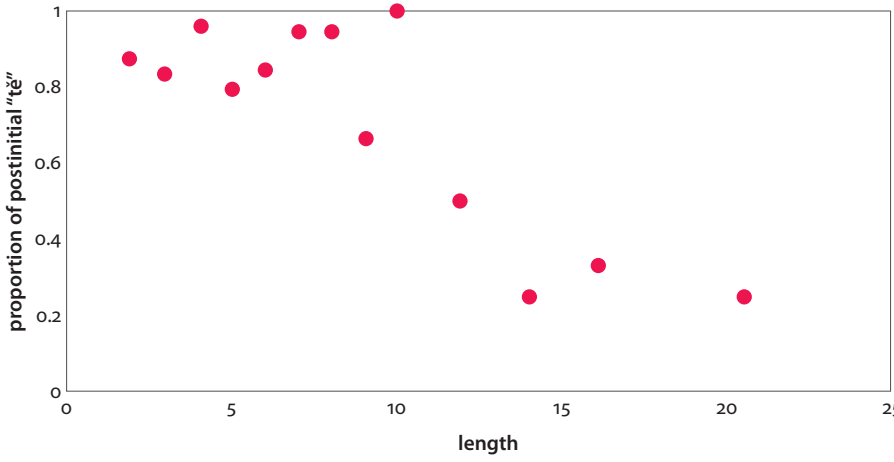


Figure 3. Proportions of post-initial positions of enclitic *tě*. The length of the initial phrase is measured by the number of letters

Alternatively, the length of the initial phrase was measured by the number of words the phrase consists of (see Table 2 and Figure 4–6). For all enclitics, we can see that the general tendency is in accordance with the hypothesis. Further, there is again an obvious difference between the behaviour of *sě*, on the one hand, and *mi* and *tě*, on the other. Namely, the proportions of the post-initial positions of *sě* in the two shortest phrases are smaller in comparison to the respective proportions of *mi* and *tě*. Optically, we can see that the shapes of the distributions differ: *sě* displays a more gradual tendency than *mi* and *tě* which have a minimal difference between ALi = 1 and ALi = 2. However, this optical observation must be taken as very preliminary and only a proper mathematical modelling can reveal if it reflects some tendency or not. In this study, the sample size, unfortunately, does not allow applying either the measurement of a correlation coefficient or of mathematical distributional models.

Table 2. Proportions of post-initial positions (pp) of enclitics *sě*, *mi*, and *tě* in the analyzed books in the Olomouc Bible. The length of the initial phrase is measured by the number of words. ALi means the average weighted length of the post-initial phrase, fp the frequency of the particular enclitic in post-initial position, fn the frequency of the particular enclitic in non-post-initial position

sě				*mi*				*tě*			
ALi	**fp**	**fn**	**pp**	**ALi**	**fp**	**fn**	**pp**	**ALi**	**fp**	**fn**	**pp**
1	229	112	0.67	1	189	8	0.96	1	91	13	0.88
2	139	119	0.54	2	46	6	0.88	2	58	9	0.87
3	16	85	0.16	3	3	4	0.43	3	4	10	0.29
4	3	53	0.05	5.14	1	6	0.14	4.33	3	6	0.33
6.14	1	21	0.05								

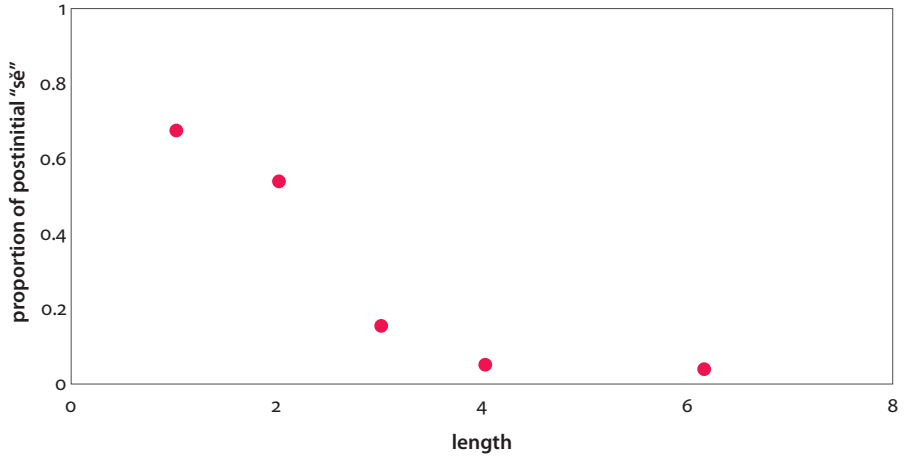


Figure 4. Proportions of post-initial positions of enclitic *sě*. The length of the initial phrase is measured by the number of words

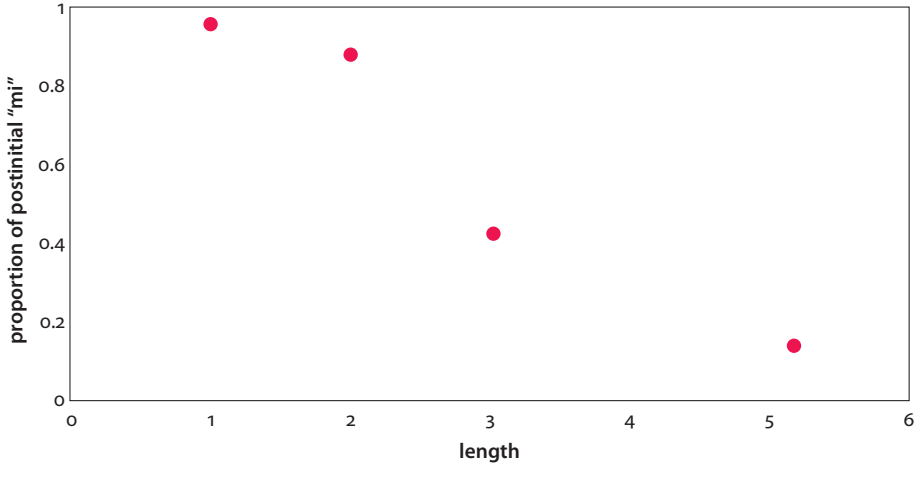


Figure 5. Proportions of post-initial positions of enclitic *mi*. The length of the initial phrase is measured by the number of words

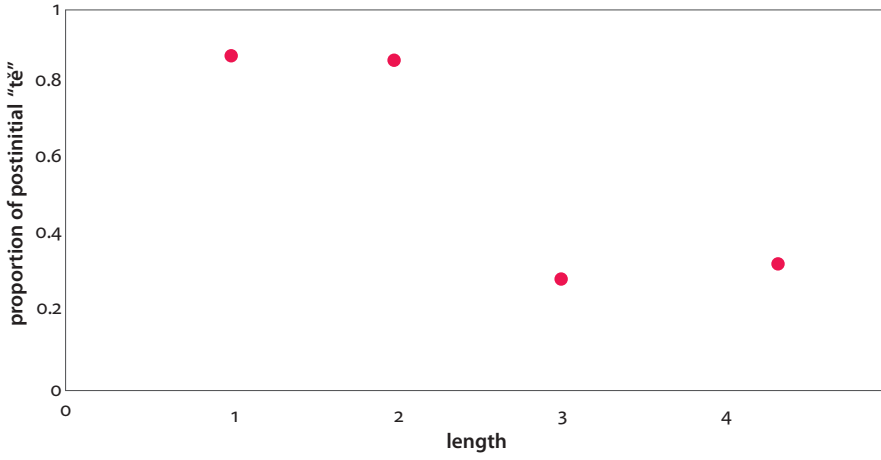


Figure 6. Proportions of post-initial positions of enclitic *tě*. The length of the initial phrase is measured by the number of words

6. Conclusion

In this study, the impact of the length of the initial phrase on the position of four enclitics in Old Czech was analyzed. The hypothesis predicting the negative correlation between the length of this phrase and the proportions of enclitics in the post-initial position was not rejected. Moreover, the study reveals that the general tendency, i.e., the longer the phrase, the lower the proportion of enclitics in post-initial position, manifests differently in the case of *sě*, on the one hand, than in the case of *mi* and *tě*, on the other. We suppose that this difference can be caused by the following factors. First, both *mi* and *tě* represent 'pure' pronouns whose function is to denote an object, while *sě* is grammatically multifunctional. Specifically, *sě* occurs as (1) a pronoun, (2) a grammatical morpheme, which performs various grammatical functions (primarily deagentization and intransitivization), (3) a discrete part of lexical items (reflexiva tantum). This language unit shares functions of a grammatical morpheme as well as of a pronoun and, thus, 'lies' on the border of these functions. Second, the results can be influenced by a word order position of a syntactically superordinate verb in a clause. Our preliminary findings (Kosek et al. 2018b) show that a fine-grained classification of word order can shed light on some characteristics of the observed phenomena. To sum up, despite all differences, there seems to be a clear tendency in accordance with the hypothesis.

Funding

This study was supported by the project *Development of the Czech pronominal (en)clitics* (GAČR GA17-02545S).

References

Ćavar, Damir & Chris Wilder. 1999. "Clitic Third" in Croatian. In Henk van Riemsdijk (ed.), *Clitics in the languages of Europe*, 429–467. Berlin: de Gruyter. https://doi.org/10.1515/9783110804010.429

Ertl, Václav. 1924. Příspěvek k pravidlu o postavení příklonek. *Naše řeč* 8(9), 257–268; 8(10), 293–309.

Kosek, Pavel, Olga Navrátilová, Radek Čech & Ján Mačutek. 2018a. Word order of reflexive 'sě' in finite verb phrases in the first edition of the Old Czech Bible translation (Part 1). *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 135. 177–188. https://doi.org/10.4467/20834624SL.18.017.8853

Kosek, Pavel, Olga Navrátilová & Radek Čech. 2018b. Slovosled prominálních enklitik mi, si, ti, ho, mu v Bibli kralické. Paper presented at Diachronní setkání v Řevnicích, Řevnice, October 31–November 2.

Kosek, Pavel, Olga Navrátilová & Radek Čech. 2019. *The development of the Czech pronominal (en)clitics*. http://www.cechradek.cz/enclitics/home.html

Kyas, Vladimír. 1997. *Česká Bible v dějinách národního písemnictví*. Prague: Vyšehrad.

Kyas, Vladimír (ed.). 1981. *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. I. Evangelia*. Prague: Academia.

Kyas, Vladimír (ed.). 1985. *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století s částmi Bible litoměřicko-třeboňské. II. Epištoly. Skutky apoštolů. Apokalypsa*. Prague: Academia.

Kyas, Vladimír (ed.). 1988. *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. III. Genesis–Esdráš*. Prague: Academia.

Kyas, Vladimír, Věra Kyasová & Jaroslava Pečírková (eds.). 1996. *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. IV. Tobiáš–Sirachovec*. Padeborn: Schöningh.

Pečírková, Jaroslava, Hana Sobalíková, Markéta Pytlíková, Milada Homolková, Vladimír Kyas & Věra Kyasová (eds.). 2009. *Staročeská Bible drážďanská a olomoucká s částmi Proroků rožmberských a Bible litoměřicko-třeboňské. V/1 Izaiáš–Daniel, V/2 Ozeáš–2. kniha Makabejská*. Prague: Academia.

Radanović-Kocić, Vesna. 1996. The placement of Serbo-Croatian clitics: A prosodic approach. In Aaron Halpern & Arnold Zwicky (eds.), *Approaching second: Second position clitics and related phenomena*, 429–445. Stanford, CA: CSLI Publications.

Šlosar, Dušan. 1967. Poloha enklitik jako kritérium k hodnocení staročeské interpunkce. *Listy filologické* 91(3). 251–258.

Toman, Jindřich. 2000. Prosodické spekulace o klitikách v nekanonických pozicích. In Zdena Hladká & Petr Karlík (eds.), *Čeština – univerzália a specifika* 2, 161–166. Brno: Masarykova univerzita.

Trávníček, František. 1956. *Historická mluvnice česká 3. Skladba*. Prague: SPN.

Uhlířová, Ludmila, Petr Kosta & Ludmila Veselovská. 2017. Klitikon. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. https://www.czechency.org/slovnik/KLITIKON (30 December, 2018)

Vintr, Josef. 2008. *Bible (staroslověnský překlad, české překlady)*. In Luboš Merhaut (ed.), *Lexikon české literatury, 4/II U–Ž, Dodatky A–Ř*, 1882–1887. Prague: Academia.

Zwicky, Arnold. M. 1994. What is a clitic? In Joel A. Nevis, Brian D. Joseph, Dieter Wanner & Arnold M. Zwicky (eds.), *Clitics. A comprehensive bibliography 1892–1991*, 12–20. Amsterdam: Benjamins. https://doi.org/10.1075/lisl.22

Term distance, frequency and collocations

Lars G. Johnsen
National Library of Norway

In this paper I study two co-occurrence measures, local to a particular corpus, for constructing collocations or relevance relations between words or terms. One is a distance measure, while the other uses different co-occurrence windows, one contained in the other. Both are discussed with respect to the common method of comparing co-occurrence measures within a particular corpus to those of a reference corpus. A practical consequence of these measures is that they may relieve the burden of computing a reference statistic, which may incur a high computational cost. We also believe that distance, as a measure in itself, has a theoretical interest. Being different from frequency, it may add something new to collocation analysis.

Keywords: collocation, term distance, frequency, Bayes, probability, concordance

1. Introduction

This paper considers uses of quantitative measures such as frequencies and distance in constructing relations between words. We build upon the line of research, starting from Firth (1957), that a word can be characterized by its closely surrounding words, and Halliday (1992), that capture patterns in language use frequencies as probabilities. In a temporary setting, these ideas about the quantitative connection between words, their contexts and their grammar or meaning, are summed up in the distributional hypothesis. In the formulation of Piper (2018: 13), it takes this form:

> At base, the distributional hypothesis assumes four things: (a) a word's meaning is tied to how often it occurs; (b) a word's meaning is tied to how often it occurs with other words within a given context; (c) these relationships are entirely contingent upon the scale of analysis; and (d) these relationships can be rendered spatially to capture the semantic associations between them.[1]

1. Thus, distributional semantics is somewhat derivative of the relation between a word and its relation to the external world. Formal textual relationships may reflect on the external relations between what words signify.

https://doi.org/10.1075/cilt.356.02joh

Together with a way of scoring using frequencies, I introduce a distance metric between words as a relevance measure. In both cases the starting point is a window, or context, containing a certain number of words around a target word. For frequencies, the statistic is word frequencies that arise from counting words in the context. For distance, the statistic is the average distance between a word and the target. The logic for scoring or ranking the collocates is that the resulting values from the statistics are compared to expected outcomes, expectations that will be made clear below. These two methods are evaluated by comparing the rankings they produce with a standard way of computing collocation scores.

Our perspective here is partly theoretical and partly practical. When doing collocation analysis in practice, frequencies need to be compared to some other reference frequencies, which may incur a high cost in terms of time and resources. For frequencies, we will look at using two or more collocation data, one contained in the other, as explained below. This way of using locally available data makes finding comparisons a bit simpler. For distance as a metric, where all the computations are done within the obtained collocation data without any reference corpus, scoring and selecting collocates for a target word should carry low computational cost.

2. Δ-score and Pointwise Mutual Information

A collocation is taken to be a bag of words equipped with a score. When this score consists of frequencies, it can be used as the basis for association measures. Different algorithms can be used to compute such associations, see, e.g., Kolesnikova (2016 for a wide range of such measures. In this work our focus is mainly on sources of information, and we will use Pointwise Mutual Information (PMI), introduced in Church & Hanks (1989), for computing association values. One interesting feature of that measure is its close connection to the Bayesian inversion, and its interpretation as a difference, or Δ-score.

Consider the case of bigrams, where the mutual information between two words x and y is written as in equation 1.

Equation 1

$$pmi(x,y) = log\,(\frac{p(x,y)}{p(x)p(y)})$$

The probability expression in the numerator is taken to be the relative frequency of the bigram, and the expressions in the denominator is the relative frequency of each word.

For the purpose of this article, the logarithm is not essential,[2] which leaves us with the inner formula which can be rewritten, using standard probability calculus, as:

Δ-score

$$\Delta(x,y) = \frac{p(x,y)}{p(x)p(y)} = \frac{p(x|y)}{p(x)}$$

The right-hand side of this last equation warrants the interpretation of PMI as a difference. Higher (or lower) Δ-score means that the frequency of x in the context of y differs from its frequency without the context. The larger (or smaller) the ratio, the bigger the mutual information. A value of around 1 means there is no difference, and the context for the target x makes no difference for its frequency.

The formulation as a difference helps in making a formal interpretation of the Δ-score. Following the discussion in Jaynes (2003), x is relevant for y when the following holds.

$$p(x) \neq p(x|y)$$

Then the score measures the degree of relevance. What kind of relevance, whether it is semantic or grammatical, or some other relation, does not follow, and is up to the interpretation of the actual relationship, and the elements that go into it. Relevance gives us a handle to go on.

The probability concept we use treats probability equal to proportions, measured as relative frequency of a word in a text or corpus. However, a note may be in order. When we talk about words, a distinction must be made between type and token, where tokens can be taken to be the smallest units of the text under the probability model, or the units that are assigned base probabilities. A word type is then taken to be a collection of tokens or occurrences, which in turn are composed of letters, i.e., smaller graphical units. A statement like "the probability of seeing the word *and* in the text T is 0.04" is formulated in terms of tokens, and will look like this on a formal probabilistic notation:

$$p(W \in \text{and} \mid W \in T) = 0.04$$

Here, the variable W stands for the chosen token word, or particular occurrence. For simplicity, we will shorten this to the following for readability:

$$p(\text{and}|T) = 0.04$$

2. The binary logarithm provides an order-preserving mapping from probabilities to information units. The objects assigned probabilities will be sorted the same way. Our focus lies more on the ordering, not so much on the space in which the items are ordered.

An interesting feature of the Δ-score is that it is the likelihood factor in the formula describing Bayes inversion, here using C as context variable:

Equation 2 Bayes inversion

$$p(C|x) = \frac{p(x|C)}{p(x)} p(C)$$

If C is taken to be the frequency list for a document, equation 2 amounts to a naive Bayesian classifier. Each word x contributes its evidence towards C irrespective of what other words are in C. Thus, the Δ-score provides a kind of naive Bayesian weight.

This raises the question, for instance in the case of generation keywords for documents, if there are relationships between the keywords themselves. Keyword production will only produce one word vector weighted by the Δ-score. Suppose a set of keywords are computed from a class of books, so that C in the formula above is an aggregation of the class. Several methods can be used to look at the co-occurrences, in order to bring back interdependence between words: a weighted graph can be constructed, and from there, using for example Louvain-clustering, as in Blondel et al. (2008), Johnsen (2016) uses similar weights for constructing clusters as communities within the graphs. The Δ-score gives rise to a closeness relation which can serve as the basis for clusters derived using methods in Moisl (2017).

For a class of documents, keywords can be extracted by evaluating the term frequency (tf) ratios of terms from the class, with terms from a reference set, typically drawn outside the documents. For example, in a library setting, a document is a representation of subject headings or a Dewey decimal class, and the reference is the totality of books, in its simplest form (Barnbrook et al. 2013; Rockwell & Sinclair 2016).

3. Data and technical method

The two methods described here are tested on the text collection of digitized texts from the Norwegian National Library, see Birkenes et al. (2015) and Johnsen (2016). The collection makes it possible to create corpora using textual data and metadata using an API (Application Programming Interface) for querying the collection. The underlying software for this functionality is found in the Github repository in Johnsen (2019).

All the collocations used here are taken from a set of one thousand books each of which contain at least one occurrence of the target word. From this set, collocations are then extracted using as parameters the number of words before and

after the target word. The data sets and how they are obtained are documented in Johnsen (2020), which contains the software as a set of Jupyter notebooks, as well as the collocation data used below.

All the digitized books, at the time comprising about 440,000 books, have been aggregated and are used as a reference, or expected frequency for words. For any corpus C and any word x, the relative frequency of x in C is computed as p(x|C), and the unconditional probability, p(x), is the relative frequency in the reference corpus.

4. Collocations

In this section I present two experiments: one, collocations based on frequency, and two, based on a distance metric. Using frequencies, a collocation analysis for a word x is performed by comparing the frequencies of words that co-occur with x with that of a reference. The probability logic is to measure the difference between the observed frequency with the expected frequency. As noted above, we will use the term collocation broadly, covering both the sense of fixed phrases and the more general relevance concept, even if they are several words apart. Collocations taken in this more general sense make them suitable for distributional semantics for modelling semantic and syntactic as well as discourse phenomena.

4.1 Frequency and context enlargement

A problem of both practical and theoretical interest in collocation analysis is to find an appropriate reference corpus to perform a comparison. For a general reference for comparison in the Norwegian examples, I will use the average relative frequency of words in almost all the published books, e.g., Birkenes et al. (2015) and Johnsen (2016). This reference will be used to evaluate other contexts.

In terms of Δ-score, there are two contexts to consider when computing a collocation score for word x in the collocation of y. One is the actual collocation, written C(y), and one is a larger general context U. The larger context may be just a tad bigger than the collocation context itself, or it could be global comprising the whole text collection.

The Δ-equation is then more accurately rendered as

$$\Delta(x,y) = \frac{p(x|C(y))}{p(x|U)} \text{ where } C(y) \subset U$$

The result of an analysis may vary with the choice of reference corpus.

Constructing the collocation, a word x is part of C(y) if it is found within a window of y, for instance *n* words before and *m* after. The values of *n* and *m* depend on the goals of the analysis. For U we choose a slightly larger window within the collocation corpus itself, so that C is a proper subset of U. If one used the pair (*n, m*) (*n* words before and *m* after) when constructing C, the construction of U is done with the pair (*k, o*) where *k* is larger than *n*, or equal to *n*, and *o* is larger than *m* or equal to *m*. Thus, the analysis is made from two collocations, one contained in the other. The point here is that the reference used to evaluate the relevance of x in C(y) is constructed locally from the source of the collocation.

Our proposal is to sort the possible collocates based on how much their frequency increases as the collocation window increases. For instance, as U is a larger context than C(y), one may expect that all words will have an increase in absolute frequency, but that their relative frequency is stable or lower. Since a larger context also means a larger amount of word types, we expect in general that all the relative frequencies drop.

The underlying idea is that if one selects collocates from, say, a window of five words after a particular target word, what can we say about the distribution if the window is widened to ten? We expect then that all words that are not relevant to the target will approximately double, while words that are connected will increase less, i.e., most of them already accounted for by occurring close to the target. Thus, we get a measure: sort the result on the words that increase the least.

This method will then need a couple of parameters, parameters that control the selection based on the absolute frequency, if there is any change at all that needs to take into account that there is an actual increment.

The case at hand consists of an example using only the right hand context for the infinitival verb *skrive* ("write"). For *skrive*, two contexts to the right are collected, one with length 10, called C, and one with length 20, called U, in the experiment the size parameters are (0, 10) and (0, 20). At the same time a reference R is provided with which the computations can be compared. In the computations R is the average over all books in the collection. For each word x in the collocation for *skrive* we have two counts, one within C(*skrive*), one within U(*skrive*), and one with R.

The numbers computed are the following, all of which measure mutual information in some way:

$$\text{ratio}(x,y) = \frac{p(x|C(y))}{p(x|U(y))}$$

$$\text{ref-large}(x,y) = \frac{p(x|U(y))}{p(x|R)}$$

The value of the ratio(x,y) measures the difference between the larger collocation context and the smaller, while ref-large(x,y) measures the difference between the larger context and the reference. The difference between the two is analyzed in two rounds. One looks at the top score for both, and the other analyses how much the two sets match, using a Jaccard-score of the sets:

$$Jaccard(A,B) = \frac{|A \cup B|}{|A \cup B|}$$

A and B are set like this, subject to a cutoff, i.e., not all of A or B. The sets are sorted and limited in some way.

$$A = \{x|\text{ratio}(x, skrive)\} \text{ and } B = \{x|\text{ref-large}(x, skrive)\}$$

First, we make a subjective assessment of the top 10 words, scored by the two measures above. Below is a table conditioned on the requirement that for any word in the collocation, its absolute frequency should increase by at least one when going from the small to the larger context. This removes some low frequency words, which we do not consider in this analysis. In Table 1 below the numbers are highlighted according to order, first one sorted by column 'ratio':

Table 1. Right context collocations of Norwegian *skrive* ("write") sorted by ratio

	trans	ratio	ref-large
lapp	note	1.81757	35.6453
testamente	wills	1.78151	16.7366
erklæring	statement	1.75867	14.9674
tale	speech	1.75867	1.05746
sammendrag	summary	1.75867	36.9752
dagbok	diary	1.7541	61.8082
skuespill	play	1.74413	13.9143
artikkel	article	1.73288	17.8474
redigere	edit	1.72669	43.805
innlegg	post	1.70538	10.4382

Sorting by the reference column gives a slightly different result, although there is a certain overlap, as seen in Table 2.

Table 2. Collocations of Norwegian *skrive* ("write") sorted by ref-large

	trans	ratio	ref-large
stiler	writings	1.37039	91.1005
tastaturet	keyboard	1.00976	89.0458
memoarer	memoirs	1.39531	87.9952
skrivemaskin	typewriter	1.03306	85.4577
diktat	dictation	1.4922	84.3947
dagbok	diary	1.7541	61.8082
låter	songs	1.67873	59.9943
reserverte	reserved	1.09631	55.1342
markøren	cursor	0.839365	48.1295
arket	sheet	0.992352	45.8756

While the index column shows that all of these are relevant words whether selected by the collocation difference or the total, the two tables differ from each other in several ways. The reference sort appears to include more words connected to writing as such, while the table sorted by ratio-column had more syntactic objects.

The question now is how to get a grip on the general discrepancy. We use the Jaccard similarity, as defined above, so that they go into the comparison with no weight attached. However, this way of doing it gives us a certain grip of what is going on. A score of about 0.3 would mean that for equal sized sets they share approximately half of their content.

The graph below (Figure 1) shows how the Jaccard-score changes with the size of the sets. So for each collocation, take the top 5, then top 10, top 15 and so on, and compute the score using these sets. Each point on the horizontal axis is the size of the set, and the vertical axis is the Jaccard-score.

The best score is between the sets shown in the tables above, when the sets consist of the top ten words, a value that holds for the top twenty, before the score settles at between 0.15 and 0.20 up to around fifty words.

A third comparison is to look at some high frequency words or tokens, like punctuation marks, prepositions and coordinations. Expectedly, punctuation marks and coordinations double in count as the window doubles, which gives them a ratio of approximately 0.5. These words, part of the so-called stop words, are more or less automatically filtered out based on sorting because of this. Their ratios between

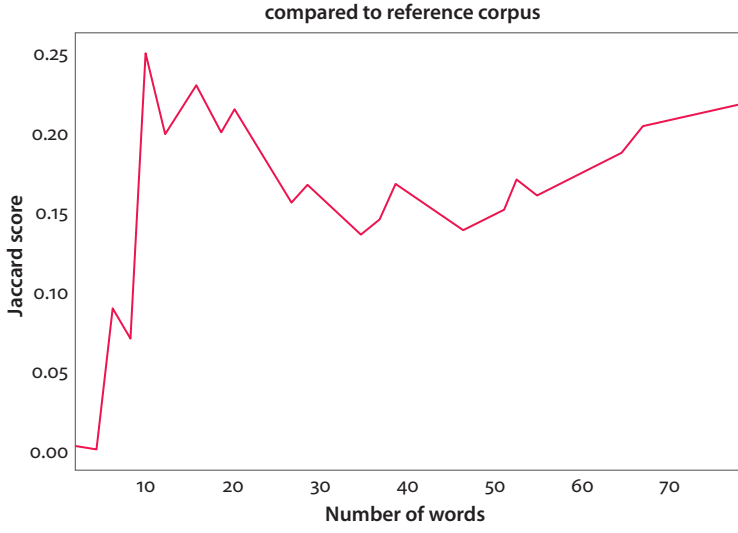


Figure 1. Jaccard-score comparing collocations from ratio with reference

large and small samples, as well as compared to the reference, hover around 1, indicating that the relative frequency is stable across different samples.

Even though the actual values differ a bit, both the ratio score and the reference score gives almost the same ranking of these words; the top two of these high frequency words are *eller* ("or") and *på* ("on"), where the latter goes together with the verb, as in *write on something*.

Table 3. Collocation columns ratio and ref-large agree on high frequency words

translate		ratio	ref-large
,	,	0.987992	0.985015
.	.	0.99972	1.20565
og	and	0.957841	1.30357
eller	or	1.12353	2.09581
i	in	0.921231	1.18278
på	on	1.0604	2.00771

4.2 Distance

The second local information source is the average distance between the target and its collocations. Although distance, or position, of a word with respect to another is indirectly used via size of context window, we are unaware of distance used as a collocation metric in itself.

The informational basis for using the average distance for selecting relevant words stems from the following observation. Any window of words, say a 9 words long contiguous sequence, when sampled from a corpus, will not dictate the position of any of its members. A word in this sequence may occur in any position from 1 to 9, with an average position at approximately 5, for a sample of appropriate size. However, specifying conditions on this sampling may disturb the distribution. For example, if the sequence is required to follow a transitive verb like *eat* (as a target of a collocation for example), some words will occur closer, like *dinner, breakfast* and *meat*, with an average distance well below 5. All the while other words, or tokens, that are irrelevant should fall around 5, for example punctuation marks or other irrelevant words. Thus, distance carries the same relation between what we observe and what we expect.

How close a word can get to another word is partly dictated by syntactic constraints, like that an English verb occurs with following noun phrases, particles and prepositional phrases, so the average position may indicate a certain role, or relation to the target word. In this way, average distance may say something about a verb's syntactic properties, as well as something about its topic structure. Spelling conventions also place some constraints, like capitalization.

For this experiment, in order to get numbers for interpretation, the windows for a target consist either of the left preceding word sequences or the right following, where both are analyzed separately, once for a verb, and once for a noun.

One effect of using distance to measure relevance is that the measure is independent of a reference corpus. As we did in the experiment above, we will compare the rankings using the distance metric with that computed from the reference. The procedure is the same as above: we look at the Jaccard similarity as well as a subjective evaluation.

Frequency is still used in order to assess the average distance, being used to adjust the closeness numbers. At the outset all averages are assumed to be in the middle of the collocation window. The number is adjusted to approach the observed average if there are enough occurrences. For example, a word occurring once will not move away from the prior average.

The method is illustrated with two cases, once with the verb *skrive* ("write"), same as above, and once with a noun, *kaffe* ("coffee"). For the verb, the right context is investigated, for the noun, the left context. For each word occurring in a context, the following is computed.

1. its average distance, and the adjusted distance
2. the raw frequency of the collocation
3. the reference score, computed same way as ref-large above

These values are put into a table for comparison.

4.2.1 *The verb*

The verb is the same as used above, the infinitival form *skrive* ("write"). In the table below, the column labelled 'dist_' is the adjusted distance; for high frequency words this is very close to the average distance. The reference column is the Δ-score. Note the occurrence of zeros in that column, which is due to the selection of words in the reference. If a word is not over a certain frequency, it is not in the reference. Even though this violates the subset assumption, it can be accommodated by adding the collocation to the reference.

Table 4. Collocates for *skrive* ("write") based on distance

	translate	freq	dist_	reference
referat	writings	7	1.64	21.7365
ned	keyboard	313	1.74	7.16901
dikt	memoirs	72	1.82	38.9332
kjærlighetsbrev	typewriter	7	1.9	0
søknader	dictation	5	1.9	12.9925
hovedoppgave	diary	8	1.9	34.3942
postnavnet	postal name	5	1.9	0
låter	songs	14	1.93	77.9857
programmene	programs	5	1.97	16.0921
romaner	novels	10	2.01	19.5996

One rather interesting feature of distance is its ability to pick up high frequency words like *ned* ("down"), as it occurs as an adverb in *write down something* or *write it down*. Its average distance is 1.7. Of all the words deemed high in the top ten list, all have a fairly high Δ-score, apart from the particle *ned*. Subjectively, all these can be seen as good collocates. The same holds if we compare to the reference column (ref-large) in the section above, that distance seems to prefer syntactic objects in addition to the particle.

Given the sensitivity for low frequency words, the Jaccard similarities are computed using words with an absolute frequency above ten. Depending on the size

of the number of words used to compare, the Jaccard score goes quickly up to 0.2 and then makes another jump when the size is around 50 words. The last five up to rank eighty are *musikk* ("music"), *program* ("program"), *skrev* ("wrote"), *ferdig* ("finished") and *verk* ("work"). We are still looking at good words.

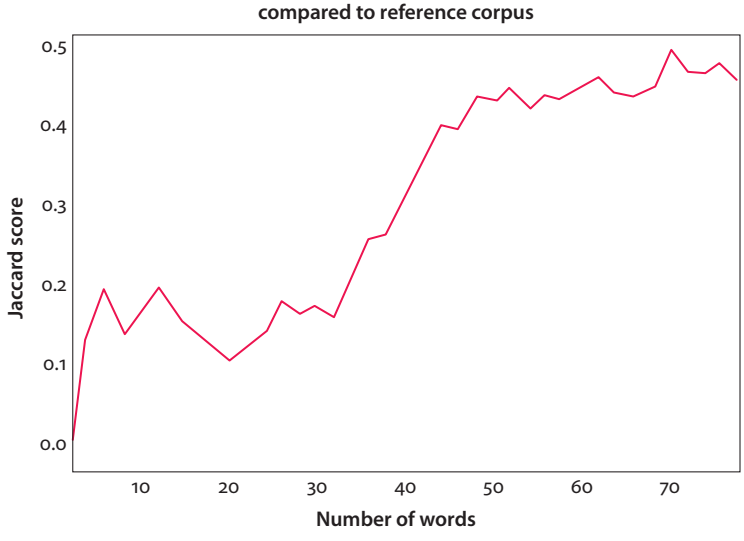


Figure 2. Jaccard similiarity of distance vs. frequency based collocates for *skrive* ("write")

How good is the distance measure in capturing stop words? The following table, structured as that above, gives the answer. Note that 'dist_' is best from low to high. The values in this table corroborate distance as a useful relevance measure.

Table 5. Distance for capturing high frequency words

translate	freq	dist_	reference
,	2649	5.38	0.753565
.	3757	5.33	0.933303
og	1671	5.39	0.966833
eller	342	4.84	1.82332
i	1235	5.75	0.843717
på	1080	4.68	1.64853

As expected, the distance for these high frequency words lies around 5.5, halfway into the window (as the numbers go from 1 to 10, the mid point is 5.5). This is the position doomed for less relevant words. These words are also close to one in the reference comparison. Note that as with the collocation extension method in the previous section, the coordinator *eller*, and preposition *på* appears a little bit closer.

4.2.2 *The noun*

For an analysis of the left context, we look at the noun *kaffe* ("coffee"), and go through the same analysis. The context used for the noun is a window of ten preceding words. The computations are otherwise similar to that in the preceding section. The distances are negative, signaling that the collocating words come before the target.

The top ten words sorted by closeness are shown in the table below.

Table 6. Left distance collocates for *kaffe* ("coffee")

	translate	freq	dist_	reference
nytraktet	freshly drawn	15	−1.3	0
nytrukket	freshly drawn	14	−1.32	0
kopp	cup	850	−1.33	1119.41
Svart	black	12	−1.38	59.5522
nykokt	freshly boiled	12	−1.38	0
slurk	sip	49	−1.43	118.642
pund	pound	13	−1.44	17.8394
svart	black	107	−1.47	32.0186
Mer	more	9	−1.5	6.77116
nylaget	freshly made	9	−1.5	0

As with the verb, missing entries for the reference is abundant, almost half of the words are not given a reference score, although they are easily recognized as relevant. These are also words that occur tight, and with (a what may be considered) high raw frequency.

The list sorted by Δ-score using the reference differs a bit from the above, but shows a high degree of relevance. Note, however, that a lot of those words are at the edge of what would be considered relevant by the distance score. In this respect, the term *kakao* ("cocoa") is interesting, being in the family of hot drinks, it has a high Δ-score, but is not so tightly connected, although fairly close.

Table 7. Collocates sorted on values computed from reference corpus

	translate	freq	dist_	reference
kopp	cup	850	−1.33	1119.41
kanne	jug	58	−1.78	581.517
skjenket	poured	163	−2.74	351.469
mineralvann	mineral water	25	−2.94	327.723
rykende	smoking	37	−2.24	257.746
Viktigste	main	25	−6	255.119
kakao	cocoa	35	−3.74	245.102
serverte	served	43	−2.26	240.712
kopper	cups	87	−2.23	220.739
vafler	waffles	13	−3.12	216.32

If we look at a selection of hot drinks, *te* ("tea") is the highest, while the word *coffee*, although kind of relevant, is on the other side.

Table 8. Comparing how other hot drinks collocate with *kaffe* ("coffee")

	Freq	Dist_	Reference
kakao	35	−3.74	245.101930
kaffe	157	−6.68	87.791997
te	120	−3.24	16.054192
sjokolade	19	−4.41	54.396609
buljong	1	−5.50	9.346958

The collocation sets are assigned a Jaccard similarity score, which, for the low frequency effects we saw above is computed for words with a frequency above 10.

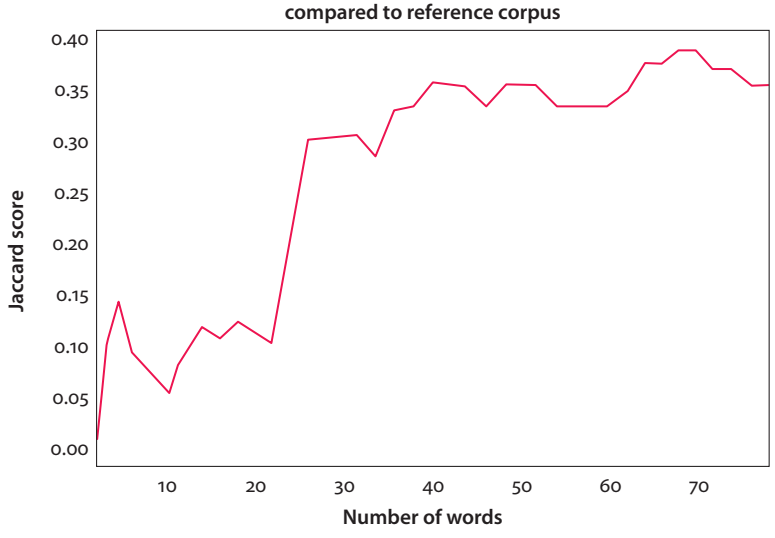


Figure 3. Jaccard similiarity of distance vs. frequency based collocates for *kaffe* ("coffee")

The graph shows the same trend as with the verb, with two noticeable rises.

5. Discussion

We set out to study how collocations can be locally computed, and used enlargement of collocation window as one such method. Another closely related method for creating a context while doing the collocation is by selecting word sequences randomly from the documents from which collocations are extracted. Adding those selections together with the collocations could then serve as the denominator for a Δ-score.

The distance metric can be used as is and requires no extra data at all. However, a hidden premise in our discussion is that distance is relevant linearly, i.e., the further away, the less relevant. Even with syntactic objects, that does not hold in general, as objects can be oblique objects, and thus get pushed at least one word further away. We leave that issue for future research.

References

Barnbrook, Geoff, Oliver Mason & Ramesh Krishnamurthy. 2013. *Collocation applications and implications*. Berlin: Springer.

Birkenes, Magnus Breder, Lars G. Johnsen, Arne M. Lindstad & Johanne Ostad. 2015. From digital library to n-grams: NB N-gram. In Beáta Megyesi (ed.), *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, 293–295. Linköping: Linköping University Electronic Press.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10. 1–13. https://doi.org/10.1088/1742-5468/2008/10/P10008

Church, Kenneth Ward & Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In Julia Hirschberg (ed.), *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, 76–83. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/981623.981633

Firth, J. R. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis (special volume of the Transactions of the Philological Society)*, 1–32. Oxford: Basil Blackwell.

Halliday, Mark. 1992. Language as system and language as instance: The corpus as a theoretical construct. In Jan Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82 Stockholm, 4–8 August 1991*, 61–78. Berlin: de Gruyter. https://doi.org/10.1515/9783110867275.61

Jaynes, Edwin. T. 2003. *Probability theory: The logic of science*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790423

Johnsen, Lars G. B. 2016. Graph analysis of word networks. In *CEUR workshop proceedings*, Vol-2021. urn:nbn:de:0074-2021-3.

Johnsen, Lars G. B. 2019. Modules, Github repository. https://github.com/Yoonsen/Modules

Johnsen, Lars G. B. 2020. Collocations, data and software. https://doi.org/10.5281/zenodo.3783742

Kolesnikova, Olga. 2016. Survey of word co-occurrence measures for collocation detection. *Computación y Sistemas* 20(3). 327–344. https://doi.org/10.13053/cys-20-3-2456

Moisl, Hermann. 2017. *Cluster analysis for corpus linguistics*. Berlin: de Gruyter.

Piper, Andrew. 2018. *Enumerations*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226568898.001.0001

Rockwell, Geoffrey & Sinclair Stéfan. 2016. *Hermeneutica: Computer-assisted interpretation in the humanities*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/9522.001.0001

A method for the comparison of general sequences via type-token ratio

Vladimír Matlach, Diego Gabriel Krivochen and Jiří Milička
Palacký University / University of Oxford / Charles University

This article proposes a new method for analyzing and comparing general linear sequences with the minimum prior knowledge on the sequences needed. Sequence analysis is a broad problem studied by various fields from sociology and computer security to linguistics or biology. The method presented here applies the simplest quantitative linguistic tools in order to achieve methods transparency and easily interpretable results. The results form a vector describing the sequence and allow their clustering, machine learning and simple visualizations by line charts or multidimensional methods as MDS or tSNE. For completeness, artifacts and several formal models are derived to describe methods behavior in both common and extreme cases.

Keywords: sequence analysis, sequence clustering, randomness test, n-gram, type-token relation, type-token ratio, confidence intervals

1. Introduction: Sequence analysis and its importance

Contemporary linguistics is often interested in text analyses which may be dependent on methods of analyzing strings of symbols of variable nature (words and morphemes, glyphs, nucleotides, etc.). Such situations may include cases where there is no prior knowledge about the proper segmentation of the units of the language in which the text is written, and when the only conceivable units are single characters. Sequence analysis may help with such analysis and offer means for sequence source comparison, complexity testing, estimation of units, etc. (see Mikros & Macutek 2015 for more applications from the perspective of quantitative linguistics).

Sequence analysis methods are also established and used in many other scientific fields. In biology, studying DNA is comparable to analyzing texts which consist of 4 letters: A, C, T, G. Triplets of these encode 20 various amino acids transcribed as 20 letters, where sequence start- or end-points may not be known in advance.

https://doi.org/10.1075/cilt.356.03mat

From this viewpoint, molecular biology and linguistics share many text and sequence analyzing methods such as edition distances, Markov chain modelling or latent topic modelling (see Blei et al. 2003 on natural language and Pritchard et al. 2000 on genetics; for more on linguistic methods for genetic sequence analysis, see Bolshoy et al. 2010).

Sequence analysis is also established methodology in social sciences: methods such as Markov chains, *n*-grams, edit distances and more are used to analyze human interactions encoded as linear symbolic sequences. This practice allows us to identify cultural patterns and further serve as the basis for finding similarities, clustering and quantitative inference (see Cornwell 2015 for details).

Computer security also uses sequence analysis in order to verify the quality of randomness generators assessing the quality of their output, which serves as cryptographic session keys protecting Internet users (see Stuttard & Pinto 2011 for an example of attacking poor-quality random generators and Govindan et al. 2018 for a critique of random generators which pass standard randomness tests). The quality of random generators and their output is thus critical.

In this chapter, we propose a general-purpose sequence analysis method based on the following key requirements:

1. The method should allow for the characterization of quantitative properties of a sequence of discrete symbols of any type. This analysis of digitized, linear, and discrete sequences depends only on the sequence being composed of distinguishable discrete units – no other prior knowledge (such as word boundaries or the existence of other units) should be necessary.
2. The method should use the simplest, but effective, quantitative linguistics concepts to ensure transparency, also allowing for the enumeration of method-inherent properties as artifacts and implied biases and guaranteeing simplicity in the interpretation of the results.
3. The results of the method must be suitable for visualizations and vectorizable for use in machine learning and simple similarity measures.

Regarding the requirements sketched above, the method proposed in this article is based solely on the classical and simplest methods of quantitative linguistics, which will guarantee both simplicity of interpretation and transparency. As we will see below, this decision allows us to formalize its numerous features and a random sequence statistical test.

2. Method: Quantitative linguistics and the most basic methodology

There are many possibilities for quantifying properties of texts and sequences using descriptive values based on distributions as Shannon Entropy (Shannon 1948), Gini coefficient (Gastwirth 1972) and other indices revealing important information. However, quantifying properties of sequences solely from the distribution of its units is not enough to determine their complexity. We can illustrate this problem on a trivial example of texts A, B and C, each with the same number of symbols 'a' and 'b':

(1) Text A:
a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b a b
Text B:
a a a a a a a a a a a a a a a a a a a a a a a a a b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b b
Text C:
a b a b a b b a b a b a b b a b a a b a b b a b b b a a b a b a a a b b a b a b a b a a b a b a

Indices such as Entropy, Gini coefficient, or the simplest Type-to-Token Ratio (TTR) yield the same results for all three texts when applied to single symbols (due to their equal probability distribution) even though we clearly see their qualitative complexity difference. The missing or unmeasured part is the 'context', which must be considered in order to properly describe a sequence.

The simplest context sensitive analysis of sequences can be carried out by means of *n*-gram analysis. *N*-grams (the *n*-tuple of all succeeding units shifting by one unit every time) are commonly used in cases when a context of a specific length is needed or when unit boundaries are not known (in DNA analysis, see Bolshoy et al. 2010: 61; in cryptanalysis, see Lasry 2018: 20, Jain & Chaudhari 2015; in anomaly detection, see Hamid et al. 2005). However, using *n*-grams requires specifying the value of $n \in \mathrm{N}$, which often comes from a priori domain knowledge (e.g., in bioinformatics $n = 3$ since three nucleotides encode one amino-acid). However, in the case of a universal method for sequence analysis, no prior inference about *n* can be made and all possible (but still reasonable) sizes of *n* should be used. For such a set of obtained *n*-grams for a given *n* and a given sequence *S*, many descriptive indices can be calculated. As we want to avoid extensively complex algorithms, we choose to use a very simple measure – the Type-to-Token Ratio (TTR; see below).

The complete method consists of five steps. The aim of the first two steps is to normalize analyzed sequences into a binary sequence thus allowing reasonable comparison between sequences using various alphabets. The third step creates and collects the *n*-gram data. The fourth step quantifies the *n*-grams' properties and the last, fifth, step is visualizing and interpreting the results.

Step 1: Normalization of the alphabet

In order to compare sequences consisting of various alphabets and their various sizes, we standardize the whole sequence by transforming each symbol into an arbitrary binary code: to each type of symbol contained in the sequence we assign an arbitrary (random) unique integer $x_{Symbol} \in [0;|A|]$, where A is set of all symbols in the relevant sequence. The x_{Symbol} is then converted into binary format, which we then express as a string with zero padding from the left such that all binary encodings have the same length. A new, normalized sequence, which is then passed for further analysis, is created by substituting each original symbol to its binary encoding. Such conversion should not change the quantitative properties of the texts in a case when most of the bits are used (see Schenkel et al. 1993). An example of the normalization of a sequence S follows:

(2) S = *"ABCA"*
A = *{A,B,C}*

Unique integer encoding:	A = 1, B = 0, C = 3
Binarized and padded form:	A_b = 01, B_b = 00, C_b = 11
Result:	$S_{Normalized}$ = "01001101"

There are two further benefits from this standardization step: a binary alphabet allows us to define mathematical models for the method more easily (see below), and it allows us to measure the length of the sequence in a standardized unit of 'bits'.

Step 2: Length normalization

After normalizing the alphabet of the sequences, the size of the sequences must be normalized to the same number of bits. This step is also necessary to ensure reasonable sequence comparison. Any substring of bit length $k \in \mathrm{N}$ (sampled randomly from the sequence or taken by successive application of sliding window) suffices. Our method has empirically shown acceptable performance on sequences as short as 6000 bits (which, in terms of the binary encoding, means around 1200 characters or approximately 300 words of English text). Random sampling of the sequence may be also used for estimating a confidence interval.

Step 3: Contextual data collection from *n*-grams

The key expectation of this method is its applicability for comparison of any type of sequences without prior knowledge of their higher structures. This expectation imposes demands on choosing the proper n-gram size to carry the test (as shown with the case of DNA above). However, the general method must reflect each possible and reasonable values of n for all sequences and thus use all $n \in [1;z]$ where z

$< k; z \in \mathrm{N}$. Empirical results have shown that for sequences of $k = 6.000$ *bits*, $z = 50$ the method yields results for general sequence comparison. For each n-gram size n the method yields a list of n-bit tokens d of v distinct types.

Note: Choosing sequence length k and n-gram size bounding
The sequence length k and the upper bound of n-gram size z can be tweaked. In order to obtain a more accurate result, higher k and z are recommended, for faster but less accurate results, z can be set lower.

Step 4: Quantification of properties

Quantifying properties of each n-gram group can be performed by any suitable measure, such as the aforementioned Shannon Entropy or Gini coefficient. Nevertheless, by selecting a measure that considers distributions we complicate the formalization of the method's properties. Instead, we choose the simplest possibility which is calculating type-to-token ratio (TTR), which can be calculated for a given number of observed types V and number of tokens N:

(3) $TTR = V/N$, where $V, N \in \mathrm{N}$.

Results of TTR are bound into the interval $[\epsilon,1]$. The result of $TTR = \epsilon$, where ϵ is an infinitesimal number, is the case of only one type ($V = 1$) being repeated infinitely in an infinite long sequence ($N = +\infty$), i.e., maximum repetition. A result of $TTR = 1$ is a situation where no type is ever repeated. The benefits of TTR are its implementational and interpretational simplicity, and a normalized interval of possible results.

Results: TTR vectors
Calculating the TTR values for each n-gram size $1 \leq n \leq z$ results in a list (or a 'vector') of TTR values bound to their n by their order:

(4) $<TTR_1, TTR_2, TTR_3, \ldots, TTR_z>$

Such a vector might appear, at a first glance, to be rather complicated for purposes of interpretation, given the number of various TTR values. But at least three graphical methods have proven their usefulness for the results interpretation and sequence comparison. We present these in the following step.

Step 5: Interpretation, visualizations, and beyond

For the purpose of illustrating the various visualization methods, and their advantages alongside their disadvantages, we use the proposed method on a dataset of various texts of various types, described, along with the number of sequences per type of source, in Table 1. Random samples of $k = 6{,}000$ *bits* were taken.

Table 1. Various text/sequence types used for method illustration

Source/Type	Count
True-random sequences (see Haahr 2018)	100
Natural language: Czech	100
Natural language: 100 various language Bibles*	100
Monkey-Typed sequences**	21
Trivial repetitions	3
Source codes in C	100
Coding DNA sequences	100
Voynich manuscript (samples)***	100

* Only one sequence sample was used for each language.
** Monkey-typed sequences were created by 21 authors by freely hitting the keyboard.
*** The Voynich manuscript is presumably from the 15th century, written in an unknown script and of unknown nature. It has escaped translation for several centuries, see d'Imperio (1978) for a detailed survey.

3. Visualization methods

3.1 Basic line-chart

The simplest method for visualizing the resulting TTR vectors, easy to interpret, explore, and compare results of one or multiple sequences, is to draw a common spline chart where the *x* axis corresponds to *n*-gram size and the *y* axis to the measured TTR value. An example can be seen in Figure 1 and a summary displaying only mean averages per sequence type, in Figure 2. This type of visualization has many advantages, such as direct access to the TTR values, and simple comparison based on comparing curve behavior; this allows for a quick identification of identical sequences and behaviors common to all sequences (such as methods artifacts and models of the method, see further). Describing what we see in the chart from the very left to the right we notice that: truly random sequences are clustered on top of themselves with slight variation, to the right of these are DNA sequences suggesting their less random (arbitrary) nature, variety of pseudo-random 'monkey-typed' texts which contain many various patterns and further to the right from the least structured natural language texts (fiction books) to more structured texts (Bibles), programming languages to the right bottom, copying *x* axis for trivial repetitions. The Voynich manuscript shows a behavior which is not similar to that of any of the other texts. The main disadvantage of this type of visualization is, however, that overcrowding impairs readability.

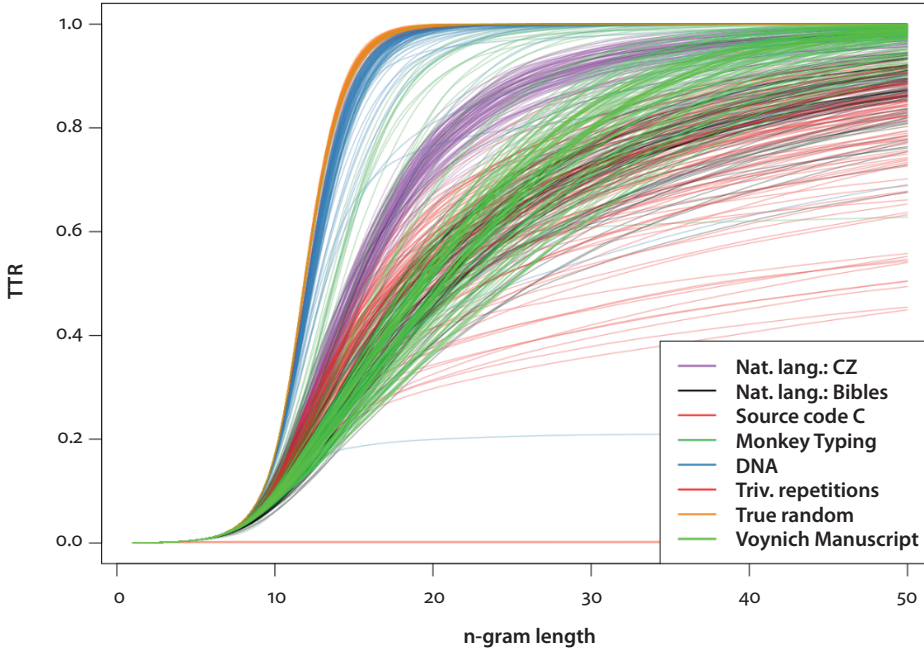


Figure 1. Line chart showing TTR vectors against *n*-gram length

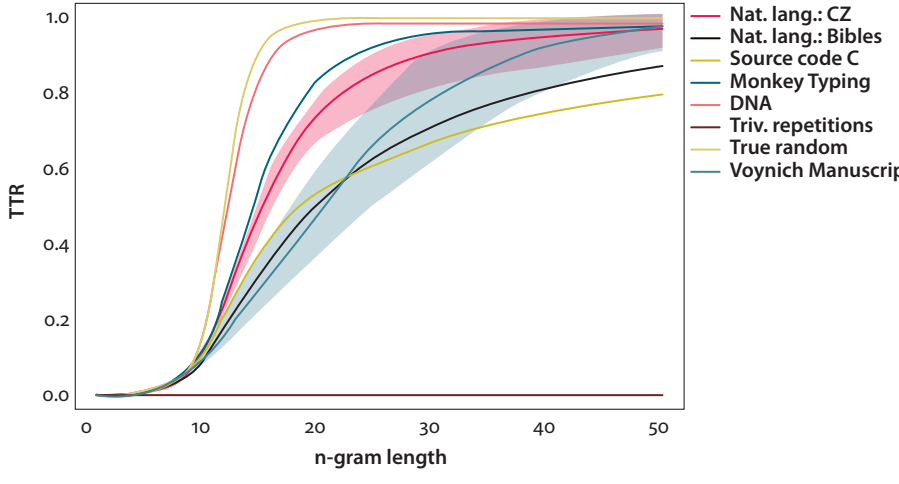


Figure 2. Line chart visualizing means of resulting TTR vectors for each sequence type with 95% confidence intervals for (illustratively chosen) Czech language texts and Voynich manuscript sequences

3.2 Classical Multidimensional Scaling (MDS)

An alternative to a line-chart for visualizing resulting TTR vectors is Classical Multidimensional Scaling (MDS; Torgerson 1958). MDS is typically used to reconstruct a 2D 'map' from pairwise distances of m-dimensional points. We can calculate pair-wise distances between all resulting TTR vectors by their Euclidean distance. The resulting chart then captures the distances (and thus the similarities) of all analyzed sequences at once, see Figure 3 for an example (axes have latent meaning enabling interpretation). The resulting chart is clearly more readable, from left to right: random sequences are clustered nearly on top of each other, neighboring with DNA sequences, pseudo-random monkey-typed texts, and blending into natural languages; the chart expands with programming language to the right towards trivial repetitions. Again, the Voynich manuscript shows a unique behavior. The main disadvantage of this method is possible overcrowding or wrong distance reconstruction stemming from reduction of high-dimensional spaces. MDS charts can 'hide' fine-graded details or distort some of the information in order to preserve larger variances (measurable by 'goodness-of-fit').

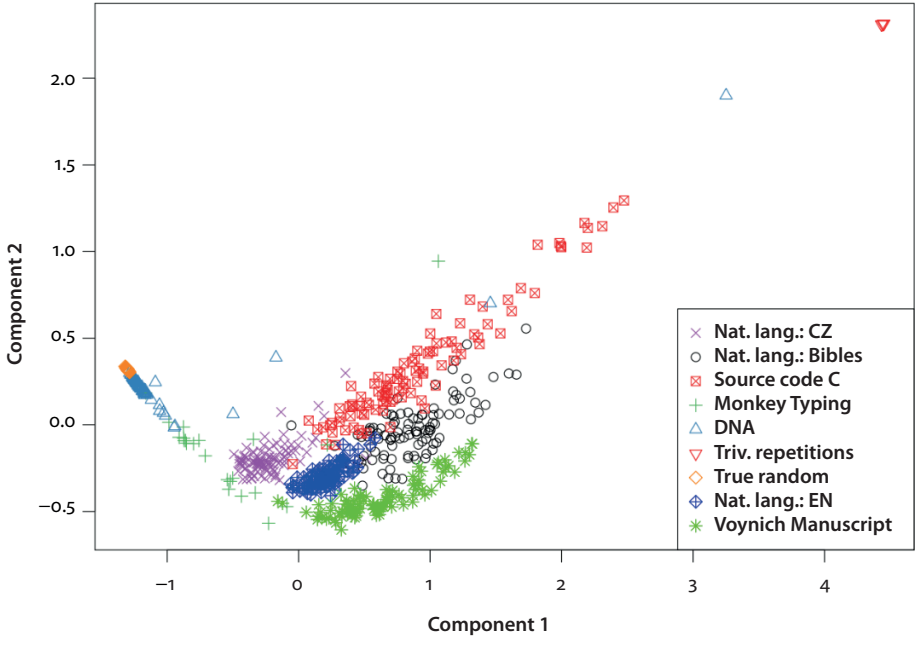


Figure 3. Classical metric MDS of resulting TTR vectors, distances calculated by Euclidean distance

3.3 t-Distributed Stochastic Neighbor Embedding

An alternative to MDS that copes with the possible overcrowding and displaying a more detailed view in exchange for losing global neighborhood relations is t-Distributed-Stochastic-Neighbor-Embedding (tSNE; Maaten & Hinton 2008). The main difference between MDS and tSNE in terms of interpretation is that the axes in tSNE do not follow any measurement, and only local neighborhoods (i.e., the most similar sequences) are placed together. This allows for a better view of details rather than of the full context, which is its main advantage. The visualization of TTR vectors using tSNE is depicted in Figure 4 with a very similar interpretation to that of MDS.

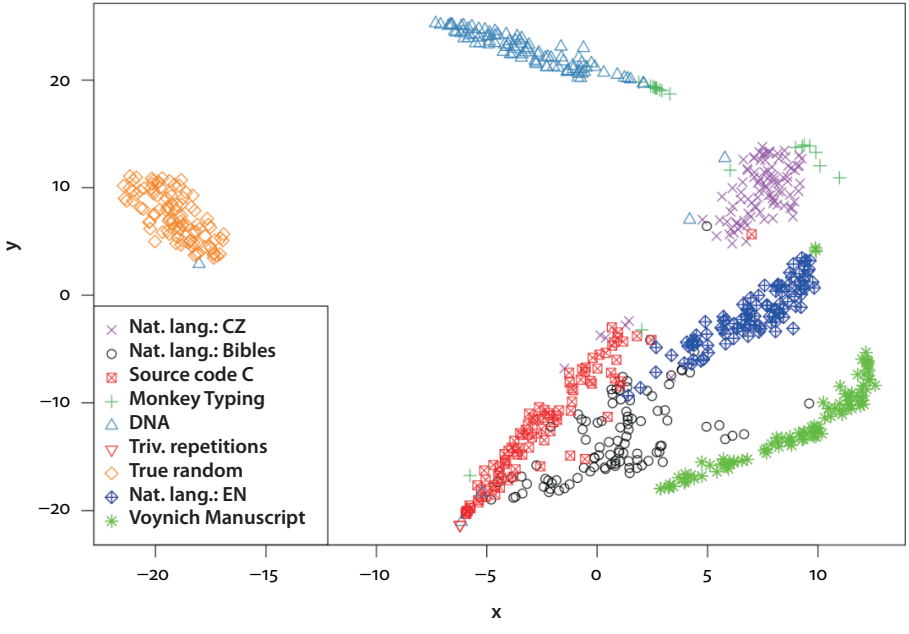


Figure 4. Visualization of TTR vectors by tSNE

4. Models of sequence behavior in our method

The method presented here has several inherent properties that can be used to define at least three models of sequence behavior: (i) truly random sequences, (ii) trivially repetitive sequences and (iii) systematic placement sequences. Moreover, we can define and explain some of the common behaviors of sequences that are caused by the method itself ('method artifacts') which should be known prior to interpreting

results. The models can be defined easily due to the simple essence of the method: relying on binary alphabet, having words of fixed length n, having a fixed sequence size k bits, and TTR. Practically all the models can also be displayed in the resulting charts as interpretative guidelines.

4.1 Model of truly random sequences

It is possible to formalize the expected behavior of truly random sequences and their TTR values respective for each tested n-gram size. Such formalization allows us to define confidence intervals up to a given level α and further allows us to test whether a given sequence is likely to be truly random by meeting α % expected behavior.

In order to formalize models, several other formalizations must be derived. First, we need to obtain the number of n-grams D in sequence of length k by (5):

$$D(n, k) = k - n + 1 \tag{5}$$

The maximum number of various binary n-grams (a size of potential vocabulary) is (6):

$$V(n) = 2^n \tag{6}$$

The average number of types for a given n-gram size n and a sequence length k can be calculated (analogously to Conroy 2018: 21–23) as (7):

$$V_{avg}(n,k) = V(n)\left(1 - \left(1 - \frac{1}{v(n)}\right)^{D(n,k)}\right) \tag{7}$$

Finally, we give a model of average TTR behavior of infinite truly random sequences with respect to n-gram size n for sequence of length k bits (8):

$$TTR_{avg}(n,k) = \frac{V_{avg}(n,k)}{D(n,k)} \tag{8}$$

For the purpose of statistical testing we derive a probability mass function of observing given number of types V in given sequence length k bits of $K = D(n, k)$ words (n-grams) of size n and size of the potential vocabulary $N = V(n)$ by (9); (Riedel 2018, modified; where $\left\{ {K \atop V} \right\}$ is a Stirling number of the second kind):

$$p(K,N,V) = \frac{1}{N^K}\binom{N}{V}V!\left\{ {K \atop V} \right\} = \frac{1}{N^K}\left\{ {K \atop V} \right\}\prod_{i=N-V+1}^{N} i \tag{9}$$

Calculating (9) by means of a program is rather problematic due to the large results of factorials. Thus, an alternative in terms of recurrence relations is suggested in (10):

$$(10)\quad p(K,N,V) = p(K-1,N,V)\,\frac{V}{N} + p(K-1,N,V-1)\left(1-\frac{V-1}{N}\right)$$
$$p(1,N,1) = 1$$
$$p(K,N,0) = 0 \text{ if } K > 0$$
$$p(0,N,V) = 0$$

From (9) or (10) we obtain the confidence interval [L; R] by (11):

$$(11)\quad CI_{Vavg}(n,k,\alpha) = [L;R]$$
$$K = D(n,k)$$
$$N = V(n)$$
$$N_{max} = \min(K,N)$$

where for $L, R \in \mathrm{N}$ applies $L \le R \le N_{max}$ and:

$$(12)\quad \sum_{v=1}^{L} p(K,N,v) \approx \alpha/2$$
$$\sum_{L \le v \le N_{max}}^{R} p(K,N,v) \approx \alpha/2$$

Testing true randomness given a sequence S of length k bits on level α is thus based on whether its results $TTR_n \in CI_{Vavg}(n,k,\alpha)$ for all examined n-gram sizes Z. Violating this condition for any $n \in Z$ rejects the hypothesis that S may be random. Any appropriate correction for multiple comparisons (e.g., Bonferroni's correction) should be employed for α with respect to $|Z|$. This test can accompany those from NIST (see Rukhin et al. 2001). However, rigorous analyses have shown that this test alone is not powerful enough to reject some of the pseudo-random generators and should be thus considered only as a necessary but insufficient condition to test true random sequences (Matlach 2018: 57–67).

Two different extreme behaviors models follow – one for the minimum and one for the maximum TTR values. These models set limits that cannot be surpassed.

4.2 Model of minimal TTR

The minimal value of TTR that can be observed for given n-grams and a sequence of length k can be trivially obtained by (13):

$$(13)\quad TTR_{min}(n,k) = 1/D(n,k)$$

Sequences following the minimal TTR model are trivial repetitions, where just 1 type of a given n-gram repeats itself.

4.3 Model of maximal TTR

The maximum number of *n*-gram types that a sequence of length *k* can produce is calculated by (14):

(14) $V_{max}(n, k) = min(V(n), D(n, k))$

and finally, the maximum TTR we can observe for any given *n*-gram size and a sequence of length *k* is obtained by (15):

(15) $TTR_{max}(n, k) = \frac{V_{max}(n, k)}{D(n, k)}$

Sequences following the model of maximal TTR are interesting: for each *n*-gram size *n*, the maximum of possible binary variations are observed in the sequence. However, to fulfill this condition, none of the tokens realized can repeat (in case the potential vocabulary contains enough types – see § 5.7). Generators of sequences following this behavior thus must have a memory built-in in order to intentionally evade the type repetition in cases where repetitions should occur naturally. For any sequences converging up to this model, the method presented thus suggests to us possible mechanics of the theoretical generator and thus allows us to make hypotheses about its strong generative power.

5. Method artefacts and specific *n* intervals

In addition to the three preceding models, we can map a general behavior of the method. As noticed in simple line-chart visualization in Figure 1 – nearly all sequences share TTR behavior in some of the intervals. Further we explain and formalize such behavior as formal consequences of the methods principle and mark this behavior as artifacts, which should be considered while interpreting the resulting data. A common reason for such similar behavior is based on the relation of the sequence length *k*, the size of *n*-grams, which induces potential vocabulary of 2^n various words, and the measure of TTR.

5.1 Interval of exhausting vocabulary Q

The first similar sequence behavior (Figure 1) can be noticed for *n*-grams of lengths 1–7. This behavior is caused by too little potential vocabulary *V* of binary *n*-grams ($|V| = 2^n$) which is used (or 'exhausted') even by chance. We can thus specify the interval *Q* for *n*-gram size *n* where we expect that all possible *n*-grams will be used at least once in a sequence of length *k* bits by finding a critical *n* size *q(k)* as (16) (in analogy to the 'Coupon collector' problem, see Mitzenmacher & Upfal 2005):

$$(16)\quad q(k) = \arg\max_{n\in\mathbb{N}} n \cdot V(n) \sum_{i=1}^{V(n)} \frac{1}{i} < k$$

From (16) we obtain interval $Q = [1; q(k)]$. For example, for a sequence of $k = 6000$ bits, the interval, where using all possible n-grams at least once is expected and thus the TTR values of any (same length) nontrivial sequences, should be principally the same, that is $Q = [1;q(6000)] = [1;7]$.

5.2 Interval of vocabulary saturation C

In contrast to an exhaustive interval Q is a situation for n where the potential vocabulary is so huge that we do not expect that *any* of the words would repeat and thus we expect $TTR = 1$ for a truly random sequence. An interval C of n-gram lengths can be specified for any given sequence of bit length k by finding the critical length n size $c(k)$ as (17):

$$(17)\quad c(k) = \arg\min_{n\in\mathbb{N}} k < n \cdot e^{V(n)}\, V^{-V(n)}\, \Gamma\big(V(n)+1, V(n)\big).$$

However, due to large numbers in $V(n)$, the following recurrent variation is more viable for implementation (18):

$$(18)\quad E(r) = 1 + \frac{r}{v(n)}\, E\,(r-1)$$
$$E(0) = 1$$
$$c(k) = n \cdot E(V(n))$$

The final interval C is defined as $C = [c(k); +\infty]$. For example, for $k = 6000$, the interval $C = [c(6000);+\infty]=[17; +\infty]$. Since $n = 17$, we expect for the random sequence that their TTR values should converge to 1.

5.3 Interval of maximum variance H

In between the two intervals described, Q and C is an interval where n-grams are neither large enough for never being repeated randomly nor too short for being picked averagely at least. It is thus an interval of n sizes allowing maximum TTR values variation in a given sequence of length k. Such an interval H may be defined by (11) and (14) as $H = [q(k) + 1;c(k)-1]$. For example, for $k = 6000$ we obtain $H = [q(6000) + 1;c(6000)-1] = [8;16]$ where we should observe the highest variance between all TTR results for tested sequences.

6. Comparison to other, similar purpose methods

We may find several methods algorithmically similar to the one presented here, as they share a similar idea on quantifying repetitions of binary combinations. In general, compression algorithms are commonly based on this type of analysis (see Ziv & Lempel 1978). Hamano & Yamamoto (2010) present a method of *t*-complexity devoted only to testing sequence randomness. However, the most similar method along with general text comparison aims is method presented in Rao et al. (2009) and Rao (2010), which was also heavily criticized in Sproat (2010) and further in Sproat (2014) for lacking formalization of the model and explanation of parameters. The similarities between the two methods are based on calculating the entropy of gradually increasing *n*-grams and plotting the results into a line-chart. However, the most problematic property of Rao's method is lacking an effective alphabet size normalization, thus implying different results for sequences created by the same generator but using different alphabets (e.g., the same random sequence in a binary format and encoded by base64 code which is treated as the same by the method presented). Also, entropy results are not normalized/bound to any interval (in comparison to TTR bound to [0; 1]) which further complicates comparison of sequences.

7. Conclusions

The method presented in this chapter provides a simple and effective way to describe the combinatorial behavior of sequences, vectorize them and use it for their comparison, clustering and estimating the quality of their generators.

We have shown that the method can be used for clustering texts by language or type (random sequences, DNA, or source-codes). The resulting sequence vectors are suitable for plotting and interpreting in the simplest manner as line-charts or by multidimensional visualization techniques as MDS or tSNE. The resulting embeddings may also be used as input features for any machine learning algorithms and thus allow automating sequence type detection.

Although we have explored and formalized several of the method's artifacts, a possible impact of the alphabet size is still possible. Repeating the step of randomized binary alphabet assignment and following recalculations for TTR should be considered for more robust results. The other possibility to encode the symbols is to use Huffman coding (see Huffman 1952).

It is also critical to understand that the method does not determine the type of the sequence but provides information on similarity to the other tested/compared

sequences. Any type of measured proximity does not ensure that the sequences are of the same type or come from the same source.

The proposed statistical test for true random sequences does not have enough statistical power to reject pseudorandom sequences and the test should be considered as necessary, but insufficient, condition for true randomness (in contrary to Hamano & Yamamoto 2010).

Using normalized entropy (or any other distribution reflecting measures) instead of TTR may be beneficial by adding information and power to the method. Considering fuzzy *n*-gram matching or skip-grams would also be beneficial, e.g., for DNA sequence analysis. Such beneficial changes would, however, complicate the formalization of the methods artifacts and the other models.

Funding

Statement: VM acknowledges the financial support of the Faculty of Arts of Palacký University Olomouc in 2019 from the Academic Research Support Fund via the project grant Metody lingvistické analýzy v Digital Humanities, no. IGA_FF_2019_019.

Statement: This work was supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (4–5). 993–1022.

Bolshoy, Alexander, Zeev (Vladimir) Volkovich, Valery Kirzhner & Zeev Barzily. 2010. *Genome clustering from linguistic models to classification of genetic texts*. Berlin: Springer.

Conroy, Matthew M. 2018. A collection of dice problems. https://www.madandmoonly.com/doctormatt/mathematics/dice1.pdf (16 August, 2018.)

Cornwell, Benjamin. 2015. *Social sequence analysis: Methods and applications*, (Structural analysis in the social sciences 37). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316212530

d'Imperio, Mary E. 1978. *The Voynich manuscript: An elegant enigma*. Fort George G. Meade, MD: National Security Agency/Central Security Service.

Gastwirth, Joseph L. 1972. The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics* 54(3). 306–316. https://doi.org/10.2307/1937992

Govindan, Vidya, Rajat Subhra Chakraborty, Pranesh Santikellur & Aditya Kumar Chaudhary. 2018. A hardware Trojan attack on FPGA-based cryptographic key generation: Impact and detection. *Journal of Hardware and Systems Security* 2. 225–239. https://doi.org/10.1007/s41635-018-0042-5

Haahr, Mads. 2018. *True random integer generator, RANDOM.ORG: True Random Number Service*. Randomness and Integrity Services Ltd.

Hamano, Kenji & Hirosuke Yamamoto. 2010. Randomness test based on T-complexity. *Communications and Computer Sciences* E93-A(7). 1346–1354. https://doi.org/10.1587/transfun.E93.A.1346

Hamid, Raffay, Amos Johnson, Samir Batta, Aaron Bobick, Charles Isbell & Graham Coleman. 2005. Detection and explanation of anomalous activities: representing activities as bags of event *n*-grams. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1. 1031–1038. https://doi.org/10.1109/CVPR.2005.127

Huffman, David A. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE* 40(9). 1098–1101. https://doi.org/10.1109/JRPROC.1952.273898

Jain, Ashish & Narendra S. Chaudhari. 2015. A new heuristic based on the cuckoo search for cryptanalysis of substitution ciphers. In Sabri Arik, Tingwen Huang, Weng Kin Lai & Qingshan Liu (eds.), *Neural Information Processing* (Lecture Notes in Computer Science 9490). 206–215. Dordrecht: Springer. https://doi.org/10.1007/978-3-319-26535-3_24

Lasry, George. 2018. *A methodology for the cryptanalysis of classical ciphers with search metaheuristics*. Kassel: Kassel University Press.

Maaten, Laurens Van Der & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(9). 2579–2605.

Matlach, Vladimír. 2018. Aplikace kvantitativní lingvistiky na analýzu sekvencí. Olomouc: Palacký University Olomouc PhD dissertation. https://theses.cz/id/15yzyh/disertace_matlach_.pdf (5 December, 2019.)

Mikros, George & Jan Macutek (eds.). 2015. *Sequences in language and text*, Volume 69. Berlin: Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110362879

Mitzenmacher, Michael & Eli Upfal. 2005. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511813603

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2). 945–959. https://doi.org/10.1093/genetics/155.2.945

Rao, Rajesh P. 2010. Probabilistic analysis of an ancient undeciphered script. *Computer* 43(4). 76–80. https://doi.org/10.1109/MC.2010.112

Rao, Rajesh P., Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari & Iravatham Mahadevan. 2009. Entropic evidence for linguistic structure in the Indus script. *Science* 324. 1165. https://doi.org/10.1126/science.1170391

Riedel, Marko. 2018. *Probability of throwing exactly V distinct sides on N sided dice by K rolls*. https://Math.Stackexchange.Com/Q/2857744 (20 June, 2018.)

Rukhin, Andrew, Juan Soto, James Nechvatal, Miles Smid & Elaine Barker. 2001. *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. Booz-Allen and Hamilton Inc Mclean VA.

Schenkel, Alain, Jun Zhang & Yi-Cheng Zhang. 1993. Long range correlations in human writings. *Fractals* 1(1). 47–55. https://doi.org/10.1142/S0218348X93000083

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Sproat, Richard. 2010. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36(3). 585–594. https://doi.org/10.1162/coli_a_00011

Sproat, Richard. 2014. A statistical comparison of written language and nonlinguistic symbol systems. *Language* 90(2). 457–481. https://doi.org/10.1353/lan.2014.0031

Stuttard, Dafydd & Marcus Pinto. 2011. *The web application hacker's handbook: Finding and exploiting security flaws*. Indianapolis: Wiley.

Torgerson, Warren S. 1958. *Theory and methods of scaling*. New York: Wiley.

Ziv, Jacob & Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* 24(5). 530–536. https://doi.org/10.1109/TIT.1978.1055934

Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian

Biljana Rujević[1], Marija Kaplar[2], Sebastijan Kaplar[2], Ranka Stanković[1], Ivan Obradović[1] and Ján Mačutek[3,4]
[1]University of Belgrade / [2]University of Novi Sad / [3]Mathematical Institute of Slovak Academy of Sciences / [4]Constantine the Philosopher University in Nitra

Ten chapters from a Russian novel and its translations into Croatian, Serbian, and Ukrainian are automatically syllabified following the same approach in all four languages. Syllable frequencies and syllable length are modelled by probability distributions which are commonly used for frequency and length of words (the Zipf-Mandelbrot distribution and the Dacey-Poisson distribution, respectively). We show that Zipf's law of brevity, according to which the more frequent words tend to be shorter, can be extended to syllables. We suggest a generalization of the Menzerath-Altmann law, a relation between word length and the mean syllable length. The generalized version of the law is valid for both word types and word tokens.

Keywords: syllable, rank-frequency distribution, length, law of brevity, Menzerath-Altmann law, Slavic languages

1. Introduction

The syllable, being one of the basic linguistic units, is notoriously difficult to define. Citing Haugen (1956: 213), "[w]hile sooner or later everyone finds it convenient, no one does much about defining it". More than 50 years later, Cairns & Raimy (2011: 1) say – on the very first page of their Handbook of Syllable – that "matters are hardly better now than they were then". In such a situation, different scholars apply different definitions, which makes their results incomparable (see also a short discussion on the problem in Radojičić et al. 2019). This is also the case of the book by Zörnig et al. (2019), where the syllable structure in many languages is investigated, but the authors rely on "prescriptions written for the given language by

https://doi.org/10.1075/cilt.356.04ruj

linguists" (Zörnig et al. 2019: 12), i.e., they follow language specific syllabification rules (which can be, and often are, also theory-specific).

This chapter focuses on quantitative analysis of some properties of syllables in four Slavic languages: Croatian and Serbian (South Slavic), and Russian and Ukrainian (East Slavic). We use a syllable definition which combines the maximum onset principle and the sonority sequencing principle (see § 2 for details). Our choice of the syllable definition is pragmatic (it is applicable for the four languages under study and thus allows us to compare results), we do not have any ambition to provide answers to open questions or to settle disputes in this area. Kelih (2012) presents an overview of several established approaches to the syllable, together with an attempt to integrate the syllable into the synergetic model of language units and their properties (see, e.g., Köhler 2005, 2011).

The analysis is performed on several levels: (1) syllable frequency (it can be modelled by the Zipf-Mandelbrot distribution for all languages under study), (2) syllable length (the Dacey-Poisson distribution is a well-fitting model), (3) the relation between syllable frequency and length (more frequent syllables tend to be shorter), (4) the Menzerath-Altmann law, which is a functional relation between word length (measured in syllables in this paper) and the mean syllable length (measured in phonemes here). This chapter can be considered a continuation of the pilot study by Radojičić et al. (2019), where results on syllable frequency, syllable length, and their mutual relation in Serbian are presented.

2. Methodology and language material

The first ten chapters of the Russian socialist realist novel *Kak zakalyalas' stal'* ("How the Steel Was Tempered") by N. Ostrovsky are used as the language material (Radojičić et al. 2019 used a complete Serbian translation of the novel). The choice is motivated by the fact that a parallel corpus, consisting of the first ten chapters of the novel and their translations into all standard Slavic languages (except for Lower Sorbian), is available (Kelih 2009), which make it possible to conduct typological studies on the level of the syllable when texts from all Slavic languages are syllabified and investigated.

Words were divided into syllables automatically. As every syllable contains a nucleus (a vowel or a syllabic consonant), we first had to solve the problem of zero-syllabic words (each of them consisting of one non-syllabic consonant) which occur in all the languages we work with (prepositions, and in the case of Ukrainian also two particles and a conjunction). We decided to attach them to the words with which they form one phonological unit (i.e., all prepositions to the next word, and,

in Ukrainian, the conjunction to the next word and the particles to the preceding word) – e.g., *s njim* "with him" in Croatian is rewritten as *snjim*. This approach was adopted by Antić et al. (2006) in the context of word length research.

The software tool that was used (http://kempelen.dai.fmph.uniba.sk/slabiky) follows the principles described in Radojičić et al. (2019) – i.e., the maximum onset principle and the sonority sequencing principle. In the first step, all syllables are deemed to end after their nuclei, i.e., after a vowel or a syllabic consonant. The maximum onset principle is 'blindly' respected in this step, and thus, preliminarily, all syllables (the last syllable in a word being a possible exception, as it must contain its coda if there is one) are kept open. If, after the first step, consonant clusters occur in intervocalic positions, the borders between syllables are reconsidered taking into account the sonority sequencing principle (according to this principle, "[b]etween any member of a syllable and the syllable peak, a sonority rise or plateau must occur" in the syllable onset, see Blevins 1995: 210). We admit possible exceptions at the beginnings of words. If a word begins with a consonant cluster that violates the sonority sequencing principle, i.e., if the sonority decreases in the onset of the first syllable (examples from different languages are presented in Clements 1990: 288), we take these onsets as they are.

As far as the sonority scale is concerned, the simplest one is used – only sonorants (approximants and nasals) and obstruents (all the other consonants) are distinguished. There are other, more fine-grained scales (see, e.g., Clements 1990 and Zec 1995), but the one we use is "sufficient to capture the most common subdivisions of segments with respect to sonority" (Zec 1995: 86).

In the following, we present those aspects of the graphemic representations of phonemes in the languages under study which are applied in the syllabification algorithm:

1. Croatian (Barić et al. 1997) has a phoneme inventory consisting of five vowels (a, e, i, o, u) and 25 consonants – eight sonorants (j, l, lj, m, n, nj, r, v) and 17 obstruents (b, c, č, ć, d, dž, đ, f, g, h, k, p, s, t, z, ž). The consonant r is syllabic if it occurs between two other consonants and the one which follows r is not j (i.e., *crni* "black" is syllabified as *cr-ni*, but *strjelica* "arrow" as *strje-li-ca*).
2. Serbian (Piper & Klajn 2013) uses both the Latin and Cyrillic alphabets (the source of our data is written in Cyrillic). Apart from this orthographic aspect, it shares the phonological properties relevant to the purposes of this paper with Croatian, i.e., it has five vowels (a/а, e/е, i/и, o/о, u/у) and 25 consonants, out of which eight are sonorants (j/ј, l/л, lj/љ, m/м, n/н, nj/њ, r/р, v/в) and 17 are obstruents (b/б, c/ц, č/ч, ć/ћ, d/д, dž/џ, đ/ђ, f/ф, g/г, h/х, k/к, p/п, s/с, t/т, z/з, ž/ж). The consonant r is syllabic if it occurs between two other consonants.

3. Syllable nuclei in Russian are represented by graphemes а, и, о, у, ы, э (each of them is pronounced as one vowel), and also by е, ё, ю, я (in some circumstances, the pronunciation of these graphemes includes the voiced palatal approximant <j> and a vowel). The consonant phonemes are written as й, л, м, н, р (sonorants) and б, в, г, д, ж, з, к, п, с, т, ф, х, ц, ч, ш, щ (obstruents), see, e.g., Avanesov (1956). Most (although not all) consonants form soft (palatalized) and hard (non-palatalized) pairs, but the palatalization does not affect the consonant's status of being a sonorant or an obstruent. The Russian alphabet also contains the so-called hard sign ъ and soft sign ь. They themselves do not correspond to any phonemes, but they indicate the palatalization status of the consonant which precedes them.
4. In Ukrainian (Ponomariv 2001), syllabic nuclei are represented by а, е, и, і, о, у (each of them is pronounced as one vowel) and by є, ї, ю, я (in some cases pronounced as the voiced palatal approximant <j> and a vowel). Sonorant consonants are represented by в, й, л, м, н, р and obstruents by б, г, ґ, д, ж, з, к, п, с, т, ф, х, ц, ч, ш, щ. Some consonants occur in both palatalized and non-palatalized variants, which does not change their sonority (or the lack thereof). The Ukrainian orthography makes use of a non-vocalic letter, the soft sign ь, and the apostrophe ' not considered a part of the alphabet. Both of them determine the palatalization status of the consonant they follow.

The outputs of automatic syllabification were checked manually (they contained a few errors, mostly caused by OCR, which were corrected; but the texts contained also some 'non-words', such as abbreviations, which were deleted).

3. Results

3.1 Syllable frequency

The automatic syllabification of the texts described above yields 119,601 syllable tokens and 2314 syllable types for Croatian, 118,702 tokens and 2213 types for Serbian, 107,860 tokens and 3607 types for Russian, and 109,458 tokens and 3441 types for Ukrainian. The ten most frequent syllables in each language are presented in Table 1. The complete data can be found on http://rgf.rs/projekti/bil/sk/SK-SRB-2016-0021.html#results.

Table 1. Ten most frequent syllables in Croatian, Serbian, Russian, and Ukrainian (i – rank, f_i – frequency)

i	Croatian		Serbian		Russian		Ukrainian	
	Syllable	f_i	Syllable	f_i	Syllable	f_i	Syllable	f_i
1	o	5360	о	5379	е	2580	на	2393
2	je	4311	је	3563	о	2347	по	2122
3	u	2945	у	2920	на	2273	го	1971
4	i	2722	и	2779	по	2260	не	1716
5	na	2524	да	2550	и	2206	ли	1662
6	da	2478	на	2498	го	1965	за	1653
7	po	2258	се	2491	не	1698	ти	1648
8	se	2255	по	2242	за	1549	ко	1439
9	ko	2108	ко	2206	у	1521	до	1434
10	li	1905	ли	1811	я	1465	ва	1368

For the ranked frequencies of syllables, the Zipf-Mandelbrot distribution (Wimmer & Altmann 1999: 666),

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

achieves a good fit (it is evaluated – as is common in quantitative linguistics – in terms of the discrepancy coefficient C, with $C < 0.02$ indicating a satisfactory fit, see Mačutek & Wimmer 2013). Parameter values which achieve the best fit are presented in Table 2 (the distribution has two free parameters a and b; n is the number of syllable types; k is a normalization constant which ensures that the probabilites sum to one, and it is determined by the values of a, b and n).

Table 2. Syllable ranked frequencies modelled by the Zipf-Mandelbrot distribution

	a	*b*	*n*	*C*
Croatian	1.725	22.521	2314	0.0166
Serbian	1.761	24.452	2213	0.0168
Russian	1.486	22.468	3607	0.0103
Ukrainian	1.626	32.993	3441	0.0095

The same model is often used for word frequencies (e.g., a chapter in the book by Popescu et al. 2009: 127–193), and also for ranking purposes in many other scientific disciplines (see, e.g., Izsák 2006 and references therein). For all four languages under study, the conclusions from Radojičić et al. (2019) remain true: while the hypothesis from Strauss et al. (2008), according to which the ranked frequencies of

syllables can be modelled by the same distribution as those for words, is corroborated, the parameter values for syllables from Table 1 differ quite dramatically from those for words (see Popescu et al 2009: 137–138).

3.2 Syllable length

In all four languages considered, the frequencies of syllable lengths (measured by the number of phonemes they consist of) can be modelled by the Dacey-Poisson distribution (Wimmer & Altmann 1999: 111; here shifted to the right by 1), with

$$P_x = \frac{(1-a)\lambda^{x-1}e^{-\lambda}}{(x-1)!} + \frac{a(x-1)\lambda^{x-2}e^{-\lambda}}{(x-1)!}, \quad x = 1,2, \ldots$$

The distribution has two free parameters, a and λ. The fitted parameter values are presented in Table 3.

Table 3. Frequencies of syllable length modelled by the Dacey-Poisson distribution

Length	Croatian	Serbian	Russian	Ukrainian
1	12164	12187	7155	4255
2	74944	75377	58852	65194
3	28018	27032	34136	34077
4	4252	3882	6553	5581
5	219	221	722	353
6	4	3	42	18
λ	0.358	0.344	0.518	0.455
a	0.855	0.856	0.890	0.939
C	0.0032	0.0029	0.0092	0.0123

Syllables behave analogously to words also with respect to this property. The Poisson distribution or one of its many generalizations or modifications (see Wimmer & Altmann 1999: 493–504) is a model which fits the word length distribution well in many languages, see Best (2005), Popescu et al. (2013), and also the bibliography by Karl-Heinz Best at http://wwwuser.gwdg.de/~kbest/litlist.htm. We remind the reader that Radojičić et al. (2019) suggested that we model the frequencies of word lengths in the whole text of the Serbian translation of *Kak zakaljalas' stal'* by another 'relative' of the Poisson distribution, namely the hyper-Poisson distribution.

3.3 Relation between syllable frequency and syllable length

Following the approach from Radojičić et al. (2019), we explored also another aspect of syllables, namely the relation between syllable frequency and syllable length. We found a statistically significant relation between these two syllable properties (the Spearman correlation coefficients are −0.422 for Croatian, −0.423 for Serbian, −0.404 for Russian, and −0.431 for Ukrainian, with the p-values less than 0.001 in all four cases). Thus, the conjecture from Radojičić et al. (2019), according to which there is a negative correlation between frequency and length for syllables, is corroborated on other languages.

This relation is an extension of the famous Zipf's law of brevity (Zipf 1949), which states that frequent words tend to be short (for recent results see works by Hernández-Fernández et al. 2016 and by Casas et al. 2019; an extensive study of 1262 texts from 986 languages by Bentz & Ferrer-i-Cancho 2016 deserves a special mention). To the best of our knowledge, this paper (together with the one by Radojičić et al. 2019) is the first report on this relation for syllables.

The general tendency – more frequent units are shorter – is the same for both syllables and words, but the correlation for syllables seems to be stronger (see also Radojičić et al. 2019). Ferrer-i-Cancho & Hernández-Fernández (2013) computed the Spearman correlation coefficients between the word frequency and word length (which they measured by the number of letters) in seven languages. The correlation is −0.269 for Croatian, the only language which is included both in this study and in the paper by Ferrer-i-Cancho & Hernández-Fernández (2013). This value represents the strongest correlation in their study. Admittedly, we have a very modest sample (five texts – if we consider the whole novel and its part to be different texts – from four closely related languages), but we allow ourselves to formulate a hypothesis that the stronger correlation is related to the (much) lower inventory of syllables (in comparison to words). Naturally, the fact that Ferrer-i-Cancho & Hernández-Fernández (2013) expressed word length in letters, and not in syllables or morphemes, which are immediate components of words, can also play an important role,[1] and a re-analysis of their data, with word length expressed in syllables, could lead to a stronger correlation.

1. Köhler (2012: 108) writes, albeit in a slightly different context, that "an indirect relationship … is a good enough reason for more variance in data and a weaker fit" – in this case for a weaker correlation.

3.4 Menzerath-Altmann law

While the results on syllable frequency, syllable length, and their mutual relation could be expected (first, they are analogous to the behaviour of word properties; second, we analyze four Slavic languages; i.e., they are relatively closely related to Serbian, for which these results were reported by Radojičić et al. 2019), it is quite surprising that the commonly used formula for the Menzerath-Altmann law (see Cramer 2005) does not achieve a good fit.

The Menzerath-Altmann law states that "the mean size of constituents is a function of the size of the construct" (Mačutek et al. 2019: 67, see also a slightly less general formulation in Altmann 1980). The mathematical model which so far has been considered general enough is function

$$(1) \quad y(x) = ax^b e^{-cx},$$

where $y(x)$ is the mean size of constituents if the size of the construct is x; a, b, c are parameters. In our context, words are constructs with their size measured by the number of syllables, while syllables are constituents (their size is measured by the number of phonemes they contain). The appropriateness of the model is expressed by the determination coefficient R^2. If $R^2 > 0.9$, the fit is usually considered good, see Mačutek & Wimmer (2013).

Papers on the relation between word length and syllable length focus mostly on word types (Menzerath 1954 in German; Roberts 1965 in English; Altmann & Schwibbe 1989 in Indonesian; Grzybek 1999 in Croatian; Kelih 2010 in Serbian). There are two exceptions which investigate the relation taking into account word tokens. Mikros & Milička (2014) work with a Modern Greek corpus and Mačutek et al. (2019) examine spoken Czech. In both cases, the fit is problematic (or even clearly not good enough) for some texts.

Our data are presented in Table 4. We follow the approach of Mačutek & Rovenchak (2011) – only word lengths which occur at least 10 times in the text were taken into account (we did not take into consideration five eight-syllable and six nine-syllable words in Croatian, seven eight-syllable and six nine-syllable words in Serbian, one nine-syllable word in Russian, and six eight-syllable and one nine-syllable words in Ukrainian). The abovementioned function does not fit our data (see Table 4) well enough.

The tendency of syllable length to decrease with increasing word length can be observed from length 2 further on, but the data are not 'smooth enough' to be fitted by function (1).

There are at least two reasons which can explain the failure of the well-established model. First, a novel is usually considered to be too long to constitute a homogeneous text. Popescu et al. (2009: 3) suggest 10,000 words as the upper limit of a

Table 4. The Menzerath-Altmann law on the word-syllable level

Word length	Croatian	Serbian	Russian	Ukrainian
1	2.06	2.06	2.32	2.37
2	2.31	2.29	2.53	2.49
3	2.24	2.22	2.41	2.41
4	2.17	2.16	2.33	2.31
5	2.11	2.09	2.27	2.22
6	2.08	2.07	2.25	2.22
7	2.08	2.11	2.21	2.27
8			2.20	

homogeneous text. Admittedly, it is an ad hoc limit, but long texts are necessarily mixtures (of particular chapters, of replicas in dialogues, etc.). Second, the possibility that function (1) is not general enough to fit data obtained from word tokens cannot be excluded (we remind the reader that Mikros & Milička 2014 and Mačutek et al. 2019 also report the goodness-of-fit problem). Most probably, Zipf's law of brevity (which, naturally, does not impact the level of types) comes into play again. If we focus only on monosyllable words, the law predicts that shorter words (coinciding with syllables in this case) occur more frequently than longer ones (and thus the mean syllable length of short words should be greater for types than for tokens). This tendency is evident if we compare the mean length of monosyllables. Kelih (2010) reports the values of 3.09 for types in a corpus of Serbian literary prose, while the mean length of monosyllables in the Serbian text under study with tokens taken into account is 2.06.

A generalization of (1), namely, function

$$(2) \quad y(x) = ax^{b + c \log x}\, e^{-dx},$$

provides an excellent fit to all our data as well as to those from Mikros & Milička (2014). The fit of its slight modification,

$$(3) \quad y(x) = y(1)\, x^{b + c \log x}\, e^{-d(x-1)},$$

with parameter a replaced with $y(1)$, i.e., with the mean syllable length in one-syllable words (see also Kelih 2010) remains very good. In such a case the model has, as in function (1), three free parameters. Data from the Russian text fitted by functions (1) and (3) are presented in Figure 1 (computations were performed by NLREG software). We chose Russian as the example here because function (1) achieves a better – albeit not satisfactory – fit for this language than for the other three (see Table 4 for the data).

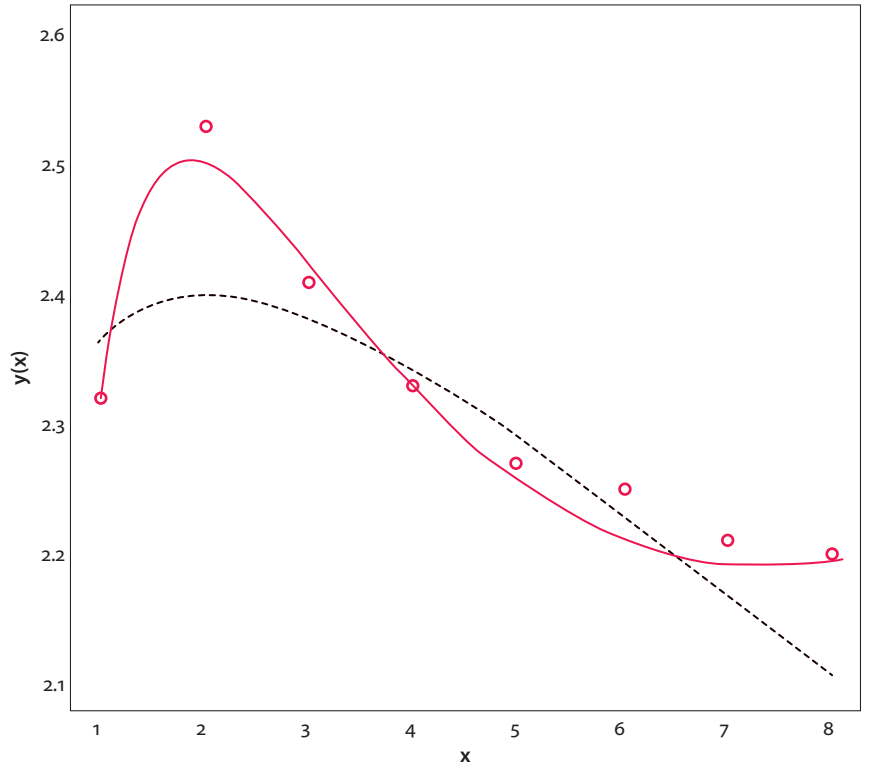


Figure 1. The relation between word length and the mean syllable length in the Russian text. Solid line – function (3), with $b = 0.10$, $c = -0.32$, $d = -0.16$, $R^2 = 0.98$. Dashed line – function (1), with $a = 2.46$, $b = 0.08$, $c = -0.04$, $R^2 = 0.74$

Questions of whether function (3) is a general model of the Menzerath-Altmann law, and, if yes, how its parameters can be interpreted, remain open for the time being.

4. Conclusions

This paper brings some new progress in understanding syllables. The hypotheses of Strauss et al. (2008) and Radojičić et al. (2019) that syllable frequency and syllable length behave analogously to the respective properties of words are corroborated on further languages. Moreover, there is a negative correlation between syllable frequency and syllable length. A generalization of the mathematical formula for the Menzerath-Altmann law is suggested. The generalized version of the law can also model the relation between word length and the mean syllable length in texts (and not only in dictionaries).

Two out of three branches of the Slavic languages (South and East Slavic) are represented in this paper. The results of our analyses indicate that syllable properties could serve also for typological purposes. In Tables 2 and 3, the parameter values for the South Slavic languages and for the East Slavic languages seem to form two groups (the same is true for mean syllable length, which is 2.21 for Croatian, 2.20 for Serbian, 2.40 for Russian, and 2.38 for Ukrainian; the differences among the means were not tested, because virtually any null is rejected in terms of p-values for such large samples, see Mačutek & Wimmer 2013). In addition, one can observe that the parameter values for Croatian and Serbian are closer to each other than the ones for Russian and Ukrainian (admittedly, the picture could change when syllable properties in the other Slavic languages are examined).

Funding

This research was supported by research projects APVV SK-SRB-2016-0021 (B. Rujević, M. Kaplar, R. Stanković, J. Mačutek) and VEGA 2/0096/21 (J. Mačutek).

References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn (ed.), *Glottometrika 2* (Quantitative Linguistics 3), 1–10. Bochum: Brockmeyer.

Altmann, Gabriel & Michael H. Schwibbe. 1989. *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Georg Olms.

Antić, Gordana, Emmerich Kelih & Peter Grzybek. 2006. Zero-syllable words in determining word length. In Peter Grzybek (ed.), *Contributions to the science of text and language. Word length studies and related issues* (Text, Speech and Language Technology 31), 117–156. Dordrecht: Springer.

Avanesov, Ruben I. 1956. *Fonetika sovremennogo russkogo literaturnogo jazyka*. Moscow: Izdatel'stvo Moskovskogo universiteta.

Barić, Eugenija, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević & Marija Znika. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Bentz, Christian & Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*. Tübingen: University of Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558. (17 January, 2020.)

Best, Karl-Heinz. 2005. Wortlänge. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics. An international handbook* (Handbooks of Linguistics and Communications Science 27), 260–273. Berlin: de Gruyter.

Blevins, Juliette. 1995. The syllable in the phonological theory. In John Goldsmith (ed.), *The handbook of phonological theory*, 206–244. Oxford: Blackwell.

Cairns, Charles & Eric Raimy. 2011. Introduction. In Charles E. Cairns & Eric Raimy (eds.), *Handbook of the syllable* (Brill's Handbooks in Linguistics 1), 1–30. Leiden: Brill.

Casas, Bernardino, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-Cancho & Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. *Computer Speech & Language* 58. 19–50. https://doi.org/10.1016/j.csl.2019.03.007

Clements, George N. 1990. The role of the sonority cycle in core syllabification. In John Kingston & Mary E. Beckman (eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*, 283–333. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.017

Cramer, Irene M. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics. An international handbook* (Handbooks of Linguistics and Communications Science 27), 659–688. Berlin: de Gruyter.

Ferrer-i-Cancho, Ramon & Antoni Hernández-Fernández. 2013. The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory* 4(1). 45–55. https://doi.org/10.1524/glot.2013.0004

Grzybek, Peter. 1999. Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen. In Branko Tošović (ed.), *Die grammatischen Korrelationen*, 67–77. Graz: Institut für Slawistik.

Haugen, Einar. 1956. The syllable in linguistic description. In Morris Halle, Horace G. Lunt, Hugh McLean & Cornelis H. van Schooneveld (eds.), *For Roman Jakobson: Essays on the occasion of his sixtieth birthday*, 213–221. The Hague: Mouton.

Hernández-Fernández, Antoni, Bernardino Casas, Ramon Ferrer-i-Cancho & Jaume Baixeries. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. In Pavel Král & Carlos Martín-Vide (eds.), *Statistical language and speech processing* (Lecture Notes in Computer Science 9918), 19–29. Cham: Springer. https://doi.org/10.1007/978-3-319-45925-7_2

Izsák, János. 2006. Some practical aspects of fitting and testing the Zipf-Mandelbrot model. A short essay. *Scientometrics* 65. 107–120. https://doi.org/10.1007/s11192-006-0052-x

Kelih, Emmerich. 2009. Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In Emmerich Kelih, Viktor Levickij & Gabriel Altmann (eds.), *Methods of text analysis*, 106–124. Chernivtsi: ČNU.

Kelih, Emmerich. 2010. Parameter interpretation of Menzerath's law: Evidence from Serbian. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and language. Structures, functions, interrelations, quantitative perspectives*, 71–78. Vienna: Praesens.

Kelih, Emmerich. 2012. *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation* (Specimina Philologiae Slavicae 168). Munich: Otto Sagner. https://doi.org/10.3726/b12003

Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics. An international handbook* (Handbooks of Linguistics and Communications Science 27), 760–775. Berlin: de Gruyter.

Köhler, Reinhard. 2011. Laws of language. In Patrick C. Hogan (ed.), *The Cambridge encyclopedia of the language sciences*, 424–426. Cambridge: Cambridge University Press.

Köhler, Reinhard. 2012. *Quantitative syntax analysis* (Quantitative Linguistics 65). Berlin: de Gruyter. https://doi.org/10.1515/9783110272925

Mačutek, Ján, Jan Chromý & Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics* 26. 66–80. https://doi.org/10.1080/09296174.2018.1424493

Mačutek, Ján & Andrij Rovenchak. 2011. Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In Emmerich Kelih, Victor Levickij & Yuliya Matskulyak (eds.), *Issues in quantitative linguistics 2* (Studies in Quantitative Linguistics 11), 136–147. Lüdenscheid: RAM-Verlag.

Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20. 227–240. https://doi.org/10.1080/09296174.2013.799912

Menzerath, Paul. 1954. *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Mikros, Georgios & Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Gabriel Altmann, Radek Čech, Ján Mačutek & Ludmila Uhlířová (eds.), *Empirical approaches to text and language analysis* (Studies in Quantitative Linguistics 17), 180–189. Lüdenscheid: RAM-Verlag.

Piper, Predrag & Ivan Klajn. 2013. *Normativna gramatika srpskog jezika*. Novi Sad: Matica srpska.

Ponomariv, Oleksandr D. (ed.) 2001. *Sučasna ukrajins'ka mova*. Kyjiv: Lybid'.

Popescu, Ioan-Iovitz, Peter Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová & Matummal N. Vidya. 2009. *Word frequency studies* (Quantitative Linguistics 64). Berlin: de Gruyter.

Popescu, Ioan-Iovitz, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Haruko Sanada, Anja Overbeck, Reginald Smith, Radek Čech, Panchanan Mohanty, Andrew Wilson & Gabriel Altmann. 2013. Word length: aspects and languages. In Reinhard Köhler & Gabriel Altmann (eds.), *Issues in quantitative linguistics 3* (Studies in Quantitative Linguistics 13), 224–281. Lüdenscheid: RAM-Verlag.

Radojičić, Marija, Biljana Lazić, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović, Ján Mačutek & Lívia Leššová. 2019. Frequency and length of syllables in Serbian. *Glottometrics* 45. 114–123.

Roberts, Aaron H. 1965. *A statistical linguistic analysis of American English* (Janua Linguarum Series Practica 8). The Hague: Mouton. https://doi.org/10.1515/9783112416426

Strauss, Udo, Fengxiang Fan & Gabriel Altmann. 2008. *Problems in quantitative linguistics 1* (Studies in Quantitative Linguistics 1). Lüdenscheid: RAM-Verlag.

Wimmer, Gejza & Gabriel Altmann. 1999. *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Zec, Draga. 1995. Sonority constraints on syllable structure. *Phonology* 12. 85–129. https://doi.org/10.1017/S0952675700002396

Zipf, George K. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Zörnig, Peter, Kamil Stachowski, Anna Rácová, Yunhua Qu, Michal Místecký, Kuizi Ma, Mihaiela Lupea, Emmerich Kelih, Volker Gröller, Hanna Gnatchuk, Alfiya Galieva, Sergey Andreev & Gabriel Altmann. 2019. *Quantitative insights into syllable structure* (Studies in Quantitative Linguistics 30). Lüdenscheid: RAM-Verlag.

N-grams of grammatical functions and their significant order in the Japanese clause

Haruko Sanada
Rissho University

The present study investigates the statistically significant order of grammatical functions in Japanese clauses by employing *n*-gram frequency data of grammatical functions. There are broad rules for the order of grammatical functions, though Japanese is an agglutinative SOV language and complements can be elliptic. I conclude that the time and the place appear between the subject and object with statistical significance. The occasion takes a position before the subject, between the subject and object, or after the object. Therefore, the occasion shows that Japanese is a free word order language. The subject and object play the role of 'anchors' in the clause. By using the 'two-sample test for equality of proportions without continuity correction data', the study introduces a descriptive verification method of implicit speaker-hearer knowledge.

Keywords: valency, sentence structure, *n*-gram, frequency, grammatical functions, position in the sentence, Japanese, Synergetic Linguistics

1. Aim of the study

The present study investigates the statistically significant order of noun phrases in clauses in Japanese based on their syntactic roles (including complements as well as adjuncts) by employing *n*-gram frequency data of noun phrases whose syntactic roles are marked or unmarked but recoverable. There are broad rules for the order of grammatical functions in clauses, though Japanese is an agglutinative SOV language and complements can be elliptic. In an earlier study (e.g. Sanada 2018a), I found that there are common patterns of neighbouring grammatical functions. Referring to the 'full valency' study offered by Čech et al. (2010), the dependent in the present study is defined as the level between the clause and the morpheme, which corresponds to the subject, the complement and the adjunct.

Based on earlier work, we expect a kind of 'bracket structure' of the valency in Japanese, in which the subject and the object sandwich other grammatical functions

https://doi.org/10.1075/cilt.356.05san

like brackets in the clause. In the present study, we statistically confirm the order of neighbouring grammatical functions in the common patterns and how broad the rules for the order of grammatical functions in clauses are. It is expected that a sentence with the verb *meet* contains the subject, the object, and the predicate. The other grammatical functions, i.e., adverbials of time, place, and occasion ('occasion' in this paper also includes the reason for or the manner of the 'meeting'), should be located between the subject and the object as adjuncts. The subject and the object can be elliptic, but the object should be elliptic less often than the subject.

Figure 1 shows the order of a common pattern of neighbouring grammatical functions obtained in Sanada (2018a). The subject often takes a position at the beginning of the clause, while the object takes the place before the predicate (both can be elliptic). Between the subject and the object, grammatical functions of time, place, and occasion can be allocated and their order is flexible. The time, the place, and the occasion can be also elliptic. The predicate must be placed at the end of the clause in the written text. The predicate can be elliptic in literary texts or in colloquial speech. However, we have no such examples because the source of our data is newspaper texts. Figure 1 shows the starting point and the ending point of the clause for the following sections. The subject and the object may work as 'brackets'.

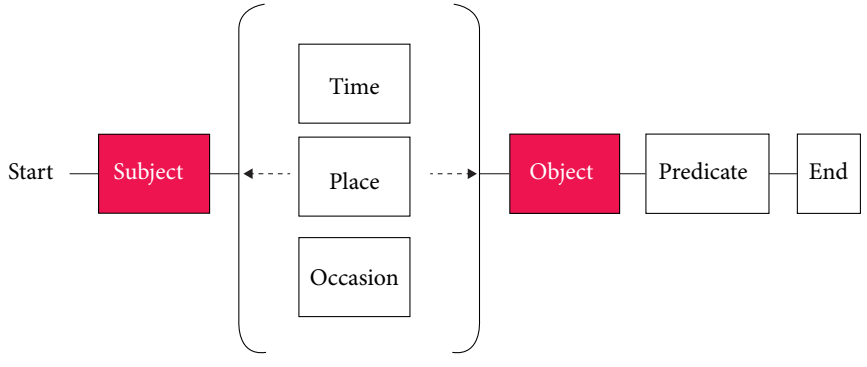


Figure 1. The order of a common pattern of neighbouring valency types and grammatical functions drawn from results of Sanada (2018a)

2. Descriptions of data and grammatical definitions

We regard the present study as one in the series of our valency studies, and we employed the Japanese valency database (Ogino et al. 2003), the same one that was employed in our previous studies (Sanada 2012, 2014, 2015, 2016, 2018b, 2019,

forthcoming). In the previous studies, 240 sentences containing the verb *meet*[1] were extracted from the valency database. These extracted sentences also include many other verbs because each sentence has one or more predicates. Three of the 240 sentences have 2 predicates with the verb *meet*.

We used the Japanese morphological analyzer *MeCab* (Graduate Schools of Informatics in Kyoto University et al. 2008) and the electronic dictionary *UniDic* (National Institute for Japanese Language and Linguistics 2008) for the extracted sentences. The software shows the boundary of the 'short unit' as a morpheme. Errors were corrected by hand. From the 240 sentences, 692 clauses, 1,889 grammatical functions and 5,626 morphemes were obtained.

In the present study, we employed the 243 clauses that contain the verb *meet*. The clauses have 765 grammatical functions and 2,365 morphemes.

Sanada (2016) posited four linguistic levels, i.e., sentence, clause, grammatical function and morpheme. Clauses must contain one predicate for each,[2] and consist of complements, adjuncts, and a predicate. The subject can be elliptic. The present study follows these definitions for consistency.

In case the clause in the sentence is divided by an embedded clause, the beginning of the first half and the second half are taken as one clause. Figure 2 shows a model of the clauses in the sentence and the embedded clause.

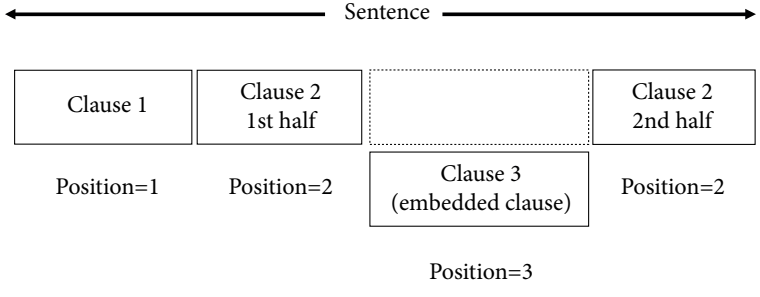


Figure 2. The position of the clause and the embedded clause

1. In a previous study (Sanada 2012) we investigated word frequencies of verbs in Japanese corpora and texts, and the same six verbs have been employed in a series of our valency studies: *au* ("meet"), *hataraku* ("work"), *yaburu* ("tear, break"), *umareru* ("be born, arise"), *ugoku* ("move"), and *ataeru* ("give"). We tentatively chose these six verbs because they do not use the auxiliary verb and have no homonyms. They also have relatively less ambiguity of meaning or usage and a high frequency in the corpora used. In the present study we employ the verb *meet* among the six verbs as the first trial because it is the simplest solution from the point of view of meaning or usage.

2. In 2 of the 692 clauses, a predicate is missing because the sentences are grammatically incorrect.

The rules used for counting the data are shown here with examples, which were also presented in our former study. A space in the example shows a morpheme boundary. A single slash mark (/) and a double slash mark (//) show the boundary of dependents and the verbs (clause elements) and boundaries of the clauses, respectively. The number of clauses, clause elements and morphemes of the example follow its English equivalents. The numbers shown with ID after the example indicate a sentence number in the database.

The 'sentence' in Japanese is optically clear, marked by a sign at the end. The notion 'morpheme' is a topic that is still being discussed. We employ the definition of the 'short unit' as a morpheme, which was developed by the National Institute of Japanese Language and Linguistics (National Language Research Institute 1964).

The notion of 'clause' is also a topic that is still being discussed in Japanese linguistics. Minami (1974, 1993) analyzed grammatically important types of the Japanese clause.[3] Here, referring to his model, we studied our data using quantitative and empirical analyses. In the present study, we defined it as a linguistic unit that has a predicate on the surface of the sentence.

The clause element is defined as the level between the clause and the morpheme. We regard the predicate and grammatical elements that are linked to the predicate as clause elements in the clause. Attributive elements, i.e., a noun and a postposition, were treated as a part of the clause element (see underlined words in (Example 1). Among the clause elements, those tagged in the Japanese valency database (Ogino et al. 2003) were defined as the complement, the rest of the clause elements, except the predicate and the subject, are defined as the adjuncts. Conjunctions, except postpositions, which belong to the clause were also regarded as a member of the clause in the present study, and categorized as an adjunct. This definition should be discussed in our future studies.

Example 1.

Yogo no Ishikawa Keiko sensei wa,/ chugaku 3 nen no shojo no hahaoya ni/ at ta.
(ID: JCO0217129)

[nurse-teacher-ATTR Mrs. Keio Ishikawa-SUBJ/ junior high school 3rd grade-ATTR girl-GEN mother-OBJ/ meet-PAST]

"Mrs. Ishikawa Keiko, a nurse-teacher, met the mother of the girl in the 3rd grade of the junior high school."

Clause = 1, Clause Element = 3, Dependents = 2, Morpheme = 16.

3. Minami's model is also employed in the software of the National Institute of Japanese Language and Linguistics to find the boundaries of some types of clause. However, his model does not cover all types of Japanese clause in the corpus because it focuses on grammatical and semantic aspects.

Japanese has no marker for relative pronouns in sub clauses (see the underlined beginning of the sentence in (Example 2) or embedded clauses that divide the upper-level clause into two parts (see the underlined words in (Example 3.).

Example 2.

Watashi ga/ at ta// chiji no hotondo wa,/ chiji shitsu ni/ nihon no ningyo ya okimono wo/ oi te i ta. (ID: JCO0138531)
[I-SUBJ/ meet-PAST// prefectural governors-ATTR most-TOPIC/ prefectural governor office-LOC/ Japan-ATTR dolls or ornamental objects-OBJ/ display-CONNECT- CONTINUOUS-PAST]
"Most of the prefectural governors whom I met displayed Japanese dolls or ornaments in their office."
Clause = 2, Clause Element = 6, Dependents = 4, Morpheme = 21.

Example 3.

Henshu cho shitsu de/ nan do ka/ at ta ga,// itsu mo/ taatorunekku no seetaa ni// zakkuri shi ta// sebiro wo/ haot te i ta. (ID: JCO0209028)
[chief editor room-LOC/ several times/ meet-PAST-CONNECT,// always/ turtleneck-ATTR sweater-CONNECT// roughly woven-PAST// jacket-OBJ/ was wearing (PROG-PAST)]
"I met him several times in the office of the chief editor, and he always wore a sweater with a turtleneck and a jacket which was roughly woven."
Clause = 3, Clause Element = 8, Dependents = 5 (**not 6, only direct dependents of the verb**), Morpheme = 25.

For more detailed definitions, e.g., definitions related to the predicate or other special cases, refer to Sanada (2016).

3. Hypotheses and methodology

We determine the frequencies of grammatical functions in our data. Figure 3 shows a number of grammatical functions of the 522 dependents from 243 clauses featuring the lexeme *meet* (Sanada 2016, corrected). The five grammatical functions are put into 13 categories defined by Čech & Uhlířová (2014). From Figure 3 we can see that the subject and the object are not obligatory in the clause, but that the object can be elliptic less often than the subject. It is confirmed by means of a chi-square test that this distribution is not homogeneous.

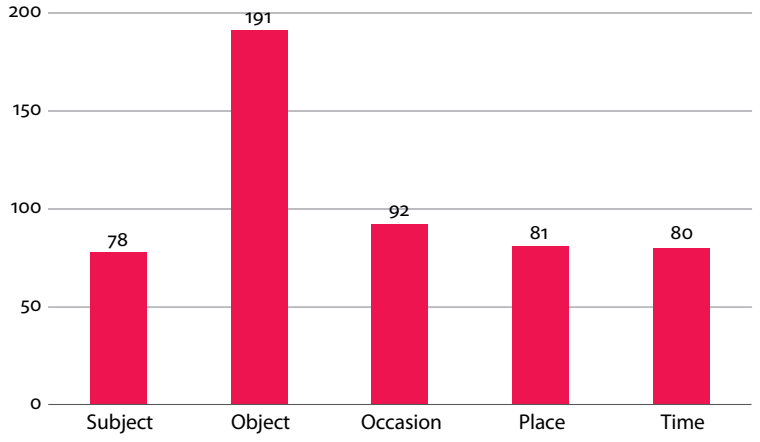


Figure 3. Number of grammatical function types of the 522 dependents from 243 *meet* clauses (Sanada 2016, corrected)

From the 522 dependents, we obtained unigrams and bi-grams of the grammatical functions (the subject, the object, the occasion, the place and the time) including markers for the beginning of the clause (the start marker) and for the end of the clause (the end marker). It should be statistically confirmed that the positions of the subject or the object must be somewhat fixed and the order of the other grammatical function types must be flexible, which is drawn in Figure 1.

To prove our 'bracket structure', we set the following hypotheses:

$H1_0$: The proportion of the grammatical function type which appears at the first position is different from the proportion of the grammatical function type which appears at the second or later position.

$H2_0$: The proportion of the grammatical function type which appears at the last position is different from the proportion of the grammatical function type which appears at the position preceding the predicate by two or more units.

$H3_0$: The proportion of bi-grams including the subject or the object is different from the proportion of bi-grams of the other grammatical function types, and the proportion of bi-grams taken from the occasion, the place, and the time must not differ significantly.

Tables 1 and 2 show the number of grammatical function types and their proportions which appear in the first position of the clause or which appear in the last position of the clause for $H1_0$ and $H2_0$, respectively. The denominators of the proportions are the total numbers for each group. For example, we calculated the number of subjects in Table 1, i.e., 61 of 243 dependents which appear in the first position of the clause, and the number of subjects, i.e., 17 of 279 dependents which

do not appear in the first position. We employed these proportions of the subject as a comparison pair. If proportions of the subject in Table 1 are significantly different, the subject has more preference to appear in the first position than in the second or later position. If proportions of the subject are not significantly different, the subject does not have any specific positional preference in the clause.

Table 1. Number of grammatical function types and their proportions that appear in the first position of the clause and those in the second or later position

Grammatical function type	Subject	Object	Occasion	Place	Time	Total
First position	61	70	39	37	36	243
	25.1%	28.8%	16.0%	15.2%	14.8%	100.0%
Second or later position	17	121	53	44	44	279
	6.1%	43.4%	19.0%	15.8%	15.8%	100.0%

Table 2. Number of grammatical function types and their proportions that appear in the last position of the clause and those preceding by two or more

Grammatical function type	Subject	Object	Occasion	Place	Time	Total
Preceding by two or more	72	34	51	47	75	279
	25.8%	12.2%	18.3%	16.8%	26.9%	100.0%
Last position	6	157	41	34	5	243
	2.5%	64.6%	16.9%	14.0%	2.1%	100.0%

We performed 'two-sample tests for equality of proportions without continuity correction data' for each pair of grammatical function types in Table 1 and Table 2. Cohen's *d* for each is calculated as the effective size, if the results of the tests are significant. Some formulae are known as Cohen's *d* for individual tests, and we employ a formula using the log odds ratio shown by Borenstein et al. (2009: 47), which converts proportions to an effect size of *d* (standard mean difference). The formula is also employed in the 'propes' function of the R package 'compute.es' (Del Re 2015: 72).

The statistical test called 'two-sample test for equality of proportions without continuity correction data' is performed as follows. Assuming that our data approximately follow the normal distribution, the *z*-statistic is calculated as (1):

$$(1) \quad z = \frac{p_A - p_B}{\sqrt{p(1-p)(1/n_A + 1/n_B)}}$$

with two proportions p_A and p_B as:

(2) $p_A = x_A/n_A,\ p_B = x_B/n_B$

and a pooled sample proportion P as:

(3) $$P = \frac{x_A + x_B}{n_A + n_B}$$

where x_A and x_B are numbers of two samples, i.e., the number of paired grammatical function types, therefore n_A and n_B are the total numbers of two groups.

Cohen's d employing the logged odds ratio *log (OR)* is obtained as follows:

(4) $$d = \frac{\sqrt{3}\log(OR)}{\pi}$$

where π is the well-known mathematical constant, with

(5) $$OR = \frac{p_A(1-p_B)}{p_B(1-p_A)}$$

and

(6) $p_A = x_A/n_A,\ p_B = x_B/n_B$

where x_A and x_B are numbers of two samples, i.e., the number of paired grammatical function types, and n_A and n_B are the total numbers of the two groups. Generally, with Cohen's d it is considered that $d = 0.2$ represents a 'small' effect size, 0.5 a 'medium' effect size, and 0.8 a 'large' effect size.

For $H3_0$ we obtained bi-grams of dependents going forward (F) and backward (B), e.g., Subject – Occasion (F) and Occasion – Subject (B). We obtained 25 (= 5 * 5) bi-grams (F) and 25 (= 5 * 5) reverse bi-grams (B) from the five grammatical function types, and obtained 25 pairs from these 50 bi-grams. The 25 bi-gram pairs and their valency type proportions are shown in the left and right columns of Table 3. The first grammatical function types of forward bi-grams (F) and the grammatical function types of backward bi-grams (B) are underlined. For example, we calculated the proportion of 12 Subject – Occasion (F) bi-grams from 72 bi-grams, i.e., 16.7%, and the proportion of 7 bi-grams of the Occasion – Subject (B) of 17 bi-grams, i.e., 41.2%. Denominators of the proportions are the total numbers for bi-grams of the subject included in the first part and the subject included in the second part. We set these proportions as comparison pairs using statistical tests. If the difference between these two proportions is statistically significant, the subject has an order preference of following the occasion to being followed by the occasion. If the difference between these two proportions is not significant, the subject has no order preference with the occasion. The underlined grammatical function type is a 'starting position' of the bi-gram to be analyzed.

Table 3. Numbers of bi-grams from five grammatical function types and their proportions

Forward bi-gram (F)	Number of bi-grams	Proportion of bi-grams	Backward bi-gram (B)	Number of bi-grams	Proportion of bi-grams
Subject – Subject (F)	1	1.4%	Subject – Subject (B)	1	5.9%
Subject – Object (F)	22	30.6%	Object – Subject (B)	0	0.0%
Subject – Occasion (F)	12	16.7%	Occasion – Subject (B)	7	41.2%
Subject – Place (F)	8	11.1%	Place – Subject (B)	1	5.9%
Subject – Time (F)	29	40.3%	Time – Subject (B)	8	47.1%
Subject (F) Total	72	100.0%	**Subject (B) Total**	17	100.0%
Object – Subject (F)	0	0.0%	Subject – Object (B)	22	18.2%
Object – Object (F)	3	8.8%	Object – Object (B)	3	2.5%
Object – Occasion (F)	18	52.9%	Occasion – Object (B)	27	22.3%
Object – Place (F)	5	14.7%	Place – Object (B)	40	33.1%
Object – Time (F)	8	23.5%	Time – Object (B)	29	24.0%
Object (F) Total	34	100.0%	**Object (B) Total**	121	100.0%
Occasion – Subject (F)	7	13.7%	Subject – Occasion (B)	12	22.6%
Occasion – Object (F)	27	52.9%	Object – Occasion (B)	18	34.0%
Occasion – Occasion (F)	6	11.8%	Occasion – Occasion (B)	6	11.3%
Occasion – Place (F)	5	9.8%	Place – Occasion (B)	5	9.4%
Occasion – Time (F)	6	11.8%	Time – Occasion (B)	12	22.6%
Occasion (F) Total	51	100.0%	**Occasion (B) Total**	53	100.0%

(*continued*)

Table 3. (*continued*)

Forward bi-gram (F)	Number of bi-grams	Proportion of bi-grams	Backward bi-gram (B)	Number of bi-grams	Proportion of bi-grams
Place – Subject (F)	1	2.1%	Subject – Place (B)	8	18.2%
Place – Object (F)	40	85.1%	Object – Place (B)	5	11.4%
Place – Occasion (F)	5	10.6%	Occasion – Place (B)	5	11.4%
Place – Place (F)	1	2.1%	Place – Place (B)	1	2.3%
Place – Time (F)	0	0.0%	Time – Place (B)	25	56.8%
Place (F) Total	47	100.0%	**Place (B) Total**	44	100.0%
Time – Subject (F)	8	10.7%	Subject – Time (B)	29	65.9%
Time – Object (F)	29	38.7%	Object – Time (B)	8	18.2%
Time – Occasion (F)	12	16.0%	Occasion – Time (B)	6	13.6%
Time – Place (F)	25	33.3%	Place – Time (B)	0	0.0%
Time – Time (F)	1	1.3%	Time – Time (B)	1	2.3%
Time (F) Total	75	100.0%	**Time (B) Total**	44	100.0%

4. Results for grammatical function types in the first position and in the second or later position

We set the following research question and the hypothesis in this section, and perform statistical tests:

RQ1: Are there any differences between the grammatical function types which appear in the first position and do they differ from the grammatical function types which appear in the second or later position?

$H1_0$: The proportion of the grammatical function type which appears in the first position is different from the proportion of the grammatical function type which appears in the second or later position.

Figure 4 shows the distributions of (1) the number of grammatical function types which appear in the first position and (2) the number of grammatical function types which appear in the second or later position. It was confirmed that the two

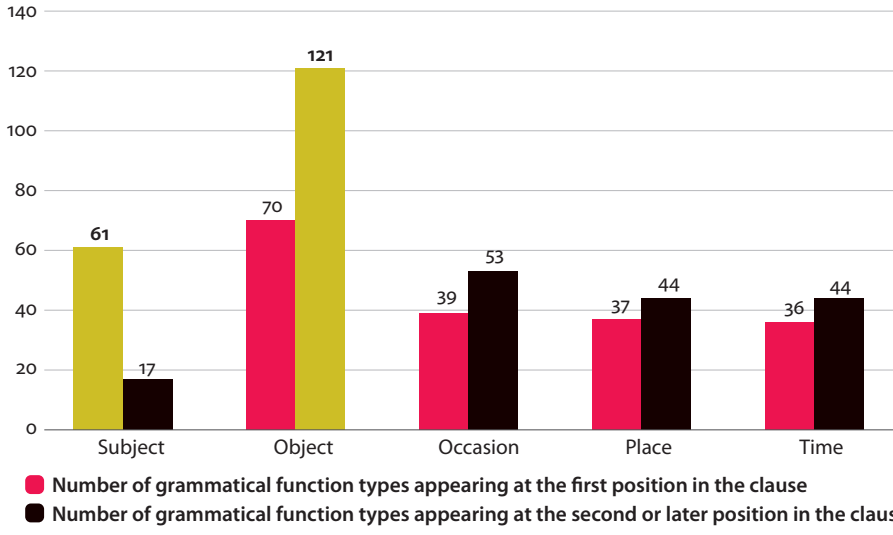


Figure 4. Paired grammatical function types that appear in the first position of the clause and those in the second or later position

distributions are significantly different, by means of a chi-square test. We performed 'two-sample tests for equality of proportions without continuity correction data' for each pair of grammatical function types. The two-sided critical *z*-values are 1.96 at the 0.025 (= 0.05/2) level, and 2.58 at the 0.005 (= 0.01/2) level. The difference of the proportion of each pair is significant and the null hypothesis is rejected if the absolute value of the *z*-statistic is greater than 1.96. The *z*-statistics and the effect sizes (Cohen's *d*) for the five pairs of grammatical function types are shown in Table 4, which appeared in the first position of the clause and those in the second or later position. The larger proportion of the significant pair is emphasized in italics in the table and yellow in the figure.

Table 4. The *z*-statistics and the effect sizes (Cohen's *d*) for pairs of grammatical function types that appear in the first position of the clause and those in the second or later position

Grammatical function type	Subject	Object	Occasion	Place	Time	Total
First position	61	70	39	37	36	243
	25.1%	28.8%	16.0%	15.2%	14.8%	100.0%
Second or later position	17	121	53	44	44	279
	6.1%	***43.4%***	19.0%	15.8%	15.8%	100.0%
z-statistics (absolute values)	6.0768	3.4455	0.8814	0.1713	0.3024	
	**	**	n.s.	n.s.	n.s.	
Effect size (Cohen's *d*)	0.9052	0.3517	0.1126	0.0229	0.0407	

Note: an asterisk (*) means '$p < .05$', two asterisks (**) mean '$p < .01$', and 'n.s.' means 'not significant'.

Among five pairs, two grammatical function types, i.e., the subject which appears in the first position and the object which appears in the second or later position, are significantly different. According to the value of Cohen's *d*, it can be interpreted that the subject has a strong preference to appear in the first position ($d = 0.9052$), and that the object has a significant preference to appear in the second or later position in the clause ($d = 0.3517$). These two are in yellow in Figure 4. The other pairs of grammatical function types, i.e., the time, the place and the occasion are not significantly different, and these grammatical function types do not have a significant preference to appear in the first position, nor in the second or later position in the clause. The interpretation is shown in Figure 5.

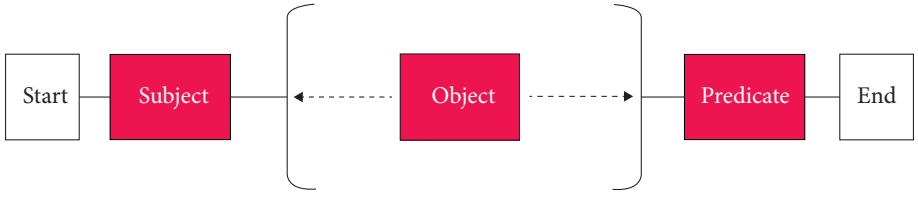


Figure 5. An interpretation of results from tests for $H1_0$

5. Results for grammatical function types in the position directly preceding the predicate and preceding it by two or more units

We set the following research question and hypothesis in this section, and perform statistical tests:

RQ2: Are there any differences between the grammatical function types which appear in the last position and the grammatical function types which appear in the position preceding the predicate by two or more units?

$H2_0$: The proportion of the grammatical function types which appear in the last position is different from the proportion of the grammatical function types which appear in the position preceding the predicate by two or more units.

Figure 6 shows distributions of (1) the number of the grammatical function types which appear in the last position and (2) the number of grammatical function types which appear in the position preceding it by two or more units. It was confirmed by means of a chi-square test that the two distributions are significantly different. We performed 'two-sample tests for equality of proportions without continuity correction data' for each pair of grammatical function types. The two-sided critical

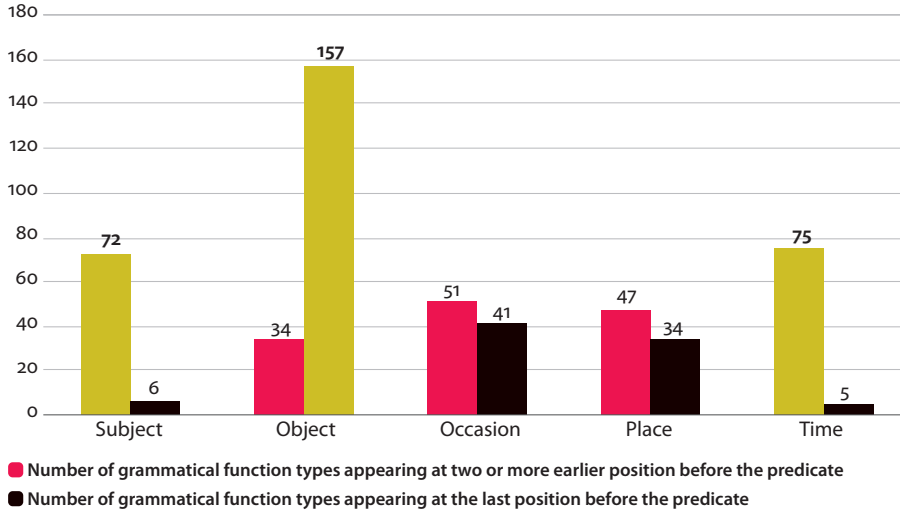


Figure 6. Paired grammatical function types appearing in the last position of the clause and those in the two or earlier positions preceding the predicate

z-values are 1.96 at the 0.025 (= 0.05/2) level, and 2.58 at the 0.005 (= 0.01/2) level. The difference in the proportion for each pair is significant and the null hypothesis is rejected if the absolute value of the *z*-statistic is greater than 1.96. The *z*-statistics and the effect sizes (Cohen's *d*) for the five pairs of grammatical function types are shown in Table 5, which appeared in the last position of the clause and in the position preceding it by two or more units. The larger proportion of the significant pair is emphasized in italics in the table and yellow in the figure.

Table 5. The *z*-statistics and the effect sizes (Cohen's *d*) for the pairs of grammatical functions appearing in the last position of the clause and those preceding by two or more units

Valency type	Subject	Object	Occasion	Place	Time	Total
Preceding by two or more units	72	34	51	47	75	279
	25.8%	12.2%	18.3%	16.8%	***26.9%***	100.0%
Last position	6	157	41	34	5	243
	2.5%	***64.6%***	16.9%	14.0%	2.1%	100.0%
z-statistics (absolute values)	7.4602 **	12.4031 **	0.4209 n.s.	0.8984 n.s.	7.8534 **	
Effect size (Cohen's *d*)	1.4446	1.4207	0.0536	0.1210	1.5780	

Note: an asterisk (*) means '$p < .05$', two asterisks (**) mean '$p < .01$', and 'n.s.' means 'not significant'.

Among five pairs, three grammatical function types, i.e., the object which appears in the last position of the clause, and the subject and the time which appear in the position preceding it by two or more units, are significantly different. According to the value of Cohen's *d*, it can be interpreted that the object ($d = 1.4207$), the subject ($d = 1.4446$), and the time ($d = 1.5780$) have a strong preference for the position in the clause, respectively. These three are in yellow in Figure 6. The other pairs of grammatical function types, i.e., the place and the occasion, are not significantly different, and these grammatical function types do not have a significant preference to appear in the last position, nor in the position preceding it by two or more units in the clause. The interpretation is drawn as Figure 7.

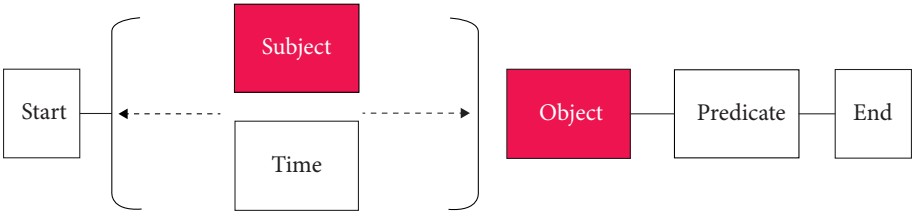


Figure 7. An interpretation of results from tests for $H2_0$

We can combine Figure 5 and Figure 7 as Figure 8.

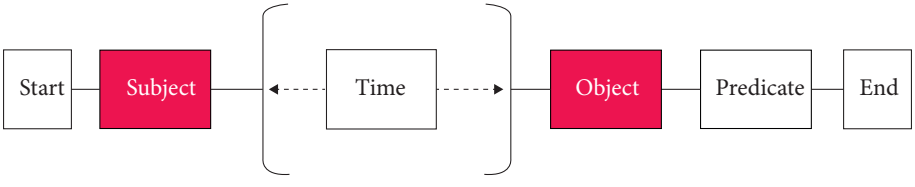


Figure 8. An interpretation of results from tests for $H1_0$ and $H2_0$

6. Results for bi-grams of grammatical function types including the subject or the object

In this section, we set the following research question and hypothesis, and perform statistical tests.

RQ3: Is there any difference between bi-grams including the subject or the object and bi-grams including the other grammatical function types?

H3$_0$: The proportion of bi-grams including the subject or the object is different from the proportion of bi-grams of the grammatical function types because the order of grammatical function types except the subject and the object must be flexible, and bi-grams with the occasion, the place and the time must not be significant.

Figure 9 to Figure 13 show paired distributions of forward and backward bi-grams going by the 'starting position' of the grammatical function types. It was confirmed by means of chi-square tests that the paired distributions are significantly different. We performed 'two-sample tests for equality of proportions without continuity correction data' for each pair of grammatical function types under the same conditions as described in § 5. The *z*-statistics and the effect sizes (Cohen's *d*) for the 25 paired bi-grams of grammatical function types are shown in Table 6. The effect size (Cohen's *d*) is not obtained if the proportions for one of the pairs are 0%. The greater proportion of the significant pair is marked by italics in the table and yellow in the figure. Significant pairs are extracted from Table 6 and Table 7.

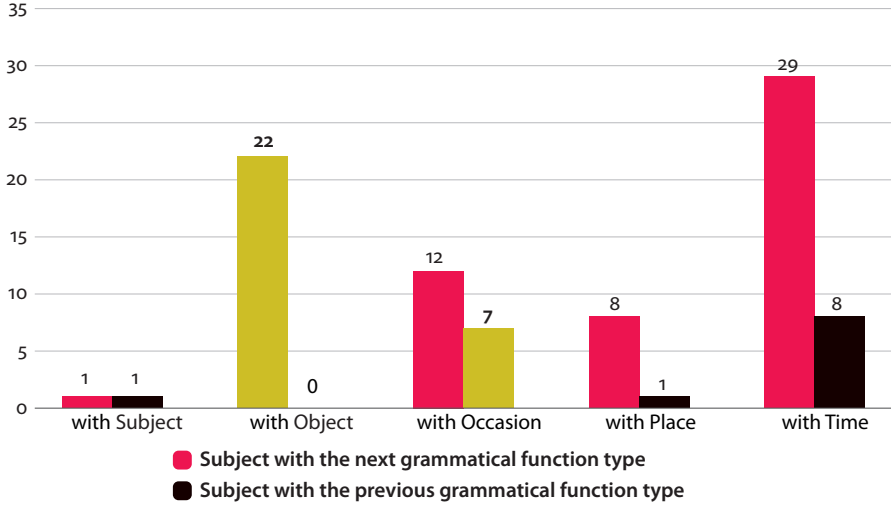


Figure 9. Numbers of forward and backward bi-grams with the subject

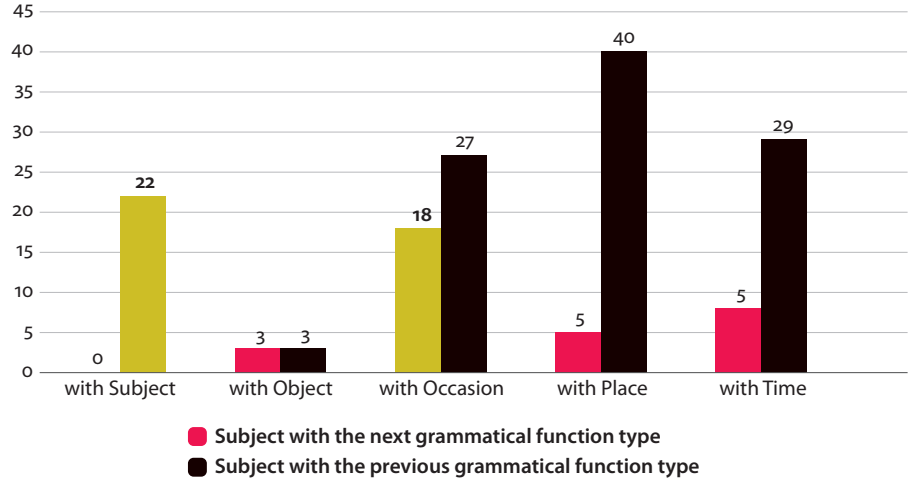


Figure 10. Numbers of forward and backward bi-grams with the object

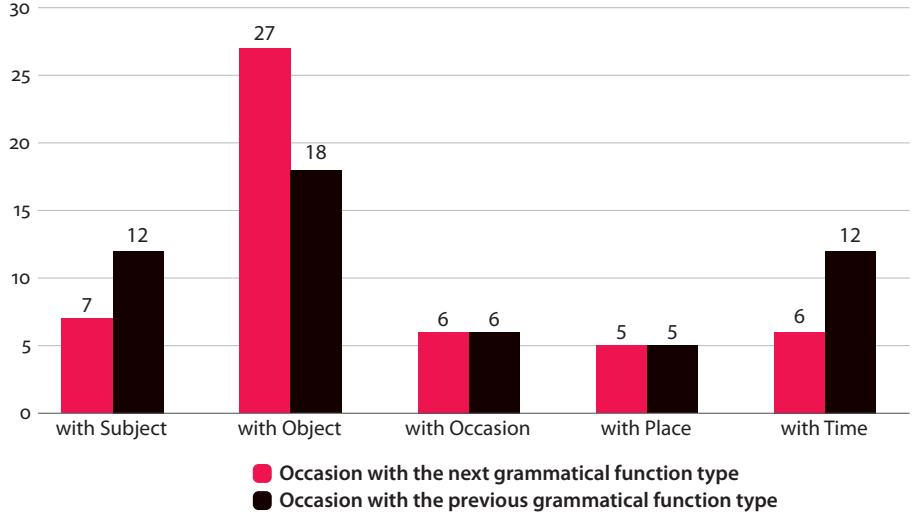


Figure 11. Numbers of forward and backward bi-grams with the occasion

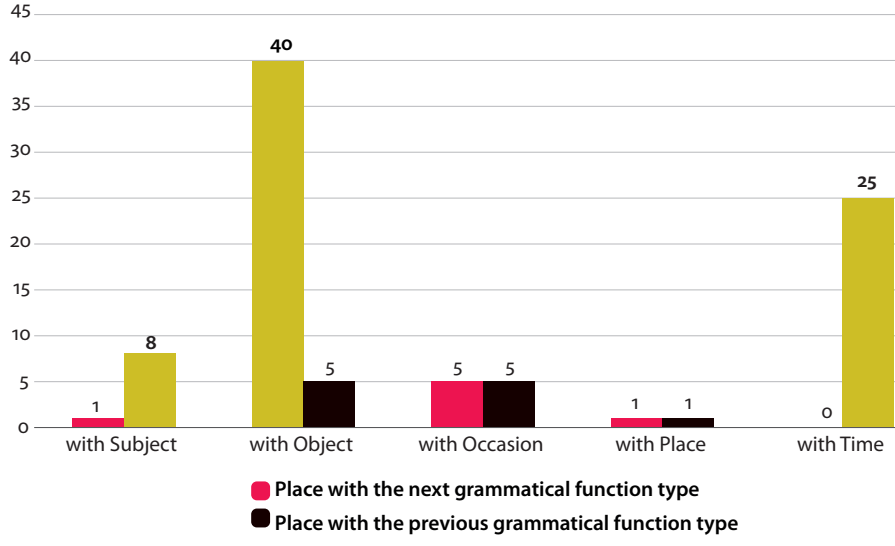


Figure 12. Numbers of forward and backward bi-grams with the place

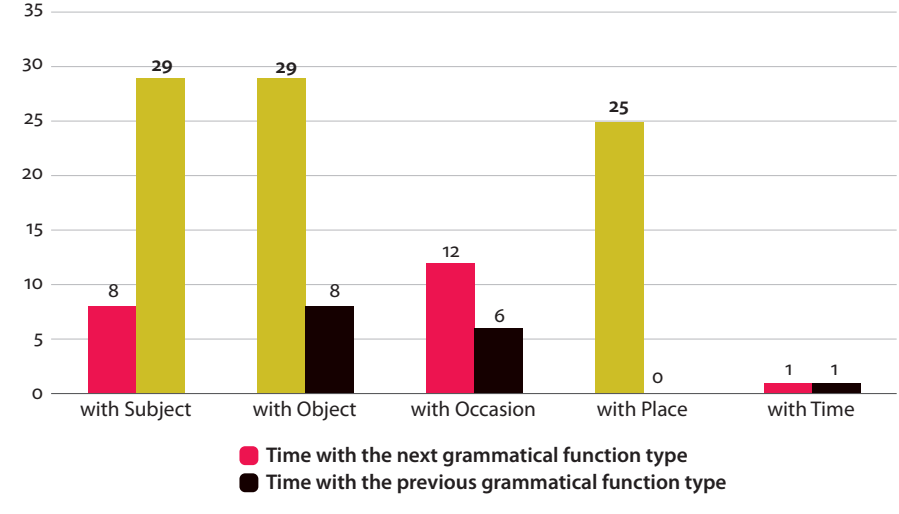


Figure 13. Numbers of forward and backward bi-grams with the time

Table 6. The *z*-statistics and the effect sizes (Cohen's *d*) for pairs of forward and backward bi-grams of five grammatical function types

Forward bi-gram (F)	Proportion of bi-grams	Backward bi-gram (B)	Proportion of bi-grams	*z*-statistics		Effect size (Cohen's *d*)
Subject – Subject (F)	1.4%	Subject – Subject (B)	5.9%	1.1243	n.s.	0.8215
Subject – Object (F)	***30.6%***	Object – Subject (B)	0.0%	2.6268	–	–
Subject – Occasion (F)	16.7%	Occasion – Subject (B)	***41.2%***	2.2182	*	0.6907
Subject – Place (F)	11.1%	Place – Subject (B)	5.9%	0.6432	n.s.	0.3822
Subject – Time (F)	40.3%	Time – Subject (B)	47.1%	0.5102	n.s.	0.1522
Subject (F) Total	100.0%	**Subject (B) Total**	100.0%			
Object – Subject (F)	0.0%	Subject – Object (B)	***18.2%***	2.6841	–	–
Object – Object (F)	8.8%	Object – Object (B)	2.5%	1.6944	n.s.	0.7370
Object – Occasion (F)	***52.9%***	Occasion – Object (B)	22.3%	3.4762	**	0.7527
Object – Place (F)	14.7%	Place – Object (B)	***33.1%***	2.0829	*	0.5802
Object – Time (F)	23.5%	Time – Object (B)	24.0%	0.0529	n.s.	0.0133
Object (F) Total	100.0%	**Object (B) Total**	100.0%			
Occasion – Subject (F)	13.7%	Subject – Occasion (B)	22.6%	1.1763	n.s.	0.3361
Occasion – Object (F)	52.9%	Object – Occasion (B)	34.0%	1.9529	n.s.	0.4316
Occasion – Occasion (F)	11.8%	Occasion – Occasion (B)	11.3%	0.0708	n.s.	0.0240
Occasion – Place (F)	9.8%	Place – Occasion (B)	9.4%	0.0640	n.s.	0.0235
Occasion – Time (F)	11.8%	Time – Occasion (B)	22.6%	1.4657	n.s.	0.4335
Occasion (F) Total	100.0%	**Occasion (B) Total**	100.0%			

Table 6. (*continued*)

Forward bi-gram (F)	Proportion of bi-grams	Backward bi-gram (B)	Proportion of bi-grams	*z*-statistics		Effect size (Cohen's *d*)
Place – Subject (F)	2.1%	Subject – Place (B)	***18.2%***	2.5636	*	1.2816
Place – Object (F)	***85.1%***	Object – Place (B)	11.4%	7.0312	**	2.0934
Place – Occasion (F)	10.6%	Occasion – Place (B)	11.4%	0.1106	n.s.	0.0409
Place – Place (F)	2.1%	Place – Place (B)	2.3%	0.0472	n.s.	0.0372
Place – Time (F)	0.0%	Time – Place (B)	***56.8%***	6.0679	–	–
Place (F) Total	100.0%	**Place (B) Total**	100.0%			
Time – Subject (F)	10.7%	Subject – Time (B)	***65.9%***	6.2849	**	1.5352
Time – Object (F)	***38.7%***	Object – Time (B)	18.2%	2.3305	*	0.5749
Time – Occasion (F)	16.0%	Occasion – Time (B)	13.6%	0.3474	n.s.	0.1034
Time – Place (F)	***33.3%***	Place – Time (B)	0.0%	4.3090	–	–
Time – Time (F)	1.3%	Time – Time (B)	2.3%	0.3848	n.s.	0.2993
Time (F) Total	100.0%	**Time (B) Total**	100.0%			

Note: an asterisk (*) means '$p < .05$', two asterisks (**) mean '$p < .01$', and 'n.s.' means 'not significant'.

Some bi-grams have a significant pair of proportions for both sides. The subject prefers being followed by the object to following the object, and the object prefers following the subject to being followed by the subject. The object has a preference to follow the place ($d = 0.5802$), and the place has a strong preference to be followed by the object ($d = 2.0934$). The time has a preference to be followed by the place, and the place has a preference to follow the time.

Some bi-grams have a significant pair of proportions for one side. The subject has a preference to follow the occasion ($d = 0.6907$), and the object has a strong preference to be followed by the occasion ($d = 0.7527$). However, the occasion has no significant bi-grams according to the tests. It can be interpreted that the occasion can appear at any position in the clause. It is possible to categorize the occasion into subgroups, like the reason or the manner for a detailed analysis. The place has a strong preference to follow the subject ($d = 1.2816$), and the time has a preference to follow the subject ($d = 1.5352$) and to be followed by the object ($d = 0.5749$).

Table 7. Significant pairs of the bi-grams of grammatical function types (extracted from Table 6)

'Starting position' of the bi-gram	Significant pairs of the bi-grams of grammatical function types	Effect size (Cohen's *d*)
Subject	Subject – Object (F) > Object – Subject (B)	–
	Occasion – Subject (B) > Subject – Occasion (F)	0.6907
Object	Subject – Object (B) > Object – Subject (F)	–
	Object – Occasion (F) > Occasion – Object (B)	0.7527
	Place – Object (B) > Object – Place (F)	0.5802
Occasion		
Place	Subject – Place (B) > Place – Subject (F)	1.2816
	Place – Object (F) > Object – Place (B)	2.0934
	Time – Place (B) > Place – Time (F)	–
Time	Subject – Time (B) > Time – Subject (F)	1.5352
	Time – Object (F) > Object – Time (B)	0.5749
	Time – Place (F) > Place – Time (B)	–

The interpretation is drawn in Figure 14. The order of the positions of the time and the occasion is not significant from both 'starting positions' of the grammatical function types.

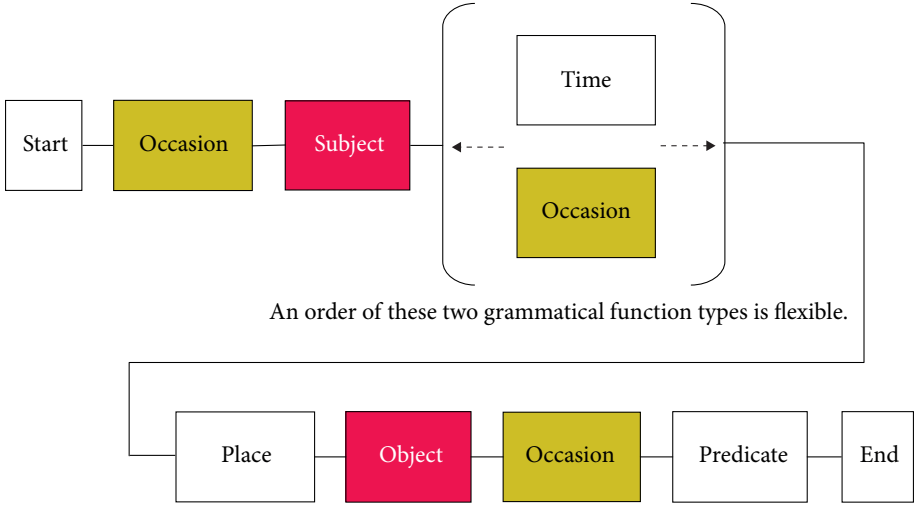


Figure 14. An interpretation of results from tests for $H3_0$

We also investigated differences of proportions of (1) 20 pairs (= 5 * (5–1)) of forward bi-grams with the reverse order of the forward bi-grams, (e.g., Time – Place (F) and Place – Time (F)), and (2) 20 pairs of backward bi-grams with their reverse order of the backward bi-grams (B) (e.g., Time – Place (B) and Place – Time (B)). The results are consistent with results shown in this section.

7. Conclusions

In the present paper, we used statistical tests for paired grammatical function types or paired bi-grams of the grammatical function types to investigate which grammatical function type has a significant positional preference in the clause, or rather which grammatical function type has a significant preference in order with another grammatical function type.

By means of our tests we have confirmed the following points:

1. The subject has a strong preference to appear in the first position in the clause.
2. The object has a strong preference to appear in the last position among the five grammatical function types, before the predicate in the clause.
3. The time has a strong preference to appear before the predicate by two or more positions.
4. The time also has a preference to appear before the object. This is consistent with 3.
5. The subject has a significant preference to appear before the object, and the object has a significant preference to appear after the subject. This is consistent with 1 or 2.
6. The object has a significant preference to appear after the place, and the place has a strong preference to appear before the object. This is consistent with 2.
7. The time has a significant preference to appear before the place, and the place has a preference to appear after the time. This is consistent with 3 and 4.
8. The place has a strong preference to appear after the subject, and the time has a strong preference to appear after the subject. This is consistent with 1.
9. The subject has a preference to appear after the occasion.
10. The object has a strong preference to appear before the occasion.
11. However, the occasion has no significant order preference with the other grammatical function types.
12. The time and the occasion have no significant order in the clause.

The present study statistically clarifies the order of the time and the place, as can be seen by comparing Figure 14 and Figure 1. It has been found that the subject has a preference to appear after the occasion, and that the occasion has no significant order preference with the other grammatical function types. The occasion can also appear between the subject and the object, and the time and the occasion have no significant order in the clause. Therefore, the behaviour of the occasion suggests that Japanese is a free word order language.

The subject and object form a kind of 'bracket structure' of the grammatical functions in the clause. However, occasion can also appear outside of the 'bracket structure', and the subject and object can be elliptic in the clause. Therefore, their role should be considered as 'anchors' in the clause.

Abbreviations[4]

ATTR	attributive	OBJ	object
CONNECT	connective form	PAST	past tense form
CONTINUOUS	continuous aspect	PERFECT, PP	postposition
COPULA, GEN	genitive	PROG	progressive form
INS	instrument	SUBJ	subject[4]
LOC	location	TOPIC	
NOM	nominalized form		

Funding

This work was partly supported by Grant-in-Aid for Scientific Research (C) [Project number 16K02741] of the Japan Society for the Promotion of Science (JSPS).

Software and digital dictionaries

Graduate Schools of Informatics in Kyoto University & NTT Communication Science Laboratories. 2008. Morphological analyzer: *MeCab*, version 0.97. (https://code.google.com/p/mecab/)

National Institute for Japanese Language and Linguistics. 2008. Digital dictionary for the natural language processing: *UniDic*, version 1.3.9. (http://www.ninjal.ac.jp/corpus_center/unidic/)

4. The Japanese postposition *ga* is a subject marker while *wa* is the marker of topic, an information on structural category. Therefore *wa* can mark the subject, when it is topicalized.

References

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins & Hannah R. Rothstein. 2009. *Introduction to meta-analysis*. Chichester, UK: Wiley. https://doi.org/10.1002/9780470743386

Čech, Radek, Petr Pajas & Ján Mačutek. 2010. Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics* 17(4). 291–302. https://doi.org/10.1080/09296174.2010.512162

Čech, Radek & Ludmila Uhlířová. 2014. Adverbials in Czech: Models for their frequency distribution. In Ludmila Uhlířová, Gabriel Altmann, Radek Čech & Jan Mačutek (eds.), *Empirical approaches to text and language analysis*, 45–59. Lüdenscheid: RAM-Verlag.

Del Re, A. C. 2015. *Manual of Package 'compute.es' (Compute Effect Sizes) of R, version 0.2-4*. https://cran.r-project.org/web/packages/compute.es/compute.es.pdf (4 March, 2018.)

Minami, Fujio. 1974. *Gendai nihongo no kozo* [The structure of the present Japanese]. Tokyo: Taishukan.

Minami, Fujio. 1993. *Gendai nihongo bunpo no rinkaku* [The outline of the grammar of comtemporary Japanese]. Tokyo: Taishukan.

National Language Research Institute. 1964. *Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki* [Vocabulary and Chinese characters in ninety magazines of today: Volume 3: Analysis of results]. Tokyo: Shuei Shuppan.

Ogino, Takao, Masahiro Kobayashi & Hitoshi Isahara. 2003. *Nihongo Doshi no Ketsugoka* [Verb valency in Japanese]. Tokyo: Sanseido.

Sanada, Haruko. 2012. Joshi no Shiyo Dosu to Ketsugoka ni Kansuru Keiryoteki Bunseki Hoho no Kento [Quantitative approach to frequency data of Japanese postpositions and valency]. *Rissho Daigaku Keizaigaku Kiho* [The quarterly report of economics of Rissho University] 62(2). 1–35.

Sanada, Haruko. 2014. The choice of postpositions of the subject and the ellipsis of the subject in Japanese. In Ludmila Uhlířová, Gabriel Altmann, Radek Čech & Jan Mačutek (eds.), *Empirical approaches to text and language analysis*, 190–206. Lüdenscheid: RAM-Verlag.

Sanada, Haruko. 2015. A co-occurrence and an order of valency in Japanese sentences. In Arjuna Tuzzi, Jan Mačutek & Martina Benešová (eds.), *Recent contributions to quantitative linguistics*, 139–152. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110420296-013

Sanada, Haruko. 2016. The Menzerath-Altmann law and sentence structure. *Journal of Quantitative Linguistics* 2 3(3). 256–277. https://doi.org/10.1080/09296174.2016.1169850

Sanada, Haruko. 2018a. Negentropy of dependency types and parts of speech in the clause. In Jingyang Jiang & Haitao Liu (eds.), *Quantitative analysis of dependency structures*, 119–144. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110573565-007

Sanada, Haruko. 2018b. Quantitative interrelations of properties of complement and adjunct. In Lu Wang, Reinhard Köhler & Arjuna Tuzzi (eds.). *Structure, function and process in texts*, 78–99. Lüdenscheid: RAM-Verlag.

Sanada, Haruko. 2019. Quantitative aspects of the clause: The length, the position and the depth of the clause. *Journal of Quantitative Linguistics* 26(4). 306–329. https://doi.org/10.1080/09296174.2018.1491749

Sanada, Haruko. Forthcoming. Length of clauses and a perspective on the three dimensional model of Synergetic Linguistics. *Rissho Daigaku Keizaigaku Kiho* (The Quarterly Journal of Rissho Economics Society), vol. 71(1).

Linking the dependents

Quantitative-linguistic hypotheses on valency

Petra Steiner
Universität Bayreuth

This chapter relates the syntactic and semantic aspects of case and valency to morphological properties and builds a model of these. Two linguistic hypotheses are derived and tested on corpus data and data of a lexicon which are built on the principles of Frame Semantics. My hypotheses are confirmed: (A) the larger the number of variables of a semantic predicate, the larger is the number of the syntactic dependents of the realized syntactic constructs, and (B) the larger the number of semantic roles, the larger is the tendency for shortening on the syntactic level.

Keywords: semantic and syntactic valency, morphological case, functional equivalents, frame semantics, FrameNet

1. Introduction

In linguistics, case is a complex and multi-layered term. It was primarily considered a set of morphological categories, such as nominative or dative, whose 'general meanings' are rather abstract (Jakobson 1936, 1984). Starting from Fillmore's (1968) seminal paper, the meaning of case changed more and more from the morphological class or class marker to the scope of syntax and semantics: from surface case to deep case, from inflection to semantic roles.

Though case and valency have become a field of larger interest in quantitative linguistics (Köhler 2012: 92–114; Steiner 2013; Jiang & Liu 2018), most studies treat frequency distributions and are restricted either to the semantic level, grammatical functions, or to syntactic realizations.

This chapter relates the syntactic and semantic aspects of case and valency to morphological case and other means of word-formation and builds a model of these. Valency and morphological case can be considered as two sides of the coding of predicate variable structures. The coding of semantic case and valency

https://doi.org/10.1075/cilt.356.06ste

can be linked to syntactic case and valency and to morphological case by different functional equivalents. This is done by a small model in § 2.

Valency is a linguistic property of verbs and other parts of speech which is mostly defined in qualitative ways. Section 3 provides qualitative and quantitative definitions for the semantic and syntactic levels of valency. In § 4, relations are derived for the number of semantic roles and the number of dependents within linguistic constructs. We assume effects of the minimization of the encoding effort as postulated by Köhler (2005: 766ff.) and expect the size of the quantitative semantic valency to boost this requirement of the synergetic system. The definitions build on Fillmore's (2007) classification for argument linking which comprises null instantiations (syntactically non-realized semantic roles) and n:1 constructs (more than one semantic roles are realized in one dependent). We derive two linguistic hypotheses for the relation between the semantic and syntactic level of valency. The data used is from the FrameNet project. Section 5 therefore introduces Frame Semantics and FrameNet annotations. For extracting the frequency data, we use a frame-annotated corpus of English language. This data is then taken for testing three statistical hypotheses in § 6. Section 7 finally presents the interpretations and conclusions.

2. The functional equivalents of coding semantic case

The notion of case is ambiguous. While 'morphological case' refers to the inflectional marking of lexemes, the notion of 'semantic case' denotes their semantic roles, which can or cannot be realized in utterances. 'Syntactic case' on the other hand can be classified by grammatical functions such as subject and object. All these instances are connected by the communicative requirement of expressing the relations and roles of semantic predicates.

For instance, the predicate MEET requires two persons or two parties, or a group who meet. Furthermore, it is presuppositional knowledge about meetings that they take place at a certain place and time. Such semantic properties can be expressed as (1) with the predicate MEET and the semantic variables PERSONA, PERSONB, PLACE, and TIME, of which some have to be expressed in their corresponding linguistic expressions and others can be omitted. Nevertheless, they all belong to the predicate. The number and kind of these variables are subsumed under the notion of 'semantic valency'.

Depending on the language and situation, different means can be chosen for the coding of the semantic cases. While the sentences in (2) and (3) are expressing two different perspectives of activity, (4) expresses a reciprocal relation. The syntactic realization can comprise a different number of dependents: two in the first three sentences, one in the last, and can have different grammatical functions: only a

subject in (4), subject and object in (2), and a prepositional object in (3). Some of the semantic roles can be omitted on the syntactic level, such as PLACE and TIME. The grammatical functions can be expressed by different phrase types, such as NPs or a PP in (3). These linguistic properties are denoted as 'syntactic valency'.

(1) MEET(PERSONA, PERSONB, PLACE, TIME, ...)

(2) *The old lady meets the boy.*

(3) *The boy meets with the old lady.*

(4) *The boy and the old lady meet.*

In English, the role assignment is mostly expressed by the sequence of the dependents, but other languages permit a more flexible arrangement by inflectional marking: the morphological case. In German, the role of PERSONB as in (5) can be expressed by an inflectional suffix and the form of the article, leading to a construct with the person who is met in first sentential position in (6), and the subject in final position.

(5) MEET(WOMAN, BOY)

(6) *Den* *Jungen* *trifft* *die* *Frau.*
the.ACC.SG boy.ACC.SG meets the.NOM.SG woman.NOM.SG
"The woman meets the boy."

Morphological case can be considered as one of various different means for marking word forms for their semantic case. Such coding of semantic case and valency is a communicative requirement which can be met by different functional equivalents. In the following, we apply Köhler's (2005: 765) synergetic model of the functional equivalents of coding to the instance of semantic case and valency, to relate these different means with each other.

Figure 1 provides an overview of functional equivalents for the coding of semantic case and valency: Morphological means (marked by M) can serve for coding semantic case by inflectional markers as in (6) or for changing the number of dependents by derivation or conversion: the semantic role of PLACE in (7) can be realized by a prepositional phrase as in (8) or incorporated into the verb as in (9).

(7) PUT(AGENT, OBJECT, PLACE)

(8) *He put the book into a box.*

(9) *He boxed the book.*

Syntactic means (S) comprise the sequence of linguistic units as in (2) vs. (3), but also syntactic constructions such as cleft sentences (10) which permit alternations of the conventional order of syntactic case.

(10) *It is the boy who the old lady meets.*

Prosodic means (P) can offer deictic possibilities for the coding of semantic case. The linguistic expression in (11) does not show morphological marking of the semantic cases. By convention, the sequence of the dependents would yield the interpretation of the semantic cases. However, stress on the first noun phrase can shift the conventional grammatical function of the first dependent from subject to object connected with a marking of focus.

(11) *Die* *Frau* *trifft* *sie.*
the.ACC.SG woman.ACC.SG meets she.NOM.SG
"She meets this woman."

Other aspects for the interpretation are semantic properties of lexemes, such as animacy or general world knowledge of what can happen, so-called frames. In general, morphological case and syntactic constructions are two ways among others for the mapping of semantic case and valency to syntactic case and valency.

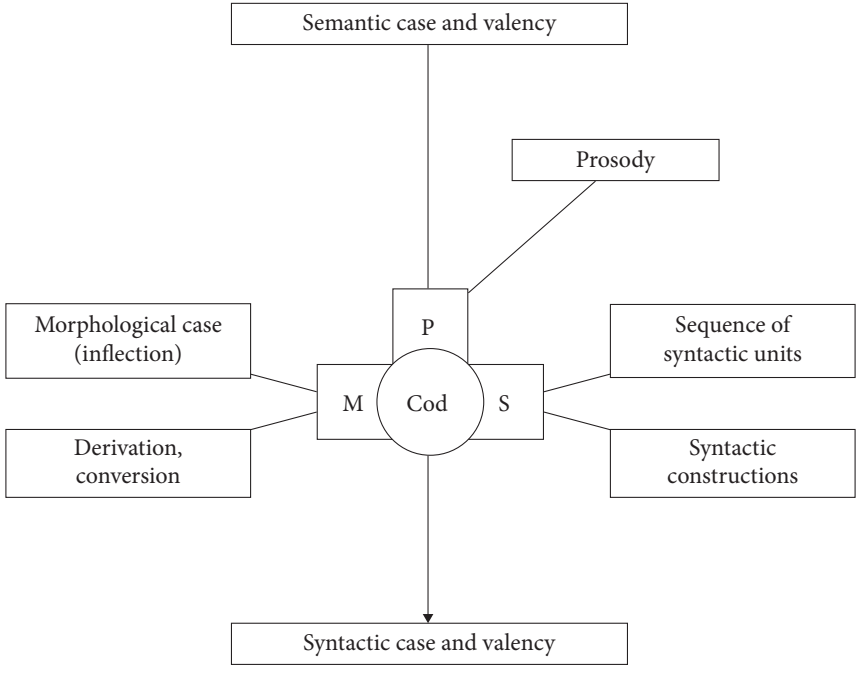


Figure 1. The functional equivalents of coding semantic case. Cod: Requirement of coding, M: Morphological means, S: Syntactic means, P: Prosodic means

3. Quantitative definitions of valency

The notion of valency in linguistics derives from the descriptive approach of dependency grammar as developed by Tesnière (1953, 1959, 1965, 2015). It denotes "[the] ability of a verb, noun, adjective, adverb or particle to open up valency slots to be filled by complements" (Herbst & Schüller 2008: 209).

In this sense, valency is a concept that comprises the kind of complements that lead to a syntactic well-formed expression. For instance, the verbs *regard* and *consider* have different restrictions (see Emons 1978: 76ff.): while both verbs can be combined with a prepositional object starting with *as* ((12)–(13)), the construction with two NPs is only syntactically well-formed for *consider* ((14)–(15)).

(12) *We regarded Bill as a friend.*

(13) *We considered Bill as a friend.*

(14) *We regarded Bill a friend.

(15) *We considered Bill a friend.* (Emons 1978: 76ff.)

In general, valency descriptions distinguish between complements and adjuncts. Complements are defined as clause constituents, which are typically associated with the government element (Herbst & Schüller 2008: 22). Complements are further differentiated into non-optional and optional, and the optional class into purely optional and contextually optional (Herbst 2003). Adjuncts are the complementary class of the former. Typical examples are the realizations of semantic roles like Time, Place, and Manner. On the syntactic level, they are defined as non-obligatory constituents, which in a strict sense do not belong to the valency. However, they play an important role in the construction of sentences and will be subject to the following investigation. Therefore, the union of complements and adjuncts is defined as:

> The dependents of a lexical unit are considered as the (complete) set of complements and adjuncts. The dependency structure of the syntactic level comprises constraints of grammatical functions (including adverbials), and the one of the semantic level constraints of the semantic dependents.

The distinction between complements and adjuncts yields manifold problems which have been discussed for decades and elsewhere (e.g., Somers 1984, 1987: 12–18; Storrer 1992: 54–95; Ágel 2000: 171–191; Herbst & Schüller 2008: 113–116; Lichte 2015: 26–32). In order to avoid these and other issues, the following quantitative measures are specified:

– Quantitative semantic valency is operationalized as the number of semantic roles as given by the frame definitions of FrameNet (see § 5).
– Quantitative syntactic valency is operationalized as the number of syntactic dependents as given by the annotated instances of the FrameNet annotation corpus (see § 5).

This is formally in line with the definition of quantitative valency by Herbst & Schüller (2008: 136) which is simply the number of complements a governing element possesses.

4. Derivation of linguistic hypotheses of valency

The relation between the semantic representation of dependency structure with its syntactic realization is called 'linking'. The main postulate of the so-called 'linking theory' is that semantic similarity of predicates results in similarities between their semantic roles and dependency structures (Levin 1993). Linking rules subsume which semantic roles tend to be realized in which sentential position or which verb classes show which syntactic realizations (Levin & Hovav Rappaport 2005). There are many exceptions, and much cross-linguistic variation. In any case, hierarchies of deep case or theta-roles are no explanations for the relationship between semantics and syntax, but mere descriptions of preferences. Rules merely "represent concepts which enable us to describe certain language phenomena" (Altmann 1978: 4). They do not explain anything (Köhler 1987); on the contrary, they necessitate explanation (Köhler 1986: 4). Contrary to this, the left part of Figure 2 shows the quantitative linking hypothesis for the quantitative valencies of the semantic and the syntactic level in connection to the syntactic functional equivalent of the requirement of coding:

Hypothesis 1: The larger the number of semantic variables of a semantic predicate, the larger is the number of syntactic dependents within the realized linguistic construct.

If the requirement of coding is met by syntactic means, this would cause the realization of semantic variables as syntactic dependents in linguistic utterances. This assumption would mean a relation of 1:1 – for each semantic role, a syntactic realization would exist.

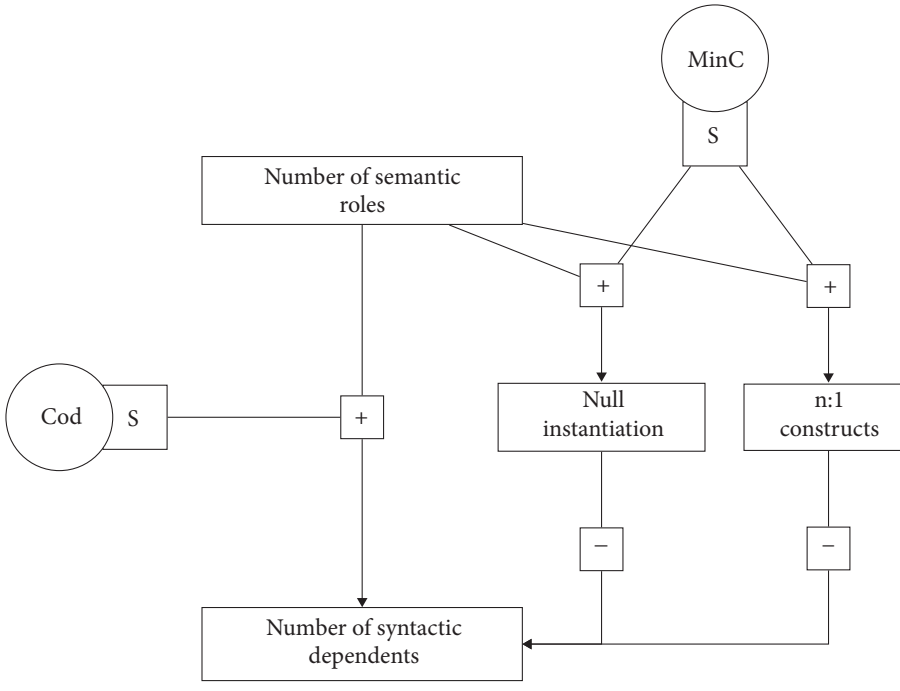


Figure 2. Relations between syntactic and semantic valency

However, in many instances this is not the case. For example, in (16), the relation is 0:1, due to a syntactic construction with an expletive *it* and no semantic role. On the other hand, often not all semantic roles are present in the realizations. (17)–(19) show instances of such 1:0 relation of semantic roles with null instantiations (Fillmore et al. 2003: 245f.). The missing object in (17) is classified as a definite null instantiation (DNI), meaning that the semantic role, the RECEIVER, can be inferred from the context. (18) contains an indefinite null instantiation (INI), which can be inferred from conventional knowledge (Ruppenhofer et al. 2016: 28f.). (19) shows a conversion which reduces the number of syntactic dependents. In (20), two semantic roles are united in one syntactic construct, meaning there is an n:1 relation.

(16) *It is important that their time should not be wasted.* (Erdmann 1988: 329)

(17) *He gave a book.*

(18) *She eats.*

(19) *He boxed the book.*

(20) *John cut her hair short.* (see Fillmore 2007: 144f.)

Shortening of linguistic units is usually caused by the requirement of the minimization of the encoding effort (MinC) (Köhler 2005: 766ff., 1986: 50f.). These shortenings can be phonological or indirect by increasing the polylexy of lexical units, as shown in the classical Köhler basic model (Köhler 1986: 74). On the syntactic level, constituents of sentences or phrases will be shortened or omitted. Null instantiations and n:1 constructs reduce the number of syntactic dependents. Moreover, the number of semantic roles will strengthen the effect of MinC, or with other words:

Hypothesis 2: The larger the semantic valency, the larger is the tendency of their non-realization on the syntactic level.

This hypothesis is antagonistic to Hypothesis 1 and shown on the right side of Figure 2.

5. FrameNet data

For testing the hypotheses, we use the FrameNet database for counts of the semantic level and the frame-annotated corpus for counts of the syntactic realization. FrameNet (Fillmore et al. 2003) is a lexicographical project founded by Charles J. Fillmore. It is based on Frame Semantics (Fillmore 1982, 1985) whose central idea is that lexical meanings are described relative to a background of coherent knowledge. The frame definitions are substantiated by instances from corpora. For example, (21) is an instance of the Travel frame with instantiations of the Frame elements (FEs) Traveler and Path. (22) comprises the FEs Traveler, Mode_of_transportation, and Frequency. While the first two FEs are considered as core (conceptual necessary) frame elements, Frequency is an extra-thematic frame element (Ruppenhofer et al. 2016: 23), which means that it does not conceptually belong to this frame. The Goal is not instantiated in this sentence but its existence can be inferred by general knowledge (INI). In (23), the FE Traveler is not realized due to the passive construction; this is termed a constructional null instantiation (CNI). Only core frame elements can be null instantiated. Peripheral (conceptually not necessary) frame elements such as Time or Manner are annotated if they are realized.

(21) [$_{\text{Traveler}}$Paul Pratt] TRAVELLED [$_{\text{Path}}$through forty-eight countries].

(22) [$_{\text{Traveler}}$I] used to TRAVEL [$_{\text{mode_of_transportation}}$by bus] [$_{\text{Frequency}}$a lot], so I had a season ticket. [$_{\text{Goal}}$INI]

(23) Days began early and ended late so that [$_{\text{Path}}$maximum distances] could be TRAVELLED.[$_{\text{Traveler}}$CNI] (FrameNet-DB 2020)

Full text annotation with frames is provided from the FrameNet project (Baker et al. 2003). For the following investigation, we take a sample of frame-annotated ANC texts and texts from the Nuclear Threat Initiative website. These consist of approximately 33,000 word tokens.

For the sake of comparability, we restrict the investigation to verbs. The XML-annotated corpus comprises 3,907 frame-annotated verbs. We extract the frame annotation sets with their frame elements. For each instance of the annotated verbs, we count the number of core, peripheral and extra-thematic syntactically realized Frame elements. Instances of null instantiations are annotated, so their number can be extracted too. 127 frame annotations had to be deleted due to missing labels, so the dependents of 3,780 verbs in the text can be further analyzed.

For the semantic valencies, the counts are drawn from the FrameNet lexical database. We count the number of all possible frame elements, which can be a lot, as peripheral and extra-thematic elements are included. For instance, the frame TRAVEL has 25 frame elements of which 7 are core FEs, 9 peripherals, and 9 extra-thematic elements. The annotation sets of frame elements from the corpus are much smaller.

6. Statistical hypotheses and testing

After the brief introduction to FrameNet and the derived data, the transfer to the statistical hypotheses becomes feasible: The number of (possible) semantic roles is determined as the size of the set of the respective frame elements from the lexical database (*SemRoles*). The number of respective syntactic dependents is operationalized as the number of the annotated syntactic dependents of an annotated frame (*SynDep*).

According to the first hypothesis, the relative change of the length of the syntactic construct is proportional to the number of (potential) semantic roles of the underlying frame. This can be formalized as

$$\frac{SynDep'}{SynDep} = \frac{a}{SemRoles} \tag{1}$$

or

$$SynDep = a\ SemRoles^{b} \tag{2}$$

with a as the proportionality factor. For the second hypothesis only the number of null instantiations is taken into account. This results in two equations: The impact of the number of semantic roles on the number of null instantiations is proportional (see Figure 2) and is represented by

$$NI = c\ SemRoles^{d} \tag{3}$$

The relation between null instantiations and the number of syntactic dependents is antiproportional:

$$SynDep = fNI^{-g} \tag{4}$$

For the method of linear regression, all equations have to become linearized by a logarithmic transformation. For testing the first hypothesis, the frequencies of all combinations of possible frame elements and annotated FEs in the corpus were counted. Table 1 shows these counts. For instance, for frames with 5 possible FEs and 2 elements in text, the corpus comprises 11 instances. For example, the verb *feel* which invokes the frame FEELING has an entry in the FrameNet lexical database with the five frame elements EXPERIENCER, EMOTION, EMOTIONAL_STATE, EVALUATION, and CAUSE. One instance with two Frame elements in text is the sentence:

(24) [EXPERIENCERI] often feel [EMOTIONAL_STATE as if I only get half the story sometimes] (FrameNet Full Text Annotation of ANC, Slate Magazine article: Entrepreneur As Madonna, 2020)

Instances of 0 syntactic dependents are possible due to constructional null instantiations, e.g., of an imperative form.

The values ≥26 were grouped and the average number of syntactic dependents for each class of semantic roles was calculated. Figure 3 shows the relationship of these variables. The R^2 value complies with the scattering. The value range is obviously small as the syntactic constructs for all SemRoles classes show their maximum at two syntactic dependents and the average increases very modestly.

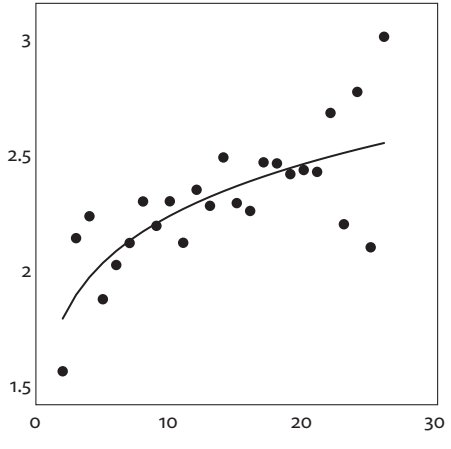


Figure 3. Average number of syntactic dependents as a function of the number of semantic roles, fitting the function $SynDep = 1.62SemRoles^{0.138}$ and resulting in $R^2 = 0.569$

Table 1. Frequencies of the numbers of syntactic dependents for frequency classes of semantic roles (SemRoles)

	Number of syntactic dependents									
SemRoles	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Σ**
2		4	5							9
3		1	18	4						23
4		2	31	3	4					40
5		3	11	1						15
6	1	6	53	9						69
7		9	107	18	3					137
8		9	84	36	6					135
9	2	17	95	38	6					158
10	4	30	184	90	17	3				328
11		81	169	71	21	2				344
12	1	43	122	87	20	4				277
13	2	28	127	66	12	2				237
14		38	227	131	46	7	1		1	451
15		73	206	122	24	7	1			433
16	2	20	53	34	5	3				117
17	1	20	68	61	12	3	1			166
18		21	88	78	13	3				203
19		16	51	38	13					118
20	3	24	138	91	27	2				285
21		11	19	24	6					60
22	1	6	27	28	12	1	1			76
23		13	13	11	5					42
24	1	4	15	11	8	2	1			42
25	1	1	5	4						11
26				2						2
32				2						2
Σ	19	480	1916	1060	260	39	4	1	1	3780

For the second hypothesis, the counts of the null instantiations of the frame-annotated corpus were extracted. Equation 3 relates the number of all semantic roles from the lexical FrameNet database to the null instantiations of the annotation sets, Equation 4 relates the number of the null instantiations to the number of the realized frame elements.

Table 2 shows the counts of the frequency classes of semantic roles and their frequencies over the numbers of null instantiation. For example, for frames with 5 possible FEs and one null instantiation in text, the corpus comprises 4 instances.

Table 2. Frequencies of the numbers of null instantiations for frequency classes of semantic roles (SemRoles)

	Number of null instantiations							
SemRoles	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Σ**
2	9							9
3	19	4						23
4	32	7	1					40
5	11	4						15
6	57	11	1					69
7	119	15	3					137
8	109	25	1					135
9	128	23	7					158
10	243	80	4	1				328
11	260	74	9	1				344
12	156	95	19	7				277
13	169	56	11	1				237
14	271	142	33	4	1			451
15	276	133	18	5			1	433
16	54	53	8	2				117
17	111	41	13	1				166
18	99	95	9					203
19	75	42	1					118
20	226	53	5	1				285
21	28	25	7					60
22	35	31	8	2				76
23	10	12	17	3				42
24	22	16	4					42
25	5	5		1				11
26	2							2
32	1	1						2
Σ	2527	1043	179	29	1	0	1	3780

The value for 2 semantic roles was excluded as there are no null instantiations annotated for such a case. Again, the values ≥26 were grouped. Then the average numbers of null instantiations for each class of semantic roles were calculated.

Figure 4 shows the relationship between the two variables. The R^2 value complies with the scattering, especially of some outliers. Again, the range of value is small, but with a growing average number of semantic roles, null instantiations become more frequent, and the hypothesis becomes substantiated.

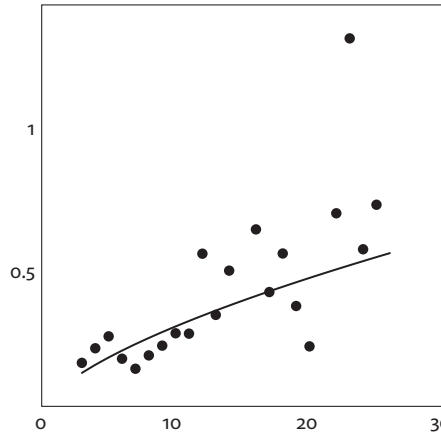


Figure 4. Average number of null instantiations as a function of the number of semantic roles, fitting the function $NI = 0.068 SemRoles^{0.644}$ and resulting in $R^2 = 0.506$

The range of null instantiation is small. However, Table 3 shows a tendency of decreasing syntactic dependents for the NI frequency classes. For the testing of the hypothesis, the last two ranks were grouped, and the average numbers of syntactic dependents for each class calculated. The domain was shifted to the right by adding 1 to permit the logarithmic transformation. Figure 5 shows that for larger frequencies of null instantiations within a syntactic construct, the number of syntactic dependents really decreases. Therefore, both parts of the second hypothesis could be confirmed.

Table 3. Frequencies of the numbers of syntactic dependents for frequency classes of null instantiations

	Number of syntactic dependents									
NIs	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Σ**
0	1	109	1352	822	208	30	3	1	1	2527
1	3	280	484	218	48	9	1			1043
2	11	77	68	19	4					179
3	4	13	11	1						29
4		1								1
6			1							1
Σ	19	480	1916	1060	260	39	4	1	1	3780

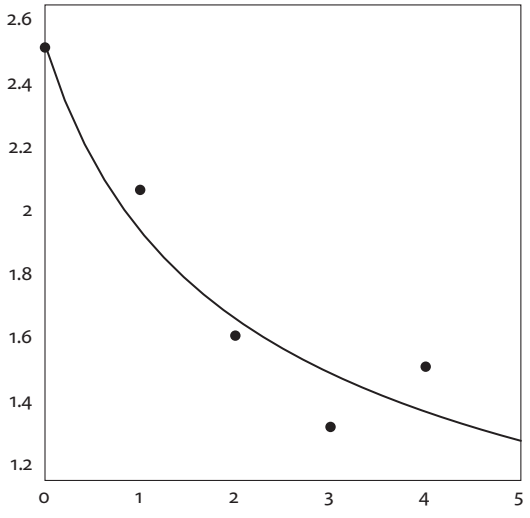


Figure 5. Average number of syntactic dependents as a function of null instantiations, fitting the function $SynDep = 2.514(NI + 1)^{-0.383}$ and resulting in $R^2 = 0.888$

7. Interpretations and conclusions

Both linguistic hypotheses with their three statistical hypotheses could be confirmed in the investigation with data based on Frame semantics. This means that the number of semantic roles influences syntax in a non-straightforward way: the more semantic roles there are, the stronger is the impact of the minimization of the encoding effort. This is a first substantiation of linking from the field of quantitative linguistics. The outliers of the data were not excluded for the sake of transparency. However, they might be results of inconsistent human annotation.

The data show other interesting aspects: most of the realizations of dependency structures in text tend towards two or three dependents even if the number of possible semantic roles from the conceptual side is large. Table 1 shows a two-dimensional distribution with the maximum in the middle of the ranges. The small domain of the different types of null instantiations seems to indicate the same phenomenon: while the average verbal construct without null instantiations comprises 2.5 dependents, this number is lower for syntactic constructions with NIs which have to be inferred either from cotext, context, or construction. The counts are too small to permit more than a working hypothesis. Nevertheless, an explanation for this might be a preference for certain valency patterns either for cognitive reasons or due to the reduction of MinC. Other investigations have to follow.

References

Ágel, Vilmos. 2000. *Valenztheorie* (Narr-Studienbücher). Tübingen: Narr.

Altmann, Gabriel. 1978. Towards a theory of language. In Gabriel Altmann (ed.), *Glottometrika 1* (Quantitative Linguistics 1), 1–25. Bochum: Brockmeyer.

Baker, Collin F., Charles J. Fillmore & Beau Cronin. 2003. The structure of the FrameNet database. *International Journal of Lexicography* 16(3). 281–296. https://doi.org/10.1093/ijl/16.3.281

Emons, Rudolf. 1978. *Valenzgrammatik für das Englische: Eine Einführung* (Anglistische Arbeitshefte 16). Tübingen: Niemeyer.

Erdmann, Peter. 1988. On the principle of 'weight' in English. In Caroline Duncan-Rose (ed.), *On language: Rhetorica, phonologica, syntactica; a festschrift for Robert P. Stockwell from his friends and colleagues*, 325–339. London: Routledge.

Fillmore, Charles J. 1968. The case for case. In Emmon W. Bach & Robert T. Harms (eds.), *Universals in linguistic theory* (A Holt international edition), 1–88. London: Holt Rinehart & Winston.

Fillmore, Charles J. 1982. Frame semantics. In *Linguistics in the morning calm*, 111–137. Seoul, Korea: Hanshin Publishing Company.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2). 222–254.

Fillmore, Charles J. 2007. Valency issues in FrameNet. In Thomas Herbst & Katrin Götz-Votteler (eds.), *Valency: Theoretical, descriptive and cognitive issues* (Trends in Linguistics. Studies and Monographs 187), 129–160. Berlin: de Gruyter. Reprint 2008. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110198775.1.129

Fillmore, Charles J., Christopher R. Johnson & Miriam N. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3). 235–249. https://doi.org/10.1093/ijl/16.3.235

Herbst, Thomas. 2003. Was soll zum Beispiel eine obligatorische Ergänzung sein? Oder ein zweivalentes Verb? – Zur Interdependenz valenzpolitischer Festlegungen. In Alan Cornell, Klaus Fischer & Ian F. Roe (eds.), *Valency in practice – Valenz in der Praxis*, 65–88. Frankfurt: Lang.

Herbst, Thomas & Susen Schüller. 2008. *Introduction to syntactic analysis: A valency approach* (Narr-Studienbücher). Tübingen: Narr.

Jakobson, Roman. 1936. Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutungen der russischen Kasus. In Praský linguistický krouek (ed.), *Études dédiées au quatrième congrès de linguistes* (Travaux du Cercle linguistique de Prague 6), 240–288. Prague: Jednota eskoslovenských matematiku a fysik. Reprint 1971. In Roman Jakobson (ed.), *Word and language* (Selected Writings II), 23–71. The Hague: Mouton. Reprint 2010. Tübingen: de Gruyter Mouton. https://doi.org/10.1515/9783110873269

Jakobson, Roman. 1984. Contribution to the general theory of case: General meanings of the Russian cases. In Roman Jakobson (ed.), *Russian and Slavic grammar: Studies 1931–1981* (Janua Linguarum. Series Maior 106), 59–103. Berlin: de Gruyter. https://doi.org/10.1515/9783110822885.59. [Translation of Jakobson 1936].

Jiang, Jingyang & Haitao Liu. 2018. *Quantitative analysis of dependency structures* (Quantitative Linguistics 71). Berlin: de Gruyter Mouton. https://doi.org/10.1515/9783110573565

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik* (Quantitative Linguistics 31). Bochum: Studienverlag Brockmeyer.

Köhler, Reinhard. 1987. System theoretical linguistics. *Theoretical Linguistics* 4(2/3). 241–257. https://doi.org/10.1515/thli.1987.14.2-3.241

Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund Genrikhovich Piotrowski (eds.), *Quantitative Linguistik – Quantitative Linguistics: Ein internationales Handbuch – An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft – Handbooks of Linguistics and Communication Science 27), 760–774. Berlin: de Gruyter. https://doi.org/10.1515/9783110155785/html

Köhler, Reinhard. 2012. *Quantitative syntax analysis* (Quantitative Linguistics 65). Berlin: de Gruyter Mouton. https://doi.org/10.1515/9783110272925

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Levin, Beth & Malka Hovav Rappaport. 2005. *Argument realization* (Research Surveys in Linguistics). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511610479

Lichte, Timm. 2015. *Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken* (Empirically Oriented Theoretical Morphology and Syntax 1). Berlin: Language Science Press. https://doi.org/10.26530/OAPEN_603338

Ruppenhofer, Josef, Michael Ellsworth, Myriam R. L. Petruck, Christopher R. Johnson, Collin Baker & Jan Scheffczyk. 2016. *Framenet II: extended theory and practice*. Berkeley, CA: International Computer Science Institute. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Somers, Harold L. 1984. On the validity of the complement-adjunct distinction in valency grammar. *Linguistics* 22(4). 507–530. https://doi.org/10.1515/ling.1984.22.4.507

Somers, Harold L. 1987. *Valency and case in computational linguistics* (Edinburgh information technology series 3). Edinburgh: Edinburgh University Press.

Steiner, Petra. 2013. Diversification of English valency patterns. In Reinhard Köhler & Gabriel Altmann (eds.), *Issues in quantitative linguistics 3* (Studies in Quantitative Linguistics 5), 369–391. Lüdenscheid: RAM-Verlag.

Storrer, Angelika. 1992. *Verbvalenz: Theoretische und methodische Grundlagen ihrer Beschreibung in Grammatikographie und Lexikographie* (Reihe Germanistische Linguistik 126). Tübingen: Niemeyer. https://doi.org/10.1515/9783110914818

Tesnière, Lucien. 1953. *Esquisse d'une syntaxe structurale*. Paris: Klincksieck.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Tesnière, Lucien. 1965. *Éléments de syntaxe structurale* (2nd edn.). Paris: Klincksieck.

Tesnière, Lucien. 2015. *Elements of structural syntax*. Translated by Sylvain Kahane & Timothy John Osborne. Amsterdam: John Benjamins. https://doi.org/10.1075/z.185. [Translation of Tesnière 1965].

Grammar efficiency and the One-Meaning–One-Form Principle

Relja Vulanović
Kent State University at Stark

The more a linguistic system departs from the One-Meaning–One-Form Principle, the less efficient it is. This notion is used to simplify the previous approach to the evaluation of grammar efficiency. The recently proposed measure of the departure from the Principle is revisited and its simplified version is included in the formula for grammar efficiency. The new and the old grammar-efficiency formulas are compared when applied to parts-of-speech systems as defined by Hengeveld. It is shown that the number of calculations is considerably reduced in the new formula, while the results obtained by the two approaches correlate well.

Keywords: One-Meaning–One-Form Principle, bijection, grammar efficiency, propositional function, parts-of-speech system, correlation

1. Introduction

A one-to-one correspondence between linguistic forms and their meaning exists to a great extent in every language, but usually there are also identical forms with different meanings and different forms with the same meaning. Anttila (1972) calls this correspondence the 'One-Meaning–One-Form Principle' (from now on, 'the Principle'). Languages that adhere more to the Principle are also referred to as 'more transparent' (Hengeveld & Leufkens 2018). Such languages represent more efficient systems (Anttila 1972: 181). This is why Miestamo (2008) includes the Principle in the criteria for determining absolute language complexity, where the measure of grammar complexity can be viewed as the reciprocal of the measure of grammar efficiency (Vulanović 2003, 2007). However, the formal measures of grammar efficiency and complexity, as defined in Vulanović (2003, 2007), do not involve any quantity that directly evaluates how much the grammatical structure under investigation violates the Principle.

https://doi.org/10.1075/cilt.356.07vul

Motivated by the need for such a quantity, Vulanović & Ruff (2018) propose a mathematical formula for measuring the degree of violation of the Principle. They consider two versions of the formula, a basic formula and a weighted one, and apply the latter to parts-of-speech (PoS) systems in the sense of Hengeveld (1992) and Hengeveld et al. (2004). The inclusion of the formula in the measure of grammar efficiency is not undertaken in Vulanović & Ruff (2018). This is done in the present paper.

The formula from Vulanović & Ruff (2018) for measuring the degree of violation of the Principle is presented and then simplified in § 2. Since the same PoS systems as in Vulanović & Ruff (2018) are used to exemplify the new grammar-efficiency formula and to compare it to the previous one, the PoS systems are recapitulated in § 3. In § 4, the number of calculations required for the old grammar-efficiency formula is illustrated by considering the Turkish PoS system, as described in Hengeveld & van Lier (2008). Section 5 deals with the new grammar-efficiency formula. It is shown how to include in grammar efficiency the newly modified measure of the extent of violation of the Principle. This is then illustrated by considering the same Turkish-PoS-system example and it is concluded that the new grammar-efficiency formula is simpler because it requires fewer calculations. At the same time, the values calculated by the old and the new formulas are comparable. This is confirmed further in § 6, where all possible basic PoS-system types (cf. Hengeveld & van Lier 2010, Vulanović 2008, 2009) are considered. Their grammar-efficiency values are found using the two formulas and a good correlation between the results is obtained. Finally, § 7 contains some concluding remarks.

2. Measures of the degree of violation of the Principle

Some notation is introduced first (cf. Vulanović & Ruff 2018). Let $|A|$ denote the number of elements in a set A (all sets in this chapter are finite and non-empty). Let Φ be a relation between two sets, X and Y, $\Phi \subseteq X \times Y$. We can think of X and Y as the sets of meanings and forms, respectively. For each $y \in Y$, we define $v_x(y)$ as the number of elements x in X such that $(x,y) \in \Phi$,

$$v_X(y) = |\{x \in X : (x,y) \in \Phi\}|.$$

Similarly, let

$$v_Y(x) = |\{y \in Y : (x,y) \in \Phi\}|.$$

We assume that $v_X(y) \geq 1$ for each $y \in Y$ and $v_Y(x) \geq 1$ for each $x \in X$, which means that all elements of X and Y are used in the relation Φ. Then, the set

$$B = \{(x,y) \in \Phi : v_X(y) = v_Y(x) = 1\}$$

contains all one-to-one pairs in Φ. We have that $B \subseteq \Phi$ and $0 \leq |B| \leq |\Phi|$. If Φ is a bijection (a one-to-one correspondence) between X and Y, then $|X| = |Y| = |\Phi| = |B|$.

Mathematically speaking, measuring how much a linguistic system departs from the Principle is the same as finding a measure, denoted by $\mu(\Phi)$, of how far Φ is from a bijection. The following definition of such a measure is proposed in Vulanović & Ruff (2018):

$$\mu(\Phi) = \mu_\theta(\Phi) := \frac{(1+\theta)|\Phi| - \theta|B|}{\min\{|X|, |Y|\}}, \tag{1}$$

where θ is a positive parameter to be chosen by the user ($\theta = 1$ is used in all calculations in Vulanović & Ruff (2018)). This is simplified here by eliminating the parameter θ and defining

$$\mu(\Phi) = \hat{\mu}(\Phi) := \frac{|\Phi \backslash B|}{\min\{|X|, |Y|\}} + 1. \tag{2}$$

Both formulas are motivated by the following guidelines.

a. $\mu(\Phi) = 1$ if Φ is a bijection, otherwise $\mu(\Phi) > 1$.
b. $\mu(\Phi)$ is greater if Φ is greater.
c. $\mu(\Phi)$ is greater if $|B|$ is smaller.
d. $\mu(\Phi)$ is greater if $|X|$ and $|Y|$ are smaller.
e. $\mu(\Phi) = \mu\Phi(^{-1})$, where $\Phi^{-1} = \{(y,x) : (x,y) \in \Phi\}$.

Properties b. and c. are satisfied by (2) because $B \subseteq \Phi$ implies that $|\Phi \backslash B| = |\Phi| - |B|$. Note also that property d. requires of $\mu(\Phi)$ to be some kind of a relative measure. Otherwise, $\mu(\Phi) = |\Phi \backslash B| + 1$ would satisfy the remaining four properties.

It is shown in Vulanović & Ruff (2018) that a weighted version of formula (1) may be more appropriate for application to PoS systems. In the same way, a generalized, weighted version of (2) is used here,

$$\mu(\Phi) = \bar{\mu}(\Phi) := \frac{||\Phi \backslash B||}{\min\{||X||, ||Y||\}} + 1. \tag{3}$$

Here, $||A|| = w_1 + w_2 + \cdots + w_n$ is the sum of weights that are assigned to each of the n elements of the set A. The weights are normalized in the sense that $\min_{i=1,2,\ldots,n} w_i = 1$. The weights depend on the set, so, for instance, $||X||$ and $||Y||$ may be different even in the case when X and Y have the same number of elements. However, since $B \subseteq \Phi$, each element of B should carry the same weight as it does in Φ. Then, $||\Phi \backslash B|| = ||\Phi|| - ||B||$ and it is easy to see that (3) satisfies properties a.–d. when $|\bullet|$ is replaced with $||\bullet||$. The only property that cannot be guaranteed is e., which is because X and Y may have different weights. If all weights in the sets Φ, B, X, and Y are set equal to 1, the formula (3) reduces to (2).

Formula (3) is used throughout the rest of the paper and $\bar{\mu}(\Phi)$ is simply referred to as μ.

3. Hengeveld's part-of-speech systems

In Hengeveld's approach to PoS systems (Hengeveld 1992; Hengeveld et al. 2004; Hengeveld & van Lier 2010), four propositional functions (syntactic slots) are considered,

P = head of predicate phrase,
p = modifier of predicate phrase,
R = head of referential (nominal) phrase,
r = modifier of referential phrase.

The four propositional functions can also be represented by the scheme in Table 1 (Hengeveld & van Lier 2010).

Table 1. The four propositional functions

	Head	**Modifier**
Predicate phrase	P	p
Referential phrase	R	r

PoS systems consist of word classes that are distinguished by the propositional functions they can fulfill. Referring to the formal description in § 2, the set of propositional functions in the PoS system is the set of meanings, X. Some PoS systems have all four propositional functions, but there are also systems without one or both modifier functions. We also consider the case $X = \{P\}$ because there are some languages that come close to this theoretical extreme (Hengeveld 1992; Hengeveld et al. 2004; Hengeveld & van Lier 2010). Therefore, $|X| = \ell$, where $\ell \in \{1,2,3,4\}$. Table 2 shows what propositional functions may be present in a PoS system.

Table 2. Possible propositional functions in a PoS system

ℓ				
4	3	3	2	1
P R r p	P R r	P R p	P R	P

The propositional function P occurs 5 times in Table 2, R – 4 times, and each r and p – 2 times. These counts are used to determine the weights assigned to the propositional functions. The smallest weight, which has to be equal to 1, is assigned to each r and p, and then, proportionally, 2 is assigned to R, and 2.5 to P. Another possible weighting system can be found in Vulanović & Ruff (2018). It is omitted here for simplicity. Therefore, because of Table 2,

$$\|X\| = 4.5+\ell-2 \text{ if } \ell = 2,3,4, \text{ and } \|X\| = 2.5 \text{ if } \ell = 1. \quad (4)$$

At the same time, the set of word classes in a PoS system is the set of forms, *Y*. Table 3 shows all theoretically possible word classes. It also includes their weights, which are discussed below. Word classes that are unattested (Hengeveld & van Lier 2010) are marked with an asterisk and most of them are unnamed. One attested word class is also left unnamed. Verbs, nouns, adjectives, and manner adverbs are the only word classes in Table 3 that are 'rigid', which means that each has exactly one propositional function. The remaining eleven word classes are 'flexible', each having more than one propositional function.

Table 3. Word classes and the propositional functions they fulfill

Word class	P	R	r	p	Weight
Verbs	V	–	–	–	1
Nouns	–	N	–	–	1
Adjectives	–	–	a	–	1
Manner adverbs	–	–	–	m	1
Heads	H	H	–	–	2
Predicatives	~~₱~~	–	–	~~₱~~	2
Nominals	–	~~₦~~	~~₦~~	–	2
Modifiers	–	–	M	M	2
*	X_1	–	X_1	–	3
*	–	X_2	–	X_2	3
Non-verbs	–	Λ	Λ	Λ	3
*Non-nouns	Z	–	Z	Z	3
	X_3	X_3	X_3	–	3
*	X_4	X_4	–	X_4	3
Contentives	C	C	C	C	4

Let the number of word classes in the system be *k*. Some relatively complicated weights for word classes are proposed in Vulanović & Ruff (2018). Here, we define those weights by referring to Table 1. Each word class can be represented as occupying a certain number of the four cells in Table 1. For instance, non-verbs Λ occupy the cells labeled R, r, and p. The weight of a word class is defined as the number of cells it occupies, except for X_1 and X_2, for which the weight is taken to be 3. The flexibility of these two word classes is penalized more because each can function both as a head and as a modifier and both as part of the predicate and referential phrases. This can be defined formally as the minimum number of cells that need to be traversed starting from a non-empty cell and going around (either in the clockwise or the counterclockwise direction) until the last non-empty cell is

reached. For the word classes other than X_1 and X_2, this count corresponds to the number of cells they occupy.

If $(x,y)\in\Phi\setminus B$, the weight assigned to this pair is defined as the product of the weights for x and y. In this way, all components of formula (3) have been defined.

We next evaluate μ for each theoretically possible basic PoS-system type (Hengeveld & van Lier 2010). A PoS system is of a basic type if each propositional function in the system is fulfilled by exactly one word class. Therefore, $k\leq l$ (the number of word classes in the system does not exceed the number of propositional functions). There are five basic PoS-system types which are rigid (they only use rigid word classes), VNam, VNaØ, VNØm, VNØØ, and VØØØ. In this notation, a PoS system is represented by the sequence of its word classes, listed in the order which indicates what word class functions as P, then as R, followed by r, and finally p. The symbol Ø means that the corresponding propositional function does not exist in the system so that there is no word class with that function. In every rigid PoS system, $k = l$ and Φ is a bijection, which implies that $\mu = 1$. The remaining basic PoS-system types are flexible (they use at least one flexible word class) and $k<l$. The relation Φ is not a bijection in any flexible PoS system, so $\mu > 1$. All flexible basic PoS-system types, whether attested or not, are presented in Table 4 with their values of μ.

Table 4. The flexible basic PoS-system types

ℓ	k	**PoS-system type**	μ
4	3	VNMM	2.000
		V~~NN~~m	2.500
		~~P~~NaP	2.750
		VX_2aX_2	2.800
		X_1NX_1m	3.100
		HHam	3.250
	2	VΛΛΛ	4.000
		~~PNNP~~, HHMM, $X_1X_2X_1X_2$	4.250
		ZNZZ	4.375
		$X_4X_4aX_4X_3X_3X_3m$	5.125
	1	CCCC	7.500
3	2	V~~NN~~Ø	3.000
		$VX_2ØX_2$	3.250
		~~P~~NØ~~P~~	3.333
		$X_1NX_1Ø$	3.625
		HHaØ, HHØm	4.000
	1	$X_3X_3X_3Ø$, $X_4X_4ØX_4$	6.500
2	1	HHØØ	5.500

The basic PoS-system types represent an abstraction that is suitable for classification purposes (Hengeveld et al. 2004). In reality, many languages have word classes with overlapping propositional functions (Hengeveld & van Lier 2008). For instance, according to the same source, Turkish has all four propositional functions and they are fulfilled by three word classes, V, M, and Λ. Therefore, in the Turkish PoS system,

$$\|X\| = 6.5 \text{ (because } \ell = 4 \text{ in (4))}, \|Y\| = 1+2+3 = 6,$$
$$\Phi = \{(\text{P, V}), (\text{R, }\Lambda), (\text{r, }\Lambda), (\text{p, }\Lambda), (\text{r, M}), (\text{p, M})\}, B = \{(\text{P, V})\},$$
$$\|\Phi\backslash B\| = 2\cdot3+1\cdot3+1\cdot3+1\cdot2+1\cdot2 = 16,$$

so that finally, $\mu = \frac{16}{6} + 1 = \frac{11}{3}$. The Turkish PoS system is considered further in the next two sections.

4. The previous grammar-efficiency formula

Most generally, the grammar is more efficient if it has fewer rules and the sentences it permits convey more information. We define absolute grammar efficiency, *AE*, first. For PoS systems, the definition is

$$AE = Q\frac{|X|}{|Y|} = Q\frac{\ell}{k}, \tag{5}$$

cf. Vulanović (2009) for instance, or more generally, Vulanović (2003, 2007). Here, *Q* is a coefficient of proportionality which decreases with the increase in the complexity of the parsing process that identifies what word classes have what propositional functions in a sentence. This is why, in the previous approach to grammar complexity, *Q* is called the 'parsing ratio'. The parsing ratio is denoted here by Q_o, where the subscript *o* stands for "old" and indicates quantities used in the previous approach to grammar efficiency. In this sense, AE_o denotes the absolute grammar efficiency in (5) with $Q = Q_o$. The parsing ratio is defined as

$$Q_o = \frac{s}{a}, \tag{6}$$

where *s* is the number of all unambiguous sentences permitted in the PoS system and *a* is the number of all parsing attempts of all permutations of each sentence in the PoS system. This is exemplified below in the Turkish PoS system.

Sentences are formally represented as strings of word-class symbols. In a PoS system with all four propositional functions, like that of Turkish, a sentence must provide information about the heads P and R, whereas the head modifiers p and r are optional. It is assumed for simplicity that p and r, when present in the sentence,

stand next to their heads. Let us only consider the basic word order in Turkish, which corresponds to the following possible orders of propositional functions: RP, rRP, RpP, and rRpP. Then, there are seven unambiguous sentences in the Turkish PoS system, thus $s = 7$. The sentences are:

$$\Lambda V,\ M\Lambda V,\ \Lambda MV,\ \Lambda\Lambda\Lambda V,\ M\Lambda\Lambda V,\ \Lambda\Lambda MV,\ M\Lambda MV. \tag{7}$$

The count s is relatively easy to find but calculating a is much more complicated. Ambiguous sentences have to be considered as well. In the Turkish PoS system, there is one such sentence, ΛΛV. Let us use this sentence to illustrate how parsing is done. Any sentence is parsed from left to right, one word at a time. The length of the sentence is not known in advance, nor are the possible orders of propositional functions. Not only the successfully completed parses need to be counted, but also those that are started and not finished successfully. Thus, ΛΛV can be interpreted as RrP, RpP, or rRP. The permitted orders of the propositional functions are considered at this stage (this is similar to the regulated rewriting of Dassow & Păun (1989)). The order RrP is eliminated and this leaves two possible interpretations, RpP and rRP, which is why the sentence is ambiguous. There is also one attempted parse when the first Λ is analyzed as p, but this parse has to be abandoned because the second Λ cannot be analyzed as P. Therefore, the sentence ΛΛV contributes the count of 4 to the total in a. By counting all parsing attempts, successfully completed or not, it is indirectly measured how far the relation Φ is from a bijection. If Φ is further away from a bijection, there are more parsing attempts, and a becomes greater, which makes Q_o and AE_o less.

Moreover, the other two permutations, ΛVΛ and VΛΛ, of the sentence ΛΛV have to be analyzed in the same way, as well as all permutations of each sentence in (7). It turns out that $a = 100$ after parsing 32 sentences. The sentence permutations are considered to measure how free the word order is in the grammar. If the word order is more restricted, the grammar has more rules and its efficiency should be less. This is achieved through Q_o. If $s = a$, the word order is free and $Q_o = 1$. Otherwise, $s<a$ and Q_o becomes less if fewer word orders are permitted, that is, if s is less.

In conclusion, for the Turkish PoS system, we have

$$AE_o = \frac{7}{100} \cdot \frac{4}{3} = \frac{7}{75} = 0.0933.$$

Relative grammar efficiency, RE, is defined within the class of grammars that all have the same value of ℓ for $|X|$ and the same value of k for $|Y|$. Let $\Gamma(\ell,k)$ denote such a class of grammars. A maximally efficient grammar in $\Gamma(\ell,k)$ has the greatest value of Q and has to satisfy certain properties. For instance, it should not permit ambiguous sentences (for other requirements, see Vulanović (2003)). Let

the greatest value of Q in $\Gamma(\ell,k)$ be denoted by Q^*. When the maximally efficient grammar exists, we define

$$RE = \frac{Q}{Q^*}. \tag{8}$$

Therefore, $RE \leq 1$ in general and $RE = 1$ only for maximally efficient grammars. If the maximally efficient grammar does not exist, we set $RE = AE < 1$.

It should be clear that finding the value of Q^*_o is a considerably involved process. The parsing ratio Q_o, as defined in (6), should be calculated for each grammar in $\Gamma(\ell,k)$. This has been done in Vulanović (2008) for all values of ℓ and k. The class $\Gamma(4,3)$ has $Q^*_o = \frac{5}{8}$. This is needed in order to find the relative grammar efficiency of the Turkish PoS system,

$$RE_o = \frac{Q_o}{Q^*_o} = \frac{7}{100} \div \frac{5}{8} = \frac{14}{125} = 0.112.$$

5. The new grammar-efficiency formula

As we have seen in the previous section, the count a within the parsing ratio Q_o is obtained through parsing, but one of its roles, to indirectly measure how far the relation Φ is from a bijection, can be simply taken over by μ. This eliminates the need for counting all parsing attempts and simplifies the evaluation of the coefficient Q in (5). The new formula for calculating Q is proposed below (the subscript n stands for 'new' and indicates the quantities which are newly defined in this paper):

$$Q = Q_n := \frac{s}{m} \cdot \frac{1}{\mu}, \tag{9}$$

where the quotient s/m is supposed to measure how free the word order is in the sentences permitted in the PoS system. When the word order is free, s/m should equal 1. Otherwise, s/m should be less than 1 and, with this, the efficiency of the grammar becomes smaller.

As in the previous section, s is the number of all permitted unambiguous sentences. The quantity m depends on the number $\hat{s}$ of all possible sentences, whether they are ambiguous or not. However, m cannot simply be equal to $\hat{s}$, as illustrated by the following example.

Consider the simple PoS system HHØØ. There is only one sentence, HH, in it, thus $\hat{s} = 1$. This sentence is ambiguous unless the order of propositional functions is restricted to either PR or RP. With such a restriction, $s = 1$, so $s/\hat{s} = 1$, which means that the word order is as free as possible. However, the value $s/\hat{s} = 1$ does not indicate that there is a rule which imposes a fixed order of the propositional

functions. This can be taken care of by setting $m = 2$, which corresponds to the two possible orders of the propositional function, PR and RP. Then, $s/m = 1/2$, which, since the value is less than 1, shows that some ordering restrictions exist. In this example, the ordering restrictions are not for the word classes (their possible orders are represented within the $\hat{s}$ count), but for the propositional functions. This motivates the definition of m as

$$m = \max\{\hat{s}, f(\ell)\},$$

where *f(ell)* stands for the maximum possible number of orders of the ℓ propositional functions in the PoS system. The above discussion shows that $f(2) = 2$. It is easy to see that $f(1) = 1$, $f(3) = 6$, and finally $f(4) = 18$. For instance, when $\ell = 4$, the orders are PR, RP, PpR, pPR, RPp, RpP, PRr, PrR, RrP, rRP, PpRr, PprR, pPRr, pPrR, RrPp, RrpP, rRPp, and rRpP (recall that the head modifiers have to stand next to their corresponding heads).

When calculating Q_n, we still need to get the count $\hat{s}$, but the sentences do not have to be parsed. This is why Q_n is easier to find than Q_o. As mentioned in the previous section, we have $\hat{s} = 32$ in the Turkish PoS system, and then, by the formula (9), we get

$$Q_n = \frac{7}{32} \cdot \frac{3}{11}$$

and

$$AE_n = \frac{7}{32} \cdot \frac{3}{11} \cdot \frac{4}{3} = \frac{7}{88} = 0.0795.$$

Compare this value to the previous $AE_o = 0.0933$.

It is also easier to find Q^*_n than Q^*_o. In the $\Gamma(4,3)$ class of grammars, it turns out that the most efficient one is the grammar for the VNMM PoS system with $s = \hat{s} = 16$ possible sentences. For this system,

$$Q_n = Q^*_n = \frac{16}{18} \cdot \frac{1}{2} = \frac{4}{9}.$$

Therefore, for the Turkish PoS system, we find that

$$RE_n = \frac{Q_n}{Q^*_n} = \frac{7}{32} \cdot \frac{3}{11} \cdot \frac{9}{4} = 0.134,$$

which can be compared to the previously calculated $RE_o = 0.112$.

6. The efficiency of basic parts-of-speech system types

In this section, we consider all basic PoS systems, the five rigid ones, and the flexible ones in Table 4. We calculate RE_o and RE_n for all of them and analyze the correlation between the values obtained by the two formulas. The value of *RE* depends on the word order permitted in the PoS system. Since this is an analysis of the general basic PoS system types, no particular word order can be assumed. We therefore consider the maximum values of *RE* (those of RE_o can be found in Vulanović (2008)). They are presented in Table 5 for the basic PoS systems that are attested according to Hengeveld & van Lier (2010).

Table 5. The maximum relative-grammar-efficiency values for the attested basic PoS-system types

PoS-system type	RE_o	RE_n
VNMM	0.914	1
V~~NN~~m	0.800	0.600
VΛΛΛ	0.728	0.797
~~PNNP~~	0.786	0.667
CCCC	0.286	0.015
V~~NN~~Ø	1	0.867
$X_3X_3X_3$Ø	1	1
HHØØ	1	1
5 rigid types	1	1

The correlation between the RE_o and RE_n values in Table 5 is very strong, with the coefficient of correlation $r = 0.960$. When all basic PoS systems, whether attested or not, are taken into account, the correlation is somewhat weaker, $r = 0.807$.

7. Conclusion

The new formula for evaluating grammar efficiency is much easier to use than the former formula from Vulanović (2003, 2007, 2008). The simplification is enabled by the inclusion of a new version of the measure of how much a linguistic system departs from the One-Meaning–One-Form Principle (Vulanović & Ruff 2018). In the new formula, the number of calculations is significantly reduced. This is particularly important for the more complicated flexible PoS systems, such as the Turkish PoS system and other systems described in Hengeveld & van Lier (2008), because now their grammar efficiency can be calculated more easily. At the same

time, the grammar-efficiency values calculated by the new formula correlate well with those obtained by the old formula. This means that the new formula can replicate most of the results and conclusions reported in papers like Vulanović (2009), where the old approach is used.

References

Anttila, Raimo. 1972. *An introduction to historical and comparative linguistics*. New York: Macmillan.

Dassow, Jürgen & Gheorghe Păun. 1989. *Regulated rewriting in formal language theory*. New York: Springer. https://doi.org/10.1007/978-3-642-74932-2

Hengeveld, Kees. 1992. Parts of speech. In Michael Fortescue, Peter Harder & Lars Kristoffersen (eds.), *Layered structure and reference in functional perspective*, 29–55. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.23.04hen

Hengeveld, Kees & Eva van Lier. 2008. Parts of speech and dependent clauses in Functional Discourse Grammar. *Studies in Language* 32. 753–785. https://doi.org/10.1075/sl.32.3.13hen

Hengeveld, Kees & Eva van Lier. 2010. An implicational map of parts of speech. *Linguistic Discovery* 8. 129–156. https://doi.org/10.1349/PS1.1537-0852.A.348

Hengeveld, Kees, Jan Rijkhoff & Anna Siewierska. 2004. Parts-of-speech systems and word order. *Journal of Linguistics* 40. 527–570. https://doi.org/10.1017/S0022226704002762

Hengeveld, Kees & Sterre Leufkens. 2018. Transparent and non-transparent languages. *Folia Linguistica* 52. 139–175. https://doi.org/10.1515/flin-2018-0003

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 23–41. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.94.04mie

Vulanović, Relja. 2003. Grammar efficiency and complexity. *Grammars* 6. 127–144. https://doi.org/10.1023/A:1026189411761

Vulanović, Relja. 2007. On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics* 20. 399–427.

Vulanović, Relja. 2008. A mathematical analysis of parts-of-speech systems. *Glottometrics* 17. 51–65.

Vulanović, Relja. 2009. Efficiency of flexible parts-of-speech systems. In Reinhard Köhler (ed.), *Issues in quantitative linguistics* (*Studies in quantitative linguistics* 5), 136–157. Lüdenscheid: RAM.

Vulanović, Relja & Oliver Ruff. 2018. Measuring the degree of violation of the One-Meaning–One-Form Principle. In Lu Wang, Reinhard Köhler & Arjuna Tuzzi (eds.), *Structure, function and process in texts*, 67–77. Lüdenscheid: RAM.

Distribution and characteristics of commonly used words across different texts in Japanese

Makoto Yamazaki
National Institute for Japanese Language and Linguistics

In this chapter, I survey the frequency distribution of commonly used words across different texts in Japanese. Using the Balanced Corpus of Contemporary Written Japanese, we examined the distribution. The results show the following. (1) The distribution draws a curve similar to Zipf's law, but the curve always begins to increase shortly before the degree of commonality reaches its maximum, (2) neither the length nor the number of the texts affects the distribution trend, (3) as the text length increases, the number of commonly used words also increases linearly, but it reaches a maximum point due to the limited number of basic words.

Keywords: distribution of commonly used words, Japanese, Zipf's law, function words, lexical balance

1. The law of distribution of words

Zipf's law (Zipf 1949) is a well-known law about the frequency distribution of words, and numerous related studies have been conducted on this law. Zipf's law is the result of lexical balance in texts (Zipf 1949: 22). Lexical balance refers to the universal nature of texts that are composed of a small number of high-frequency function words and a large number of low-frequency content words.

Zipf's law is undoubtedly not necessarily the only frequency distribution of words. As it applies to single texts, different perspectives can be applied to examine multiple texts. Suppose, for instance, there are words used in all N texts; these words can be divided into N groups from words used in only one text to words used in all N texts. Let us then calculate the frequency distribution of these N groups. Given the lexical balance of the abovementioned instance, the frequency distribution would be expected to show that the number of words used in all texts is small,

https://doi.org/10.1075/cilt.356.08yam

while the number of words used in multiple texts is large. Moreover, the number of words used in only one text is the largest. This study thus confirms the actual statistical distribution empirically. The specific research questions are as follows.

RQ1. What is the frequency distribution of words used commonly in different texts?
RQ2. Does the length of a text affect the distribution?
RQ3. Does the number of texts affect the distribution?

2. Previous studies

Several vocabulary studies have reported the distribution of words used in multiple texts. The National Institute for Japanese Language and Linguistics (1952) published the results of a survey conducted on all the newspapers printed in a month; it included a description of the number of days and times each word was used and tallied the number of verbs, adjectives, and non-conjugated words. According to the survey's graph that indicated the results, it was noticed that the number of the words decreased with an increase in the number of days for which they were used and began to rise toward the end (p. 95). The National Institute for Japanese Language and Linguistics (1983) also published the results of a vocabulary survey conducted on nine high school textbooks (four and five textbooks from social studies and science, respectively); it included a table indicating the number of the same words used in different textbooks and the number of textbooks in which those words were used. In addition to the total number of all words, the survey results also showcased the distribution of Japanese, Chinese-derived, and foreign-language derived loan words. This finding revealed that a significant difference prevailed between Japanese and Chinese-derived words in the distribution of the number of the same words used in different textbooks. The Japanese words used in all nine textbooks accounted for about 77% of all the words, and their percentage decreased remarkably thereafter. Conversely, the Chinese-derived words used in all nine textbooks accounted for about 20%, and their percentage fluctuated between 7% and 15% thereafter, with no significant decrease or increase.

Although these previous studies provide some statistical data, as they are all the results of a single survey, it is questionable whether the same results will be obtained if the surveys are repeated. However, there are considerable differences between the textbooks used in the National Institute for Japanese Language and Linguistics' (1983) survey in terms of the number of words they contain, ranging from 43,000 words in the earth science textbook, which is the smallest word count, to 93,000 words in the Japanese history textbook, the largest word count.

3. Data and method

The Balanced Corpus of Contemporary Written Japanese (BCCWJ) was employed for this study. This corpus contains about 100 million words of modern written Japanese. Completed in 2011, it encompasses 13 different genres, including books, magazines, newspapers, and blogs (Figure 1).

Publication Subcorpus	**Library Subcorpus**
35 million words	30 million words
Books, Magazines, Newspapers	Books
2001–2005	1986–2005
Special Purpose Subcorpus	
35 million words	
White papers, School textbooks, Publicity newsletters of local government, Best-selling books, Bulletin board, Blog, Poetry verses, Law, Minutes of the National Diet	
Various years between 1971 and 2008	

Figure 1. Structure of BCCWJ

BCCWJ comprised samples of 20,668 books (Maekawa et al. 2014). Moreover, the samples from these books were employed as data for this study because their quantity was deemed sufficient for its purpose. Instead of using all the book samples, *N* necessary words were extracted from the beginning of texts and used for analysis. Also, word counts did not include auxiliary symbols, such as punctuation marks, or spaces used to indicate indentation.

The degree of commonality was determined by the following procedure. First, *N* words were extracted from the beginning of multiple texts randomly selected from each genre of BCCWJ books. For the selection of texts, sampling without replacement was conducted to prevent the texts from being selected more than once. This process was followed by counting of the number of the same words used in the selected texts. For example, there are three sets of data consisting of four words (Figure 2, Table 1); in this case, as Word A is used in all the texts, the degree of commonality of Word A is 3. In the meantime, Words B and C are used in two texts, so their degrees of commonality are 2. As Words D, E, F, and G are used only

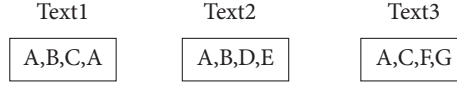


Figure 2. Sample texts

in one text, their degrees of commonality are 1; further, the frequency distribution was examined on the basis of these degrees of commonality.

Note that the *N* words extracted from the beginning of texts do not necessarily refer to the *N* words of the texts selected from the beginning of the books. This is because a certain amount of text was extracted based on the reference points arbitrarily selected from the books to be recorded in BCCWJ.

Table 1. Degree of commonality of words of the sample texts

Word	Text1	Text2	Text3	Degree of commonality
A	✓	✓	✓	3
B	✓	✓		2
C	✓		✓	2
D		✓		1
E		✓		1
F			✓	1
G			✓	1

Let us look at an actual example. Table 2 shows the degree of commonality for ten texts consisting of 100 words. There are 343 words (number of types) whose degree of commonality is 1, i.e., they are used only in one text, while 29 words are used in two texts. Table 2 indicates that the distribution is as previously expected.

Table 2. Degree of commonality of 10 texts containing 100 words

Degree of commonalty	No. of words(type)
1	343
2	29
3	14
4	9
5	3
6	2
7	2
8	4
9	4
10	2

However, as providing a single example was deemed insufficient, the extraction of texts was repeated 100 times to calculate the average. The results are shown in Table 3, indicating almost the same distribution as in Table 2.

In addition, ten types of text were extracted for the same ten texts, by changing their lengths from 100 to 1000 words in increments of 100 words. Furthermore, ten types of text were extracted by changing the number of texts to be extracted

Table 3. Degree of commonality of ten texts containing 100 words (average)

Degree of commonalty	No. of words(type)
1	370.05
2	34.86
3	11.84
4	5.75
5	3.41
6	2.12
7	2.07
8	2.33
9	3.32
10	4.38

from 10 to 100 in increments of 10. In other words, a total of 100 types of frequency distributions were obtained from the data of ten texts with 100 words to the data of 100 texts containing 1000 words.

4. Results

Figure 3 shows the distribution of the degree of commonality of words used in ten texts containing 100 words. As the degree of commonality indicated by the horizontal axis is not a constant variable, and thus it is not an appropriate method of display, a line graph is used for visibility. Figure 3 exhibits the distribution of the number of

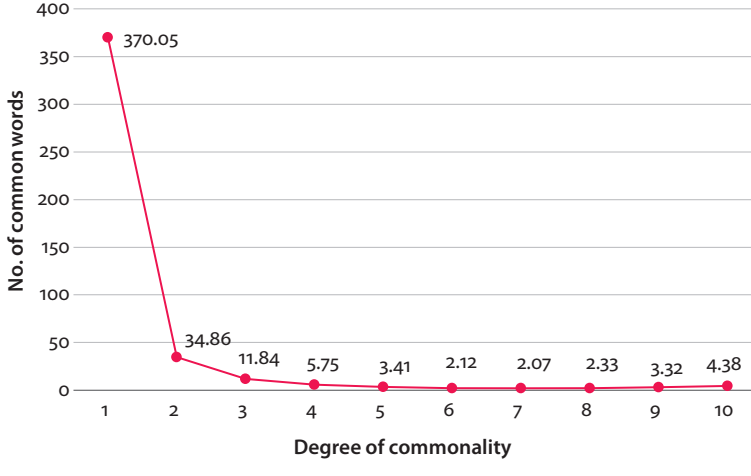


Figure 3. The average number of commonly used words (type)
Text length: 100, Number of texts: 10

types. The number of words that are used in only one text is overwhelmingly large, and the number of words used in two texts is less than one tenth of the previous one. The number of words slowly decreased thereafter and then remained at the same level. The shape of the distribution resembles Zipf's curve. One thing to be noted here is that the lowest point of the curve is not the curve's endpoint. It is at the 7th position of the degree of commonality.

Figure 4 is a boxplot that represents the data of ten texts containing 100 words. Tukey's HDS test performed on adjacent degrees of commonality in Figure 4 revealed a significant difference at the 1 percent level in each pair of 1 and 2, 2 and 3, and 3 and 4. In other words, the decrease in the number of words with degrees of commonality between 1 and 4 was statistically significant.

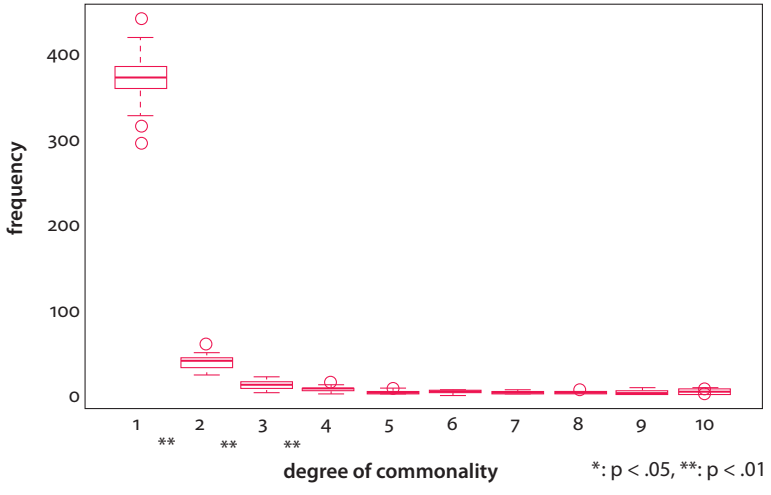


Figure 4. Boxplot of commonly used words (type)
Text length: 100, Number of texts: 10

Figure 5 shows the distribution of the degrees of commonality of words used in ten texts containing 200 words. The distribution also resembles Zipf's curve in this case. Similarly, Figure 6 is a boxplot, representing the distribution. In addition to the significant difference at the 1 percent level found between adjacent degrees of commonality between 1 and 4, Figure 5 indicates a significant difference at the 5 percent level between 9 and 10 degrees of commonality.

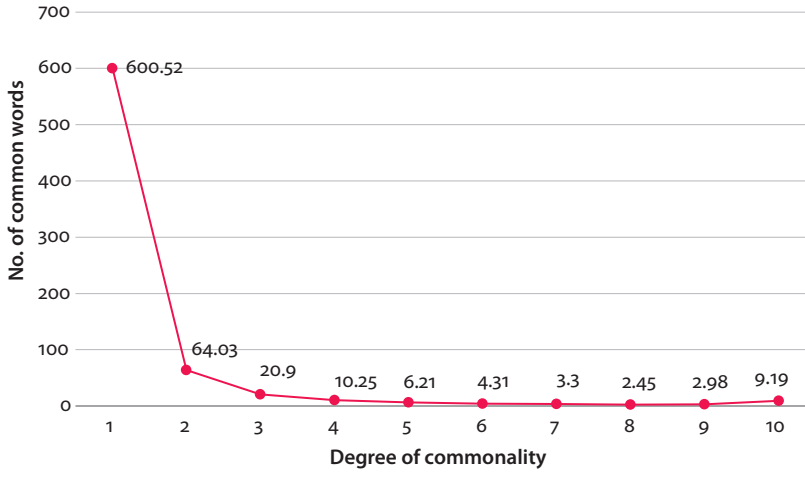


Figure 5. The average number of commonly used words (type)
Text length: 200, Number of texts: 10

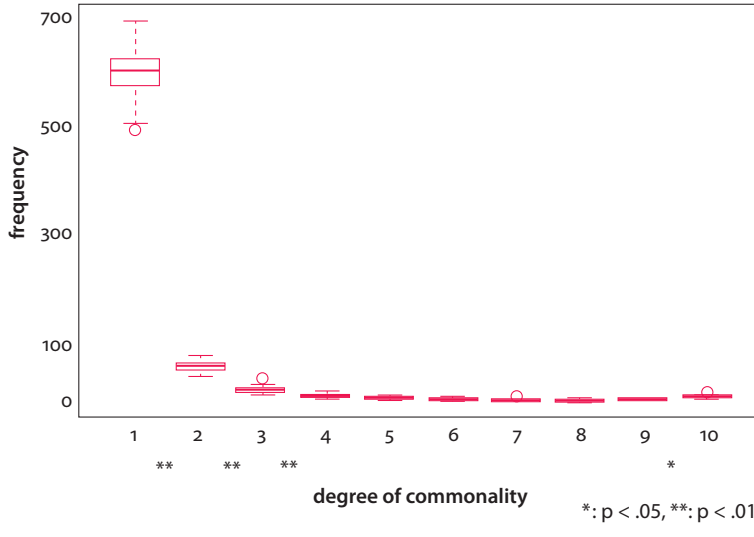


Figure 6. Boxplot of commonly used words (type)
Text length: 200, Number of texts: 10

A significant difference at the 5 percent level between 9 and 10 degrees of commonality was also observed in the case of texts with 300 words (Figure 7). However, no significant difference was observed in the case of texts containing 400 words, as indicated in Figure 8, and no significant difference was observed up to a length of 1000 words.

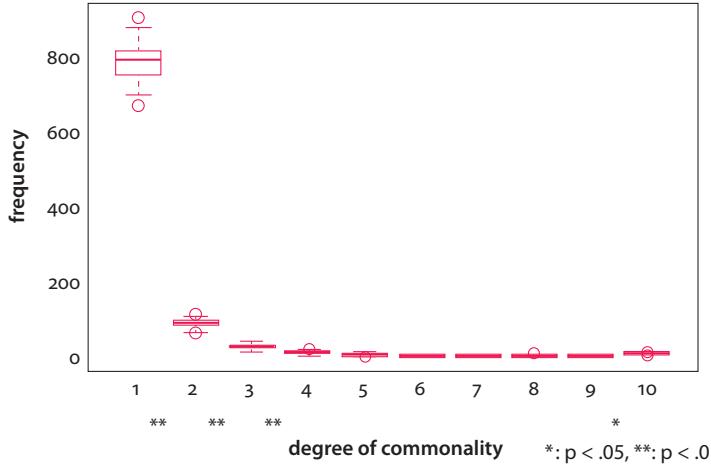


Figure 7. Boxplot of commonly used words (type)
Text length: 300, Number of texts: 10

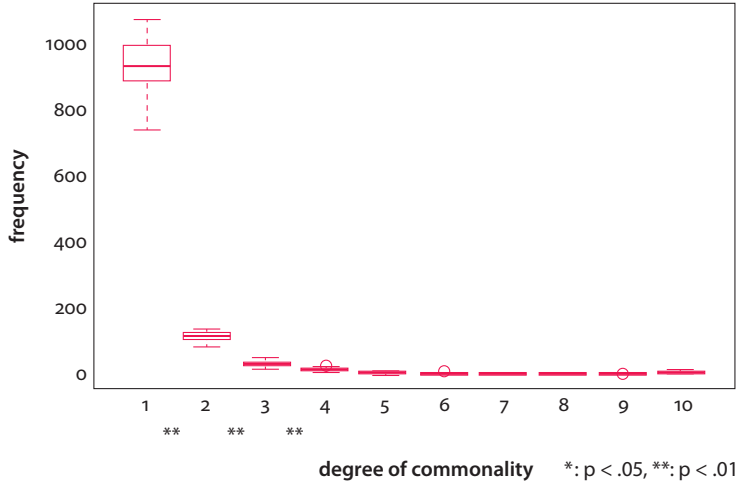


Figure 8. Boxplot of commonly used words (type)
Text length: 400, Number of texts: 10

In terms of distribution, the answer to RQ1 would be 'the distribution is similar to Zipf's curve'. Though it seems insignificant, there is a factor that should not be overlooked. In every distribution, the number of commonly used words slightly increases at the end of the graph. Let us look at this situation from another perspective. Table 4 shows the average value of commonly used words in ten texts. The length of the texts ranges from 100 to 1000 words. The shaded cells indicate the lowest values in each column. Their positions are between the cells of 7 and 9 degrees of commonality and are never in the cells of 10. The cells of 10 degrees of commonality are those with the maximum degree of commonality.

Table 4. The average number of commonly used words: 10 texts

Degree of commonality	100	200	300	400	500	600	700	800	900	1000
1	370.05	600.52	789.34	941.83	1094.71	1230.59	1356.68	1462.37	1587.97	1677.67
2	34.86	64.03	91.63	119.22	146.2	168.02	193.54	214.81	241.87	260.97
3	11.84	20.9	29.37	37.73	47.54	56.24	65.92	73.13	81.6	91.48
4	5.75	10.25	14.79	18.91	23.17	26.44	30.48	35.94	40.48	43.8
5	3.41	6.21	8.96	10.95	13.96	15.98	18.54	20.63	22.87	25.34
6	2.12	4.31	5.62	7.94	9.3	10.42	12.2	14.15	14.99	16.58
7	2.07	3.3	4.33	5.6	6.16	7.76	9.25	9.83	10.82	12.26
8	2.33	2.45	4.09	4.72	5.74	5.98	6.83	7.42	8.69	9.58
9	3.32	2.98	3.86	4.74	5.41	6.4	6.46	7.31	8.58	8.86
10	4.38	9.19	11.24	13.32	15.19	16.81	18.98	19.71	20.86	22.24

Table 5 shows the average number of commonly used words in 20 texts. As in Table 4, the cells with the lowest values in the column are shaded. Their positions are between the cells of 14 and 17 degrees of commonality and are never in the last cells (cells with the maximum degree of commonality). The table also indicates that the average value tends to increase after the cell of the lowest degree of commonality.

Table 5. The average number of commonly used words: 20 texts

Degree of commonality	100	200	300	400	500	600	700	800	900	1000
1	644.8	1006.8	1306.6	1551.3	1786.5	1978.4	2150.3	2321.0	2488.3	2623.2
2	69.5	129.6	184.5	233.5	283.5	330.2	373.9	413.4	454.1	490.3
3	23.0	41.1	61.9	82.2	100.9	118.9	135.6	154.4	170.2	185.2
4	11.6	21.4	29.7	37.9	48.7	57.1	66.1	73.8	84.1	92.5
5	6.7	13.1	17.1	22.7	28.3	33.9	39.6	44.1	51.0	54.1
6	5.2	8.6	11.5	15.1	17.5	21.2	25.1	28.5	33.3	35.8
7	3.5	6.4	8.6	10.8	13.0	14.8	17.6	20.3	22.7	24.8
8	2.6	4.6	6.9	8.5	10.3	11.5	12.4	14.7	16.3	18.4
9	2.0	3.3	5.2	6.2	7.0	8.7	10.0	11.3	12.9	13.8
10	1.5	2.6	3.7	5.7	6.2	7.1	8.1	9.4	9.9	11.3
11	1.3	2.3	2.8	4.1	5.6	6.3	6.6	7.0	8.1	9.3
12	0.9	1.9	2.5	3.1	4.2	5.0	6.1	6.2	7.1	7.3
13	0.7	1.9	2.0	2.7	3.5	4.4	5.0	5.8	6.3	6.2
14	0.7	1.4	2.1	2.5	3.0	3.6	4.3	5.0	5.2	6.0
15	0.8	1.4	2.1	2.5	2.5	3.1	3.3	4.0	4.6	5.3
16	1.1	1.3	1.9	2.3	2.4	2.8	3.0	3.7	3.9	4.4
17	1.5	1.2	1.7	2.0	2.7	3.0	3.0	3.1	3.6	3.8
18	2.0	1.5	1.8	2.4	2.7	3.0	3.8	3.7	3.8	3.9
19	2.4	2.3	2.5	2.5	3.3	4.1	4.0	4.7	5.0	5.4
20	2.7	7.4	9.6	11.2	12.6	13.3	14.6	16.2	16.9	18.1

Table 6 shows the average number of commonly used words in 30 texts. The degrees of commonality higher than 21 are displayed for visibility. The shaded cells are positioned between the cells of 22 and 27 degrees of commonality and are never in the last cells.

Table 6. The average number of commonly used words: 30 texts

Degree of commonality	100	200	300	400	500	600	700	800	900	1000
21	0.45	0.85	1.34	1.48	1.77	2.17	3.14	3.63	3.67	3.63
22	0.39	0.89	1.21	1.41	1.71	2.05	2.27	2.96	3.07	3.68
23	0.48	0.85	1.29	1.72	1.32	1.61	2.12	2.53	3.32	3.22
24	0.61	0.70	1.36	1.64	1.67	1.57	2.16	2.00	2.44	2.81
25	0.98	0.89	1.12	1.44	1.81	1.99	1.68	2.01	2.41	2.85
26	1.30	0.84	1.20	1.74	1.90	1.82	1.97	2.10	2.15	2.46
27	1.26	0.78	1.18	1.34	2.05	2.54	2.47	2.42	2.71	2.61
28	1.64	1.23	1.25	1.65	2.02	2.20	2.82	2.72	2.99	3.12
29	1.98	2.68	1.91	2.08	2.36	3.11	3.37	4.30	3.89	4.81
30	1.76	6.18	8.96	10.21	11.23	12.04	13.09	14.09	15.49	15.54

As indicated by Tables 4 and 5, the distributions of the texts with 100 to 1000 words were almost the same. It is safe to say that the length of the text does not affect the distribution. Therefore, the answer to RQ2 is 'No'.

This leads to an additional question: Does the number of words with the highest degree of commonality ever reach a point of saturation as the length of text increases? It is safe to assume words that appear in all the texts should be basic ones. Usually, as there seems to be an upper limit to the number of basic words, the words with the highest degree of commonality should stop increasing at some point. Figure 9 shows the case of 10, 20, 30, and 40 texts with the number of words increased to 8000, the number seems to stop at 7000. As it was not possible to extract the necessary amount of text with more than 8000 words from BCCWJ, it is not yet clear whether this saturation is stable. The tentative answer to the abovementioned question is that there is an upper limit at some point in the range examined. The saturation point is somewhere between 7000 and 8000 words in text length.

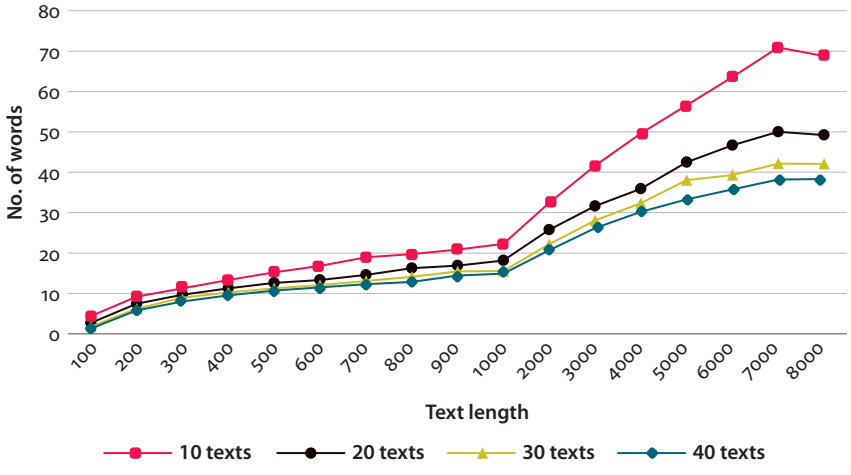


Figure 9. Number of words corresponding to the maximum commonality, Number of texts: 10 to 40

Let us now consider RQ3. Figures 10 to 12 are graphs showing the statistical data for 20, 30, and 100 texts with 100 words. As suggested by these figures, regardless of the number of texts, the shape of the distribution resembles Zipf's curve. Therefore, the answer to RQ3 is 'No'.

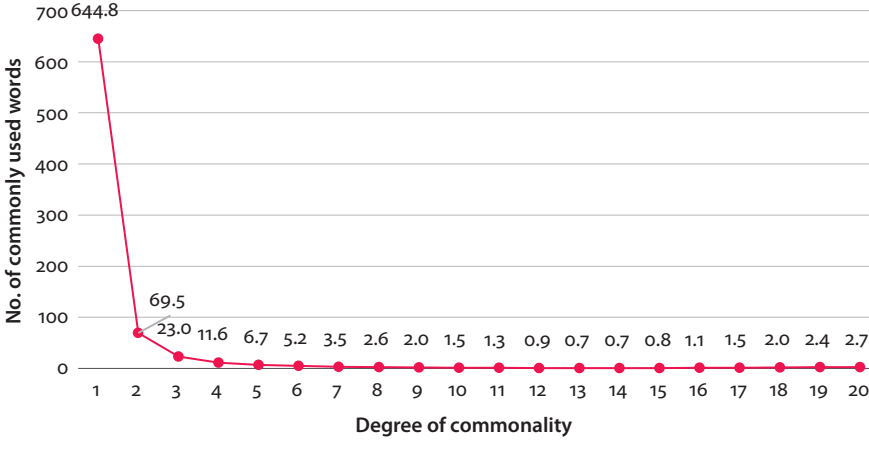


Figure 10. The average number of commonly used words
Text length: 100, Number of texts: 20

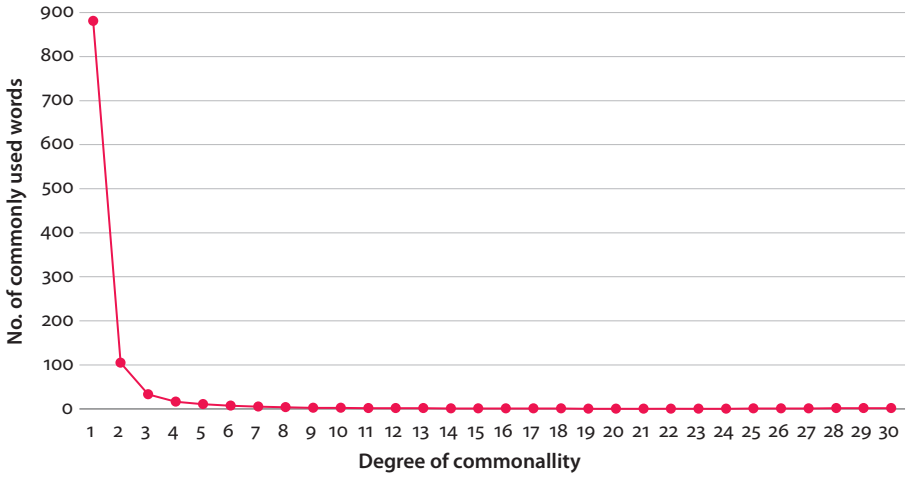


Figure 11. The average number of commonly used words
Text length: 100, Number of texts: 30

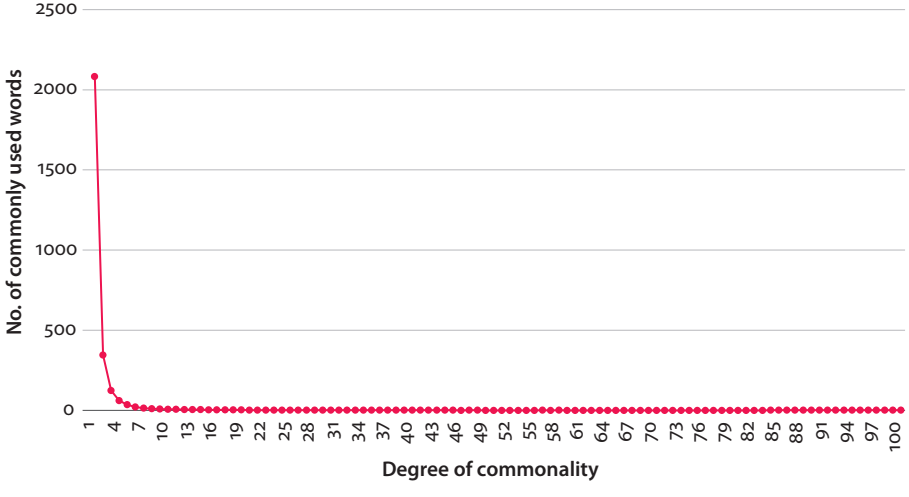


Figure 12. The average number of commonly used words
Text length: 100, Number of texts: 100

5. Interpretation

Why does the distribution of degrees of commonality of words resemble Zipf's curve? From a broad perspective, the functional properties of words seem to be closely related to the degree of the same words appearing in different texts. In other words, the stronger the functional properties of a word, the higher the possibility that the word appears in multiple texts. Table 7, which shows the ratio of each part of speech by the degree of commonality, explains the abovementioned assertion. Particles and auxiliary verbs, which are function words, increase in proportion as the degree of commonality rises, whereas nouns, which affect the topic and are less functional, decrease in proportion as the degree of commonality increases. Verbs seem to appear at a relatively constant rate, regardless of the degree of commonality, suggesting that some verbs have strong functional properties, while others do not.

Table 7. The ratio of parts of speech. Text length: 100; Number of texts: 10

Commonality POS	1	2	3	4	5	6	7	8	9	10
Noun	78.8	55.8	42.5	33.3	28.0	12.9	17.4	11.1	0.0	0.0
Verb	10.1	16.4	15.1	17.9	16.0	19.4	21.7	22.2	20.0	9.1
Adverb	2.4	4.0	3.8	1.3	0.0	0.0	0.0	0.0	0.0	0.0
Suffix	2.3	6.4	8.1	3.8	6.0	3.2	0.0	0.0	0.0	0.0
Adjective Verb	2.3	1.6	1.1	1.3	2.0	3.2	0.0	0.0	0.0	0.0
Adjective	1.3	2.9	1.6	2.6	2.0	3.2	0.0	0.0	0.0	0.0
Affix	0.6	1.1	1.1	2.6	2.0	0.0	0.0	0.0	0.0	0.0
Particle	0.5	4.4	12.4	17.9	20.0	32.3	43.5	55.6	66.7	72.7
Pronoun	0.3	1.8	3.8	5.1	6.0	3.2	0.0	0.0	0.0	0.0
Interjection	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Auxiliary	0.2	1.9	4.8	10.3	12.0	19.4	17.4	11.1	13.3	18.2
Adnominal	0.2	1.5	3.2	2.6	4.0	3.2	0.0	0.0	0.0	0.0
Conjunction	0.2	0.9	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Others	0.4	1.1	1.1	1.3	2.0	0.0	0.0	0.0	0.0	0.0
Total	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

6. Conclusion and further challenges

In this study, BCCWJ was employed to examine the distribution of words that appear in multiple texts, which had not been fully established until now. The results revealed the following.

1. The higher the degree of commonality, the smaller the number of words. The distribution draws a curve similar to Zipf's law, but the curve always begins to increase shortly before the degree of commonality reaches its maximum.
2. Neither the length nor the number of texts affects the distribution trend. Further, almost the same distribution was observed when the length of text ranged from 100 to 1000 words and the number of texts ranged from 10 to 100.

Future questions include whether this distribution trend can be observed in other languages, and how to represent mathematical models of the distribution curves. Different texts were used in this survey, and it will be interesting to study the distribution of words if these texts are divided into many parts and are treated as different texts.

Funding

This paper is one outcome of a project of the Center for Corpus Development, National Institute for Japanese Language and Linguistics. Texts included in the registers of Library Books within the BCCWJ were compiled by MEXT KAKENHI Grant Number: 18061007.

References

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka & Yasuharu Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation* 48. 345–371. https://doi.org/10.1007/s10579-013-9261-0

National Institute for Japanese Language and Linguistics. 1952. *A Research Newspaper Vocabulary*. Tokyo: Shuei Publishing Co.

National Institute for Japanese Language and Linguistics. 1983. *Studies on the Vocabulary of Senior High School Textbooks*, Volume 1. Tokyo: Shuei Publishing Co.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.

PART II

Empirical studies

The perils of big data

Sheila Embleton, Dorin Uritescu and Eric S. Wheeler
York University

The use of large amounts of data and the technologies to process them are characteristic of modern research. However, such practices come with risks of misleading the researcher. While there is much that could be said on this topic, here briefly is our cautionary tale to others, based on our direct experiences.

Keywords: big data, research practices, statistical packages, Romanian, dialects, Crișana, shibboleths

1. Motivation

A generation ago, a linguist with an interest in the quantitative study of language might have spent 'a quiet Sunday afternoon' cutting the paper copy of a text into words, and sorting and counting the bits of paper – not only dreary work, but a method that had real limitations as to how much could be processed this way and as to what might be found. With the arrival of readily available digital data sets, computers, and appropriate software, things changed radically: large volumes of text could be processed (counted, parsed, reformatted), using state-of-the-art statistical packages, to achieve a depth and range of coverage that was inconceivable only a decade or two before. The age of 'Big Data' had arrived. Importantly, it was possible to look for – and find – patterns in the data that were subtle, or unexpected, or simply unimagined, and to justify the results with substantial evidence.

Before we begin, let us make clear that we only discuss 'scientific' aspects of the use of big data. We leave aside questions of accuracy of data, and of how it is obtained and stored, as well as all the social, economic, and political questions of privacy, confidentiality, intellectual property, access, ownership, and who gets to profit commercially from the use of the stored data. These are of course important subjects, worthy of urgent debate, but that debate is for other venues and should include many players from many disciplines, primarily the social sciences. Additionally,

https://doi.org/10.1075/cilt.356.09emb

many people these days see 'Big Data' sometimes as a type of modern-day panacea, thinking the answer to almost any debate or even policy issue is somehow just to collect more data. Yet one still has to ask the right questions and know how to interpret the data. For example, in an excellent essay on the use of Big Data in higher education, Patricia McGuire (2019) speaks of the importance of bringing professional judgement to bear on the analysis, to find the meaning in the data, reminding us that "sometimes words are important to interpret statistics". She further points out that "Easy access to voluminous data allows just about anyone to extract random factoids as evidence to assail or affirm" for or against whatever point they are trying to make. These types of issue again are largely social science-oriented issues and we will not discuss them further here.

We (Embleton, Uritescu & Wheeler, hereafter EUW) have worked on dialect patterns (see the references in EUW 2004, 2007a, b, 2018 in part to give ready and flexible access to dialect data, in part to show how multidimensional scaling (MDS; see Embleton & Wheeler 1997a, b, Wheeler 2005, EUW 2008, 2011, 2013) can be useful for visualizing data that otherwise might be presented in clusters or trees, and in part to investigate the relationship between geography and language variation. We hope we have contributed both to methods in dialectometry as well as to results in the actual dialectology of the languages we have worked with (British English, Finnish, and dialects of northwest Romania so far, with others to come). We have made good use of the concept of Big Data to get results that would otherwise be invisible or inaccessible.

But, along the way, we have uncovered some problems of using Big Data, or at least, using it naively. This note, then, is a cautionary tale for the benefit of others going the same route.

2. Some background

Embleton & Wheeler began looking at multidimensional scaling as a way of visualizing linguistic differences (1997a English dialects; 1997b, 2000 Finnish dialects) that avoided some of the pitfalls of isoglosses and binary trees.

With Uritescu, we applied and further developed the techniques using a conservative dialect region of Romanian (the Crişana area in the northwest; data from Stan & Uritescu 1996, 2003 and subsequent work). The result was RODA, the Romanian Online Dialect Atlas (EUW 2002, 2007a). In particular, RODA provided the possibility of searching the data using an immensely wide range of patterns and making interpretive maps based on any such defined pattern. For example, one could make a dialect map based on only Latin-derived lexical items with syllabic or non-syllabic final vowels – a highly specialized search, but one that was useful in a subsequent discussion of Romanian phonology.

In RODA, it was possible to define the Romanian dialects in any of millions of possible ways, depending on what criteria (what 'shibboleths') were used. Big Data had allowed us to access a lot of data (the hard copy version of the atlas will be published in five volumes), and to see it in many, many ways.

Of course, the obvious thing to do was to measure the linguistic differences (location to location) over all the data sets, and then use MDS to visualize the result (using the built-in features of RODA, EUW 2009). The result showed that areas on the geographic edges (north and south) of the field-work area were clearly distinguished from the rest, which were more tightly connected, and this was consistent with Uritescu's intuition during the fieldwork about some areas and with the results of his research on the dialectal structure of northwestern Romanian (cf. Uritescu 1983, 1984a, 1984b, 1986). We have also extended the techniques to some Chinese data, and are working on Mambila languages in Nigeria and Cameroon. So, Big Data was working for us.

3. The muddle in the middle

The problem with the Romanian result was not that it was wrong, but that it did not seem to say all that could be said. As a further study, we divided the data into traditional linguistic categories (phonetic, morpho-phonemic, morphological, lexical) and applied the technique again. The results again gave credible pictures of the dialect situation, but they were not all the same picture: dialects based on one category were different from those based on another, and each of the pictures tended to separate the dialect areas more distinctly than when using all the data. Compare Illustrations 1 and 2.

Clearly, averaging all the data sets together was blurring the result. Big Data had allowed us to look at lots of data, but lots of data (used naively) was not as useful as smaller data sets, selected insightfully.

It is not just a matter of quantity; it is possible that some features are more important in the perception of dialects than others (to the typical person, perhaps vowel differences are not as noticeable as consonant differences) and the use of the unimportant features is again hiding distinctions by smoothing out the numbers.

Big Data does not substitute for meaningful analysis. Big Data may make it easier to find the evidence to support an analysis (or not support it), but deciding what is important does not come from the data; that is a decision of the analyst, and in general requires the researcher's insight and experience.

Some might argue for automatic discovery procedures, and the results that they can produce when searching large data sets. But such results are always a consequence of the data and methods applied, and the assumptions made; different methods or data can generate different results, and the choice of which is 'right' is not determined by the tools.

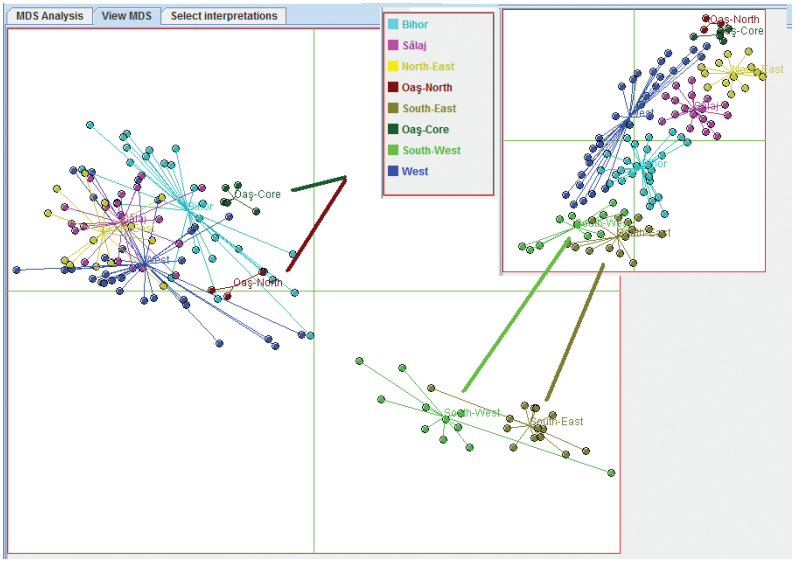


Illustration 1. An MDS picture of Crișana, in north-west Romania, using 'all' the available dialect data. The inset shows the geographic regions; lines connect some of the geographic regions to the corresponding dialect areas. Note that the south-west and south-east are clearly separate linguistically from the rest of the region. Oaș, a region in the north, is not.

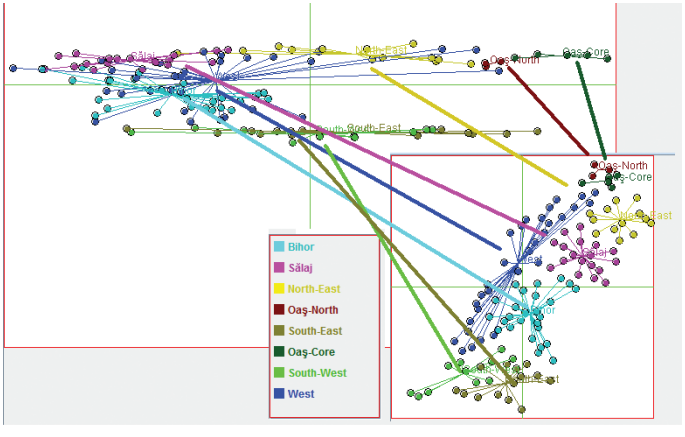


Illustration 2. An MDS picture of the Crișana region, using only phonetic data (cf. Illustration 1). Although the picture has a slightly different viewpoint, it also more clearly separates the various geographic regions in the linguistic picture, with the south-east and south-west, Oaș-core, Oaș-north, and the north-east distinct from the three central regions.

4. Faith and reason

Big Data is not just data. It is also the tools and methods that allow one to search the large databases now available.

In applying multidimensional scaling to the problem of visualizing relationships, we used software packages (such as R and SPSS) and we developed data structures and software of our own. Doing it ourselves gave us greater control over our tools, whereas using packages allowed us to leverage the work of others. Here are some lessons learned.

Our data sets are stored in simple text format, rather than as relational data bases. The database programs that were popular and current when we began (ca. 1997) do not exist today, but the text files can still be used by current or custom software packages. By keeping it simple, we avoided a dependency we did not want.

We created our own package for calculating our MDS pictures (MDS puts data points at their exact position in a high-dimensional space, using say a linguistic distance matrix; it then projects a best possible image of the data, in two or three dimensions). This gave us the flexibility to go on to create an interactive 3-D viewer as part of our 'access to data' agenda. However, unlike existing statistical packages (that we also used sometimes), our MDS does not measure 'stress' or other calculations of the goodness of our picture. In testing our software, we found there were limitations (due to matters of scale, and to rounding errors in the calculations) to our tools. They were good for creating the visualizations but they may not have been numerically accurate for the auxiliary measures; we did not want to go beyond our expertise.

Using your own software requires time and effort to create, test and maintain the work, and carries risks (especially the risk of subtle numerical calculation errors); using software packages requires faith that the package does all this well (a reasonable assumption, but an assumption all the same) and faith that the package is really doing what you think it is doing (a less reasonable assumption, but one the researcher can address with some added work). It is all too easy to accept the answers from Big Data unquestioningly, without really knowing what went into getting the answer.

5. Data, and more data

Perhaps obviously, Big Data depends on having lots of data. Today, there is a lot of data, widely available. Still, it is important to understand where the data comes from and what it does (and does not) represent.

For our work on Finnish, we digitized a hard-copy atlas (Kettunen 1940) which was a careful, scholarly work, and one that has been used in its same hard-copy format by generations of Finnish dialect researchers. We knew what we were getting, and the challenge was in the effort to create the data sets, and test the goodness and accuracy of our work. Even so, there were a large number of editorial decisions to be made as we tried to interpret what the hard-copy maps really said. For example, the hard-copy maps had several locations that seemed to be on the borders of a dialect feature; maybe the feature applied to that location, and maybe it didn't. Modern data would not help us decide what the case was over half a century earlier, and there was (and is) no more explicit data source than the one we had.

Even in the clear cases, one cannot ignore the questions a sociolinguist would ask: did everyone have this feature? or only some people? or some people sometimes? A thoughtful researcher is aware of these problems, and takes them into account. The danger with Big Data is that it can make it all too easy to accept the conclusions of the process while hiding the steps that led to it, or the assumptions underlying it, and the limitations that go with it.

6. In short

By all means, rely on Big Data: use lots of data, and use the range of sophisticated tools available for making sense of that data. It is a powerful approach to the study of any subject, especially one as complex and intricate as the human use of language.

At the same time, be aware of the responsibilities that come with Big Data.

- The data must be selected and separated into subsets with care, so that patterns in the data are clear and not obscured by sheer quantity.
- The methods, with their assumptions and limitations, must be applied knowledgeably, so that the results are indeed the results that the researchers think they have discovered.
- The data sources, especially when they are developed by others, need to be thoroughly understood, as well.

With due concern for these matters, and due attention to the potential pitfalls, Big Data can indeed give us big results.

Acknowledgements

Between the time when this paper was accepted and its final publication, Dorin Uritescu, a key part of our research team and responsible for all our Romanian data, passed away, on April 15,

2020. Our team is profoundly affected by the loss of an excellent scholar and a wonderful caring human being. May he rest in peace. Our research over several decades has been supported by SSHRC, the Social Sciences and Humanities Research Council of Canada, to whom we are deeply grateful.

References

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2002, 2007a. *Romanian Online Dialect Atlas*. http://vpacademic.yorku.ca/romanian (now at http://pi.library.yorku.ca/dspace/ under the "dialectology" community, "RODA" collection)

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2004. An exploration into the management of high volumes of complex knowledge in the social sciences and humanities. *Journal of Quantitative Linguistics* 11(3). 183–192. https://doi.org/10.1080/0929617042000314930

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2007a. Data capture and presentation in the Romanian Online Dialect Atlas. *Linguistica Atlantica* 27. 37–39.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2007b. Romanian Online Dialect Atlas: Data capture and presentation. 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*, 87–96. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110894219.87

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2008. *Digitalized dialect studies: North-Western Romanian*. Bucharest: Romanian Academy Press.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2009. Data management and linguistic analysis: Multidimensional Scaling applied to Romanian Online Dialect Atlas. In Reinhard Köhler (ed.), *Studies in Quantitative Linguistics* 5, 10–16. Lüdenscheid: RAM-Verlag.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2011. Defining dialect regions with interpretations. Advancing the multidimensional scaling approach. Paper presented at Methods In Dialectology 14 Conference, London, Canada, August 2–6.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2013. Defining dialect regions with interpretations. Advancing the multidimensional scaling approach. *Literary and Linguistics Computing* 2013. 28(1). https://doi.org/10.1093/llc/fqs048

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2018. An Expanded Quantitative Study of Linguistic vs Geographic Distance Using Romanian Dialect Data. In Lu Wang, Reinhard Köhler, & Arjuna Tuzzi (eds.), Structure, Function and Process in Texts, Proceedings of Qualico 2016, 25–33. Lüdenscheid, Germany: RAM-Verlag.

Embleton, Sheila & Eric S. Wheeler. 1997a. Multidimensional scaling and the SED data. In Viereck, Wolfgang & Heinrich Ramisch (eds.), *The Computer Developed Linguistic Atlas of England*, Volume 2, 5–11. Tübingen: Max Niemeyer.

Embleton, Sheila & Eric S. Wheeler. 1997b. Finnish dialect atlas for quantitative studies. *Journal of Quantitative Linguistics* 4. 99–102. https://doi.org/10.1080/09296179708590082

Embleton, Sheila & Eric S. Wheeler. 2000. Computerized dialect atlas of Finnish: Dealing with ambiguity. *Journal of Quantitative Linguistics* 7. 227–231. https://doi.org/10.1076/jqul.7.3.227.4109

Kettunen, Lauri. 1940. *Suomen murrekartasto* [The dialect atlas of Finland]. Helsinki: Suomalaisen kirjallisuuden seura.

McGuire, Patricia. 2019 October 27. *How higher education's data obsession leads us astray*. The Chronicle of Higher Education. https://www-chronicle-com.ezproxy.library.yorku.ca/article/How-Higher-Education-s-Data/247409. Accessed October 31, 2019.

Stan, Ionel & Dorin Uritescu. 1996, 2003. *Noul Atlas lingvistic român. Crișana* [The New Romanian Linguistic Atlas. Crișana]. Volume 1, 1996, Volume 2, 2003. București: Romanian Academy Press.

Uritescu, Dorin. 1983. Asupra repartiției dialectale a graiurilor dacoromâne. Graiul din Oaș [On the dialect structure of Daco-Romanian. The dialect of Oaș]. In Ion Gheție (ed.), *Materiale și cercetări dialectale II* [Dialectal materials and research II]. Cluj-Napoca: The University of Cluj-Napoca. 231–246.

Uritescu, Dorin. 1984a. Subdialectul crișean [The dialect of Crișana]. In Valeriu Rusu (ed.), *Tratat de dialectologie românească* [Treatise of Romanian Dialectology], 284–320, maps 78–106. Craiova: Scrisul Românesc.

Uritescu, Dorin. 1984b. Graiul din Țara Oașului [The dialect of Tara Oașului]. In Valeriu Rusu (ed.), *Tratat de dialectologie românească* [Treatise of Romanian Dialectology], 390–399, maps 171–174. Craiova: Scrisul Românesc.

Uritescu, Dorin. 1986. Theoretical problems of phonological change and the history of Romanian phonology. *Revue roumaine de linguistique* 31(3). 227–248.

Wheeler, Eric S. 2005. Multidimensional scaling for linguistics. In Reinhard Koehler, Gabriel Altmann & Rajmund G. Piotrowski (eds). *Quantitative linguistics. An international handbook*. Berlin: Walter de Gruyter. 548–553.

From distinguishability to informativity

A quantitative text model for detecting random texts

Maxim Konca, Alexander Mehler, Daniel Baumartz
and Wahed Hemati
Goethe University Frankfurt

We present a study of the distinctiveness of random and non-random texts based on text characteristics of quantitative linguistics. We additionally experiment with text features that evaluate contiguity associations among sentences by means of BERT (Bidirectional Encoder Representations from Transformers). To this end, we experiment with generative models for random texts as currently discussed in the context of neural networks. The chapter contributes to the clarification of deficits of existing random text models and of the informativeness of quantitative text features.

Keywords: random text, quantitative text characteristics, text classification, BERT

1. Introduction

Recent advancements in the study of neural networks concern natural language processing as well as natural language generation (Reiter & Dale 1997). With this program, powerful language models have been developed that allow the generation of random texts based on prior learning of syntagmatic and paradigmatic regularities from arbitrarily large text corpora. The resulting randomizations ideally reflect linguistic knowledge at least at the level of characters, words, and multi-word expressions to map both contiguity and similarity associations of these units in texts of the underlying language. In contrast to classical random text models (Biemann 2007; Kubát et al. 2014), which proceed, so to speak, 'macroscopically' by requiring that the resulting randomizations satisfy certain text laws (e.g., Zipf's first law) or quantitative text characteristics (e.g., proportion of *hapax legomena*, TTR, etc.), the recent generative models allow the representation of 'microscopic' relations of both types of association (by contiguity or similarity), starting from character sequences up to words and phrases. As a consequence, macroscopically generated random texts may seem natural from the point of view of quantitative text characteristics, while being practically unreadable, so that they can be directly identified as non-instances

https://doi.org/10.1075/cilt.356.10kon

of the respective target languages. Conversely, although microscopically generated random texts are not so easily identifiable as artificial instances of the respective target language (thus fulfilling a certain variant of a Turing test), they possibly largely disregard quantitative text characteristics, unless the underlying language model learns these characteristics 'on the fly' in order to reproduce them by means of their randomized output. In this way, we obtain a matrix of text randomization that contrasts macro- and microscopic approaches to criteria of naturalness of random texts (conformity with quantitative text characteristics, readability, etc.). Ideally, a random text model combines the best of both worlds, so that neither a speaker of the targeted language nor a text classifier can distinguish random texts from their natural counterparts, whether based on criteria of readability or comprehensibility or of quantitative text characteristics. As a matter of fact, recently proposed generative models as the GPT-2 model (Radford et al. 2019) currently fail to do so.

In this chapter, we show that, using rather classic approaches to text classification, it is easily possible to separate randomized texts from their non-randomized counterparts. To this end, we experiment with vectorized text representations whose dimensions correspond to a set of quantitative text characteristics, so that the classifier has to learn that quantitative characteristics of randomized texts differ significantly from their non-random counterparts. We show that this is remarkably easy to demonstrate. In other words, random texts are non-naturally random with respect to easy-to-calculate text characteristics as studied by quantitative linguistics. However, starting from this somewhat sobering result, which concerns the quality of current random text models, the study goes a decisive step further. It turns the tables, so to speak, against the quantitative text characteristics and asks which of them best support the distinction between random and non-random texts and which do not. In this way, we indirectly identify uninformative text characteristics that do not support the desired distinction: these features say little (compared to their more salient counterparts, see Figure 5) about the specifics of the statistical nature of natural language texts, since they are unable to distinguish them from random counterparts, or hardly distinguish them at all. In this way we ultimately gain access to a classification-driven evaluation criterion for the quality or meaningfulness of models of quantitative text linguistics. The outcome of the chapter is fourfold:

1. We experiment with methods of random text generation based on probabilistic language models and develop a set of text classifiers to distinguish them from their non-random counterparts, where the classifiers explore vectors of text characteristics that quantify the underlying random and non-random texts.
2. We offer both a broad and detailed sensitivity analysis to evaluate the validity of text characteristics with regard to the distinction between the two target classes.
3. We analyze the correlations between the characteristics and find a small group of antagonistic features with very good discrimination properties, but from very different evaluation perspectives.

4. Finally, we outline future work to improve the current random text models to overcome their problems described in this study.

The chapter is organized as follows: Section 2 describes the corpora analyzed in our study, the models used to randomize them, and the text characteristics used to quantify all texts. Section 2 also describes our data processing and evaluation procedure. Section 3 presents our findings, which are discussed in § 4. Finally, § 5 gives a conclusion and an outlook on future work.

2. Text corpora and their quantification

2.1 Quantification

According to our research agenda, we seek easy-to-calculate quantitative features that separate natural language texts from their random counterparts. To keep this scenario simple, we experiment with a set of indices as described in Kubát et al. (2014). That is, we compute well-known models of quantitative text linguistics (Kubát et al. 2014; Mehler 2005). While all these models follow the counting approach, more recent models based on neural networks allow the modeling of semantic associations between text units of arbitrary size. To reflect this approach, we additionally experiment with a subset of models that investigate the autocorrelation of sentence associations in time series of consecutive sentences, whose association probability is measured with the help of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018). The idea behind this approach is that natural language texts exhibit a kind of Markov pattern of associated sentences (in the sense of contiguity associations) as a manifestation of textual cohesion and that these patterns can be captured in the latter way. Random texts that disregard such Markovian, sometimes discontinuous, associations are then easier to identify. Interestingly, we show that the corresponding group of text features is actually informative in this sense. All in all, we study 22 indices listed as follows and whose components are given in Table 1:

Table 1. Parameters of quantitative text characteristics used in this study

Parameter	Description
fr_i	frequency of the rank r_i
h	h-point
N	total number of tokens
r	token's frequency rank
V	vocabulary size (number of types)

(1) **A** – *adjusted modulus* (Kubát et al. 2014):

$$A = \frac{h^{-1}(f_{r1}^2 + V^2)^{1/2}}{\log_{10} N}$$

where h is the h-point (see below); for f_{r1} and V see Table 1.

(2) **alpha** – *writer's view* (Popescu & Altmann 2007):

$$cos\ \alpha = \frac{-((h-1)(f_{r_1} - h) + (h-1)(V-h))}{((h-1)^2 + (f_{r_1} - h)^2)^{1/2}((h-1)^2 + (V-h)^2)^{1/2}}$$

(3) **ATL** – *average token length*.

(4) **ab1, ab2, ab3, ab4** – autocorrelation coefficients (Parzen 1963) calculated from BERT's (Devlin et al. 2018) next sentence prediction probabilities with lag 1, 2, 3, and 4.

(5) **G** – *Gini coefficient* (Popescu & Altmann 2006):

$$G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{i=1}^{V} r_i f_{r_i}\right)$$

(6) **h** – *h-point* (Hirsch 2005; Popescu 2009): the point, where the token's frequency rank and the frequency itself are equal.
If no such point exists, two neighbouring frequencies, for which $f_{ri} > r_i$ and $f_{rj} < r_j$, are used to calculate h as follows:

$$h = \frac{f_{ri} r_j - f_{rj} r_i}{r_j - r_i + f_{r_i} - f_{r_j}}$$

(7) **H** – *entropy* (Esteban & Morales 1995):

$$H = \log_2 N - \frac{1}{N}\sum_{i=1}^{V} f_{r_i} \log_2 f_{r_i}$$

(8) **hl** – *hapax legomena percentage*: the percentage of unique types.

(9) **L** – *curve length* (Popescu et al. 2011):

$$L = \sum_{i=1}^{V-1} \sqrt{(f_{r_i} - f_{r_{i+1}})^2 + 1}$$

(10) **Λ** – *lambda* (Čech 2015; Popescu et al. 2011):

$$\Lambda = \frac{L \log_{10} N}{N}$$

where L is the curve length.

(11) **Q** – *activity* (Altmann 1988):

$$Q = \frac{v}{v + a}$$

where v is the number of verbs and a the number of adjectives.

(12) **R1** – *vocabulary richness* (Kubát et al. 2014):

$$R_1 = 1 - \left(\frac{\sum_{i=1}^{h} f_{ri}}{N} - \frac{h^2}{2N} \right)$$

where h is the h-point.

(13) **RR** – *repeat rate* (Kubát et al. 2014):

$$RR = \frac{1}{N^2} \sum_{i=1}^{V} f_{ri}^2$$

(14) **RRR** – *relative repeat rate* (McIntosh 1967):

$$RRR = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}}$$

(15) **stc** – *secondary thematic concentration* (Čech et al. 2013):

$$stc = \sum_{i=1}^{2h} \frac{(2h - r_i) f_{ri}}{h(2h - 1) f_{r1}}$$

(16) **tc** – *thematic concentration* (Popescu & Altmann 2011):

$$tc = 2 \sum_{i=1}^{T} \frac{(h - r_i) f_{ri}}{h(h - 1) f_{r1}}$$

where T is the number of autosemantic words whose rank r is above the h-point.

(17) **ttr** – *type-token-ratio* (Wimmer 2005):

$$ttr = \frac{V}{N}$$

(18) **UG** – *unique trigrams*: the ratio of the number of *hapax legomena* and the total number of character-based 3-grams that can be generated by them.

(19) **VD** – *verb distances*: the average distance between verbs, measured by the number of tokens. Following Kubát et al. (2014), all verbs were considered.

2.2 Text corpora and their randomization

To test our hypotheses, we experiment with two corpora. We start with a subcorpus of the Gutenberg Project, denoted by *GC*, which consists of 3000 texts, whose lengths range from less than 500 to more than 500,000 tokens.[1] We derive a random corpus from GC named *RGC*, which contains 3000 texts resulting from uniform token sampling where 2000 texts contain 1000 tokens, 500 texts contain 4000 tokens and another 500 texts 8000 tokens.[2] To control text length, we experiment with a second corpus, named *BC* (*Book Corpus*), of 5000 texts of approximately 1000 tokens each (±20 tokens to end sentences). It was created by taking verbatim the first 5000 chunks of the book corpus that was available via http://www.cs.toronto.edu/~mbweb/ and used in Devlin et al. (2018) and Zhu et al. (2015). As before, we derive a random corpus from BC, named *RBC*, by means of uniform sampling to generate 2000 random texts.

While RGC and RBC are both based on uniform sampling from our source corpora, we created a third corpus, named *RC*, of 3500 random texts by means of a generative model. To this end, we experimented with several models and found that GPT-2 (Radford et al. 2019) is one of the few that is capable of generating 'readable' texts of considerable length without suffering from an impending mode collapse, as 'Generative Adversarial Networks' do (Metz et al. 2016).[3] Recurrent neural networks (RNN), 'Long Short-Term Memories' (LSTM) (Hochreiter & Schmidhuber 1997) in particular, another family of models we tested, are notoriously hard to train, due to their sequential nature, which inhibits the possible speed up that parallelization would provide. Moreover, RNNs are struggling to capture long-range dependencies, which makes the generation of long sequences harder.

GPT-2 is a probabilistic language model (Bengio et al. 2003), whose architecture is based on a stacked and scaled up Transformer (Vaswani et al. 2017). The Transformer uses both a decoder-encoder attention (Bahdanau et al. 2014) and a multi-head self-attention (Cheng et al. 2016) mechanism that allow it to take into account different types of relation between parts of the (text) sequence to calculate (conditional) probability distributions. The authors of GPT-2 released two pre-trained versions of the model, a large (including 345 million parameters) and a smaller one (114 million parameters). Since these models can easily be fine tuned, we trained the smaller model on the non-random corpora until acceptable levels of prediction errors were reached. After the fine tuning was done, the

1. www.gutenberg.org
2. To make the results of our analysis reproducible, we set the random state to 0.
3. Fore a demo of GPT-2 see https://www.openai.com/blog/better-language-models/.

unconditional samples (starting token <|endoftext|>) of the largest possible size (1024) were generated.[4]

All in all, our corpora consist of approximately 16,500 (random or non-random) texts and more than 265 million tokens (see Table 2). To perform our experiments and to account for their imbalance (number of texts and text size), we sampled the final random corpora from these sources as shown in Table 3. In this way, we experiment with two source corpora and two randomization methods to generate three random corpora which contribute altogether to eight classification experiments (Table 3).

Table 2. Overview of the corpora used

Corpus	Size	Text length (tokens)	Description	Random
BC	5000	1000	part of BookCorpus (Zhu et al. 2015)	0
GC	3000	500–500,000	subset of Project Gutenberg texts	0
RC	3500	1000	GPT-2 generated corpus	1
RBC	2000	1000	uniform sampling from BC	1
RGC	3000	1000–8000	uniform sampling from GC	1

Starting from these corpora, the distributions of the text characteristics of § 2.1 are shown in Figure 1. It shows remarkable differences (in terms of h and A) and commonalities between our two non-random corpora BC and GC. Figure 2 displays the 22 text characteristics as functions of each other based on the non-random texts. The heatmaps of the corresponding distance correlations (Székely et al. 2007) are shown in Figure 3. They exhibit a fairly clear trend for both corpora, BC and GC: a larger, rather uncorrelated subset of characteristics, including concentration and especially BERT autocorrelation indices, and a second group of very strongly correlated indices such as hl and ttr, so that in the present scenario at least one of them should be redundant for text classification. In the range between these two extremes, we observe partially nonlinear dependencies that make distance correlation the method of choice for assessing them. Obviously, however, rather uncorrelated (ideally orthogonal) or non-linear correlating features are more interesting for text modeling, since they tend to capture different sources of text information. According to this analysis, our feature set from § 2.1 shows a broad spectrum of dependencies of quantitative text characteristics, suggesting their usefulness for the classifications now under consideration.

4. Of course, the texts generated by our random models are artificial texts: we focus on learning to distinguish natural language texts from these random texts in order to obtain a criterion for the evaluation of models of quantitative text linguistics – within the framework of our classification experiment.

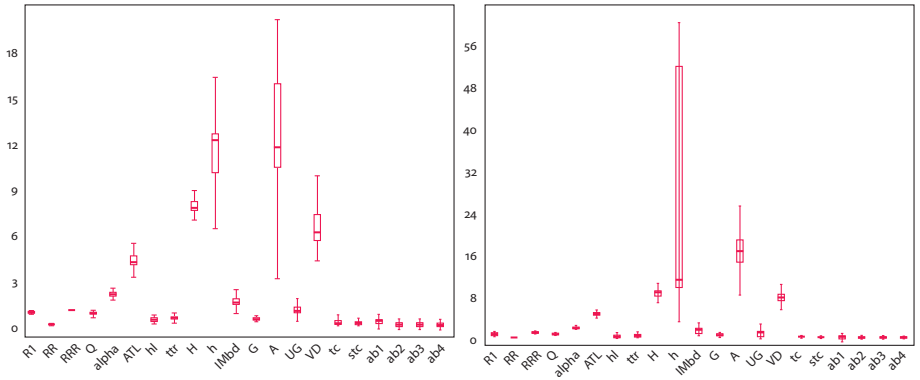


Figure 1. Overview of text indices distribution (L excluded): Left: Book corpus BC; right: Gutenberg corpus GC

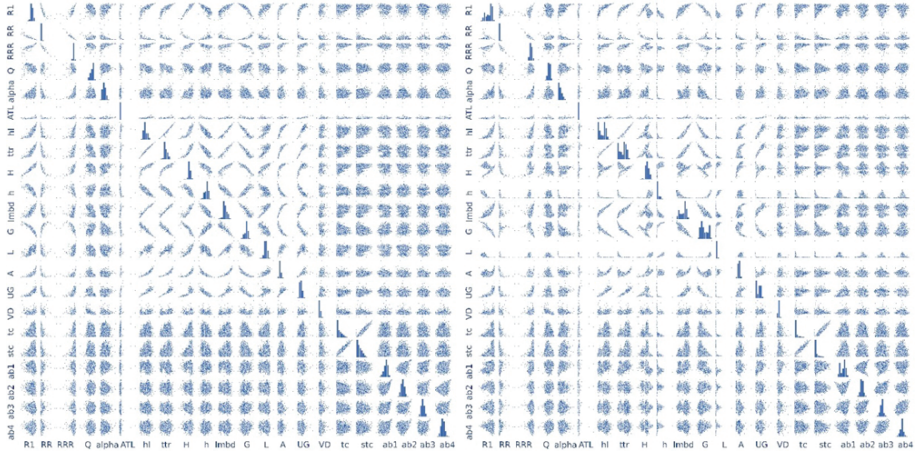


Figure 2. Functional dependencies between text indices (rows and columns): Left: Book corpus BC; right: Gutenberg corpus GC. All values are scaled with MinMaxScaler, i.e., lie between 0 and 1

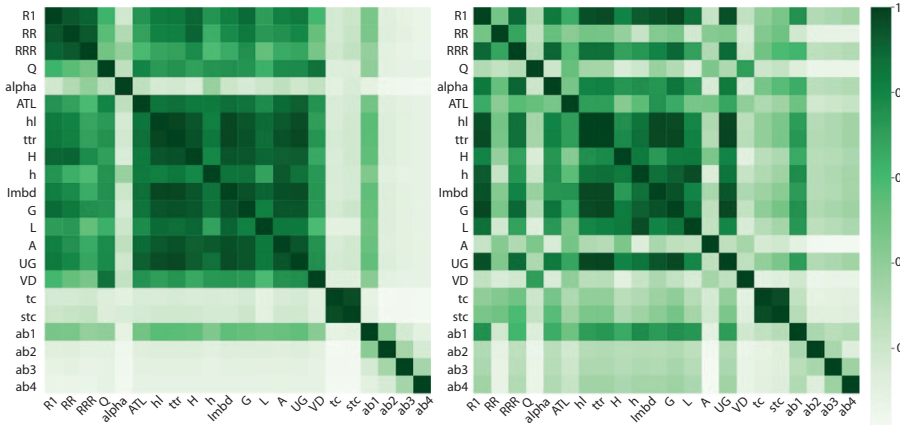


Figure 3. Distance Correlation of text features: Left: Book corpus BC; right: Gutenberg corpus GC

2.3 Classification and evaluation methods

To prepare our experiments, we performed NLP-based preprocessing including corpus cleansing, tokenization, lemmatization and PoS tagging using spaCy's en_core_web_lg model.[5] Next, each of the 22 text characteristics was calculated for all texts of all corpora from Table 2. The resulting feature vectors were then used to train two classifiers (whereby the data sets were always divided into training (75%) and test sets (25%): scikit-learn's 'Random Forest Classifier' (RFC) (Breiman 2001) and scikit-learn's 'Support Vector Machines' (SVM) (Boser et al. 1992; Smola & Schölkopf 2004; Chang & Lin 2011). Since the classifiers have access to the feature vectors as a whole, and since we are more interested in the classificatory performance of the individual features, we experiment with a number of evaluation methods (beyond F-value statistics):

1. We perform feature sensitivity analyses using the methods of Sobol (Sobol 2001; Saltelli 2002; Saltelli et al. 2010) and Morris (Morris 1991; Campolongo et al. 2007).
2. We experiment with two variants of coefficient-based iterative feature elimination algorithms: IFEf (Iterative Feature Elimination) using scikit-learn's RFC and IFEv using scikit-learn's SVMs. In each elimination round, the text characteristic of least importance in the sense of the SVM's coef – or the RFC's feature importance – is eliminated.

5. https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-2.2.5

3. We additionally experiment with two variants of error-based elimination algorithms: EBFEf (Error Based Feature Elimination) using scikit-learn's RFC and EBFEv using scikit-learn's SVM. Now, in each elimination round, the feature with the smallest impact on the error rate is eliminated, where we tested both F1-score and Cohen's kappa (Cohen 1960).

3. Results

The classificatory settings that we analyzed (see Table 3) are summarized by their F1- and κ-scores in Table 4. It provides a global picture of our sensitivity and elimination analyses, which clearly shows that random and non-random texts can be distinguished almost perfectly (rows 1, 2, 7 and 8). It also shows that SVM and RFC are almost equal (except for experiments 3 and 4) – for the corresponding confusion matrices see Figure 4. Further, Table 4 demonstrates that while natural language books are practically indistinguishable (row 3), Gutenberg texts do so much better (row 4) (possibly also because of their length variance). However,

Table 3. Overview of experiments performed

Experiment	Description
BC vs. RBC + RC	non-random vs. random with equal text lengths
GC vs. RGC + RC	non-random vs. random with widely varying text lengths
BC vs. BC	obfuscated BC, random labels assigned to 50% of the corpus
GC vs. GC	obfuscated GC, random labels assigned to 50% of the corpus
RBC vs. RBC	obfuscated BC, random labels assigned to 50% of the corpus
RBC vs. RC	uniform vs. GPT-2 generated
BC vs. RC	non-random vs. GPT-2 generated
GC vs. RC	non-random vs. GPT-2 generated

Table 4. Classification scores for eight different classification experiments

	Experiment	F1(svm)	F1(rfc)	κ(svm)	κ(rfc)
1	BC vs. RBC + RC	0.986	0.976	0.973	0.954
2	GC vs. RGC + RC	0.997	0.998	0.996	0.997
3	BC vs. BC	0.182	0.480	−0.015	−0.023
4	GC vs. GC	0.641	0.486	−0.002	−0.002
5	RBC vs. RBC	0.553	0.518	0.027	0.023
6	RBC vs. RC	0.999	0.999	0.998	0.997
7	BC vs. RC	0.999	0.994	0.998	0.985
8	GC vs. RC	0.997	0.997	0.994	0.995

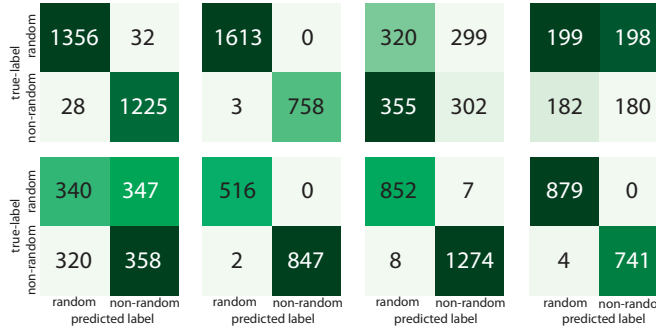


Figure 4. Confusion matrices (RFC); top row from left to right: Experiments 1, 2, 3, 4; bottom row: Experiments 5, 6, 7, 8

given the two target classes, the scores obtained are so low that this does not affect the abovementioned almost perfect classification – this is especially confirmed by the κ-values: our text characteristics are obviously valid in terms of distinguishing random and non-random texts. Further, while the two randomization methods (uniform, GPT-2) generate perfectly separable texts (row 6), this is no longer true if we sample the target classes' members from RBC (row 5).[6] Once more, this finding confirms the special role of our text feature model regarding the distinction of random and non-random texts. However, we now have to ask which of these features are actually responsible for our classification results.

Despite the fact that all texts in Experiment 1 had the same length, which should make it harder for classifiers to achieve good results, both classifiers showed remarkable accuracy, 99.9%. Although the grid search for optimal parameters was performed only for the SVM classifier, the RFC did not fall behind, which suggests that it was a trivial task to distinguish both corpora.

Perhaps due to the large differences in text lengths both classifiers reached even higher accuracies in Experiment 2, misclassifying only 3 (RFC) and 4 (SVM) texts out of 2,500. However, L, which shows noticeable sensitivity to text length, is not weighed as an important feature by either Sobol or Morris methods, that would be the case without appropriate re-scaling of the data, as we found out during our experiments with different scaling methods.

The IFEf and IFEv methods use the information from the coefficients of the RFC and SVM respectively. The values of these coefficients were increasing and decreasing between rounds, without showing any stable trend, which made it impossible to quantify the differences in importance of each indicator, i.e., the results we were able to obtain with these methods were of ordinal nature, where hl (and/or

6. Thereby capturing completely different aspects of randomization.

ttr), ATL, RR and ab1 were ranked as top features. If we did not fix the random state, these methods showed somewhat unstable results, where the rank of an indicator could jump 10 positions up and down. The most volatile were middle and low ranks, keeping high ranks relatively stable.

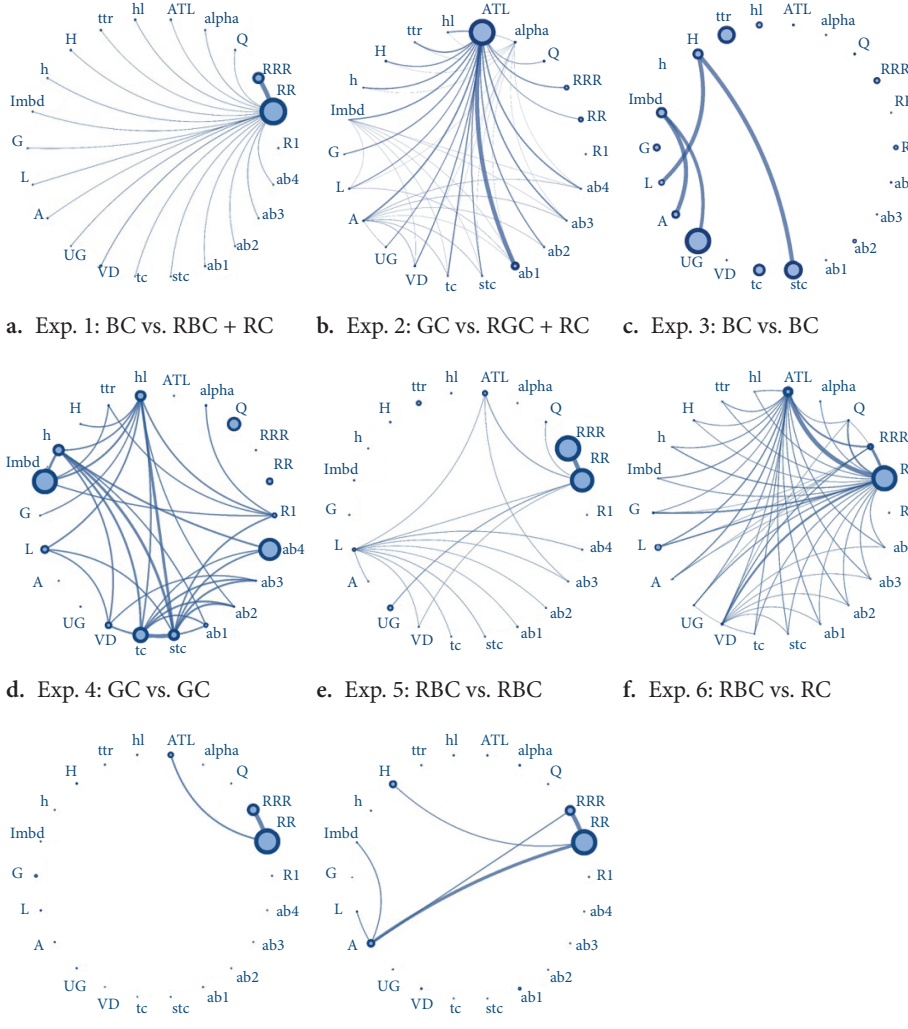
In this setting the error-based methods (EBFEf and EBFEv) showed even more volatility than IFEf and IFEv. This could be due to the fact that the differences in the error metric used as an elimination criterion were mostly very small. However, as the features were eliminated, both F1-score and κ were steadily decreasing. In Experiments 1 and 2, F1-score did not drop below 90%, κ below 83%, even with only two features remaining.

4. Discussion

Our sensitivity analysis based on Sobol (see Figure 5)[7] shows a clear trend: on the one hand, repeat rate (RR) and relative repeat rate (RRR) stand out as sensitive indices in several experiments (1, 5, 6, 7, 8), ATL is highly sensitive in Experiment 2 (and partly also in Experiment 7). On the other hand, RR and RRR have a higher distance correlation (see Figure 3) and also interact in the sense of high S2 values (Figure 5).[8] In addition, we observe in experiments 1, 2, 7 and 8 a tendency towards star graphs (in the perfect sense in experiments 1 and 7), whereby RR and ATL each become centres of interactivity: other features interact in pairs usually to a higher degree only with these indices. The two experiments on self-differentiation of our source corpora are exceptional cases: BC (Experiment 3) makes several indices sensitive to the output variable (classification score); the resulting network of indices, however, is very sparse. The opposite is true for Experiment 4: we obtain a rather dense network of closely linked indices which, according to Table 4, are associated with a significantly higher F-score than in Experiment 3. Lambda and ab4 are in this case outstanding text characteristics (this finding is reproduced by the sensitivity analysis based on Morris – see Figure 6): thus, if the classifier is to distinguish artificially formed classes based on the Gutenberg corpus, it tends to disperse its information sources, combining simple frequency statistics with more semantic indices such as ab4, while evaluating interactions of the characteristics to a large extent. RBC is easier to separate from RC (Experiment 6) than from itself (Experiment 5) (Table 4), but at the price of a more complex, interactive, less concentrated text representation model (Figure 5).

7. Courtesy of https://github.com/antonia-had/Radial convergence plot

8. The exceptional role of RR and RRR in Experiment 1 and 5 is mirrored by the sensitivity analysis based on Morris – see Figure 6.



a. Exp. 1: BC vs. RBC + RC **b.** Exp. 2: GC vs. RGC + RC **c.** Exp. 3: BC vs. BC

d. Exp. 4: GC vs. GC **e.** Exp. 5: RBC vs. RBC **f.** Exp. 6: RBC vs. RC

g. Exp. 7: BC vs. RC **h.** Exp. 8: GC vs. RC

Figure 5. Sobol sensitivity. The node size indicates the first order index (S1) per parameter, the node border thickness indicates the total order index (ST) per parameter, and the thickness of the line between two nodes indicates the second order index (S2)

What lesson do these experiments teach us? When it comes to separating random from non-random texts, very few, very simple indices dominate the variance of the target variable. In Experiment 2 we observe the additional sensitivity of ab1, a semantic feature that measures the autocorrelation of the contiguity association of adjacent sentences: apparently, these feature groups (ATL, RR, RRR on the one hand and ab1 on the other) are very heterogeneous. In other words, *it is easy to separate*

(Experiment 1, 2, 7 and 8), *but sometimes* (Experiment 2) *only in connection with a semantic feature*, which in the present case presupposes a probabilistic language model. At the same time, the effectiveness of a large number of other more or less famous frequency-based text features is low: only a few of them prevail, and mostly only in obfuscated scenarios (see experiments 3 and 4). On the basis of the results documented in Table 4 and the sensitivity analysis of Figure 5, we conclude that the randomization models examined here appear to be 'unnatural' from the point of view of certain text characteristics and are devalued accordingly, but that this does not apply to all characteristics to the same extent: informativity in this sense is a very selective variable that separates the group of text characteristics into a very small group of highly sensitive characteristics and a large group of insensitive ones.

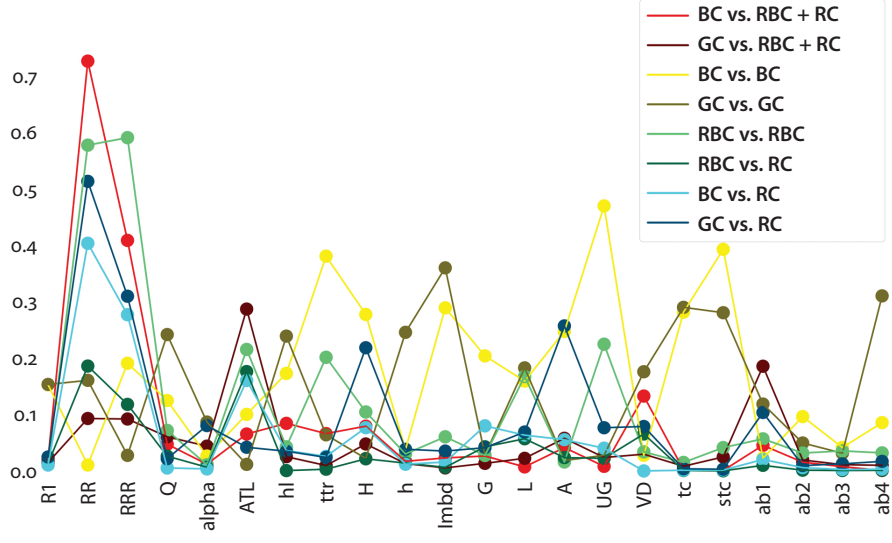


Figure 6. Morris Sensitivity for all corpora

5. Conclusion

We have presented a study which shows that current random text models still generate texts that are easily distinguishable from non-random counterparts and, secondly, that a remarkably small set of rather simple quantitative text characteristics is sufficient to show this, while most other indices investigated here are rather uninformative in this sense. This finding is of course due to the experiments of this study: in other studies, these features may prove their informativity. We also showed that rather uncorrelated quasi-semantic features that investigate contiguity associations by means of BERT are a valuable resource in some experiments.

One might now think that we have just presented another study on the possibility of automatic text classification, which once again confirms that classifiers can be found to separate any groups of text. However, such an expectation that successful text classification is generally feasible is certainly wrong and also does not correspond to the experience from a large number of text classification experiments (e.g., at word (Joachims 2002), phrase (Gabrilovich & Markovitch 2006), sentence (Baayen et al. 1996; Mehler, Hemati, Uslu & Lücking 2018) or text structure level (Mehler et al. 2007)), which usually publish positive results, but still far away from almost maximum F-scores. Moreover, such a view misses the significance of our study, which shows that the classification we aimed for was simply not possible with the majority of the text characteristics considered here: whatever these characteristics model, they are not sufficient to separate two otherwise easily separable sets of text (one natural, the other artificial). However, this is ultimately the aim of the corresponding research in quantitative text linguistics: to find quantitative text characteristics that are as meaningful as possible with regard to the quantitative structure of natural language texts (Altmann 1988). And this includes that these characteristics should help to distinguish natural language texts from certain aggregates of a very different, that is, artificial kind. We have thus effectively developed a test that allows us to check the quality of text characteristics – certainly this is only one test scenario among many possible and therefore not an exhaustive scenario. We imagine, however, that such research would be carried out in quantitative text linguistics, which should examine the validity of its models also with the help of machine learning.

Future work will focus on significantly increasing the number of text features thereby relying on a feature generation model of Mehler, Hemati, Gleim & Baumartz (2018), investigating further random text models, and finally, using the results to improve these text models.

References

Altmann, Gabriel. 1988. *Wiederholungen in Texten*. Bochum: Brockmeyer.

Baayen, Harald, Hans van Halteren & Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3). 121–131. https://doi.org/10.1093/llc/11.3.121

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent & Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3. 1137–1155.

Biemann, Chris. 2007. A random text model for the generation of statistical language invariants. In Candace Sidner, Tanja Schultz, Matthew Stone & ChengXiang Zhai (eds.), *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, 105–112. Rochester, NY: Association for Computational Linguistics.

Boser, Bernhard E., Isabelle M. Guyon & Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In David Haussler (ed.), *Proceedings of the fifth annual workshop on computational learning theory*, 144–152. New York: Association for Computing Machinery. https://doi.org/10.1145/130385.130401

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32. https://doi.org/10.1023/A:1010933404324

Campolongo, Francesca, Jessica Caribon & Andrea Saltelli. 2007. An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software* 22(10). 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004

Čech, Radek. 2015. Text length and the lambda frequency structure of a text. In George K. Mikros & Ján Macutek (eds.), *Sequences in language and text*, 71–88. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110362879-006

Čech, Radek, Ioan-Iovitz Popescu & Gabriel Altmann. 2013. Methods of analysis of a thematic concentration of the text. *Czech and Slovak Linguistic Review* 3. 4–21.

Chang, Chih-Chung & Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3). 1–27. https://doi.org/10.1145/1961189.1961199

Cheng, Jianpeng, Li Dong & Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In Jian Su, Kevin Duh & Xavier Carreras (eds.), *Proceedings of the 2016 conference on empirical methods in natural language processing*, 551–561. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1053

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1). 37–46. https://doi.org/10.1177/001316446002000104

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.

Maria Dolores Esteban & Domingo Morales. 1995. A summary on entropy statistics. *Kybernetika* 31(4). 337–346.

Gabrilovich, Evgeniy & Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the twenty-first national conference on artificial intelligence*, 2006 Jul 16 (Vol. 6, pp. 1301–1306) Boston, MA: *AAAI Press*.

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

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Joachims, Thorsten. 2002. *Learning to classify text using support vector machines*. Boston: Kluwer. https://doi.org/10.1007/978-1-4615-0907-3

Kubát, Miroslav, Vladimír Matlach & Radek Čech. 2014. *Quita. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

McIntosh, Robert P. 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology* 48(3). 392–404. https://doi.org/10.2307/1932674

Mehler, Alexander. 2005. Eigenschaften der textuellen Einheiten und Systeme [Properties of textual units and systems]. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistik. ein internationales handbuch / quantitative linguistics. An international handbook*, 325–348. Berlin: De Gruyter.

Mehler, Alexander, Peter Geibel & Olga Pustylnikov. 2007. Structural classifiers of text types: Towards a novel model of text representation. *Journal for Language Technology and Computational Linguistics (JLCL)* 22(2). 51–66.

Mehler, Alexander, Wahed Hemati, Rüdiger Gleim & Daniel Baumartz. 2018. VienNA: Auf dem Weg zu einer Infrastruktur für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprache. In Henning Lobin, Roman Schneider & Andreas Witt (eds.), *Forschungsinfrastrukturen und digitale Informationssysteme in der germanistischen Sprachwissenschaft*, Volume 6, 149–176). Berlin: De Gruyter.

Mehler, Alexander, Wahed Hemati, Tolga Uslu & Andy Lücking. 2018. A multidimensional model of syntactic dependency trees for authorship attribution. In Jingyang Jiang & Haitao Liu (eds.), *Quantitative analysis of dependency structures*, 315–348. Berlin: De Gruyter. https://doi.org/10.1515/9783110573565-016

Metz, Luke, Ben Poole, David Pfau & Jascha Sohl-Dickstein. 2016. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*.

Morris, Max D. 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics* 33(2). 161–174. https://doi.org/10.1080/00401706.1991.10484804

Parzen, Emanuel. 1963. On spectral analysis with missing observations and amplitude modulation. *Sankhyā: The Indian Journal of Statistics, Series A*, 383–392.

Popescu, Ioan-Iovitz. 2009. *Word frequency studies*, Volume 64. Berlin: Walter de Gruyter.

Popescu, Ioan-Iovitz & Gabriel Altmann. 2006. Some aspects of word frequencies. *Glottometrics* 13. 23–46.

Popescu, Ioan-Iovitz & Gabriel Altmann. 2007. Writer's view of text generation. *Glottometrics*, 15, 71–81.

Popescu, Ioan-Iovitz & Gabriel Altmann. 2011. Thematic concentration in texts. *Issues in quantitative linguistics* 2. 110–116.

Popescu, Ioan-Iovitz, Radek Čech & Gabriel Altmann. 2011. *The lambda-structure of texts*. Lüdenscheid: Ram-Verlag.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8). 9.

Reiter, Ehud & Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3(1). 57–87. https://doi.org/10.1017/S1351324997001502

Saltelli, Andrea. 2002. Making best use of model evaluations to compute sensitivity indices. *Computer physics communications* 145(2). 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1

Saltelli, Andrea, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto & Stefano Tarantola. 2010. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications* 181(2). 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Smola, Alex J. & Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14(3). 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Sobol, Ilya M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* 55(1–3). 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Székely, Gábor J., Maria L. Rizzo & Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6). 2769–2794. https://doi.org/10.1214/009053607000000505

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wimmer, Gejza. 2005. The type-token-relation. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik: Ein internationales Handbuch* [Quantitative linguistics: An international handbook], 361–368. Berlin: De Gruyter.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba & Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, 19–27. Cambridge, MA: IEEE. https://doi.org/10.1109/ICCV.2015.11

A Modern Greek readability tool

Development of evaluation methods

George Mikros and Rania Voskaki
Hamad Bin Khalifa University / Centre for the Greek Language

The aim of this paper is to develop an automatic readability analysis tool that focusses on Modern Greek as a foreign language. Based on previous work done in the Centre for the Greek Language (CGL), we offer an enhanced methodology in readability prediction for Modern Greek texts matching the adequacy level (A1 to C2) according to the Common European Framework of Languages. The proposed tool is based on several stylometric indices inspired by work done in the field of quantitative linguistics. The resulting feature vectors train a Random Forest, a robust and accurate machine learning algorithm that predicts readability in our testing dataset with 0.943 accuracy, surpassing all previous readability tools for Modern Greek. Further, analysis of the results with advanced visualization methods reveals the complex and fluid dynamics of the features used and their readability predictions.

Keywords: readability tool, corpora, annotation, evaluation methods

1. Introduction

The present research aims to investigate whether automatic readability analysis (DuBay 2004) can enhance its accuracy using lexical differentiation indices inspired by relevant research in quantitative linguistics. Moreover, we aim to improve existing readability tools for Modern Greek and develop analysis methods that are robust and at the same time accurate in classifying texts into reading difficulty levels that match the linguistic skills of people learning Modern Greek as a foreign language.

The MOGRead is one of the most accurate readability tools available for Modern Greek and it was developed in order to meet the needs of teachers of Modern Greek as a second/foreign language (L2). Given that the existing Modern Greek corpora are few and rarely updated, the need for an effective readability tool

https://doi.org/10.1075/cilt.356.11mik

is imperative, especially for less used and less taught languages, such as Modern Greek. The tool is developed by the Centre for the Greek Language (CGL) and it is available at the Portal for Teaching Modern Greek as a foreign / L2:[1] it takes as input plain text or readable file format of any size and outputs text statistics and bar graphs depicting the adequacy level (A1 to C2) according to the Common European Framework of Languages.

MOGRead is based on a text classification model trained using multinomial logistic regression using 12 linguistic features some of which are 'classic' readability features and others are features specific to the Greek language. The features employed are sentence length, 'long' words, Named Entities, text size (optional), Guiraud's R, conjunctions, pronouns, pre- and suffixes, ancient adv. types, 'easy' words, passive verbs, adjectives + participles.

MOGRead is a tool that uses 'classic' readability features and although it works reasonably well, it doesn't utilize a modern machine learning algorithm. Moreover, quantitative linguistics research on lexical diversity has offered a wide variety of textual indices which have not yet been applied to readability research. Our study focusses on using these indices along with more traditional features and combines them with a robust machine learning algorithm in order to enhance Modern Greek readability assessment for assisting teaching of Modern Greek as L2.

2. Readability analysis: A short literature review

The most common readability formula was created by Flesch (1948). The Flesch Reading Ease Readability Formula rates texts on a 100-point scale; the higher the score, the easier it is to understand the document. Most standard passages have a readability score of approximately 60 to 70. In the framework of the Flesch formula, we take into account the Average Sentence Length (ASL), that means the number of words divided by the number of sentences and the Average of Syllables per Word (ASW), that is the number of syllables divided by the number of words.

Dale & Chall (1948) in another 'classic' readability paper define readability as "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting". They designed another popular readability formula to correct certain shortcomings in the Flesch Reading Ease Formula. They listed 3000 easy words, 80% of which are known to fourth-grade readers, and they proposed a

1. http://www.greek-language.gr/certification/readability/index.html

sentence-length variable plus a percentage of 'hard' words. Following this line of research, Gunning (1952) published the Fog-Index, a readability formula developed for adults, which uses two variables, average sentence length and the number of 'hard' words for each 100 words. A hard word is defined as a word that is more than two syllables long.

Those three readability formulas mark the end of the first 30 years of classic readability studies. Since 1960, new developments accelerated the study of readability. Fry (1968) was the first to propose the use of graphs. The Fry readability score was a visual assessment of a text's grade level. For a sample of text, it plots the number of syllables per 100 words on the horizontal axis (x-axis), and the number of sentences per 100 words on the vertical axis (y-axis). The region this point falls in is an estimation of grade level.

In 1969 McLaughlin (1969) created the SMOG readability formula and defined readability as "the degree to which a given class of people finds certain reading matter compelling and comprehensible" (p. 188). According to the SMOG formula, the word length and sentence length should be multiplied rather than added. It counts the number of words of more than two syllables in 30 sentences.

The Flesch-Kincaid formula (Kincaid et al. 1975) constitutes a recalibration of the original Flesch formula, developed to rate texts on the U.S. grade school level, where the comprehension of a text corresponds to a score of eight.

More recent studies on readability consider deeper linguistic features requiring text processing and machine learning methods. The computer tool Coh-Metrix (Graesser et al. 2004) developed a wide variety of NLP modules using lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora and latent semantic analysis. It calculates many quantitative linguistic measures, including coreferential cohesion, causal cohesion, density of connectives, latent semantic analysis metrics, and syntactic complexity.

Moreover, NLP studies enable new features able to capture a wider range of readability factors and the combination of those features through machine learning algorithms (François & Fairon 2012). This new trend in readability studies, referred to as 'AI readability', uses many texts assessed by experts and utilizes them as training data in various machine learning algorithms, transforming the original readability problem from regression to a classification task. Collins-Thompson & Callan (2004) were among the early works on this new approach to statistical readability assessment. They tried to reform the problem of predicting the reading difficulty of a text in terms of statistical language modeling. They derived a measure based on an extension of multinomial naïve Bayes classification that combines multiple language models to estimate the most likely grade level for a given passage with better results compared to the classic readability formulas. Schwarm & Ostendorf (2005)

extended this method using multiple language models, NLP outputs (parse trees, PoS tags) and some 'classic' readability features (sentence length and word length). They used a linear SVM and found that trigram statistical language models combined with all the other features increase readability estimation accuracy. Pitler & Nenkova (2008) enhanced the pool of features used in readability studies by adding discourse-based features. The experiments with discourse features demonstrated promising results in predicting the readability level of text for both classification and regression approaches.

Recent advancements in deep learning and word-embedding models have also started to influence readability analysis. A number of studies have already applied various deep neural network topologies and transformer models in order to estimate automatically readability in monolingual (Martinc et al. 2018; Mohammadi & Khasteh 2019) and multilingual texts (Azpiazu & Pera 2019) with reported success.

3. Methodology

3.1 Corpus

Our main research aim is to evaluate MOGRead in terms of effectiveness and efficiency and propose novel methods that will increase the tool's accuracy. This will be obtained by testing new machine learning algorithms and simultaneously exploring the effectiveness of new features that should be language independent and easily calculated. Also, we aim for a text classification model that will not be black box, but it will be interpretable so that we can explain the way our features interact with text readability levels.

In order to assess the MOGRead readability model and develop a competitive and enhanced version, we had to retrain our new tool with texts that have been manually classified into all Modern Greek language levels and describe a fairly representative sample of texts produced in each level from learners of Modern Greek as L2. For this reason, we created two distinct corpora of plain texts: (a) a set of texts ($N = 301$), labeled according to their language level A1 to C2 following the Common European Framework for languages, available online on the Portal for teaching Modern Greek as L2, and (b) a testing corpus ($N = 35$) that we have annotated and classified to the corresponding language levels. As for the analysis of both corpora by quantitative methods, we used QUITA (Kubát et al. 2014). In order to avoid serious class imbalance biases in algorithm training, we tried to keep the number of texts per language level similar. The training corpus statistics per language level can be seen in the following table (Table 1):

Table 1. Training corpus descriptive statistics

Language Level (Broad Categories)	Language Level (Detailed Categories)	Texts	Words	SD	Min	Max
A	A1 (8–12)	32	3,767	30.4	86	213
	A1	35	4,031	22.3	84	159
	A2	72	10,728	31.2	102	257
B	B1	40	11,199	80.2	162	464
	B2	40	19,644	175.8	208	1105
C	C1	44	28,927	225.3	213	1063
	C2	38	27,455	218.4	197	1114
Total		**301**	**105,751**			

3.2 Features

Readability is a highly complex perception phenomenon which is based on many dynamically interacting factors related mainly to the text and its appearance. After many decades of research, there is a consensus (Milone 2014) that the most relevant linguistic features are related to the following two general areas of complexity: (a) complexity of the vocabulary in terms of size and difficulty measured in various ways and (b) complexity of the syntax measured again as sentence length or other more sophisticated approaches to sentence structure.

In this paper we want to use all the classic measurements in word and sentence length but also approach the vocabulary difficulty in a novel way and utilizing indices inspired by work done in the field of quantitative linguistics. More specifically, the following list of features has been calculated using specialized software for quantitative linguistic indices (Kubát et al. 2014) and custom scripts:

- *h*-point: This index represents the 'bisector' point of the rank ~ frequency distribution at which rank = frequency. It was originally proposed by Jorge E. Hirsch for scientometrics (Hirsch 2005), introduced into linguistics by Popescu (2007) and further developed by Popescu et al. (2007). Using the definition mentioned above, the *h*-point splits the vocabulary into two basic parts, namely into a class of magnitude *h* of frequent function words (synsemantics) and a much larger class of content words (autosemantics) with size *V*-*h* which are not so frequent but build the bigger part of the text's vocabulary (Popescu et al. 2009a: 19).
- Entropy (*H*): The term 'entropy' has been used in many scientific disciplines with different meanings, mainly defining quantitatively the diversity of the uncertainty of a system. In this paper, entropy is calculated using the Shannon

formula on the word frequencies of the corpus (Oakes 1998: 59). Using this definition, texts with big vocabularies and low frequencies produce high entropies while texts with controlled vocabularies and formulaic or systematic word usage exhibit lower entropy.

- Yule's characteristic *K*: A measure of vocabulary 'richness' based on the work of Yule (1944). The index measures the lexical repeat-rate and has been found to be sufficiently robust regardless of the text size from which it is calculated (Tweedie & Baayen 1998).
- Writer's view: An index proposed by Popescu & Altmann (2007) that is connected to the golden ratio ($\varphi \approx 1.618$). It is defined as the angle that is formed between the word frequency ~ rank distribution end and its top, as seen from the *h*-point. Popescu et al. (2009b: 26), commenting on its name, claim that it is "baptized in this way because one can imagine the writer "sitting" at this point and controlling the equilibrium between autosemantics and synsemantics".
- R1: An index of vocabulary richness proposed in Popescu et al. (2009a: 29–34) which is based on the *h*-point and the cumulative relative frequencies up to the *h*-point.
- Repeat Rate (*RR*): The repeat rate shows a degree of vocabulary concentration in a text measuring vocabulary richness inversely. So the higher RR is, the less vocabulary diversity a text has (Kubát et al. 2014).
- Relative Repeat Rate of McIntosh (*RRmc*): This is a normalized index of Repeat Rate so that it takes values in the interval < 0, 1 >, originally proposed by McIntosh (1967).
- Curve Length (*L*): A vocabulary richness index based on the curve of the rank ~ frequency distribution, defined as the sum of Euclidean distances between all points on the curve (Kubát et al. 2014).
- Curve length R Index (*R*): A vocabulary richness index derived from the curve length (*L*). It is defined as the ratio of the curve length above the h-point to the whole curve length (Kubát et al. 2014).
- Adjusted Modulus (*A*): A frequency structure indicator which is supposed to be independent of text length (Popescu et al. 2010).
- Gini Coefficient (*G*): A measure of statistical dispersion originally developed for econometric analysis, based on the Lorenz curve. It can be used as a measure of vocabulary richness by taking into account the rank ~ frequency distribution reversing the rank order (Popescu et al. 2009a: 54–63).

It is the first time that most of the above indices have been used as readability predictors. Moreover, 7 more 'classic' features (Standardized TTR, SD TTR, Average Word Length, SD Average Word Length, Average Sentence Length, SD Average Sentence Length, Numbers in the text) were calculated additionally for our corpus.

3.3 Machine learning algorithm: Random Forest

We experimented with various machine learning algorithms, but we selected Random Forest, an ensemble (i.e., a collection) of unpruned decision trees (Breiman 2001). Random forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees and can be used for classification or regression. It uses randomness in two levels: (a) a random sampling of training data points when building trees and (b) random subsets of features considered when splitting nodes. More specifically (Koehrsen 2018):

a. Each tree in a random forest is trained from a random sample of the data points. The samples are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree. The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest will have lower variance but not at the cost of increasing the bias. At test time, predictions are made by averaging the predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as bagging, short for bootstrap aggregating.
b. The other main concept in the random forest is that only a subset of all the features are considered for splitting each node in each decision tree. Generally, this is set to the square root of the features used for classification meaning that if there are 16 features, at each node in each tree, only 4 random features will be considered for splitting the node.

Using Random Forest as a classification algorithm has a lot of advantages, especially when we are using biased and unbalanced data as very frequently is the case with readability studies. Below are the most prominent ones (Kho 2018):

– Parallelizability: They are parallelizable, meaning that we can split the process over multiple machines to run it. This results in faster computation time. Boosted models are sequential in contrast and would take longer to compute.
– Suitability with with high dimensional data: Random forests are great with high dimensional data since we are working with subsets of data.
– Quick prediction/training speed: It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features. Prediction speed is significantly faster than training speed because we can save generated forests for future uses.

- Robust handling of outliers and non-linear data: Random forest handles outliers by essentially binning them. It is also indifferent to non-linear features.
- Handling of unbalanced data: It has methods for balancing error in class population unbalanced data sets. Random Forest tries to minimize the overall error rate, so when we have an unbalanced data set, the larger class will get a low error rate while the smaller class will have a larger error rate.
- Low bias, moderate variance: Each decision tree has a high variance, but low bias. But because we average all the trees in Random Forest, we are averaging the variance as well so that we have a low bias and moderate variance model.

As classification target, we set the broad categories of the levels of language competence (A, B, C) instead of the more detailed ones (A1, A2, B1 … C2) since the dataset was not big enough to train the model sufficiently for each detailed language competence category.

The algorithm parameters were optimized using different values for the number of features used in the repeated sampling; the evaluation of the algorithm fit was based on accuracy in 10-fold cross-validation.

4. Results

The enhanced version of our readability model achieved higher cross-validated accuracy compared to the standard MOGRead tool and an older version of the same tool that used only word and sentence length as parameters. More specifically, our approach obtained 0.943 accuracy, while the MOGRead obtained 0.914 and the older tool 0.857. Moreover, the Random Forest algorithm produced a detailed list of each feature in the model and its importance in the classification accuracy for each language level. The analysis can be seen in Figure 1.

From the inspection of the graph above, we can extract valuable information regarding the role of different features in the perception of text difficulty across different levels of linguistic competence. Average Word Length (AWL) is one characteristic example since it appears to be more important for readability prediction in the A and C levels of linguistic proficiency but not in the middle level B. On the contrary, Average Word Length standard deviation (AWL_sd) seems to follow a linear developmental path and becomes increasingly important as the language level of the text increases. This kind of meta-analysis can help us not only understand how the model fits our data but also what the features are with the most dynamic interaction with our classification categories.

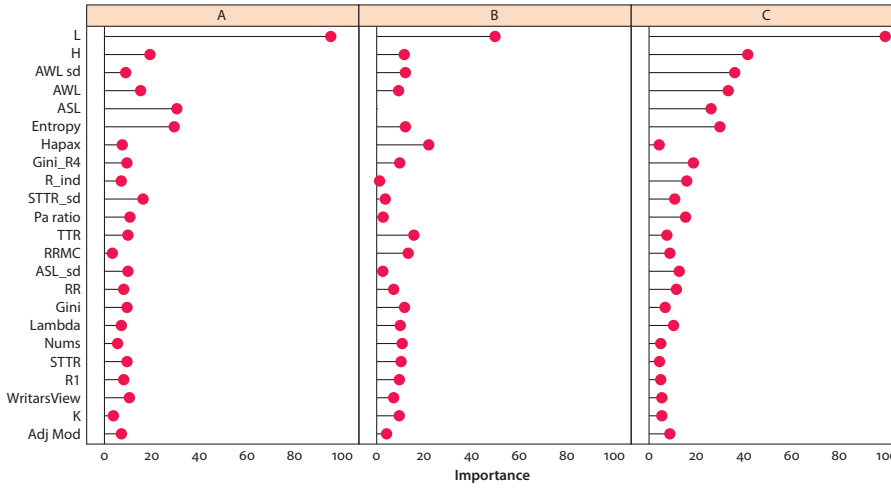


Figure 1. Feature significance plot for each class of the language level

A more detailed picture emerges as we examine the 10 most important features in each language competence level. Table 2 shows the ranking per language competence level:

Table 2. 10 most important features per language competence level

Rank	A	B	C
1	L	L	L
2	ASL	Hapax	H
3	Entropy	TTR	AWL_sd
4	H	RRMC	AWL
5	STTR_sd	AWL_sd	Entropy
6	AWL	H	ASL
7	Pa_ratio	Entropy	Gini_R4
8	ASL_sd	Gini	R_ind
9	WritersView	Nums	Pa_ratio
10	Gini_R4	STTR	ASL_sd

From the ranking above, it can be inferred that the L index is highly correlated with readability across all levels. Readability judgments in the most basic level (Level A) are also based on the Average Sentence Length (ASL), Entropy, and h index (H). In the intermediate level (Level B) Hapax Legomena, Type-Token Ratio (TTR), and the normalized Repeat Rate (RRMC) predict better readability. The most advanced language competence level (Level C) can be predicted by h index (H), average word length (AWL), and its standard deviation measurement (AWL_sd).

The different ranking of these features across language levels hides a complex and fluid interaction of their behavior regarding their predictive ability. In order to understand how these indices are related to the readability judgments, we must not only rank them in terms of their predictive power but also profile their overall behavior across their distribution. In order to explore these dynamics, we used a novel Random Forest visualization tool, Forest Floor, which moves its mapping from feature space to prediction space. This tool uses feature contributions, a method to decompose trees by splitting features, and then subsequently performing projections. The advantages of Forest Floor over partial dependence plots is that interactions are not masked by averaging. Therefore, it is possible to locate interactions, which are not visualized in a given projection (Welling et al. 2016). The resulting visualization can be seen below (Figure 2):

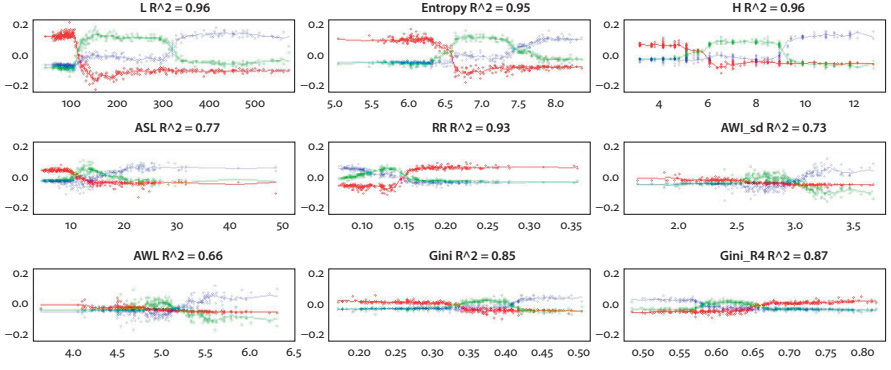


Figure 2. Interaction plots of the most important stylometric features across their distribution in the 3 language competence levels (Level A: Red, Level B: Green, Level C: Purple)

These plots contain the interaction of the most important features (measured by their R^2 value) across the 3 levels of language competence. In order to evaluate the interpretive power of these visualizations we should examine how a well-understood feature is behaving. Average Sentence Length is a characteristic case of a complex interaction with the readability and the language levels. We can see that small sentences (low values in the x axis) are correlated with high readability prediction in the beginner language learners (A level: red line). However, as sentence length increases there is a crosscut of the purple line (Level C) at the value of 20. That means that sentences over 20 words are becoming better predictors of the readability scores in higher language competence levels.

A second example can be offered in the plot of the entropy. Smaller values of lexical entropy (more predictable vocabulary) predict better readability in the

lower level of language competence (Level A). As the entropy of the text increases (gets over 6.5), it becomes a better predictor of readability in the B level. Then, as the entropy increases more and gets over 7.6, another crosscut happens and the entropy becomes a better readability predictor for the highest language competence level (Level C).

It is evident that a static analysis of the features' importance in the readability prediction cannot uncover this complex fluid dynamics hidden between the readability and the factors that are related to it. Due to this explorative visualization method we managed to understand that different entropy values relate with different language competence levels and that the increase of the text entropy associates with better prediction of text readability in higher language competence levels.

5. Conclusion

In this study we presented an enhanced version (compared to an existing tool) of a text readability analysis method for learners of Modern Greek as L2. We showed that readability modeling can be enhanced using quantitative linguistics indices of lexical differentiation and machine learning algorithms which are based on features that can capture more general quantitative properties of language and are language independent.

More importantly, it became evident that readability modeling as well as any other linguistic-related behavior should not be treated as a black box procedure. No matter how accurate the machine learning models we are developing are, we need to be able to explain our models, understand our misclassifications, and interpret how our features interact dynamically with our classification categories.

Funding

The funding of MOGRead was based in the framework of the programme "Greek language attainment: Support and qualitative promotion of teaching/learning Greek as a foreign/second language", implemented by the Division for the Promotion and Support of the Greek Language of the NSRF operational programme "Education and Lifelong Learning" of the National Ministry of Education (2011–2015).

References

Azpiazu, Ion Madrazo & Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics* 7. 421–436. https://doi.org/10.1162/tacl_a_00278

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32. https://doi.org/10.1023/A:1010933404324

Collins-Thompson, Kevyn & James P. Callan. 2004. A language modeling approach to predicting reading difficulty. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 193–200. Boston, MA: Association for Computational Linguistics.

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin* 27(2). 37–54.

DuBay, William H. 2004. *The principles of readability*. Costa Mesa, CA: Impact Information.

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology* 32. 221–233. https://doi.org/10.1037/h0057532

François, Thomas & Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In Jun'ichi Tsujii, James Henderson & Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–477. Jeju Island, Korea: Association for Computational Linguistics.

Fry, Edward. 1968. A readability formula that saves time. *Journal of Reading* 11(7). 513–578.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2). 193–202. https://doi.org/10.3758/BF03195564

Gunning, Robert. 1952. *The technique of clear writing*. New York: McGraw-Hill.

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

Kho, Julia. 2018 October 19. Why random forest is my favorite machine learning model. *Towards Data Science*. Retrieved 5 September 2020, from https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

Kincaid, Peter J., Robert P. Fishburne , Jr., Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel*. Millington, TN: Chief of Naval Technical Training Naval Air Station Memphis. https://doi.org/10.21236/ADA006655

Koehrsen, Will. 2018 August 30. An implementation and explanation of the random forest in Python. *Towards Data Science*. Retrieved 5 September 2020, from https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Kubát, Miroslav, Vladimír Matlach & Radek Čech. 2014. *QUITA: Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Martinc, Matej, Senja Pollak & Marko Robnik Šikonja. 2018. Assessing readability with deep neural language models. Paper presented at the 2nd HBP Student Conference: Transdisciplinary Research Linking Neuroscience, Brain Medicine and Computer Science, Ljubljana, Slovenia, February 14–16.

McIntosh, Robert P. 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology* 48(3). 392–404. https://doi.org/10.2307/1932674

McLaughlin, G. Harry. 1969. SMOG Grading – a new readability formula. *Journal of Reading* 12(8). 639–646.

Milone, Michael. 2014. *Development of the ATOS® Readability Formula*. Wisconsin Rapids, WI: Renaissance Learning, Inc.

Mohammadi, Hamid & Seyed Hossein Khasteh. 2019. Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957*.

Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Pitler, Emily & Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In Mirella Lapata & Hwee Tou Ng (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 186–195. Honolulu, HI: Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613742

Popescu, Ioan-Iovitz. 2007. The ranking by the weight of highly frequent words. In Peter Grzybek & Reinhard Köhler (eds.), *Exact methods in the study of language and text*, 555–565. Berlin: De Gruyter. https://doi.org/10.1515/9783110894219.555

Popescu, Ioan-Iovitz & Gabriel Altmann. 2007. Writer's view of text generation. *Glottometrics* 15. 71–81.

Popescu, Ioan-Iovitz, Karl-Heinz Best & Gabriel Altmann. 2007. On the dynamics of word classes in text. *Glottometrics* 14. 58–71.

Popescu, Ioan-Iovitz, Gabriel Almann, Peter Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová & Matummal N. Vidya. 2009a. *Word frequency studies*. Berlin: Mouton de Gruyter.

Popescu, Ioan-Iovitz, Ján Mačutek & Gabriel Altmann. 2009b. *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Popescu, Ioan-Iovitz, Ján Mačutek, Emmerich Kelih, Radek Čech, Karl-Heinz Best & Gabriel Altmann. 2010. *Vectors and codes of text*. Lüdenscheid: RAM-Verlag.

Schwarm, Sarah E. & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In Kevin Knight, Hwee Tou Ng & Kemal Oflazer *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 523–530. Ann Arbor, MI: Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219905

Tweedie, Fiona J. & Harald R. Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352. https://doi.org/10.1023/A:1001749303137

Welling, Soeren H., Hanne H. F. Refsgaard, Per B. Brockhoff & Line H. Clemmensen. 2016. Forest floor visualizations of random forests. *arXiv preprint arXiv:1605.09196*.

Yule, George Udny. 1944. *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Phonological properties as predictors of text success

Jiří Milička and Alžběta Houzar Růžičková
Institute of the Czech National Corpus, Charles University /
Institute of Phonetics, Charles University

In this study, the relation of phoneme structure and success rate of Czech texts is investigated. The study is based upon two phenomena: (1) the beauty-in-averageness effect and (2) the euphony principle. The main objective is to examine whether one or the other prevails in the perception of written text, whether they interact in some way, or whether the phoneme structure has no major effect on the text's appeal to the readers. Several phoneme groups, both vowel and consonant, are focused on.

Keywords: quantitative linguistics, corpus, phonology, phonetics, internet mediated communication, euphony, beauty in averageness, statistical study of literary vocabulary, Czech

1. Introduction

This paper deals with supply and demand in the communication process. Production and perception are two sides of the same coin – the form of language and the properties of text are products of human abilities in both. The quantitative linguistic theories (e.g., the Zipfian principle of least effort (Zipf 1949), the Altmannian theory proposal (Altmann 1978) and the Köhlerian synergetic control circle (Köhler 1986), etc.) take into account that language is a trade-off between the demands and capabilities of the text receiver and the capabilities of the producer. In other words, the producer tries not only to minimize effort, but also to maximize success of the communication and therefore accommodates their texts to satisfy the needs of the receiver.

While corpora are mostly utilized to study production of texts, there is no inherent reason why they could not be used to study perception. A typical corpus consists of natural texts. These naturally occurring texts are usually intended to be consumed by some more or less specific group of receivers ('target audience' or at

https://doi.org/10.1075/cilt.356.12mil

least 'model reader' as defined by Umberto Eco (1984: 7)) and some information about the reception of the text among the intended readers is usually available – more or less precise, more or less specific.

Once we accept that not all texts were created equal and include the meta-information describing the text success into the corpus, we can proceed to research how the capabilities of the producers meet the demands of the receivers. In this study, we focus on phonological features.

Here are the effects we expect should play a role in shaping the results.

1.1 Beauty-in-averageness effect

When evaluating texts' appeal related to their phonological properties, there is a principle that might come into play, a phenomenon referred to as the beauty-in-averageness effect (Winkielman et al. 2006).

The phenomenon of average or typical members of a given category being perceived as the most positive ones was first observed in the case of human faces: the most average faces have been found to be rated as the most attractive ones (Langlois & Roggman 1990; Trujillo et al. 2014). A similar principle has also been observed in non-human visual inputs such as birds, fish, automobiles (Halberstadt & Rhodes 2003), or even abstract patterns (Winkielman & Cacioppo 2001; Winkielman et al. 2006). These studies conclude that average or typical stimuli are cognitively easier to process, which leads to more positive responses by the perceiver.

With that said, it should be mentioned that in the field where this phenomenon was first observed, it apparently does not apply fully. According to the findings of DeBruine et al. (2007), the most average faces are, in fact, not the most attractive ones. Although averageness has been found to be one component of attractiveness, there is another dimension defining faces' attractiveness which is independent of the effects of averageness: "moving away from average in one direction along this dimension will increase attractiveness, while moving away from average in the opposite direction will decrease attractiveness" (DeBruine et al. 2007: 1428). This effect applied to the phonological level of a literary text is traditionally called 'euphony'.

1.2 Effect of euphony

Phonemic properties of texts have been widely worked with in poetry, employing the concept of euphony, based on the idea that some phonemes are more pleasant than others; it is the repetition of such phonemes that makes a text euphonic (in contrast to cacophony which is caused by higher frequency of unpleasant phonemes). According to Perrine (1972: 258), who briefly summarizes the long tradition of euphony in poetry since Dionysius of Halicarnassus, euphonic phonemes are for example vowels, which are considered more euphonic than consonants, "with

longer vowels also being preferred to shorter", or liquids, nasals, and semi-vowels /l m n r v w/ in comparison to other consonants (In this context, Perrine (1972) classifies /v/, or rather the segment represented by the grapheme *v*, as a semi-vowel, not specifying in which language. However, McMahon (2002) categorizes the English /v/ as a fricative.)

From the large body of sound iconicity literature, we call attention to the classic work of Thorndike (1945), who mapped occurrence of individual phonemes in words of "pleasant and unpleasant meanings" in six European languages. Based on his findings, Thorndike argues that "in each of the six languages investigated there is an association of certain sounds with pleasant meanings and of other sounds with unpleasant meanings" (Thorndike 1945: 145). He hypothesizes that there is a general connection between the pronunciation ease of the sounds and their pleasantness; however, he also finds language-specific effects.

The pleasantness of individual phonemes has also been investigated experimentally, e.g., by Whissell (1999, 2000) who focused on the correlation of phonemes' frequency and the emotion of a text sample in English. Her results show that some phonemes, such as /l/, high vowels or front vowels occur more often in more pleasant texts using soft or tender language, while other phonemes, such as /r/, velar plosives, low vowels or back vowels had higher frequency in less pleasant texts. Looking for the reason why some phonemes appear as more pleasant than others, Whissell argues that it lies in the nature of their articulation:

> […] facial feedback theorists claim that smiling, even without intention or awareness, will make a person happier because smiling is the muscular pattern associated with happiness and the brain interprets emotion on the basis of expression. By the same argument, pronouncing the phoneme /iy/[1] might make a person fractionally happier because the motor act of producing it imitates a smile in many respects […]. […] sounds that are produced towards the back of the throat (including /k/ and /g/) share some of the muscular responses characteristic of the negative and active emotions of disgust and anger. Whissell (1999: 43)

1.3 Spoken data vs. written corpora

A question arises: to what extent, if at all, is it possible to expect any effect of phonological structure in written texts? Traditionally, studies focusing on both production and perception of phonemes' realizations have investigated their phonetic properties in spoken material.

In written texts, it is not possible to measure phonetic properties of individual phoneme realisations; therefore such data may seem unfit for phonetic domain

1. /iː/ in the IPA

research. On the other hand, the advantage of written corpora in general is that they can offer huge quantities of data of different kinds – not only occurrences of individual units, but also readers' reactions to them, such as 'likes' or 'views' on websites. Thanks to these data, it is possible to study not only texts' own properties (production), but also what impression the texts give to the reader (perception). In traditional research on phonetic phenomena, studying both speech production and perception must cope with very limited data; getting individual speakers' recordings of reasonable quality, as well as performing perception tests under suitable conditions, is very time-consuming, compared to obtaining analogous data using written language. That is one reason why trying to observe phonetic or rather phonological features of written texts appears to be an interesting subject. Another reason is finding parallels between spoken and written language; does a reader perceive a written and spoken text similarly?

2. Data

The corpus consists of blogs downloaded from blog.idnes.cz. The blogging site belongs to the mainstream Czech news publisher MAFRA which to some extent determines its content and readership. Politically, mostly mainstream or slightly conservative texts are about politics and everyday life topics. The site does not target any specific generation but the topics are suitable rather for the adult audience.

The corpus covers the time span from 2/5/200712:00 to 28/5/2017 19:38 and comprise 344,000 texts, (181,450,572 tokens) which were viewed 517,819,018 times. The texts are almost exclusively written in the Czech language. The corpus is unfortunately not balanced as for gender: there are ca. 2787 female authors (68,280 texts) and 5655 male authors (213,602 texts). The texts assigned to institutions and blog masters were sorted out.

All texts were automatically transcribed to a simplified phonological transcription.

3. Method

3.1 Observed units

We examined the individual phonemes' or phoneme classes' relation to the text success. We selected some of the phoneme categories mentioned in the existing studies (see § 1) to compare our results with findings acquired based on different language or genre material.

In our study, phonemes are treated as abstract units related to some general mental representations of speech sound categories. Although we employ distinctive features as cues to categorize phonemes, we do not consider them the sole characteristics of phonemes (unlike, for example, Generative Phonology or Prague School; see Dresher 2011). Instead, we approach phonemes from the Exemplar Theory point of view, i.e., as complex and detailed units formed by individual experienced realizations (Pierrehumbert 2000). With such a premise, we expect the readers' mental representations of phonemes to be activated during perceptual processing of written texts.

To test whether principles concerning vowel appeal described by Perrine (1972) and Whissell (1999, 2000) apply in Czech prosaic material, we examined the vowel ratio in proportion to all phonemes. Diphthongs were considered single phonemes. As the Czech phonological system involves vowel quantity as a distinctive feature, the long vowel ratio in proportion to all vowels was examined as well; diphthongs were left out of this analysis, as their classification as either long or short vowels is not unequivocal. We also focused on individual vowel quality classes: front vowels /ɪ iː ɛ ɛː/, back vowels /u uː o oː/, close vowels /ɪ iː u uː/, mid vowels /ɛ ɛː o oː/ and open central vowels /a aː/ in proportion to all vowels. Once again, diphthongs were not a part of this analysis.

We also observed text success relation to some consonantal phonemes' frequency; namely, we decided to examine sonorants, divided into several categories. First, the individual phonemes /r/ and /l/ in non-syllabic positions were analyzed, to see how their influence on text success in Czech differs from Whissell's (1999, 2000) findings based on English material. Here we need to point out that the English and Czech /r/ properties differ: in English (British or American), /r/ is defined as a retroflex approximant, which can be not realized in some word positions (in non-rhotic accents); it can also be realized as an alveolar tap or rarely also as an alveolar trill (McMahon 2002), whereas Czech /r/ is classified as an alveolar trill in all positions (Šimáčková et al. 2012). In Czech, /r/ and /l/ can also be syllabic; we focused on such cases as well, in comparison to vowel syllabic nuclei.

Furthermore, the phonemes /r l m n/, which are considered the most euphonic consonants by Perrine (1972), were analyzed together, in proportion to all consonant phonemes; Perrine (1972) also names /v/ and /w/, but the Czech language lacks the /w/ phoneme and Czech /v/ is not a semi-vowel as it is described by Perrine (1972). As the 'euphonic' phonemes belong among sonorants, we also decided to analyze this class, containing the phonemes /l r m n ɲ j/, as a whole.

The relative frequencies of these units were observed, i.e., we were concerned with the ratio of vowels to all phonemes, ratio of open vowels to all vowels, sonorant ratio to all consonants, non-syllabic /r/ ratio to all non-syllabic consonants, syllabic /r/ and /l/ ratio to all syllabic nuclei, etc.

3.2 Success rate

Number of likes per view (abbreviated as LpV) was chosen as a main text success metric. LpV is not directly accessible to the readers of the blogging site (they can see only the number of views and 'karma', a metric which is straightforwardly derived from the number of likes). Nevertheless, the authors can see LpV statistics, and the trending texts on the homepage are sorted according to this number so it can play a role as the target metric for the authors.

3.3 Resampling

We measured the dependency of the text success rate on the various linguistic properties of the texts. As we were concerned not only with mere dependency function, but also with the confidence intervals of this function, we made 10,000 resamples of our corpus.

In accordance with the classical bootstrapping methodology (Efron 1987), each resample consisted of 344,000 texts chosen randomly with repetition from our corpus, i.e., the resample contained the same number of texts as the original corpus. Then, a continuous chunk of text of a certain length was randomly chosen from each resampled text, the properties were measured on the chunk, and the results were mapped to the success rate of the text.

The chunk length was set to be 500 statistical units (e.g., when studying the ratio of vowels, all phonemes are considered statistical units). Only a small number of texts in the corpus had fewer units and those texts had to be omitted.

4. Results

The results for each phoneme or phoneme class observed can be seen in the figures below. Each figure describes two different variables: (1) likes per view (LpV) represented by a line chart and (2) distribution of texts with a given proportion of observed units expressed by a histogram. The two variables use different units; nevertheless they share the same scale.

The LpV line chart in each figure is accompanied by a 95% confidence interval. At the edges, the intervals are wider and the line charts are less reliable, as there is less data available.

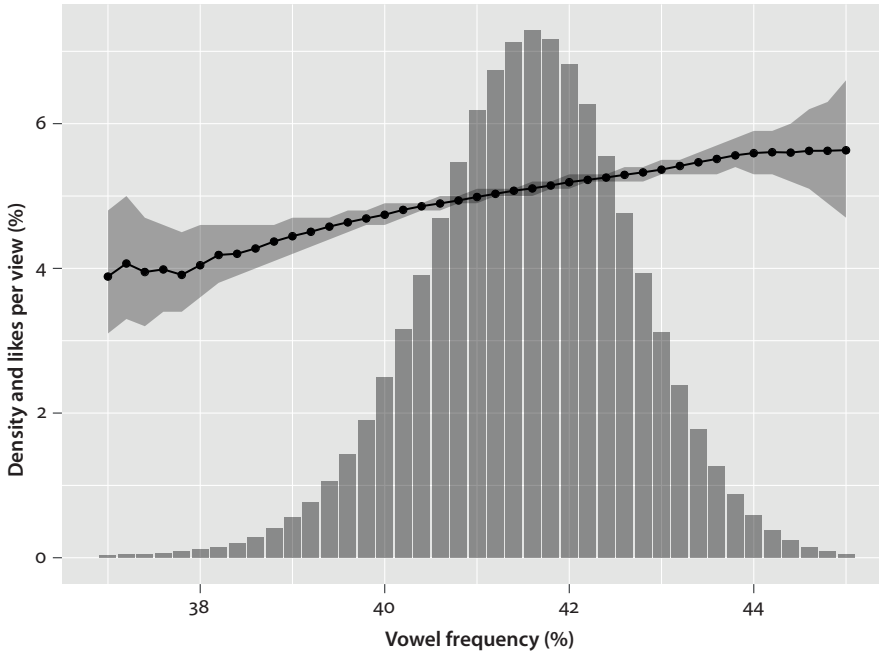


Figure 1. Dependency of LpV on the vowel ratio alongside the vowel ratio distribution

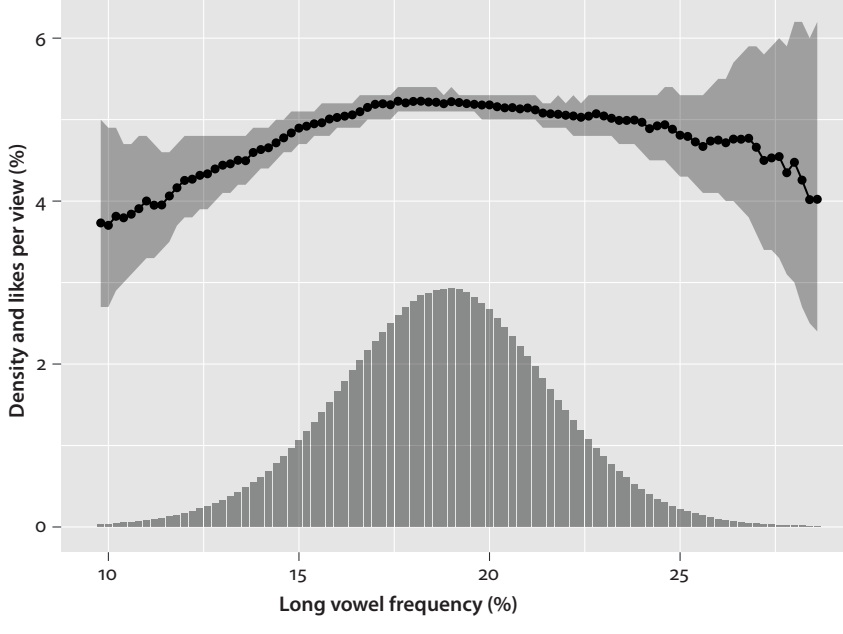


Figure 2. Dependency of LpV on the long vowel ratio alongside the long vowel ratio distribution

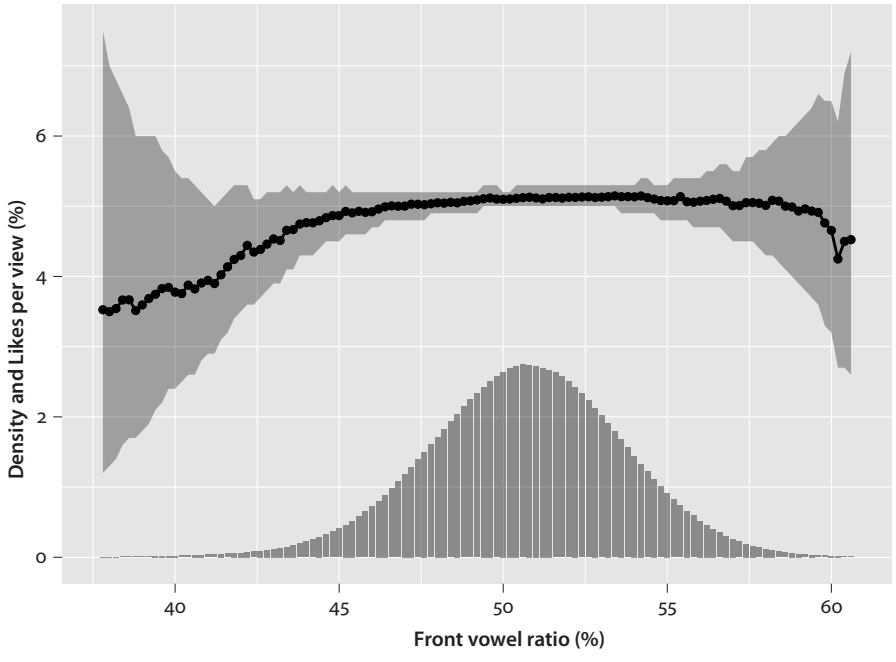


Figure 3. Dependency of LpV on the front vowel ratio alongside the front vowel ratio distribution

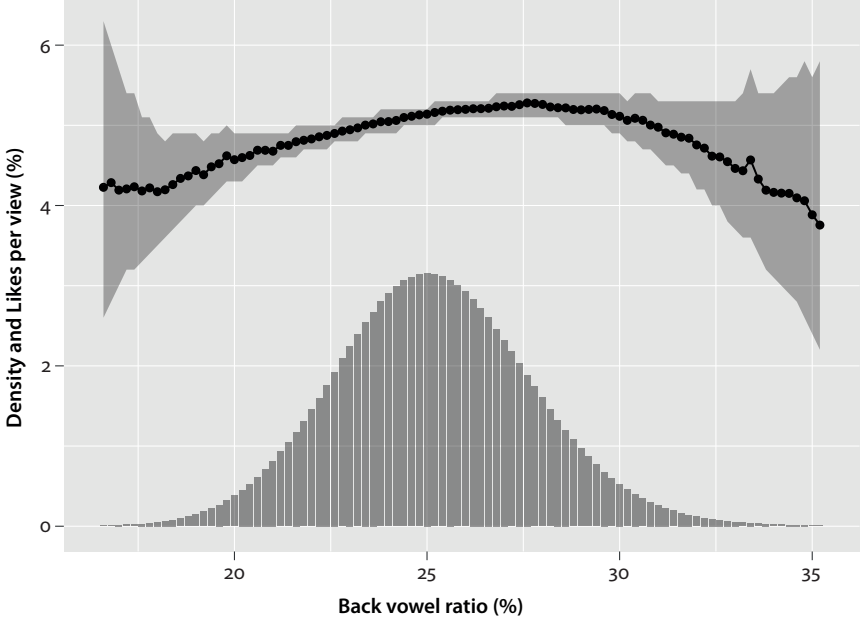


Figure 4. Dependency of LpV on the back vowel ratio alongside the back vowel ratio distribution

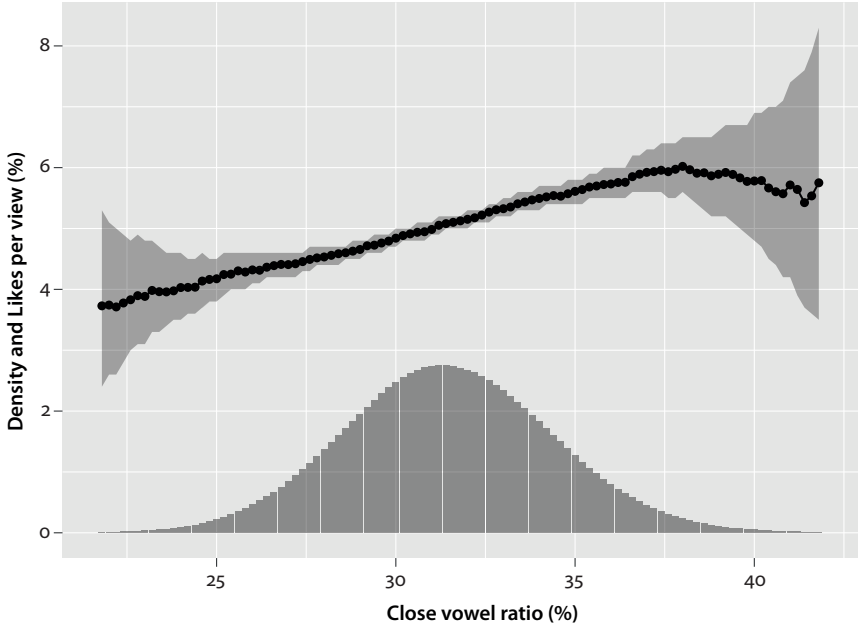


Figure 5. Dependency of LpV on the close vowel ratio alongside the close vowel ratio distribution

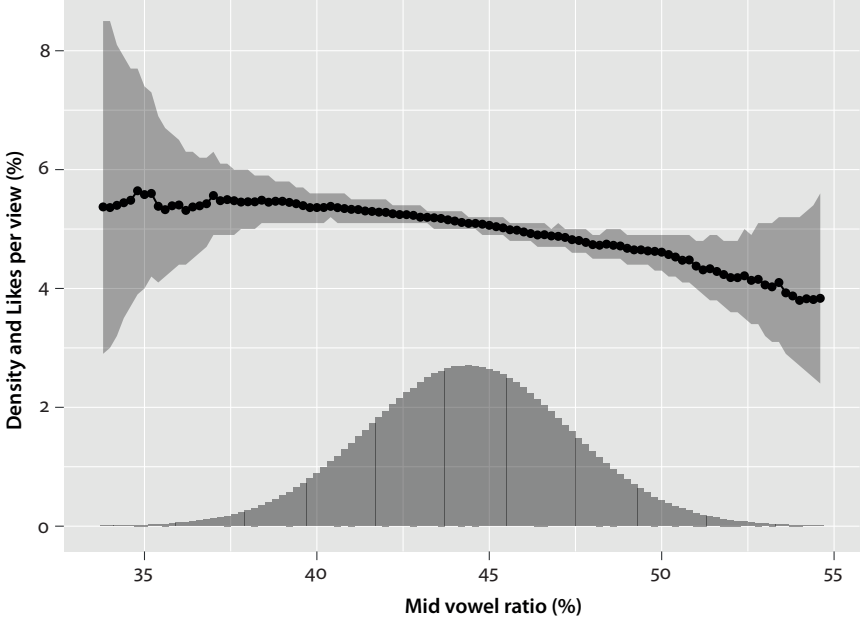


Figure 6. Dependency of LpV on the mid vowel ratio alongside the mid vowel ratio distribution

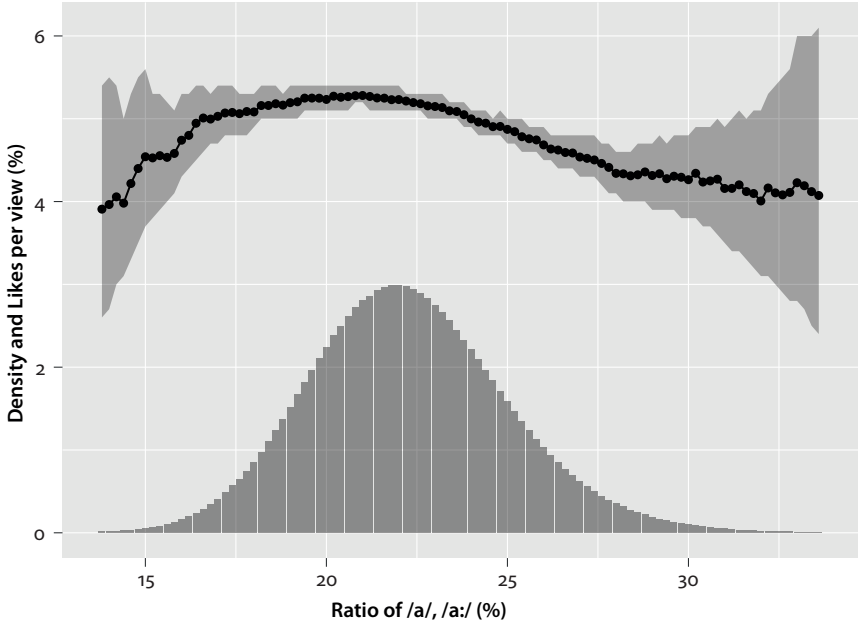


Figure 7. Dependency of LpV on the ratio of /a/ and /a:/ alongside the ratio of /a/ and /a:/ distribution

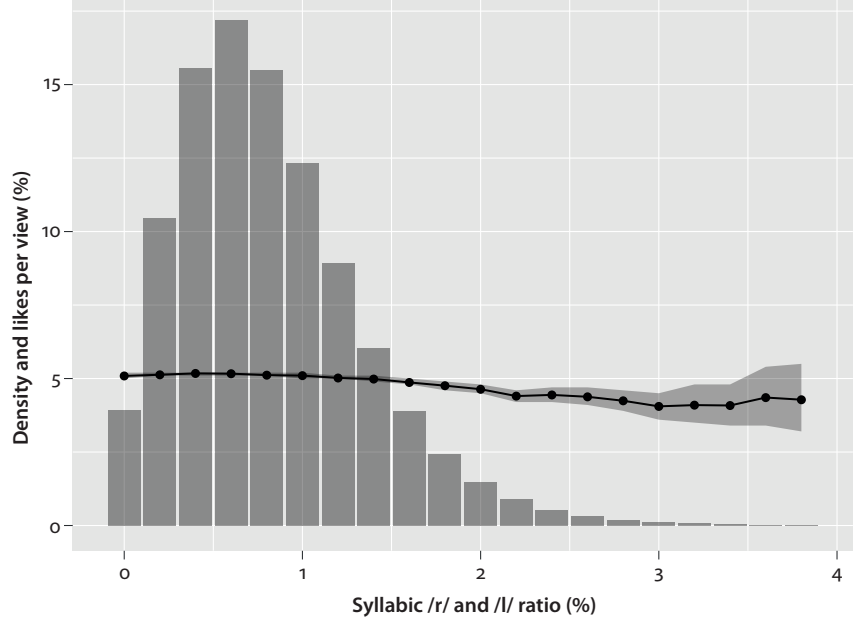


Figure 8. Dependency of LpV on the syllabic /r/ and /l/ ratio alongside the syllabic /r/ and /l/ ratio distribution

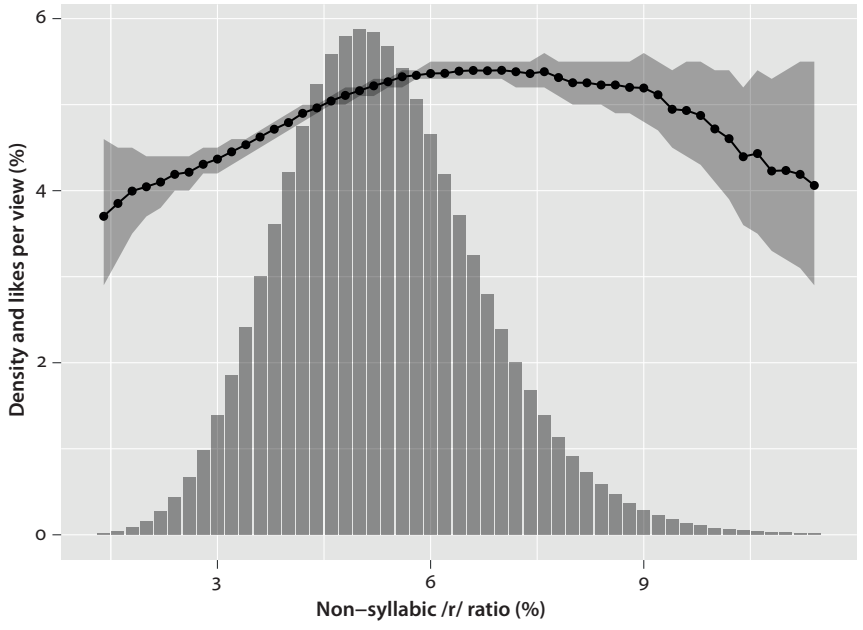


Figure 9. Dependency of LpV on the non-syllabic /r/ ratio alongside the non-syllabic /r/ ratio distribution

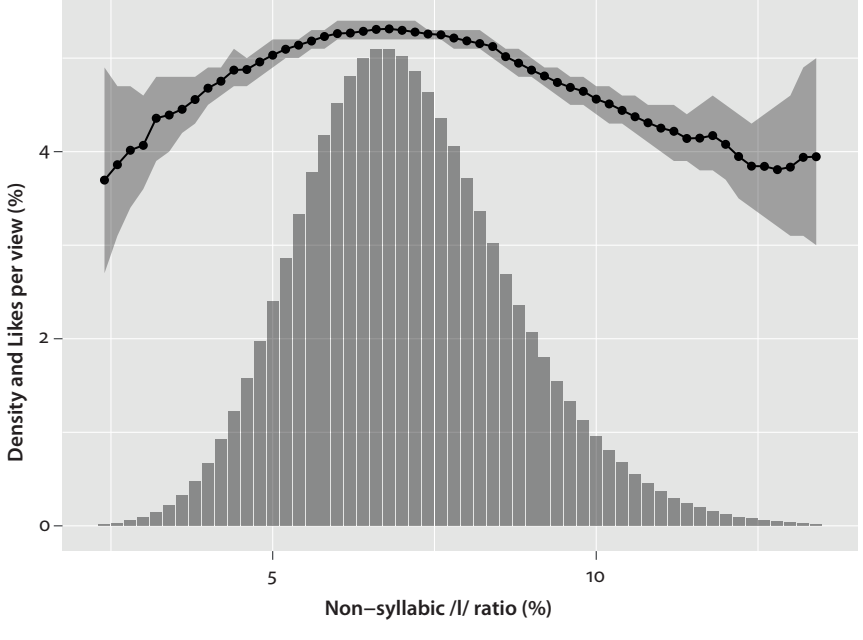


Figure 10. Dependency of LpV on the non-syllabic /l/ ratio alongside the non-syllabic /l/ ratio distribution

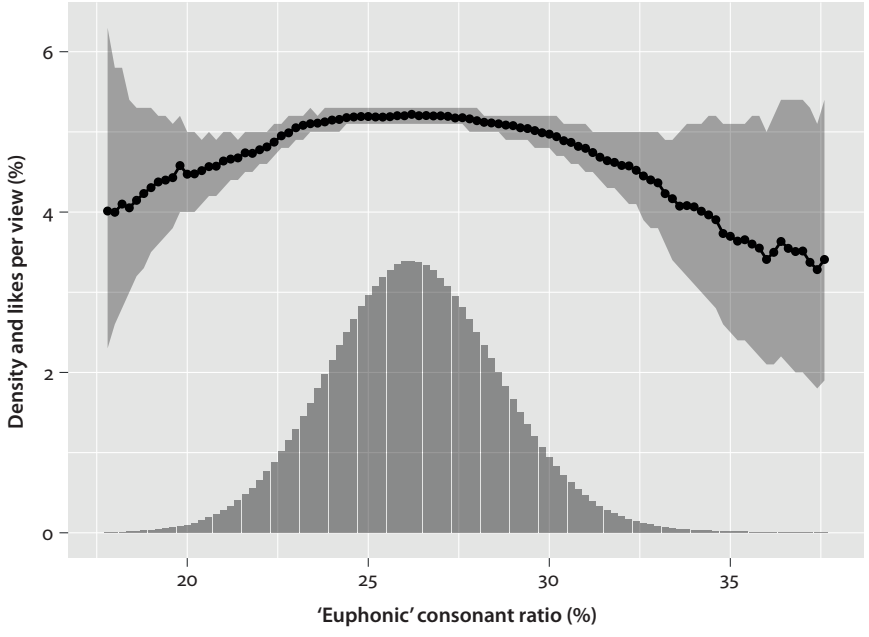


Figure 11. Dependency of LpV on the 'euphonic' consonant ratio alongside the 'euphonic' consonant ratio distribution

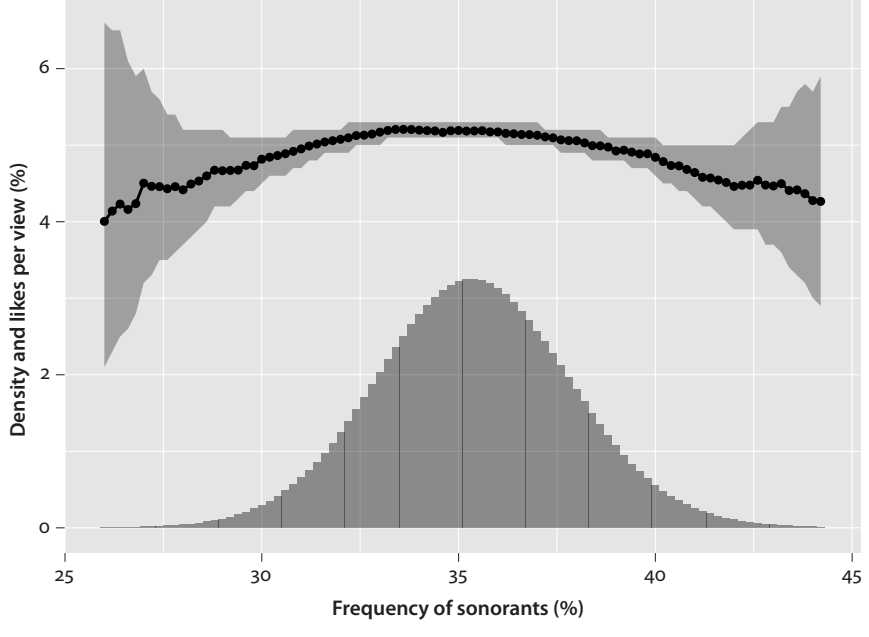


Figure 12. Dependency of LpV on the sonorant ratio alongside with the sonorant ratio distribution

Following principles described in the § 1 we can observe four main tendency types in our results: the majority of phonemes observed actually follow the euphony (or cacophony) principle, while several other phonemes' relation to the text success could be assigned to the beauty-in-averageness effect. In two phonemes, the two principles mentioned apparently come into an interaction. Only in one case, there is no apparent dependency of likes per view ratio on the phoneme frequency.

The euphony principle is quite prominent in the overall vowel ratio (Figure 1); the more vowels there are, the more successful the text is; in the figure, we can see a rising tendency from about 37.5% of vowels among all phonemes and there is no apparent falling tendency in higher vowel percentage. Moreover, the actual effect size is considerable: raising the vowel rate by 10% increases the average number of likes approximately by 20%. We can observe a similar tendency in back vowels (which are also rounded: Figure 4), close vowels (Figure 5) and non-syllabic /r/ (Figure 9) – in these three cases, the success plot appears to be falling in higher percentages of the segments; however, due to the smaller amount of data for such occurrences, the confidence intervals are too wide for the falling tendency to be affirmed with certainty. The phonemes (or phoneme classes) mentioned above can therefore be considered euphonic. An inverse principle, i.e., cacophony, appeared as well, namely in the case of mid vowels (Figure 6) – apparently, the fewer mid vowels appearing in the text, the more appealing it is to the reader. This tendency can be also observed in /r/ and /l/ in syllabic positions (Figure 8); even though the falling function is not as prominent in the figure, according to the narrow confidence intervals it is statistically significant.

The beauty-in-averageness effect, i.e., the text distribution mode meeting the peak of LpV plot, applies in the cases of /l/ in non-syllabic positions (Figure 10) and so called 'euphonic' consonants (Figure 11) as well as sonorants in general (Figure 12); in the two latter figures, according to the LpV plot, it could be assumed that its peak is shifted slightly towards the lower ratio of the observed phonemes compared to the texts' density, however, that cannot be said with absolute certainty, considering the confidence intervals.

The beauty-in-averageness effect and the cacophony (rather than euphony) principle appear to interact in long vowels (Figure 2) and open central vowels /a aː/ (Figure 7), where we can see the rising tendency to a certain point, where the tendency starts to fall again. The peak of the LpV plot, however, does not copy the text density mode, but it is significantly shifted to the lower phoneme ratio.

Only in the case of front vowels (Figure 3) is there no apparent influence of the phoneme ratio on text success. The LpV plot appears to follow the beauty-in-averageness principle; however, considering the confidence interval, the tendency cannot be considered significant.

5. Discussion

In this study, we examined the relationship of phoneme structure and success rate for Czech texts. We based our study upon two phenomena: (1) the beauty-in-averageness effect and (2) the euphony principle, and we examined whether one or the other prevail in perception of written texts, whether they interact in some way, or whether the phoneme structure has no major effect on the texts' appeal to the readers. We focused on several phoneme groups, both vowel and consonant. All of the four outcomes mentioned occurred in our results.

In the overall vowel rate, as well as in back and close vowels and non-syllabic /r/, the principle of euphony can be observed: the more represented they were in the text, the higher was the text's success (i.e., the more likes per view the text had). On the contrary, in mid vowels and syllabic /r/ and /l/, an opposite tendency can be observed: the fewer representations of those phonemes, the more successful the text; that can be described as the cacophony principle. The appeal of vowels to the reader corresponds with the assumptions of traditional literary studies which describe them as more euphonic than consonants (Perrine 1972). However, in the case of other phonemes we examined, these traditional assumptions differ from our findings (see below). As for the vowel quality and /r/, our findings partly differ from those of Whissell's (1999, 2000) studies, who found that in English, high vowels are perceived more positively by the readers, while /r/ and back vowels are less pleasant. However, our results suggest that Czech texts with a higher ratio of all those categories are more successful among readers.

In the case of non-syllabic /l/, 'euphonic' consonants, and sonorants, the beauty-in-averageness effect applied: the more typical their ratio was, the better the text was accepted by the readers. According to Whissell's (1999) findings, /l/ is linked to more positive emotion; therefore we could expect the texts with a higher /l/ ratio to be more successful – that, however, is not the case in the Czech written material. A similar observation can be made regarding the so-called 'euphonic' phonemes according to Perrine (1972), as well as sonorants in general.

Interaction of the euphony (or rather cacophony) principle and the beauty-in-averageness effect occurred in the case of long vowels and open central vowels. Perrine (1972) describes long vowels as being more euphonic than the short ones, but our results suggest that readers prefer texts with rather fewer long vowels than typical. According to Whissell (1999), low vowels tend to appear more in less pleasant texts. That corresponds with our finding, i.e., that texts with fewer open central vowels than typical are favoured by readers, but unlike mid vowels, the cacophony principle does not apply on its own – it interacts with the preference for typicality.

No apparent relationship of phoneme ratio and text success was found in front vowels, even though according to Whissell's (1999) findings, it occurs more often in more pleasant texts, so it could be assumed that readers would prefer texts with a higher ratio.

Based on our results, we can conclude that phoneme structure is related to the text success among readers; both the euphony principle and the beauty-in-averageness effect can be observed. Our study was performed on Czech prose; individual phonemes' relation to the text success can differ according to the language and genre. As Thorndike (1945) points out, pleasantness of individual sounds is language specific; that corresponds with the fact that our results differ from for example Whissell's (1999, 2000) findings based on English material.

Many aspects of our methodology were chosen arbitrarily; we selected phonemes and phoneme classes based on previous studies and the traditional phoneme categorization. For further research, it would be suitable to run corpus driven heuristics: analyses of each phoneme individually and trying all possible phoneme groupings, searching for interesting results automatically. This would also solve the problem of intersections of chosen classes as some phonemes may fall into several observed categories that show opposite tendencies, therefore we can expect some interference of those tendencies to appear.

Another phonological phenomenon we did not examine in this study was consonant clusters. We intend to study the relation of consonant cluster ratio and text success, taking into account what types of phoneme they consist of. In further research, we also want to examine the syllabic structure and its relation to text success, comparing, for example, open and close syllables. The question of syllabification is very complex; several approaches exist in this area (for more see Šturm 2017).

Only a limited number of possibly infinite methodological variations and operationalizations could be employed in our study: language and genre choice, success metric, window size, etc. Repeating the study several times with different settings is necessary to make more confident and more general conclusions.

Acknowledgement

This study was supported by the programme Progres Q08 "Czech National Corpus" implemented at the Faculty of Arts, Charles University. We are also grateful to Veronika Nováková for proofreading and Pavel Šturm and anonymous reviewers for various valuable insights.

References

Altmann, Gabriel. 1978. Towards a theory of language. *Glottometrika* 1. 1–26.

DeBruine, Lisa M., Benedict C. Jones, Layla Unger, David R. Feinberg & Anthony C. Little. 2007. Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance* 33. 1420–1430.

Dresher, B. Elan. 2011. The phoneme. In Marc van Oostendorp, Colin J. Ewen, Elizabeth V. Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, 241–266. Oxford: Blackwell. https://doi.org/10.1002/9781444335262.wbctp0011

Eco, Umberto. 1984. *The role of the reader: Explorations in the semiotics of texts*. Bloomington: Indiana University Press.

Efron, Bradley. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82(397). 171–185. https://doi.org/10.1080/01621459.1987.10478410

Halberstadt, Jamin & Gillian Rhodes. 2003. It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review* 10. 149–156. https://doi.org/10.3758/BF03196479

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Langlois, Judith H. & Lori A. Roggman. 1990. Attractive faces are only average. *Psychological Science* 1. 115–121. https://doi.org/10.1111/j.1467-9280.1990.tb00079.x

McMahon, April. 2002. *An introduction to English phonology*. Edinburgh: Edinburgh University Press.

Perrine, Laurence. 1972. Euphony. In Alex Preminger, Frank J. Warnke & O. B. Hardison, Jr. (eds.), *Princeton encyclopedia of poetry and poetics*, 258. Princeton, NJ: Princeton University Press.

Pierrehumbert, Janet B. 2000. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.45.08pie

Šimáčková, Šárka, Václav Jonáš Podlipský & Kateřina Chládková. 2012. Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association* 42. 225–232. https://doi.org/10.1017/S0025100312000102

Šturm, Pavel. 2017. Determining syllable boundaries in Czech. Prague: Charles University dissertation.

Thorndike, Edward L. 1945. The association of certain sounds with pleasant and unpleasant meanings. *Psychological Review* 52. 143–149. https://doi.org/10.1037/h0055510

Trujillo, Logan T., Jessica M. Jankowitsch & Judith H. Langlois. 2014. Beauty is in the ease of the beholding: A neurophysiological test of the averageness theory of facial attractiveness. *Cognitive, Affective & Behavioral Neuroscience* 14. 1061–1076. https://doi.org/10.3758/s13415-013-0230-2

Whissell, Cynthia. 1999. Phonosymbolism and the emotional nature of sounds: Evidence of the preferential use of particular phonemes in texts of differing emotional tone. *Perceptual and Motor Skills* 89. 19–48. https://doi.org/10.2466/pms.1999.89.1.19

Whissell, Cynthia. 2000. Phonoemotional profiling: A description of the emotional flavour of English texts on the basis of the phonemes employed in them. *Perceptual and Motor Skills* 91. 617–648. https://doi.org/10.2466/pms.2000.91.2.617

Winkielman, Piotr & John T. Cacioppo. 2001. Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology* 81. 989–1000. https://doi.org/10.1037/0022-3514.81.6.989

Winkielman, Piotr, Jamin Halberstadt, Tedra Fazendeiro & Steve Catty. 2006. Prototypes are attractive because they are easy on the mind. *Psychological Science* 17. 799–806. https://doi.org/10.1111/j.1467-9280.2006.01785.x

Zipf, George K. 1949. *Human behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Calculating the victory chances

A stylometric insight into the 2018 Czech presidential election

Michal Místecký
University of Ostrava

The goal of this chapter is to determine stylometric features and keywords of the selected texts produced by the candidates for the 2018 Czech presidential election, and to interpret whether these may have had any impact upon the final results. The stylometric indexes researched include MATTR (moving-average type-token ratio), ATL (average token length), TC (thematic concentration), STC (secondary thematic concentration), Q (activity), and VD (verb distances); finally, a keyword analysis for two chosen candidates' programmes is carried out. The outcomes of the analyses show that each candidate adopts a special strategy to influence his electorate and that this strategy can be captured via stylometric methods.

Keywords: stylometry, quantitative linguistics, political discourse, presidential election, keyword analysis, Czech

1. Introduction

The chapter focuses on performing a set of stylometric analyses on corpora of texts connected to the Czech presidential election, which took place in January 2018. The goal is to find stylistic and thematic intersections among them, and to assess the efficiency of the rhetoric they use to influence the electorate. The 2018 election was the second of the direct type in the history of the Czech Republic, and both sparked multiple national debates (cf. Tait 2018). This research, therefore, can be of help for many professions that deal with the sphere of politics (marketers, rhetoric advisors, ghost-writers, politicians themselves, etc.), and may arouse public interest as well. A similar analysis has already been carried out for Czech, Italian, and American presidents' speeches (Čech 2014; Kubát & Čech 2016; Zörnig & Altmann 2016; Rimkeit-Vit & Gnatchuk 2016).

https://doi.org/10.1075/cilt.356.13mis

2. Material and methods

The materials were assembled to represent each of the nine candidates who participated in the election (their brief characteristics are given in Table 1, together with their results in the two rounds). There are two separated sets of texts studied – first, there are eight new-year speeches that the candidates wrote at the request of Czech Radio, complemented by the incumbent president's 2017 Christmas Address; second, I contrast the political programmes of two candidates, Marek Hilšer and Michal Horáček, for reasons stated later. As all these texts are Czech, if quoted in the study, their English translations are provided; the Czech originals are given in the footnotes. The overview of the corpus is presented in Table 2.

Table 1. The nine presidential candidates and their characteristics

Candidate	Features	First-round rank (percentage)	Second-round rank (percentage)
Jiří Drahoš	An academic worker; a strongly apolitical candidate	2 (26.6%)	2 (48.63%)
Pavel Fischer	An ambassador to France and a former advisor of Václav Havel	3 (10.23%)	–
Petr Hannig	A singer, producer, and talent-seeker	8 (0.56%)	–
Marek Hilšer	A doctor and civil activist	5 (8.83%)	–
Michal Horáček	A bookmaker, a lyric-writer, a Velvet Revolution symbol	4 (9.18%)	–
Jiří Hynek	A businessman in the weapon industry	7 (1.23%)	–
Vratislav Kulhánek	A businessman connected with Škoda cars and Czech ice hockey	9 (0.47%)	–
Mirek Topolánek	A former right-wing politician and manager	6 (4.3%)	–
Miloš Zeman	A left-wing politician and economist; the incumbent president	1 (38.56%)	1 (51.36%)

I analyze the texts on the basis of multiple stylometric indexes, which are intended to measure various properties of these texts (e.g., richness of lexis, degree of information and intellectuality, topic focus, narrative / descriptive character, or syntactic complicatedness). The counts are standard in contemporary quantitative linguistics (Čech et al. 2014; Kubát et al. 2014; Kubát 2017; Andreev et al. 2018; Melka & Místecký 2020), and have been proved to be independent of text length. Besides these quantifications, keyword calculations will be used, to clarify relations among the individual candidates.

Table 2. An overview of the corpus

Text	Candidate	Length (in words)
Czech Radio new-year speeches	Jiří Drahoš	521
	Pavel Fischer	296
	Petr Hannig	470
	Marek Hilšer	556
	Michal Horáček	362
	Jiří Hynek	611
	Vratislav Kulhánek	321
	Mirek Topolánek	645
	Miloš Zeman	1,255
Programmes	Marek Hilšer	2,612
	Michal Horáček	6,969

First, vocabulary richness will be assessed via MATTR (moving-average type-token ratio; Covington & McFall 2010). The index is based upon the ratio of types (the individual words appearing in a text) and tokens (all the words in a text), which is, however, dependent on the text length; in order to avoid this influence, the text is divided into a sequence of sections ('windows') the type-token ratios (TTRs) of which are counted. The final figure is the average of the partial TTRs. The length of the window depends on the researcher; if a window is, for example, five words, the first one will include the sequence from the first word of the text to its fifth one, the second window the sequence from the second word to the sixth one, etc. The aforementioned can be expressed formally as

$$MATTR\ (L) = \frac{\Sigma_{i=1}^{N-L} V_i}{L \times (N - L + 1)},$$

L standing for the length of the window, *V* for the types, and *N* for the text length in words. The index will be calculated using the MATTR software (Covington 2007), which works with word-forms.

Second, the index of ATL (average token length) will be employed to measure lexical complexity of the vocabulary. Its mathematical core is as follows:

$$ATL = \frac{1}{N} \sum_{i=1}^{N} p_i,$$

p_i is the length of a word in graphemes and *N* the length of the text in words.

Third and fourth, the thematic words of the texts will be identified via the counts of thematic concentrations (cf. Čech 2016). These calculations may be used when the length of a text is in the interval of < 200; 6,500 > (Čech & Kubát 2016),

which excludes from the analysis the programme of Michal Horáček; this will, together with Marek Hilšer's programme, be researched on the basis of keywords. The formula of the thematic concentration is based upon the *h*-point, an approximate divide between the synsemantic and autosemantic sections of the rank-frequency word distribution. If words of a text are ranked according to their decreasing frequencies, the *h*-point is identical with the spot where the rank of the word equals its frequency, namely

$$r = f(r);$$

if such a word is not to be found, the formula for the *h*-point is

$$h = \frac{f(i) \times r_j - f(j) \times r_i}{r_j - r_i + f(i) - f(j)},$$

where r_i stands for the highest rank for which $r_i < f(i)$, and r_j is the lowest rank for which $r_j > f(j)$ (cf. Čech & Kubát 2016: 8). If an autosemantic word penetrates into the region over the *h*-point, it is considered thematic; its thematic weight is counted in the following way:

$$TW = 2 \times \frac{(h - r') \times f(r')}{h \times (h - 1) \times f(1)}.$$

In the formula, *r′* signifies the rank of the autosemantic word, *f*(*r′*) its frequency, and *f*(1) the frequency of the top-scoring word (= of rank 1). The denominator serves to minimize the impact of text length. The thematic concentration of the whole text is the sum of all the thematic weights, i.e.,

$$TC = \sum TW.$$

The calculation of thematic concentration is one of the ways to detect words of prominent importance in a text, and to calculate their 'positions' in it with exactitude. Alternatively, it is possible to employ the doubled value of the *h*-point, i.e., *2h*. Such is the manner of counting secondary thematic concentration; the formula of the secondary thematic weight of a word thus reads

$$STW = 2 \times \frac{(2h - r') \times f(r')}{h \times (2h - 1) \times f(1)},$$

with the total secondary concentration equalling

$$STC = \sum STW.$$

The advantage of STC is that it usually finds more salient words in a text; its disadvantage is that it is not well methodologically founded, as doubling the *h*-point has no justification in the rank-frequency distribution of words.

Fifth, the study will focus on assessing the active / descriptive character of a sample. A simple way of researching it is to use the Busemann coefficient (Busemann 1925), which takes into account the number of verbs and adjectives. Its formula is

$$Q = \frac{V}{A + V},$$

V meaning the number of verbs, and *A* the number of adjectives. The ratio makes possible a trichotomous division of texts – into active ($Q > 0.5$), neutral ($Q = 0.5$), and descriptive ones ($Q < 0.5$).

Sixth, the index measuring syntactic complexity will be calculated – verb distances (VD). It concerns the average number of words occurring in between two verbs; mathematically

$$VD = \frac{1}{D}\sum_{i=1}^{D} d_i,$$

with *d* signifying the number of words in between two verbs, and *D* the total of distances in a sample.

The aforementioned indexes (ATL, TC, STC, Q, and VD) will be calculated using the QUITA (Quantitative Index Text Analyzer) software (Kubát et al. 2014). It is to be noted that in the case of TC and STC, the programme works with word lemmata, and as to the activity measurement, it does not take into account static and modal verbs, such as *have, be*, or *can*. On the other hand, in case of VD, all the verbal forms are counted as separate units. The QUITA lemmatizer and part-of-speech tagger for Czech is Majka+Corpus, which works with an 80-percent success rate.

Next, keywords will be searched for in a confrontation of the programmes by Horáček and Hilšer. The procedure is carried out by the LancsBox software (Brezina et al. 2015). The analysis is based upon the simple math count (Kilgarriff 2009), which calculates the ratio of the relative frequencies of the studied word in the two corpora. The results are the keywords typical of the investigated corpus (hereinafter referred to as the plus keywords), those occurring eminently in the reference corpus (the minus keywords), and the lockwords – the ones that occur in both the texts. This way, the study tries to uncover differences between the programmes as well as their similarities.

3. Results and interpretations

I present the outcomes of the research in two separate sections – first, the results of the stylometric analysis of the new-year addresses are commented upon; second, the study concentrates on the programmes of Horáček and Hilšer.

3.1 The new-year addresses: A stylometric perspective

The overall results are presented in Table 3; the top figures are marked with asterisks (the TC results are not analyzed, as there are too few numbers available). As to MATTR, there are not considerable differences between the candidates; the highest number of Pavel Fischer may have been attributable to his intellectual stylization and visionary approach to the presidential function. His way of writing will be exemplified in the following excerpt.

> With the previous year ending, one section of our journey always ends. It is good to keep in memory that we can make a new start in the year to come, with a new confidence and new hope. This year, as a presidential candidate, I am doing this with a vision of a renewal of the state and society. The political opinions and social statuses may divide us, but desire for calm life in safety and for economic prosperity joins us together.[1] (*iRozhlas.cz*, 27 December 2017)

Table 3. The results of the stylometric analysis of the new-year addresses

Candidate	MATTR	ATL	TC	STC	Q	VD
Jiří Drahoš	0.87	4.70	0.04	0.05	0.57*	5.05
Pavel Fischer	0.89*	4.81	0	0.09	0.46	6.36
Petr Hannig	0.88	5.24*	0	0.04	0.34	6.78*
Marek Hilšer	0.83	4.88	0	0.02	0.47	6.13
Michal Horáček	0.86	4.83	0	0	0.56	4.81
Jiří Hynek	0.80	4.65	0	0.03	0.57*	4.93
Vratislav Kulhánek	0.87	4.77	0	0.03	0.58*	4.23
Mirek Topolánek	0.88	5.06	0.09	0.11*	0.46	5.42
Mirek Zeman	0.83	4.73	0	0.01	0.51	4.49

1. "S koncem uplynulého roku se vždy symbolicky uzavírá jeden úsek naší cesty. Je dobré mít na paměti, že do roku nového můžeme vykročit jinak, s novým přesvědčením a novou nadějí. Letos, jako kandidát na prezidenta, tak osobně činím s vizí obnovy státu i společnosti. Politické názory a sociální postavení nás mohou rozdělovat, touha po klidném životě v bezpečí a ekonomické prosperitě nás ale spojuje."

Pavel Fischer's speech tries to transcend a mere political message by addressing the abstract topics of hope, renewal, and life; this goes hand in hand with using a wide range of expressions, and produces a philosophical timbre typical of his style. This approach seems to be effective with voters, as Fischer came third in the first round of the election.

A different kind of intellectualization was adopted by Petr Hannig, who scores highest in ATL and VD. These two indexes are indicators of lexical and syntactic complexities, which points out Hannig's attempt at employing sophisticated vocabulary and rich sentence structures. Both can be shown in this extract.

> Interwar Czechoslovakia was, thanks to its founding fathers, not only a democratic island among the despotic states of that-time central Europe, but also, thanks to the hard Czechoslovak crown, one of the ten economically most successful states of the world.[2] (*iRozhlas.cz*, 27 December 2017)

Hannig's way of presenting reminds one of academic writing, making use of enriching by-thoughts, a parallel sentence structure ("not only … but also"), and rational reasoning; his vocabulary draws upon American-like expressions ("founding fathers") and strives to achieve scholarly precision. In this manner, he tries to create a trustworthy image of a self-assured educated person with balanced views of the Czech national spirit.

Next, there are several top scorers in the domain of activity, such as Vratislav Kulhánek, Jiří Drahoš, or Jiří Hynek. As Kulhánek and Hynek did not succeed in the first round of the election, our attention is focused on a passage from the address by Jiří Drahoš.

> 100 years will elapse since the period when we started to write down the history of our renewed state sovereignty. Let us remember that at the beginning of the longest period of oppression within this era, in 1946, there was also an election, which took place, however, with a limited choice of parties.[3]
>
> (*iRozhlas.cz*, 27 December 2017)

Drahoš's style is straightforward and electorate-directed; he uses a system of energetic verbs ("elapse", "start", "remember", "take place") to infuse the text with dynamism and a personalized tone. This is emphasised by the first-person imperative

---

2. "Meziválečné Československo bylo díky otcům zakladatelům nejenom demokratickým ostrovem mezi autoritářskými státy tehdejší střední Evropy, ale i díky tvrdé československé koruně i jedním z deseti hospodářsky nejúspěšnějších států světa."

3. "Uplyne 100 let od doby, kdy se začaly psát dějiny naší obnovené státní samostatnosti. Připomeňme si, že na počátku nejdelšího období nesvobody v této éře, v roce 1946, stály rovněž volby, i když už se odehrály s výběrem omezeným."

employed in the second sentence. As Drahoš's public presentation was sometimes characterized as stiff and too passive (cf. Musil 2018), the high value of activity seems to be an intentional product of his marketing strategy.

Last but not least, the study focuses on thematic words. First, the secondary thematic concentration figures will be commented upon, as the thematic concentration ones are too few to be of any interpretative value. The speech by Mirek Topolánek features the highest score, with the emphasis being put upon political terms (see later). The handling of secondary thematic words in Topolánek's text will be explained based on this excerpt.

> In the home political situation, the greatest problem was the breakdown of the political scene. Due to its fatal division, we have got a minority GOVERNMENT here. Such a cabinet is not stable and cannot push through laws and decisions of the citizens' problems. At the same time, there have always been a number of variants of majority democratic GOVERNMENTS. … If the one-party minority GOVERNMENT doesn't win the confidence of the Chamber of Deputies, the second attempt should be used so that a solid majority GOVERNMENT composed of democratic parties be formed.[4] (*iRozhlas.cz*, 27 December 2017)

Topolánek's tactics are based upon clear argumentation and repetition of the core notions. This stylistic feature enables him to maintain a person's attention, and serves as an emphatic cohesive device of his speech. To a certain extent, this strategy is reminiscent of advertising.

Next, the secondary thematic words of all the candidates will be presented (see Figure 1; the candidates' names are in blue, the words are capitalized). As seen, the expressions interconnect the texts into a network of family resemblance (cf. Wittgenstein 2001 [1953]); this means that although there is not a single topic that would be shared by all the candidates, they 'communicate' through those shared by pairs (e.g., "life" is shared by Drahoš and Fischer, "liberty" by Hannig and Hynek, etc.).

The thematic expressions uncover the candidates' orientations. Topolánek, for instance, confirms his position of a political candidate, focusing on "state", "government", "president", "Czech", and "year". A rather opposite perspective is Fischer's, who speaks about "life", "year", and "new", seeking a more highbrow electorate.

---

4. "V domácí politice bylo největším problémem štěpení politické scény. Kvůli její fatální rozdělenosti tu máme menšinovou VLÁDU. Takový kabinet není stabilní a nemůže prosazovat zákony a řešení problémů občanů. Přitom existovala a stále existuje řada variant většinových demokratických VLÁD. […] Pokud jednobarevná menšinová VLÁDA nedostane důvěru Poslanecké sněmovny, měl by být druhý pokus využit k tomu, aby vznikla solidní většinová VLÁDA složená z demokratických stran."

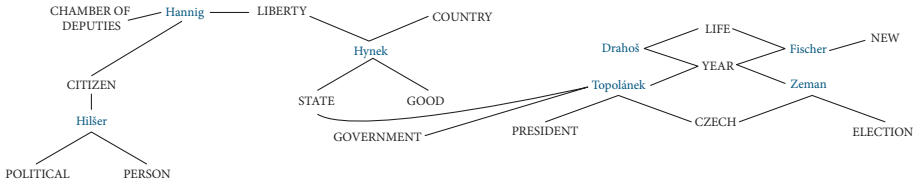


Figure 1. The system of the secondary thematic words in the presidential candidates' speeches

Hynek and Hannig, as businesspeople, cherish "liberty", and Marek Hilšer reinstates the term of "citizen", connected to the legacy of Václav Havel. Via the very words, then, it is possible to determine the cardinal semantic points of the speeches investigated.

To conclude, it is of importance that the new-year speech proposal by Michal Horáček does not manifest any (secondary) thematic words, so he is absent from Figure 1. As this situation is peculiar and may be an outcome of various tactics, it is going to be researched in § 3.2.

3.2 Marek Hilšer vs. Michal Horáček: The keyword analysis

The aforementioned absence of thematization in Michal Horáček's speech will be paid more attention in this section. In order to determine the roots of this particularity, his presidential-election programme will be compared to the one written by Marek Hilšer. Hilšer's was chosen as his statement is rather long and compact; other candidates either scatter their opinions around over many documents, or provide only brief accounts of their main visions. The texts will be compared on the basis of the keyword analysis, as introduced in § 2.

The results are presented in Table 4. Where needed, the morphological abbreviations indicating cases (GEN: genitive; DAT: dative; ACC: accusative; LOC: locative), number (SG: singular), genders (MASC: masculine; FEM: feminine; NEU: neuter), animacy (IN: inanimate; AN: animate), and parts of speech (ADJ: adjective; VERB: verb; ADV: adverb) are given. In case two categories for one translation are presented, they are separated by a semi-colon (e.g., MASC; GEN). If a single-word equivalent of the Czech word does not exist in English, the translation is hyphenated (e.g., "of-quality", "if-it-is"); if two translations are possible, they are divided by a slash (e.g., "just / exactly"). The "numbers" written in brackets mean that the thousand-order zeroes appeared in the analysis (e.g., "000"). The original Czech expressions are listed in Table 5.

Table 4. An overview of the keyword situation in the programmes by Michal Horáček and Marek Hilšer

Plus keywords	Lockwords	Minus keywords
care	was [FEM]	his
healthcare	ten	years [GEN]
of-quality [ADJ]	to / towards	vital [NEU]
castle	better [ADV]	less
role [ACC, LOC]	except	then
research [GEN, DAT, LOC]	power [GEN, DAT, LOC] / can [VERB]	just / exactly
to find	modern	this [FEM; SG; ACC] / here
if-it-is	over	these [FEM; NOM / FEM; ACC / MASC IN; NOM / MASC IN; ACC / MASC AN; ACC]
him [ACC]	afterwards	[numbers]
appointing	our [MASC; GEN]	when [ADV]

Table 5. An overview of the keyword situation in the programmes by Michal Horáček and Marek Hilšer (in Czech)

Plus keywords	Lockwords	Minus keywords
péče	*byla*	*jeho*
zdravotnictví	*deset*	*let*
kvalitní	*ke*	*nezbytné*
hrad	*lépe*	*méně*
roli	*mimo*	*pak*
výzkumu	*moci*	*právě*
najít	*moderní*	*tu*
bude-li	*nad*	*ty*
ho	*nato*	[*čísla*]
jmenování	*našeho*	*kdy*

In general, the programmes show a situation which is very close to the one in the new-year speech proposals. Compared to Horáček's, Hilšer's text is more topic-focused, speaking about high-quality healthcare (logically enough, as he is a doctor by profession), the castle (meaning the Prague Castle, the seat of the president), and appointment of judges, since this is a duty of the head of state. On the other hand, the words that are prominent in Horáček's programme (and rare in Hilšer's) are mostly of a functional, hedging, or rhetorical nature ("just", "this", "vital"); at most, they indicate the time flow ("years", "when"), or focus on giving facts (the role of numbers, but a close reading of the programme has revealed that even these are quite marginal). In the pre-election surveys, Michal Horáček was

considered one of the powerful rivals of Miloš Zeman (cf. *euro.cz*, 8 November 2017); it is thus a question whether his final loss – he came fourth – was not due to the lack of thematic expressivity in both his programme, and his proposed new-year speech.

4. Conclusions

The outcomes of the research can be summarized in the forthcoming points.

1. The candidates appear to have used distinctive strategies to achieve the support of the electorate. There seem to be noticeable tendencies of intellectualization (Pavel Fischer and Petr Hannig, both in their own ways), energization (Jiří Drahoš), and thematic focalization (Mirek Topolánek). These choices correspond to the characters of the candidates and the messages they would like to communicate (vision, historical knowledge, easy-goingness, decisiveness).
2. As to their topics, the election participants are connected on the basis of family resemblance – it would be difficult to identify one particular core of their speeches, but the individual themes link the candidates together. Roughly, the most weight was put on political notions, but each person running for office seems to keep his own ideas and tactics (such as "citizen" in case of Hilšer, absence of political expressions in the speech by Fischer, the notion of liberty in the texts of Hannig and Hynek, etc.). No thematic words were detected in Michal Horáček's speech, which has fuelled another part of the investigation.
3. The second part of the research – the keyword analysis of the programmes of Michal Horáček and Marek Hilšer – has confirmed the absence of word-expressed thematization in the former. The words that he, contrary to Hilšer, used were mostly hedging devices ("his", "just", "this", etc.). I have hypothesized that this stylistic feature may have led to Horáček's result in the election.

Overall, the tools that have been employed for the analysis have proved able to provide solid grounds for interpretations of the results. The original idea of the study is thus to be confirmed: stylometric measurements could be useful for professional aides of political figures, as well as for analysts, language consultants, and other occupations linked to the area.

References

Andreev, Sergey, Michal Místecký & Gabriel Altmann. 2018. *Sonnets: Quantitative inquiries*. Lüdenscheid: RAM-Verlag.

Brezina, Vaclav, Tony McEnery & Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2). 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Busemann, Adolf. 1925. *Die sprache der jugend als ausdruck der entwicklungsrhythmik* [Language of youth as an expression of development rhythm]. Jena: Fischer.

Covington, Michael. 2007. *MATTR, version 2.0 for Windows*. http://ai1.ai.uga.edu/caspr/. (27 April, 2020.)

Covington, Michael & Joe McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100. https://doi.org/10.1080/09296171003643098

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

Čech, Radek. 2016. Tematická koncentrace textu v češtině [Thematic concentration of text in Czech]. Praha: Ústav formální a aplikované lingvistiky.

Čech, Radek & Miroslav Kubát. 2016. Text length and the thematic concentration of text. *Mathematical Linguistics* 2(1). 5–13.

Čech, Radek, Ioan-Iovitz Popescu & Gabriel Altmann. 2014. *Metody kvantitativní analýzy (nejen) básnických textů* [Methods of quantitative analysis of (not only) poetic texts]. Olomouc: Univerzita Palackého v Olomouci.

euro.cz, 2017 November 8. Volební model: Favority prezidentské volby jsou Zeman a Drahoš [The election model: Zeman and Drahoš are the favourites]. https://www.euro.cz/prezidentske-volby-2018/volebni-model-favority-prezidentske-volby-jsou-zeman-a-drahos-1381845. (27 April, 2020.)

iRozhlas.cz, 2017 December 27. Novoroční poselství kandidátů na prezidenta: jak hodnotí uplynulý rok a co si přejí v tom novém? [New-year messages from the presidential candidates: How do they sum up the previous year and what do they wish in the new one?]. https://www.irozhlas.cz/zpravy-domov/novorocni-poselstvi_1712270610_kno. (27 April, 2020.)

Kilgarriff, Adam. 2009. Simple maths for keywords. In Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds.), *Proceedings of corpus linguistics conference CL2009*. Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009/. (27 April, 2020.)

Kubát, Miroslav, Vladimír Matlach & Radek Čech. 2014. *QUITA. Quantitative index text analyzer*. Lüdenscheid: RAM-Verlag.

Kubát, Miroslav & Radek Čech. 2016. Quantitative analysis of US presidential inaugural addresses. *Glottometrics* 34. 14–27.

Kubát, Miroslav. 2017. *Kvantitativní analýza žánrů* [A quantitative analysis of genres]. Ostrava: Ostravská univerzita.

Melka, Tomi & Michal Místecký. 2020. On stylometric features of H. Beam Piper's *Omnilingual*. *Journal of Quantitative Linguistics*, 27(3). 207–243. https://doi.org/10.1080/09296174.2018.1560698

Musil, Michal. 2018 January 27. Proč Zeman vyhrál a Drahoš prohrál [Why Zeman won, and Drahoš failed]. *reportermagazin.cz*. https://reportermagazin.cz/a/i28SV/proc-zeman-vyhral-a-drahos-prohral. (27 April, 2020.)

Rimkeit-Vit, Lyubov & Hanna Gnatchuk. 2016. Euphemisms in political speeches by USA presidents. *Glottometrics* 35. 16–21.

Tait, Robert. 2018 January 25. Czech presidential election on a knife-edge as challenger cries foul. *theguardian.com*. https://www.theguardian.com/world/2018/jan/25/czech-presidential-election-knife-edge-milos-zeman-jiri-drahos. (27 April, 2020.)

Wittgenstein, Ludwig. 2001 [1953]. *Philosophical investigations*. Hoboken, NJ: Blackwell.

Zörnig, Peter & Gabriel Altmann. 2016. Activity in Italian presidential speeches. *Glottometrics* 35. 38–48.

Topological mapping for visualisation of high-dimensional historical linguistic data

Hermann Moisl
Newcastle University

This paper addresses an issue in visualization of high-dimensional data abstracted from historical corpora whose importance in quantitative and corpus linguistics has thus far not been sufficiently appreciated: the possibility that the data is nonlinear. Most applications of data visualization in these fields use linear proximity measures which ignore nonlinearity, and, if the data is significantly nonlinear, can give misleading results. Topological mapping is a nonlinear visualization method, and its application via a particular topological mapping method, the Self-Organizing Map, is here exemplified with reference to a small historical text corpus.

Keywords: Historical linguistics, nonlinearity, high-dimensional data, topological mapping, clustering

1. Introduction

Discovery of the chronological or geographical distribution of collections of historical text can be more reliable when based on multivariate rather than on univariate data because, assuming that the variables describe different aspects of the texts in question, multivariate data necessarily provides a more complete description. Where the multivariate data is high-dimensional, however, its complexity can defy analysis using traditional philological methods. Increasingly, the first step in interpreting such complexity is data visualization because it gives insight into latent structure, thereby facilitating hypotheses which can then be tested using a range of other mathematical and statistical methods (Moisl 2015).

The present discussion addresses an issue in data visualization whose importance in quantitative and corpus linguistics has thus far not been sufficiently appreciated: the possibility that the data is nonlinear. Most visualization applications in these fields use linear proximity measures which ignore nonlinearity, and, if the data is significantly nonlinear, can give misleading results.

https://doi.org/10.1075/cilt.356.14moi

The discussion is in three main parts: the first part outlines the nature of nonlinearity in data generally and in linguistic data specifically, the second shows why nonlinearity is a problem for linear visualization methods, and the third shows how topological mapping can be used to visualize high-dimensional data in a way that takes nonlinearity into account.

2. Nonlinearity

2.1 Nonlinearity in natural processes

In natural processes there is a fundamental distinction between linear and nonlinear behavior. Linear processes have a constant proportionality between cause and effect. If a ball is kicked x hard and it goes y distance, then a $2x$ kick will appear to make it go $2y$, a $3x$ kick $3y$, and so on. Nonlinearity is the breakdown of such proportionality. In the case of our ball, the linear relationship increasingly breaks down as it is kicked harder and harder. Air and rolling resistance become significant factors, so that for, say, $5x$ it only goes 4.9 y, for $6x$ 5.7y, and again so on until eventually it bursts and goes hardly any distance at all. Such nonlinear effects pervade the natural world and gives rise to a wide variety of complex and often unexpected – including chaotic – behaviours (Strogatz 2000; Bertuglia & Vaio 2005).

2.2 Nonlinearity in data

Data is a description of objects from a domain of interest in terms of a set of variables such that each variable is assigned a value for each of the objects. Given m objects described by n variables, a standard representation of data for computational analysis is a matrix M in which each of the m rows represents a different object, each of the n columns represents a different variable, and the value at $M_{i,j}$ describes object i in terms of variable j, for $i = 1..m$, $j = 1..n$. The matrix thereby makes the link between the researcher's conceptualization of the domain in terms of the semantics of the variables s/he has chosen and the actual state of the world, and allows the resulting data to be taken as a representation of the domain based on empirical observation.

M is linear when the functional relationships between all its variables, that is, the values in its columns, conform to the mathematical definition of linearity. A linear function f is one that satisfies the following properties, where x and y are variables and a is a constant (Lay 2010):

– Additivity: $f(x + y) = f(x) + f(y)$ – adding the results of f applied to x and y separately is equivalent to adding x and y and then applying f to the sum.

– Homogeneity: $f(ax) = af(x)$ – multiplying the result of applying f to x by a constant is equivalent to multiplying x by the constant and then applying f to the result.

A function which does not satisfy these two properties is nonlinear, and so is a data matrix in which the functional relationships between two or more of its columns are nonlinear.

Matrices have a geometrical interpretation. For each row vector of M:

– The dimensionality of the vector, that is, the number of its components n, defines an n-dimensional Euclidean space.
– The sequence of n numbers comprising the vector specifies the coordinates of the vector in the space.
– The vector itself is a point at the specified coordinates

The set of row vectors in M defines a configuration of points in the n-dimensional space called the data manifold. Linear manifolds are shapes consisting of straight lines and flat planes and represent linear data, whereas nonlinear manifolds consist of curved lines and surfaces and represent nonlinear data; examples are given in Figure 1.

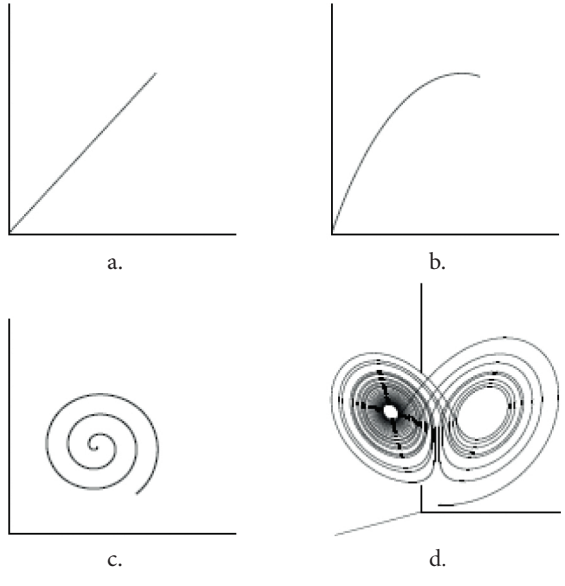
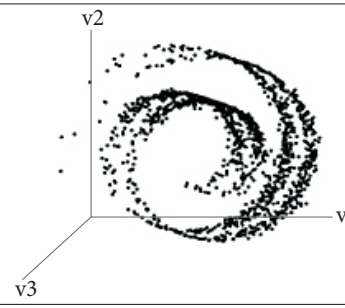


Figure 1. Linear and nonlinear manifolds in two- and three-dimensional space

An unbounded range of nonlinear manifolds is possible in any dimensionality. Figure 2 gives another example of a nonlinear manifold in three-dimensional space.

	V1	V2	V3
1	5.214	7.087	6.729
2	0.421	8.426	7.896
3	6.134	5.686	6.088
…	…	…	…
1600	0.945	11.501	10.867

a: 1600 3 -dimensional vectors



b: The vectors in (a) plotted in 3-dimensional space

Figure 2. Nonlinear manifold in three-dimensional space

2.3 Nonlinearity in linguistic data

Data abstracted from a natural process known to be linear is itself guaranteed to be linear. Data abstracted from a known nonlinear process is not necessarily nonlinear, but may be. The human brain – the generator of language – is a nonlinear dynamical system that exhibits highly complex physical behaviour in which nonlinearity arises on account of latency and saturation effects in individual neuron and neuron assemblies. One must, therefore, always reckon with the possibility that data abstracted from speech or text will be nonlinear.

3. The problem

The problem that nonlinearity poses for cluster analysis of high-dimensional multivariate data is easily seen. A metric space M(V,*d*) is a vector space V on which a metric *d* is defined in terms of which the distance between any two points in the space can be measured. Numerous distance metrics exist (Deza & Deza 2009: Chapters 17, 19). For present purposes these are divided into two types:

1. Linear metrics, where the distance between two points in a manifold is taken to be the length of the straight line joining the points, or some approximation to it, without reference to the shape of the manifold.
2. Nonlinear metrics, where the distance between the two points is the length of the shortest line joining them along the surface of the manifold and where this line can but need not be straight.

This categorization is motivated by the earlier observation that manifolds can have shapes which range from perfectly flat to various degrees of curvature. Where the manifold is flat, as in Figure 3a, linear and nonlinear measures are identical. Where it is curved, however, linear and nonlinear measurements can differ to varying degrees depending on the nature of the curvature, as shown in Figures 3b and 3c.

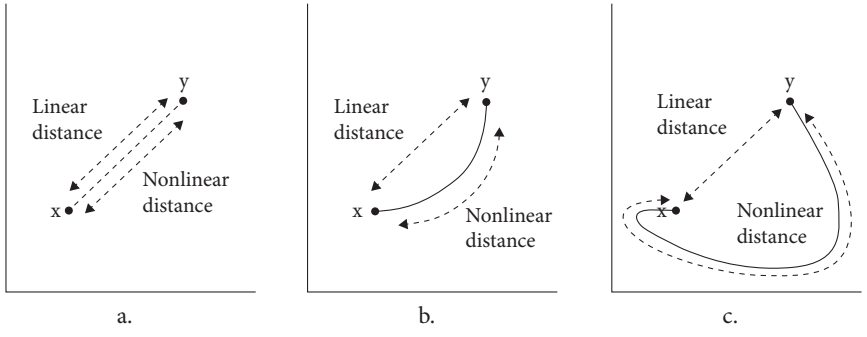


Figure 3. Linear and nonlinear distances

For Figure 2b, the linear distance is shown in Figure 4 between two points; the nonlinear distance follows the surface of the curve, and is obviously much greater.

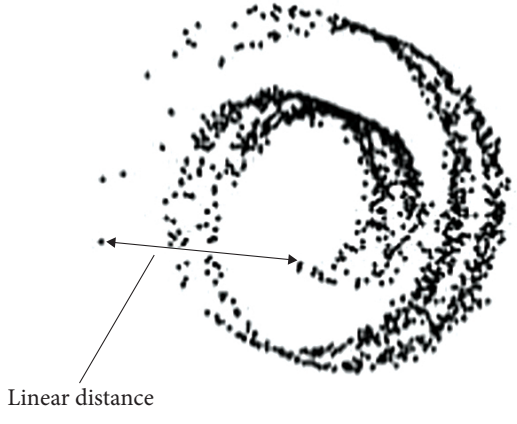


Figure 4. Linear distance between points on a nonlinear manifold

Commonly used visualization methods such as principal component analysis for projection into two- or three-dimensional space, or hierarchical cluster analysis using proximity measures like the Euclidean, are linear: they take no account of any curvature in the manifold, and can thereby introduce distortions into visualization results in some proportion to the degree of nonlinearity in the manifold.

This is shown in Figure 4 for three-dimensional data, but the situation extends to any dimensionality.

One way of precluding nonlinearity errors is to use a nonlinear distance measure (Moisl 2015). The alternative is to use a topological method, outlined in what follows.

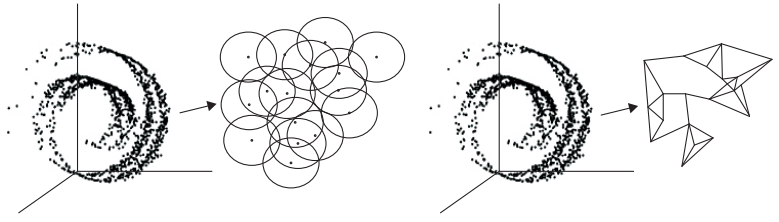
4. Topological mapping

4.1 Topology

Topology (Munkres 2000; Reid & Szendroi 2005; Sutherland 2009; Lee 2010) is an aspect of contemporary mathematics that grew out of metric space geometry. Its objects of study are manifolds, but these are studied as spaces in their own right, topological spaces, without reference to any embedding metric space and associated coordinate system. Topology would, for example, describe a manifold embedded in the metric space of Figure 5a independently both of the metric defined on the space and of the coordinates relative to which the distances among points are calculated, as in Figure 5b. Topology replaces the concept of metric and associated coordinate system with relative nearness of points to one another in the manifold as the mathematical structure defined on the underlying set; relative nearness of points is determined by a function which, for any given point p in the manifold, returns the set of all points within some specified proximity to p. But how, in the absence of a metric and a coordinate system, is the proximity characterized?

Figure 5. A manifold embedded in a three-dimensional coordinate system and as a topological object

The answer is that topological spaces are derived from metric ones and inherit from the latter the concept of neighbourhoods. In a metric space, a subset of points which from a topological point of view constitutes a manifold can itself be partitioned into subsets of a fixed size called neighbourhoods, where the neighbourhood of a point p in the manifold can be defined either as the set of all points within some fixed radius ε from p or as the k nearest neighbours of p using the existing metric and coordinates; in Figure 6 a small region of the manifold from Figure 5 is magnified to exemplify these two types of neighbourhood.



a. Neighbourhoods of diameter ε

b. Neighbourhoods for k = 3 nearest neighbours

Figure 6. Neighbourhoods in a magnified fragment of a geometric object in metric space

In Figure 6a the neighbourhood of every point is the other points within a radius of ε, shown as circles within the magnification rectangle; in 6b a neighbourhood of any point is the k nearest points irrespective of distance, shown for $k = 3$ as lines connecting each point to the three nearest to itself. Once a manifold has been partitioned into neighbourhoods and thereby transformed into a topological space, the frame of reference is discarded and only the neighbourhoods defined in terms of the metric are retained. In this way, manifolds of arbitrary shape can be conceptualized as being composed of metric subspaces; if the original metric is Euclidean, for example, the manifold in Figure 5b can be understood as a patchwork of locally-Euclidean subspaces. Intuitively, this corresponds to regarding the curved surface of the Earth as a patchwork of flat neighbourhoods, which is how most people see it.

4.2 Projection of topological structure into low-dimensional space

High-dimensional manifolds can be visualized as low-dimensional ones by means of projection in which the topology of the high-dimensional manifold, that is, the neighbourhood structure, is preserved in the low-dimensional one, so that points close to one another in high dimensions are close to one another in the low-dimensional projection. This can be conceptualized as in Figure 7, where a three-dimensional manifold is projected onto a two-dimensional surface.

Figure 7. Projection from three to two dimensions

4.3 Preservation of nonlinearity

The set of neighbourhoods which constitutes the topology of a manifold by definition follows the surface of the manifold, whatever its shape. Because a projection preserves the topology, that shape is preserved – in other words, nonlinearity is preserved in the projection.

4.4 Example

The aim of this section is to show how topological mapping can be used to discover structure in high-dimensional multivariate data abstracted from a multi-document corpus. It does this by using a particular topological mapping method, the Self-Organizing Map (SOM), to infer the relative chronology of a collection of Old English, Middle English, and Early Modern English texts from spelling data abstracted from them.

4.4.1 *The text collection*

The list of texts comprising the example corpus is given in Table 1. Because the aim is methodological – to exemplify the application of a visualization method to historical linguistic data rather than to generate novel results about the history

of English – this discussion assumes that the textual accuracy offered by current critical editions is unnecessary, and that readily available online texts of diverse provenances suffice. To this end, the Old English texts were downloaded from the *Sacred Text Archive* (https://www.sacred-texts.com/), the Middle English ones from the *Corpus of Middle English Prose and Verse* (https://quod.lib.umich.edu/c/cme/), and the Early Modern English ones from *Corpora of Historical English* (http://davies-linguistics.byu.edu/personal/histengcorp.htm); in a few cases more conveniently formatted texts were downloaded from other sites.

Table 1. The example corpus C

Old English	Middle English	Early Modern English
Exodus	Sawles Warde	King James Bible
Phoenix	Henryson, Testament of Cressid	Campion, Poesie
Juliana	The Owl and the Nightingale	Milton, Paradise Lost
Elene	Malory, Morte Darthur	Bacon, Atlantis
Andreas	Gawain and the Green Knight	More, RichardIII
Genesis A	Morte Arthure	Shakespeare, Hamlet
Beowulf	KingHorn	Jonson, Alchemist
	Alliterative Morte Arthure	
	Bevis Of Hampton	
	Chaucer, Troilus	
	Langland, Piers Plowman	
	York Plays	
	Cursor Mundi	

4.4.2 *Spelling data*

Spelling is used as the basis for inference of the relative chronology of the above texts on the grounds that it reflects the phonetic, phonological, and morphological development of English over time. The variables used to represent spelling in the texts are letter pairs: for 'the cat sat', the first letter pair is (t,h), the second (h,e), the third (e,<space>), and so on. All distinct pairs across the entire text collection were identified, and the number of times each occurs in each text was counted. The fragment of the resulting data matrix M in Table 2 exemplifies this.

Table 2. Fragment of the frequency matrix M abstracted from C

	1. hw	2. we	3. fe	…	841. jm
Exodus	35	149	125	…	0
Sawles Warde	52	147	45	…	0
…	…	…	…	…	…
King James	0	42	36	…	0

M was normalized to compensate for variation in document length and truncated to the most important 100 letter pairs, yielding a new matrix M' used in the analysis that follows. Details of normalization and truncation are available in (Moisl 2015: Chapter 3).

4.4.3 *The Self-Organizing Map*

The Self-Organizing Map (SOM) is a topological mapping method. It is an artificial neural network that was originally invented to model a particular kind of biological brain organization, but can also be used without reference to neurobiology as a way of visualizing high-dimensional data manifolds by projecting and displaying them in low-dimensional space. It has been extensively and successfully used for this purpose across a wide range of disciplines. Figure 8 shows the architecture of the SOM.

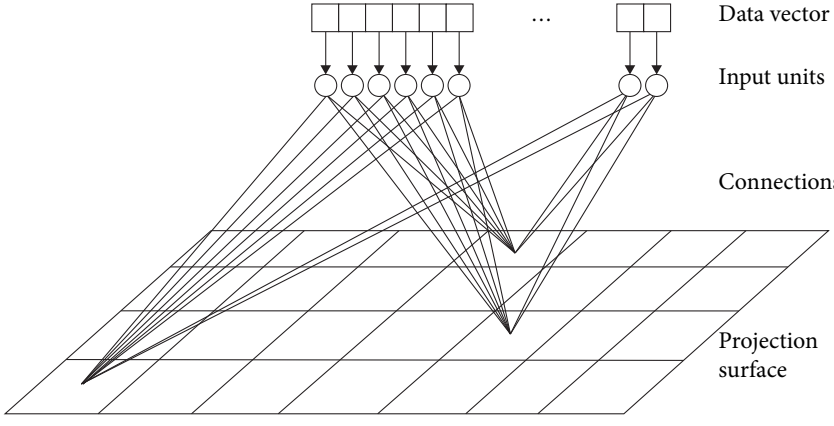


Figure 8. SOM architecture

An *n*-dimensional data vector is loaded into the input units. These values are propagated along the connections in such a way that the vector is assigned to one of the cells on the two-dimensional projection surface; only a selection of connections is shown, and in reality every input unit is connected to every projection surface cell. Once every input vector has been projected onto the surface, the topology of the *n*-dimensional data manifold has been mapped onto the two-dimensional projection space, where it is available for visual inspection. The relative strengths of the connections linking the input units to the projection surface are fundamental to the SOM's mapping, and, as with other artificial neural network architectures, these are iteratively learned from the set of input vectors in any given application. The learning procedure is quite complex and so is not rehearsed here; details are given in the following citations. The standard work on the SOM is Kohonen (2001). Shorter accounts are Haykin (1999: Chapter 9), Van Hulle (2000), Lee & Verleysen (2007: Chapter 5), Izenman (2008: Chapter 12.5), Xu & Wunsch

(2008: Chapter 5.3.3), Moisl (2015: Chapter 4); collections of work on the SOM are in Oja & Kaski (1999) and Allinson et al. (2001). For overviews of applications of the SOM to cluster analysis and data analysis more generally see Kohonen (2001: Chapter 7) and Vesanto & Alhoniemi (2000).

An intuition for how the SOM works can be gained by looking at the biological brain structure it was originally intended to model: sensory input systems (Van Hulle 2000; Kohonen 2001: Chapters 2, 4). The receptors in biological sensory systems generate very high-dimensional signals which are carried by numerous nerve pathways to the brain. The retina of the human eye, for example, contains on the order of 10^8 photoreceptor neurons each of which can generate a signal in selective response to light frequency, and the number of pathways connecting the retina to the brain is on the order of 10^6 (Hubel & Wiesel 2005). At any time t, a specific visual stimulus $s_{(t)}$ to the eye is transmitted via the nerve pathways to the visual cortex, and this generates a pattern of retinal activation $a(t)$ which is in turn transmitted to the rest of the brain for further processing. It is the response of the visual cortex to retinal stimulation which is of primary interest here. The visual cortex is essentially a two-dimensional region of neurons whose response to stimulation is spatially selective: any given retinal activation $s(t)$ sent to it activates not the whole two-dimensional cortical surface but only a relatively small region of it, and subsequent stimuli $s(t+1)$, $s(t+2)\ldots s(t+n)$ activate other regions whose distances from $a(t)$ and from one another on the cortical surface are proportional to the relative similarities of the $s(t)\ldots s(t+n)$. If, say, seven activations $s(1)\ldots s(7)$ are input in temporal succession, and if the members of each of three sets $\{a(1), a(3), a(7)\}$, $\{a(2), a(5)\}$,and $\{a(4), a(6)\}$ are similar to one another but different from members of the other two sets, then the 'cortex' is sequentially activated, and these successive activations, when superimposed as in Figure 9, show a cluster structure. This is the basis for the SOM's use as a projection method.

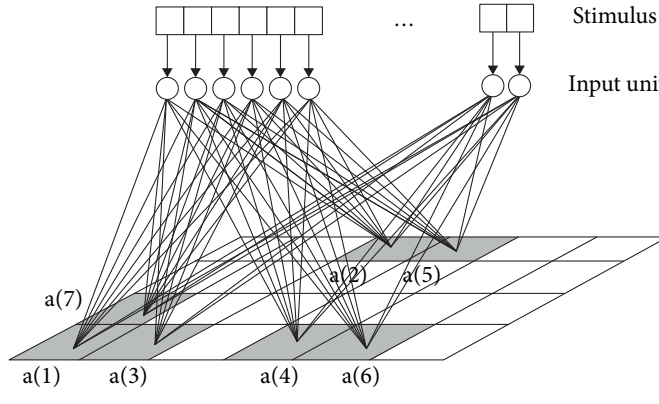


Figure 9. Selective activation of the projection surface in response to stimuli

The mathematical model corresponding to the above physical one has three components together with operations defined on them:

- An n-dimensional input vector R, for some arbitrary n, which represents the retina
- A $p \times q$ output matrix M which represents the sensory cortex, henceforth referred to as the lattice
- A $p \times q \times n$ matrix C which represents the connections, where $C_{(i,j,k)}$ is the connection between the neuron at $M_{(i,j)}$ (for $i = 1\ldots p$, $j = 1\ldots q$), and the one at $R_{(k)}$ (for $k = 1\ldots n$)

For data projection a SOM works as follows, assuming an $m \times n$ data matrix D is given. For each row vector D(i) (for $i = 1\ldots m$) repeat the following two steps: (1) Present D(i) as input to R, and (2) Propagate the input along the connections C to selectively activate the cells of the lattice M; in mathematical terms this corresponds to the inner product of R with each of the connection vectors at C(i,j). The result of the inner product R.C(i,j) is stored at M(i,j): M(i,j) = R.C(i,j). Once all the data vectors have been processed there is a pattern of activations on the lattice M, and this pattern is the projection of the data matrix D.

There is a strong temptation to interpret the lattice activation pattern spatially, that is, to interpret any groups of adjacent, highly activated units as clusters, and the distance between and among clusters as proportional to the relative distances among data items in the high-dimensional input manifold. That temptation needs to be resisted. The SOM differs from the distance-based clustering methods in that the latter try to preserve relative distance relations among objects on the data manifold, whereas the SOM tries to preserve the manifold topology – cf. Kaski (1997), Verleysen (2003), Lee & Verleysen (2007: Chapter 5). To see the implications of this, it is necessary to understand that, when it is said that a SOM preserves the topology of the input manifold on the output lattice, what is meant is that it preserves its neighbourhood structure: all the vectors in a given neighbourhood are mapped to the same lattice cell, and the vectors in the adjoining neighbourhoods are mapped to nearby cells. The result is that the vectors which are close to one another in the input manifold in the sense that they are in the same or nearby neighbourhoods will be close on the SOM output lattice. The problem, though, is this: just because active cells are close together on the SOM lattice does not necessarily mean that the vectors which map to them are topologically close in the input manifold. This apparently paradoxical situation arises for two reasons – see discussion in, for example, Ritter et al. (1992: Chapter 4) and Moisl (2015: Chapter 4).

1. The topology of the output manifold on the lattice to which the SOM maps the input one must be fixed in advance. In the vast majority of applications the SOM output topology is a two-dimensional plane, that is, a linear manifold, with rectangular or hexagonal neighbourhoods which are uniform across the lattice except at the edges, where they are necessarily truncated. There is no guarantee that the intrinsic dimensionality of the input manifold is as low as 2, and therefore no guarantee that the output topology will be able to represent the input manifold well. In theory, the SOM is not limited to two-dimensional linear topology, and various developments of it propose other ones, but where the standard one is used some degree of distortion in the lattice's representation must be expected – cf. Verleysen (2003), Lee & Verleysen (2007: Chapter 5); the projection is optimal when the dimensionality of the lattice is equal to the intrinsic dimensionality of the data.
2. The dynamics of SOM training do not at any stage use global distance measures. The mapping from input to output space depends entirely on local neighbourhood adjacency. As such, the SOM cannot be expected consistently to preserve proportionalities of distance between individual vectors and vector neighbourhoods. As a result, the SOM may squeeze its representation of the input topology into the lattice in such a way that units associated with neighbourhoods which are far apart on the input manifold may nevertheless be spatially close to one another on the lattice.

In view of (1) and (2), how can a SOM lattice be interpreted so as to differentiate cells which are spatially close because they are topologically adjacent in the input manifold, and cells which are spatially close on account of the above distorting factors but topologically more or less distant? The answer is that it cannot be done reliably by visual inspection alone; interpretation of a SOM lattice by visual inspection is doubly unreliable – a subjective interpretation of an ambiguous data representation. This is a well known problem with SOMs (Kohonen 2001: 165), and a variety of ways of achieving the required differentiation exists. A frequently used one is the U-matrix (Ultsch & Siemon 1990; Ultsch 2003), which graphically shows the boundaries between areas of the lattice which are genuinely close topologically by means of peaks and troughs, or by colour coding, or by a combination of the two. The U-matrix is used here, and exemplified below.

4.4.4 *Result*

The result of the SOM projection of M' is shown in Figure 10.

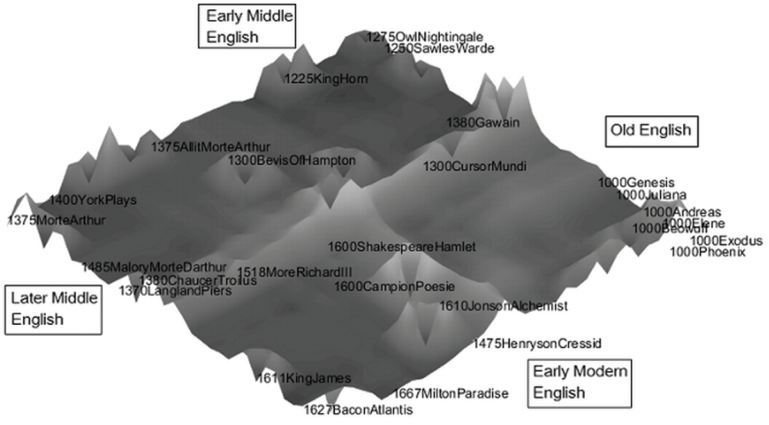


Figure 10. U-matrix representation of the two-dimensional SOM projection of the 100-dimensional frequency matrix M' abstracted from the example corpus C

The lattice shows four main topological areas separated by peaks, with sub-regions within each. The main regions correspond to the four chronological stages of English; text-labels are anchored on the left, that is, '1600ShakespeareHamlet' is, for example, located at the initial '1' on the map. As can be seen, the projection from the 100-dimensional data matrix onto a two-dimensional surface has clustered the texts in accordance with what is independently known of their dates.

5. Conclusion

Topological mapping is widely applicable to data abstracted from multi-text historical linguistic corpora:

- Where the characteristics of the corpus language are well known, as for the well-studied European languages, topological mapping can be used to bestow the fundamental scientific characteristics of objectivity and replicability on them.
- Where they are less well known, as for corpora in non-European languages, it can be used to identify objective, replicable geographical and relative chronological distributions.

References

Allinson, Nigel, Hujun Yin, Lesley Allinson & Jon Slack (eds.). 2001. *Advances in self-organising maps*. Berlin: Springer. https://doi.org/10.1007/978-1-4471-0715-6

Bertuglia, Cristoforo & Franco Vaio. 2005. *Nonlinearity, chaos, and complexity: The dynamics of natural and social systems*. Oxford: Oxford University Press.

Deza, Michel & Elena Deza. 2009. *Encyclopedia of distances*. Berlin: Springer. https://doi.org/10.1007/978-3-642-00234-2

Haykin, Simon. 1999. *Neural networks. A comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall International.

Hubel, David & Torsten Wiesel. 2005. *Brain and visual perception: The story of a 25-year collaboration*. Oxford: Oxford University Press.

Izenman, Alan. 2008. *Modern multivariate statistical techniques. Regression, classification, and manifold learning*. Berlin: Springer.

Kaski, Samuel. 1997. Data exploration using Self-Organizing Maps. Helsinki: Helsinki University of Technology PhD thesis.

Kohonen, Teuvo. 2001. *Self-Organizing Maps* (3rd edn.). Berlin: Springer. https://doi.org/10.1007/978-3-642-56927-2

Lay, David. 2010. *Linear algebra and its applications* (4th edn.). London: Pearson Education International.

Lee, John. 2010. *Introduction to topological manifolds* (2nd edn.). Berlin: Springer.

Lee, John & Michel Verleysen. 2007. *Nonlinear dimensionality reduction*. Berlin: Springer. https://doi.org/10.1007/978-0-387-39351-3

Moisl, Hermann. 2015. *Cluster analysis for corpus linguistics*. Berlin: de Gruyter. https://doi.org/10.1515/9783110363814

Munkres, James. 2000. *Topology* (2nd edn.). London: Pearson Education International.

Oja, Erkki & Samuel Kaski. 1999. *Kohonen maps*. Amsterdam: Elsevier.

Reid, Miles & Balasz Szendroi. 2005. *Geometry and toplogy*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511807510

Ritter, Helge, Thomas Martinetz & Klaus Schulten. 1992. *Neural computation and Self-Organizing Maps*. Boston: Addison-Wesley.

Strogatz, Steven. 2000. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry and engineering*. New York: Perseus Books.

Sutherland, Wilson. 2009. *Introduction to metric and topological spaces* (2nd edn.). Oxford: Oxford University Press.

Ultsch, Alfred. 2003. *U∗-Matrix: a tool to visualize cluster in high-dimensional data*. Technical report 36. Marburg: Department of Computer Science, University of Marburg.

Ultsch, Alfred & Peter Siemon. 1990. Kohonen's self-organizing feature maps for exploratory data analysis. *Proceedings of the International Neural Network Conference, INNC '90*, 305–308. Paris: Springer.

Van Hulle, Marc. 2000. *Faithful representations and topographic maps*. Hoboken, NJ: John Wiley and Sons.

Verleysen, Michel. 2003. Learning high-dimensional data. In Sergey Ablameyko, Marco Gori, Liviu Goras & Vincenzo Piuri (eds.) *Limitations and future trends in neural computation*, 141–162. Amsterdam: IOS Press.

Vesanto, Juha & Esa Alhoniemi. 2000. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks* 11. 586–600. https://doi.org/10.1109/72.846731
Xu, Rui & Don Wunsch. 2008. *Clustering*. Hoboken NJ: Wiley.

Book genre and author's gender recognition based on titles

The example of the bibliographic corpus of microtexts

Adam Pawłowski[1], Elżbieta Herden[1] and Tomasz Walkowiak[2]
[1]University of Wrocław / [2]Wrocław University of Science and Technology

The subject of this chapter is the application of automatic taxonomy methods to the corpus of microtexts, consisting of book titles. We test two hypotheses. The first one claims that simply on the basis of a book title one can automatically recognize its genre (writing species). The second assumes the possibility of recognizing the author's gender on the basis of the book's title. FastText and word2vec methods were applied. The analyses give a positive (and rather astonishing) result: with properly chosen *n*-grams more than 70% of titles could be correctly assigned a writing species, while the accuracy of the gender recognition of the author was almost 80%. Both values significantly exceed the levels of random recognition. The research was conducted on the corpus of titles derived from the Polish national bibliography.

Keywords: corpus linguistics, automatic taxonomy, gender recognition, book genre, fastText, word2vec, bibliography, Polish

1. The problem

Bibliographies have become in recent years the subject of quantitative and qualitative research conducted within the framework of digital humanities. Their volume, counted in millions of records available in digital form, is extensive enough to use NLP methods, statistics and text mining. NLP techniques allow the text to be processed at the morphosyntactic level (including, among others, lemmatization). Statistical tools enable the creation of full, quantitative descriptions of bibliographic corpora, whereas text mining methods are used to create advanced data representations and search tools, based on, among others, machine learning techniques such as word2vec, topic modelling, statistical classifiers (e.g., linear soft-max classifier) or

https://doi.org/10.1075/cilt.356.15paw

neural networks. The content of large bibliographies in the case of big data research represents a unique cognitive value from the point of view of cultural anthropology and also to some extent of scientometrics. Titles, although they belong to the class of microtexts, synthesize information important from the point of view of the author. When analyzed in large quantities, they reflect the general state of knowledge in society, its preferences, broad civilizational trends, and intellectual fashions. These qualities can be fully utilized thanks to a specific feature of bibliographic structures, not found in literary or applied texts, the presence of meticulously and methodically prepared metadata, which allows the verification of the effects of empirical research, conducted on those fields of bibliographic records that contain information in text format (primarily titles).

The presence of metadata indicating, among other things, the date and place of publication, the genre of the text and the gender of the author, plays a special role in the broader context of quantitative research into language and corpus linguistics using automatic methods. The Achilles heel of these studies is the limited possibility of verifying the results obtained, combined with a multitude of text-mining methods. This is the result of several factors:

- instability of structural features of the text (e.g., difficulty of measuring unit length, the questionable issue of text segmentation into smaller parts);
- the fluidity of semantic categories assigned to individual words or multi-word structures (meanings are constructed ad hoc in the process of sender-recipient interaction, the phenomenon of polysemia, homography, and word correlations are widespread);
- a practically unlimited number of potential results of certain text processing operations (this applies in particular to classification, but also to the generation of semantic clusters by topic modelling).

While the first two limitations are well known in empirical linguistics, the third has only in recent years been recognized as a problem. The mass availability of electronic text and the development of NLP methods has led to a situation where clustering of texts (and microtexts) is relatively easy, but evaluation of the quality of the result obtained is often impossible as the number of potential divisions of classified objects into clusters (sets) is practically unlimited (it depends on the selection of relevant features of these objects and the metrics of similarity applied). In natural language text research, testing (evaluation) of results can be carried out only by people, or by application of formal measures (e.g., minimization of the system perplexity). The situation is completely different in bibliographic corpora, where the reference system for potential tests exists in the form of metadata. The above arguments (and counterarguments) indicate the possibility of effectively processing large bibliographies as text corpora with automatic methods.

2. Data and research hypotheses

The subject of our research is a corpus of approximately 1,850,000 bibliographic records extracted from the national bibliography of the Polish National Library via the API interface.[1] The records are stored in a strongly redundant and slightly archaic MARC format, which is used in most of the bibliographic databases of the world. The fields containing the title of the work (only titles in Polish were included), the authors' data, key words and genology data (writing genre) were considered relevant for automatic analyses. The database includes almost exclusively literature published in the 20th and 21st centuries without taking into account the actual year that the text was written.

Two hypotheses were formulated in preparation for the research. The first one concerns the efficiency of recognizing the literary genre of the text solely on the basis of the title and assumes that the method of machine learning can effectively attribute specific texts to the writing genre to which they actually belong. The problem of genology in this case was a complex one due to the number of genres and inconsistencies in the description of classification traits. Working solely on metadata, we distinguished as many as 7000 writing species (or genres). This number was then reduced to the set of 52 most salient genres because the smallest classes were eliminated (many of them were the outcome of human error). Further steps of text processing included the elimination of genres equivalent in terms of subject matter of publication but bearing different names. The final list obtained in this way consisted of 31 items and allowed for an effective verification of the automatic classification results (cf. § 4.1).

The second hypothesis is riskier but worth consideration if only to eliminate possible assumptions about its validity. Recently, literature on the subject includes studies devoted to automatic recognition of the 'cultural gender of the text'. These studies focus primarily on belles-lettres (cf. Rybicki 2016; Walkowiak & Piasecki 2018), but also on microtexts available in social media, i.e., tweets and text messages (Mikros 2013; Mikros & Perifanos 2013; Schwartz et al. 2013; Silessi et al. 2016). Having a corpus of titles at our disposal, we assumed that it would be possible to automatically recognize the gender of the author, especially in the case of genres that allowed the author relative freedom in composing the title (mainly in belles-lettres and biography). It was assumed that the hypothesis would be positively verified if the attribution obtained automatically would be significantly better than the purely statistical probability of gender attribution. As an additional working hypothesis, we posited that the effectiveness of recognizing 'gender' would be directly proportional

1. http://data.bn.org.pl/

to the length of the title (for example, it is doubtful that a good result will be obtained on the basis of one-word titles). Both hypotheses can be verified based on the metadata contained in the records, which allows for virtually flawless validation.

3. Methodology

In our research, we used a recently developed deep learning package called 'fastText' (Joulin et al. 2017). It consists of two different approaches: supervised and unsupervised. The first, that we call here 'supervised fastText', is based on the representation of documents (doc2vec) as an average of word embedding (word2vec) and uses a linear soft-max classifier (Goodman 2001) to assign the doc2vec representation to one of a range of known classes. This hidden representation is used by a linear classifier for all classes (i.e., literary genres and the author's gender), allowing information about word embedding learned for one class to be used by others. 'Supervised fastText' by default ignores word order, much like the classical bag of words (BoW) method (Harris 1954). The main idea behind 'supervised fastText' is to perform word embedding and classifier learning in parallel (simultaneously). Since 'supervised fastText' forms the linear model, it is very effective for training and achieves solutions faster by several orders of magnitude compared with other competing methods (Joulin et al. 2017). During classification, words that do not exist in the embedding model (because they do not exist in the training corpus) are omitted from the averaging. 'FastText' allows us to build the embeddings not only for single words but also for word *n*-grams. It allows local word order to be taken into account.

The second approach, referred to here as the 'fastText language model', is an extension of the word2vec (Le & Mikolov 2014) model built on Common Crawl and Wikipedia by 'fastText' in unsupervised mode. The models were trained by CBOW (Continuous Bag of Words model) with position weights and subword information (Grave et al. 2018).[2] The word representation is constructed as the sum of the character *n*-grams embeddings (for *n*-grams appearing in the word). It allows for generation of word embedding for words not seen in a training corpus and for working with inflected languages such as Polish. Since the fastText language model provides vector representations of individual words, we have represented documents as an average of these vectors (Mikolov et al. 2013). The vector representations were classified by the Multi-Layer Perceptron (MLP) based on a model trained by the back error propagation method on the learning set (Hastie et al. 2013).

2. https://fasttext.cc/docs/en/crawl-vectors.html

4. Experiments and results

4.1 Recognizing the literary genre of the text

In order to test the first hypothesis, the number of classes had to be limited based on the criterion of the number of elements. It was initially assumed that effective classification was possible if the class had no fewer than 5000 titles, while the less numerous classes were rejected (in such cases the training corpus would be too small to be effective). This allowed us to identify 52 writing genres, which included a total of about 570,000 titles. The data were randomly divided into training and testing set in the proportion of 2 to 1. This proportion was used throughout all experiments performed. The primary results of automatic classification using the 'supervised fastText' method (with word unigrams) obtained an accuracy rate of 54% on the test set. Detailed analysis showed that classes not recognized correctly by the algorithm had in fact a large similarity to certain classes recognized correctly (for example, sub-genres of different types of textbooks, novels, and stories were clustered together). This means that mistakes in attribution are actually due to polysemy in the language, where 'collections' and 'anthologies', for example, are treated as separate categories (although in reality they can describe the same objects).

As a result, the number of genre classes was reduced to 31, namely: academic textbooks (Pol. *podręczniki akademickie*), novels (Pol. *powieści*), anthologies (Pol. *antologie*), conference materials (Pol. *materiały konferencyjne*), popular publications (Pol. *wydawnictwa popularne*), guides (Pol. *poradniki*), biographies (Pol. *biografie*), albums (Pol. *albumy*), diaries (Pol. *pamiętniki i wspomnienia*), textbooks for vocational schools (Pol. *podręczniki dla szkół zawodowych*), stories (Pol. *opowiadania i nowele*), children's literature (Pol. *literatura dla dzieci*), travel guides (Pol. *przewodniki turystyczne*), textbooks for primary schools (Pol. *podręczniki dla szkół podstawowych*), developing (Pol. *poradniki rozwoju osobistego*), Polish journalism (Pol. *publicystyka polska*), support materials (Pol. *dokumenty towarzyszące*), comics (Pol. *komiksy*), children's poetry (Pol. *poezja dla dzieci*), youth novel (Pol. *powieść młodzieżowa*), religious considerations and meditations (Pol. *rozważania i rozmyślania religijne*), historical literature (Pol. *literatura historyczna*), publications for children (Pol. *wydawnictwa dla dzieci*), textbooks for high schools (Pol. *podręczniki dla szkół ponadgimnazjalnych*), statistical data (Pol. *dane statystyczne*), analyses and interpretations (Pol. *analizy i interpretacje*), commemorative books (Pol. *księgi pamiątkowe*), exercises and tasks (Pol. *ćwiczenia i zadania*), bibliography (Pol. *bibliografia*), encyclopaedias (Pol. *encyklopedie*), compendia and indexes (Pol. *kompendia i repertoria*).

Table 1. Accuracy of the attribution of titles to their literary genres

Classification method	Accuracy
supervised fastText, unigrams	64.9%
supervised fastText, bigrams	71.1%
supervised fastText, trigrams	**71.8%**
supervised fastText, fourgrams	71.7%
supervised fastText, fivegrams	71.5%
fastText language model+MLP	66.2%

The results of automatic classification using the 'supervised fastText' method with word embeddings built on single words, word bigrams, trigrams etc., and verified on the basis of the metadata included in the database records, turned out to be surprisingly satisfactory. The relationship between the length of word *n*-grams and the accuracy of the attribution (Figure 1) prompted the observation that the most effective assignments can be carried out on tri-word titles. However, even single word titles allow us to achieve a nontrivial level of recognition of a literary genre. The accuracy in this case was at 64.9%, while random attribution accuracy for the most frequent class of academic textbooks in a test set would be at 19.6%. Although tri-word titles are the most effective in absolute numbers (71.8%), in fact, a caesura occurs between single- and multi-word titles. Interestingly, an increase in the length of the title (4-grams, 5-grams etc.) does not lead to an increase in the accuracy of assignment of titles to their proper classes of literary genre. This fact can be interpreted as an effect of significant limitations of the title naming system and its semantic saturation. Despite appearances, this system does not allow for free

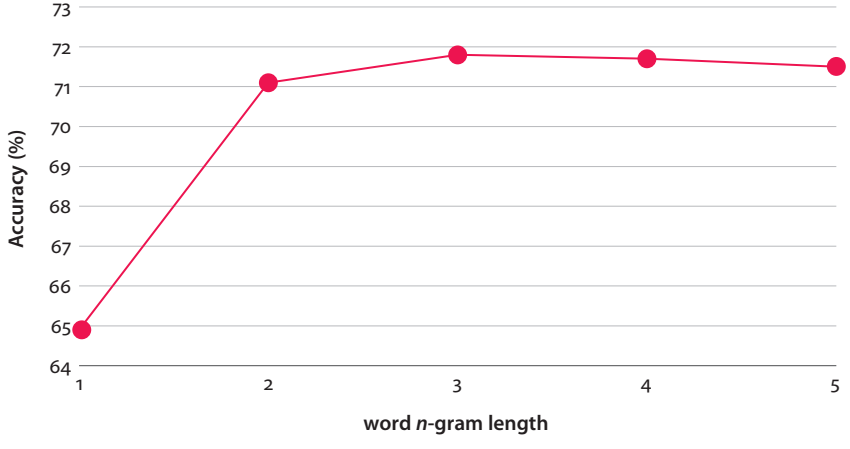


Figure 1. Accuracy of the literary genre attribution for the 'supervised fastText' method as a function of word *n*-grams length

choice of the linguistic and/or stylistic measures. Therefore, the automatic method based on a well-trained algorithm allows for effective attributions of even very short microtexts, which would be less probable in the case of free speech samples or other text species.

We also conducted a detailed analysis of the attribution accuracy for selected literary genres (treated here as classification categories). First, we measured the number of correct decisions from all assignments made to a target specific class (precision), and the number of correct decisions from all assignments expected to a specific class (recall). Table 2 gives the precision and recall values for the first four genre classes (about 50% of all titles). In all cases, the quality of automatic matching of titles to the appropriate classes proved to be very high.

Table 2. Quality of writing genre attribution by automatic method (first 4 genres)

Writing genre	Precision	Recall	Support
Academic textbooks	88.2%	82.1%	19.6%
Novels	81.9%	67.8%	14.3%
Anthologies	68.5%	64.5%	11.1%
Conference materials	82.4%	80.6%	7.77%

The investigation of false decisions of the algorithm applied also proved interesting. Some text genres were intensely confused, which could suggest a method error. Table 3 shows that, for example, in the class indexed manually as 'studies' (Pol. *opracowania*) only 14.3% of titles were attributed correctly, while 20.6% were assigned to 'academic textbooks'; 28.8% of the titles indexed manually as 'children's poetry' were correctly recognized, but 21.1% were assigned to the 'anthology' class. Similarly, the algorithm assigned 24.7% of titles indexed manually as 'youth novel' to the 'novel' class, and part of 'support materials' was recognized as 'academic textbooks'. But on closer examination, it turns out that decisions based on the 'supervised fastText' algorithm are correct, because at the stage of manual indexing of documents, due to human errors, some titles of works apparently similar in terms

Table 3. Genres most frequently misclassified (selected 4 genres)

Writing genre (A)	Classified correctly (precision)	Misclassified as (B)	Percentage of genre A classified wrongly as genre B
Studies (Pol. *opracowania*)	14.3%	Academic textbooks	20.6%
Children's poetry	28.8%	Anthology	21.1%
Youth novel	61.3%	Novels	24.7%
Support materials	59.5%	Academic textbooks	19.1%

of content were placed in different classes. In fact, children's poetry collections are also anthologies, a youth novel is a novel, and 'studies' or 'support materials' are also academic textbooks. And however paradoxical this may sound, it can be claimed that the 'supervised fastText' algorithm, trained on a sufficiently large database, is capable of correcting errors or inconsistencies in human decisions.

Our research shows that the decision to choose an analysis algorithm is not obvious and there are several possible options here. Since the 'supervised fastText' method is not able to take into consideration words not present in the training set, we also tested the 'fastText language model' (see § 3) trained on a huge corpus of Polish texts capable of calculating vector representations of words for unseen words (due to a usage of character *n*-grams). The averaged word vectors were classified by the MLP (Multi-Layer Perceptron) classifier, achieving accuracy of 66.2%. The results (see Table 1) were better than for the 'supervised fastText' method, in that usage of word unigrams had a lower success rate compared with usage of longer word *n*-grams.

4.2 Automatic recognition of the author's gender

Another important aim of the research was to automatically recognize the author's gender and evaluate the result obtained. This task was carried out in four stages. At the beginning all titles of multi-author works were eliminated, so that the author's gender category was unambiguous. The second phase of the experiment included automatic recognition of the author's gender and indexing database records with 'M' or 'F'. This operation was necessary because metadata do not provide this information (apparently considered obvious). This task turned out to be relatively easy because Polish as an inflected language is characterized by a very useful feature, which is the almost common occurrence of the female first name suffix *-a* ([cons.] *-a*). Male names have different endings, where consonants and sonants (l, r) dominate. The gender of non-Polish authors of translated books, as well as some rare names, were recognized semi-automatically. The third phase of the experiment consisted in eliminating those literary genres where titles are constructed according to more or less strict rules preventing the free linguistic expression of gender. This applies primarily to texts related to science, education and administration. Genres giving more freedom of titling, defined broadly as belles-lettres, poetry, drama, biography, or some 'how-to' guides, were left in the corpus. We obtained in this way about 280,000 well profiled and annotated titles where 71% of authors were male and 29% female. The last phase consisted in the application of the 'fastText' algorithm to perform automatic attribution of author's gender using only titles. The task was carried out separately for titles composed of one word, two words,

etc. The best result was obtained for four- and five-word titles (79%). Yet, as in the case of literary genre attribution (see above), a slightly lower value for unigrams can be observed, followed by a steep increase and stable values for longer titles (Table 4 and Figure 2).

Table 4. Accuracy of the author's gender attribution based on the title for selected genres

Classification method	Accuracy
supervised fastText, unigrams	75.7%
supervised fastText, bigrams	78.2%
supervised fastText, trigrams	78.9%
supervised fastText, fourgrams	79.0%
supervised fastText, fivegrams	79.0%
supervised fastText, sixgrams	78.9%

Estimating this result is not easy. In the case of a proportional distribution of authors, it should be assumed that any value significantly higher than 50% will be considered satisfactory (random attribution in such a corpus of texts should yield a 50% success rate). The 79% value is therefore significantly higher and could be considered satisfactory. However, in this case the distribution of authors is not balanced and it is much more advantageous to estimate the gender attribution separately for male and female authors. The result of this attribution (Table 5) turned out to be very good, as both precision and recall values were about 20 percentage points

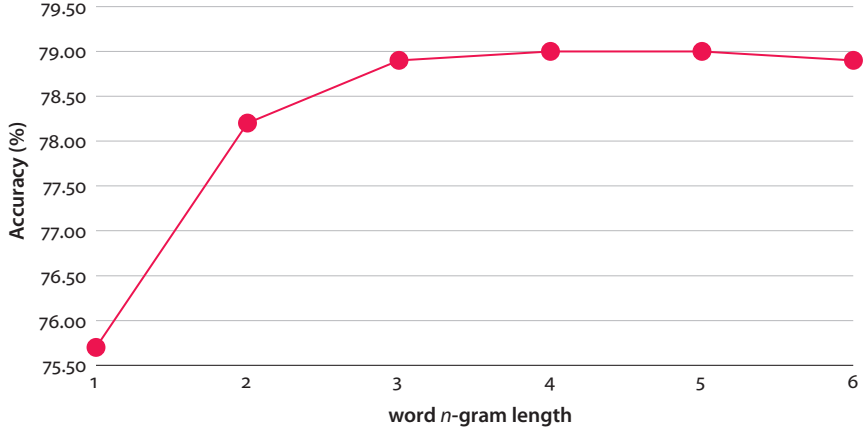


Figure 2. Accuracy of the gender attribution for 'supervised fastText' method in a function of word *n*-grams length

higher than the values of random attribution. Of course, the question remains open as to whether a 100% attribution is possible. Given the complexity of the human mind and the impact of culture on gender, this seems impossible – at least for very short and standardized microtexts, such as titles (it would be probably feasible in the case of free text, e.g., in social media).

Table 5. Author's gender attribution based on four-word book titles for selected genres

1 Gender	2 Precision	3 Recall	4 Expected random attribution (support)
Male	90.8%	81.8%	71%
Female	49.6%	68.4%	29%

Using this result, a test of a subsidiary hypothesis was also carried out, which was taken for granted at the outset, namely that there are genres that allow for the free shaping of titles and those that impose the structure and vocabulary of the titles. Table 6 contains the results of the author's gender recognition accuracy test across the entire corpus of titles (over 855,000 items). The result turned out to be surprising, as the quality not only did not decrease, but even increased slightly. The results for different word *n*-grams in titles are presented in Table 6. Table 7 lists the precision and recall values obtained for word 4-grams.

Table 6. Accuracy of the author's gender attribution based on the title for all genres

Classification method	Accuracy
supervised fastText, unigrams	76.5%
supervised fastText, bigrams	80.3%
supervised fastText, trigrams	81.4%
supervised fastText, fourgrams	81.9%
supervised fastText, fivegrams	82.0%
supervised fastText, 6-grams	82.0%
supervised fastText, 7-grams	82.0%
supervised fastText, 8-grams	81.9%

Table 7. Author's gender attribution based on four-word book titles for all genres

Gender	Precision	Recall	Expected random attribution (support)
Male	91.4%	85.7%	76%
Female	50.9%	65.0%	24%

Does this result mean that titles of scientific or law books can be considered an expression of the author's gender? Of course not. The high values of these parameters in the most standardized genres probably stem from the fact that certain subject

areas are gender-specific. It is therefore not the case that the lexemes such as 'law', 'medicine', 'administration' or 'physics' connote masculinity or femininity. Rather, publications on certain topics are (or have been) more often written by women or men and this principle was learned by the 'supervised fastText' algorithm. In fiction and other 'free' genres there is no such dependence, which makes the previously obtained assignments attributes valuable from a cognitive point of view.

5. Conclusions

The research conducted strongly supports the view that automatic taxonomy of microtexts as short as book titles is possible and gives positive results. Previous studies in microtext taxonomy and gender recognition were also successful but they were carried out on tweets which, despite being considered short, seem extremely elaborate and rich in content compared to book titles (cf. Mikros & Perifanos 2013). Available research in book title automatic processing is rather scarce: it concerns only 'book genres' and was conducted on smaller corpora derived from Amazon databases (Ozsarfati et al. 2019; Chiang et al. 2015). In our study an effective recognition of the book genre (called also writing species) as well as the author's gender was conducted on the basis of two-, three- and four-word sequences. We worked on an extensive bibliographic corpus in a flexional language (Polish). The result obtained is unexpected and intuitively not obvious, as artificial intelligence in application to language data derived from great bibliographies has apparently exceeded human capabilities. Nevertheless the study seems difficult to challenge, as it is methodologically sound and based on solid empirical material.

Unfortunately, we do not have the results of similar tests carried out on human respondents. But our intuition and many years of scientific experience suggest that an artificial intelligence system should give 'human-like' results in recognizing a genre of a book or of a scientific paper based just on its title. This is due to the fact that meanings of words or expressions in titles are more or less universally reproduced in the minds of persons speaking a given language and computer algorithms can only replicate this competence. Yet recognizing the author's gender solely from the titles would be difficult for human respondents as it requires something more than just semantic knowledge. Researchers undertaking this task should be familiar with the entire knowledge base developed in the course of machine learning from the training set of data – in this case a bibliographical corpus consisting of hundreds of thousands of items. This explains, why it may seem that a computer has a linguistic competence that exceeds that of a human being. The source of this success, apart from the narrowly understood technology (memory capacity, processing speed etc.), is most likely the philosophy of artificial intelligence algorithms

(here 'fastText'), which assume extensive training on large and reliable data in order to build a knowledge base. This makes it possible to effectively process new items, such as words, clauses, sentences etc., created within the same communication system in a given language.

At the end of these reflections, it is worthwhile to ask a more general question: is AI the future of linguistics? The matter seems delicate, as humans have always felt a subconscious fear of 'thinking' machines. This archetype is deep-rooted, as it dates back to biblical times, where the beginning of the story of artificial beings threatening humans is found (e.g., Golem). But today the facts are such that the number of texts artificially generated by algorithms is growing (it is hard to say, for example, who in the Amazon databases quoted above created the book metadescriptions – humans or computers). In addition, we can observe a massive increase in machine-readable texts that people are no longer able to acquire or analyze. These reasons make automatic, complex and seemingly opaque algorithms of language analysis and creation the inevitable future of linguistics.

Funding

This research was supported by Polish National Science Center (NCN) under grant 2016/23/B/HS2/01323 "Methods and tools of corpus linguistics in the research of a bibliography of Polish book publications from 1997 to 2017".

The NLP tools used in this study were developed by the CLARIN-PL consortium.

References

Chiang, Holly, Yifan Ge & Connie Wu. 2015. *Classification of book genres by cover and title*. http://cs229.stanford.edu/proj2015/127_report.pdf (7 September, 2020).

Goodman, Joshua. 2001. Classes for fast maximum entropy training. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, I-561–I-564. Salt Late City, UT: IEEE. https://arxiv.org/pdf/cs/0108006.pdf (7 September, 2020.) https://doi.org/10.1109/ICASSP.2001.940893

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Nicoletta Calzolari et al. (eds.), *Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1550.pdf (7 September, 2020).

Harris, Zellig S. 1954. *Distributional structure. WORD* 10(2–3). 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2013. *The elements of statistical learning: Data mining, inference and prediction* (Springer series in statistics). New York: Springer.

Joulin, Armand, Edouard Grave, Piotr Bojanowski & Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Mirella Lapata, Phil Blunsom & Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/E17-2068.pdf (7 September, 2020). https://doi.org/10.18653/v1/E17-2068

Le, Quoc & Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Eric P. Xing & Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196. Bejing: JMLR. http://proceedings.mlr.press/v32/le14.pdf (7 September, 2020).

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé , III & Katrin Kirchhoff, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics, 746–751. https://www.aclweb.org/anthology/N13-1090.pdf (7 September, 2020).

Mikros, George K. 2013. Systematic stylometric differences in men and women authors: a corpus-based study. In Reinhard Köhler & Gabriel Altmann (eds.), *Issues in Quantitative Linguistics 3: Dedicated to Karl-Heinz Best on the Occasion of His 70th Birthday*, 206–223, Lüdenscheid: RAM–Verlag. https://www.academia.edu/3429459/Systematic_stylometric_differences_in_men_and_women_authors_a_corpus-based_study (7 September, 2020.)

Mikros, George K. & Kostas Perifanos. 2013. Authorship attribution in Greek tweets using author's multilevel n-gram profiles. In Eduard Hovy, Vita Markman, Craig Martell & David Uthus (eds.), *AAAI Spring Symposium: Analyzing Microtext*. https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/viewFile/5714/5914. (7 September, 2020.)

Ozsarfati, Eran, Egemen Sahin, Can J. Saul & Alper Yilmaz. 2019. Book genre classification based on titles with comparative machine learning algorithms. In *IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 14–20. Singapore: IEEE Press. https://doi.org/10.1109/CCOMS.2019.8821643

Rybicki, Jan. 2016. Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities* 31(4). 746–761. https://doi.org/10.1093/llc/fqv023

Schwartz, Roy, Oren Tsur, Ari Rappoport & Moshe Koppel. 2013. Authorship attribution of micro-messages. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu & Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA: Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1193.pdf (7 September, 2020.)

Silessi, Shannon, Cihan Varol & Murat Karabatak. 2016. Identifying gender from SMS text messages. In *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 488–491. Anaheim, CA: IEEE. https://doi.org/10.1109/ICMLA.2016.0086

Walkowiak, Tomasz & Maciej Piasecki. 2018. Stylometry analysis of literary texts in Polish. In Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz & Jacek M. Zadura (eds.) *Artificial Intelligence and Soft Computing* (Lecture notes in Artificial Intelligence 10842), 777–787. Cham: Springer. https://doi.org/10.1007/978-3-319-91262-2_68

Quantitative analysis of bibliographic corpora

Statistical features, semantic profiles, word spectra

Adam Pawłowski, Krzysztof Topolski and Elżbieta Herden
University of Wrocław

The subject of this chapter is bibliographic corpus analysis, with data from the Polish national bibliography from the period 1801–2019. The research allowed us to discover and compare quantitative characteristics of the bibliographic corpus and of the reference corpus of general language. It was shown that the two corpora differ significantly. In particular, differences in the share of particular parts of speech and of the frequency distribution of lexemes were demonstrated. The statistical distributions of word spectra were also studied. The best fit was obtained for generalized inverse Gauss-Poisson and Zipf-Mandelbrot distributions. The analysis of parameters of both distributions for bibliographic and reference corpora also revealed differences between them. The best perspective for future research on bibliographic corpora is, apart from quantitative linguistics, semantic analysis and text-mining.

Keywords: quantitative linguistics, corpus linguistics, word spectra, statistical distributions, MARC, bibliography, Polish, book titles

1. Large-scale bibliographies as text corpora

Originally created as registers of literary output, large-scale bibliographies have to date served more practical purposes, in particular for publication searches, their archivization, and/or evaluation of scientific output of institutions, disciplines, and individuals. However, over time, the cognitive potential they have developed has been recognized, allowing them to be treated not as auxiliary tools but as fully-fledged research objects. The conditions for undertaking research on large-scale bibliographies were numerous and, it seems, all of them have now been met. First of all, the critical mass, i.e., the volume of data needed to draw valuable conclusions and generalizations of a scientific nature, has long since been exceeded. Secondly, this data has become available to researchers in digital form. Thirdly, methods of automatic natural language processing have emerged which allow quick analysis of fields encoded in

https://doi.org/10.1075/cilt.356.16paw

text format, moving away from the bibliography as storage place. In particular, the processing of large collections consisting of text fields of bibliographic records has been made possible through the use of corpus and quantitative linguistics.

It is noteworthy that the application of quantitative methods in 'library science' is not a completely new phenomenon. Such a view would unjustifiably depreciate the achievements of past researchers, who were not inferior to contemporary researchers, and often put forward great ideas, but who did not have computer tools to implement innovative ideas. One of the first methodical applications of statistics in the study of bibliographic inventories is the work of Gustav Schwetschke (Schwetschke 1850; Schwetschke 1877), who used the catalogues of the bookfairs in Frankfurt am Main and Leipzig from 1564–1846 (the German national bibliography for that period did not exist at the time) to present the geographical and numerical distribution of the German bookselling industry from the 16th to 19th century.

Another example of the use of traditionally understood statistics in bibliography research is research conducted in France in the 1950s by Lucien Febvre & Henri-Jean Martin (1958), founders of the *Annales* school. Their groundbreaking work *L'apparition du livre* argued that books were not only carriers of ideas, but also material objects and as such commodities subject to the laws of economics. From more recent studies it is worth mentioning the monumental *The Cambridge History of the Book in Britain* (CHBB 1999–2019), covering the history of the book in the British Isles from 400 to the end of the 20th century. The compendium uses rich bibliographic material to recreate the processes of the creation, production, distribution and reception of the book in the British Isles.

Newer studies apply more advanced methods of statistics and methodology developed within the framework of digital humanities. This type of work consists in analyzing old, digitized bibliographies and comparing them with the resources of modern incunabula or old print databases to determine the state of preservation of old printing or publishing production (Green et al. 2011). In this context, it is worth noting the interdisciplinary research conducted by the Helsinki Centre for Digital Humanities. Researchers associated with the Centre document the history of book production in Finland and Sweden based on the analysis of large collections of bibliographic metadata (databases of retrospective national bibliographies) (Tolonen et al. 2019a; Lahti et al. 2019). They introduced the term 'bibliographic data science' for this new research paradigm (BDS)[1] (Tolonen et al. 2019b).

This review indicates that the development of bibliography research in recent years is a fact, but also reveals the one-sidedness of the approach taken so far. In

1. NB, by introducing the concept of statistical analysis based on bibliographic data, Finnish researchers are actually using a bibliographic method that is well established in the 19th century. Its transfer to the digital sphere obviously creates new research opportunities.

particular, there is a noticeable lack of application of quantitative linguistics methods that is likely to reveal new, cognitively valuable aspects of bibliographies understood as a specific informative text genre. The advantages of bibliographic corpora, important from the perspective of quantitative linguistics, include size (they consist of hundreds of thousands and even millions of records), careful preparation (data are 'clean', because they are input by competent employees and not by unprepared users), systematic approach (there are exact dates of each entry, and missing dates can be easily reproduced) and repeatability of structures (each record is assumed to have the same structure). The weakness of bibliographies, when compared to real text corpora, is the lack of longer discursive fragments that convey more complex information than just titles, keywords or generic qualifiers. This fact, however, is not an obstacle to effective quantitative and text-mining tasks. The change in approach to large-scale bibliographies is due to the fact that computer systems allow for automatic extraction of complete sets of data from selected fields (e.g., title or keywords) and their comprehensive analysis by NLP methods. Thanks to this, these large data resources, initially created as databases, become valuable research material for corpus and quantitative linguistics.

2. Data and hypotheses

The subject of this analysis is a collection of 553,000 records extracted from the resources of the Polish National Library. They represent contemporary texts from the years 1997–2017 and a small number of re-editions of works written earlier (mainly in the 19th century). Records are stored in MARC21 format, used worldwide in database systems of libraries. This format contains a large number of fields and is difficult to process automatically because it is highly redundant (the same information may be repeated in different fields).

From a formal point of view (irrespective of repetitions due to redundancy), the list of fields suitable for linguistic research is as follows:

- author (field 100, subfield 'a', if subfield 'e'='autor' or field 700, subfield 'a', if subfield 'e'='redakcja' or 'e'= 'autor');
- title (field 245),
- subject according to the list of subject headings list of the Polish National Library (various fields starting with the digit six – 6xx);
- genre (field 655).

In our case, the title field 245 was selected for the study because it contains the most complete text data in the form of sentence equivalents or full sentences. Although it is difficult to speak of a corpus analysis in the full sense of the word here because

in a bibliographic corpus there is no category of a text as a larger, compact whole, the titles fulfil basic communication functions and have such a complex structure that the use of corpus tools is fully justified. Other fields in text format ('subject' or 'genre') intended for machine-processing, on the other hand, contain only one-word terms that can be machine-processed with corpus or information retrieval tools.

Since the large corpora of titles have not been yet analyzed from the perspective of statistical linguistics, the research conducted was primarily exploratory in nature. Basic quantitative parameters of the corpus were calculated (average lengths of words, sentences, titles, as well as histograms of distributions of these parameters). Frequency lists of words in titles were also prepared and analyzed, as well as the distribution of PoS (part of speech) shares in the corpus, which made it possible to create a morphological and a semantic profile of this type of textual resource and to present it against the background of the general language. The nature of these tasks required lemmatization of the text, which was carried out with the use of the WCRF Tagger.[2]

The research hypothesis was that the corpus of titles will be significantly different from the corpus of the general language, which should be reflected, among others, in the values of its quantitative parameters. In order to verify this hypothesis, statistical distributions of vocabulary (so-called word spectra) were generated for the corpus and for the data from the National Corpus of Polish Language. Their parameters were estimated and their values were compared. Apart from that, comparisons of the POS in the two above mentioned corpora were made, and the relationship between the length of the title and the mean length of the word was also analyzed.

3. Research method

Data for research were obtained through the programming interface (API) of the National Library BN Data.[3] It allows us to extract large sets of records or download files with databases containing BN resources representing bibliographical records from longer periods of time. Excerption of the fields containing textual data from the MARC records was performed with our own programming tools. Pre-processing of data for the study required lemmatization, which was performed with the use of WCRFT2 morphosyntactic analyzer, available in the CLARIN-PL

2. http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki/

3. http://data.bn.org.pl/

infrastructure.[4] This tool also enabled automatic recognition and annotation of POS in the text. Estimation of probability distributions and other statistical tests were conducted with the use of package R. The general methodological principle of corpus linguistics was adopted, which requires that special corpora be compared with reference ones. Following this principle, results generated on the basis of the bibliographic corpus were compared with the National Corpus of Polish (NKJP).[5] Since not only single words but also sentences were analyzed in the reference corpus, it was necessary to use the corpus manually segmented into sentences. Language is a complex phenomenon and automatic segmentation of Polish texts into sentences gives a high percentage of errors; there are furthermore different ways of calculating sentence length (numbers, units of measurement, auxiliary symbols, etc. can be noted differently). For this reason, we did not use the entire NKJP resource, but only the manually annotated 1-million-word subcorpus.[6] This was not an obstacle to conclusions, because in this case the increase in the volume of the corpus did not significantly increase the quality of estimation of the parameters examined.

4. Results: An overview

The results gave an interesting picture of the corpus of titles, confirming the advanced hypotheses. For general statistics on the two corpora, see Table 1. As one can see, the statistical profile of the title corpus differs significantly from that of the reference corpus. Words in titles are on average longer (there are most probably fewer function words), while titles are on the average shorter than sentences in the general language.

Table 1. Basic statistical parameters of the bibliographic and general corpus

Unit	Bibliographic corpus		General language	
	Mean length	Stand. deviation	Mean length	Stand. deviation
Word (letters)	$m = 6.49$	$d = 3.81$	$m = 5.78$	$d = 3.36$
Sentence / clause (words)	–	–	$m = 13.06$	$d = 11.57$
Title (words)	$m = 7.32$	$d = 5.53$	–	–

4. http://ws.clarin-pl.eu/tager.shtml

5. http://www.nkjp.uni.lodz.pl/

6. http://www.nkjp.pl/index.php?page=14&lang=1

The analysis of the histograms of the length of titles and their deviations in both corpora gave quite an interesting picture, unheard of in the case of the general language. As can be seen in Figure 1, short titles (from 2 to 4 words) dominate in contemporary publications, while the decrease in the number of longer titles is (surprisingly) almost linear. In order to verify this result, a similar analysis was carried out on the reference corpus. It was assumed that the equivalent of the titles in the bibliographic corpus would be sentences in the general language (represented here by the National Corpus of Polish). Sentences are not the same as titles, just as bibliographic corpora do not show all communication functions of language. In both cases, however, these units are basic carriers of meaning, they are semantically and syntactically closed, and also their length is, at least seemingly, similar. The result of the comparison was in this case difficult to predict, so the approach used was necessarily purely exploratory. Due to the specificity of the data, titles of up to 10 words and sentences of up to 20 words were considered.

Figure 2 shows that titles and sentences have generally similar distributions but differ due to the more complex structure of the general language. What makes them similar is the existence of an extremum to which the values increase and then decrease. This extremum indicates the most common, and therefore communicatively optimal, length of respective units. As one can see, this optimum is in a different place for both corpora. Communication by means of titles takes place through book covers. It requires great conciseness and simplicity: a good title should be captured in the blink of an eye. Normal communication by means of texts or speech is not subject to such pressure: nothing has to be understood at a glance. The curve's shape in Figure 2 corresponds to the natural process of perception, which is determined

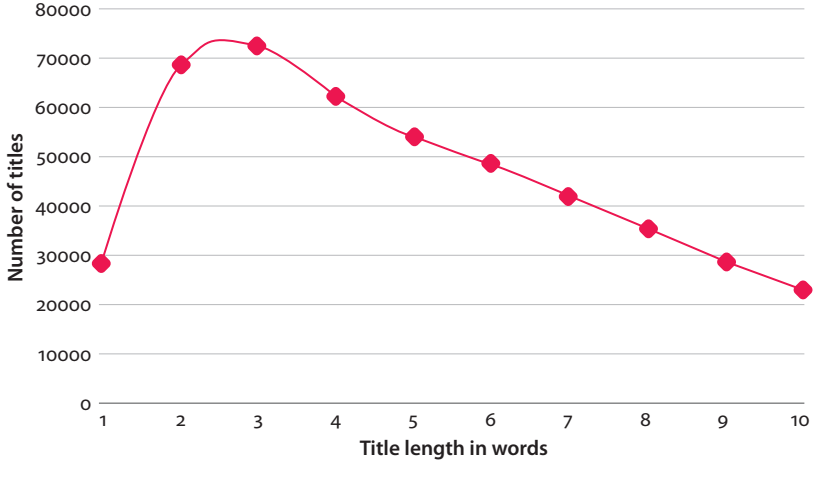


Figure 1. Histogram of title lengths in the bibliographic corpus

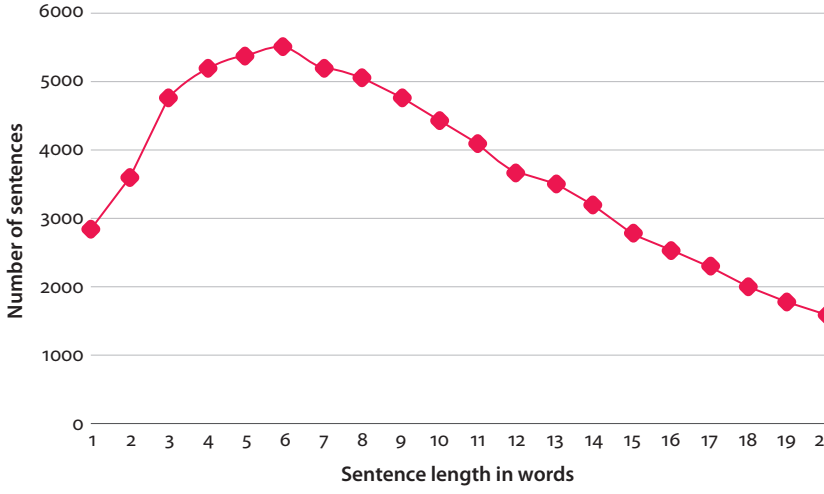


Figure 2. Histogram of sentence lengths in the general language (National Corpus of Polish)

by Zipf's forces minimizing the effort required to transfer / acquire a given amount of information. This natural sentence length would be close to 7–8 units (thus much longer than a title length).

An even more unusual result is the histogram of the average word length in the title plotted against the title length measured by the number of words (Figure 3). Here, in the case of three- and four-word titles (the most frequent), an anomaly in the form of a sudden drop appears. This anomaly results from the morphosyntactic

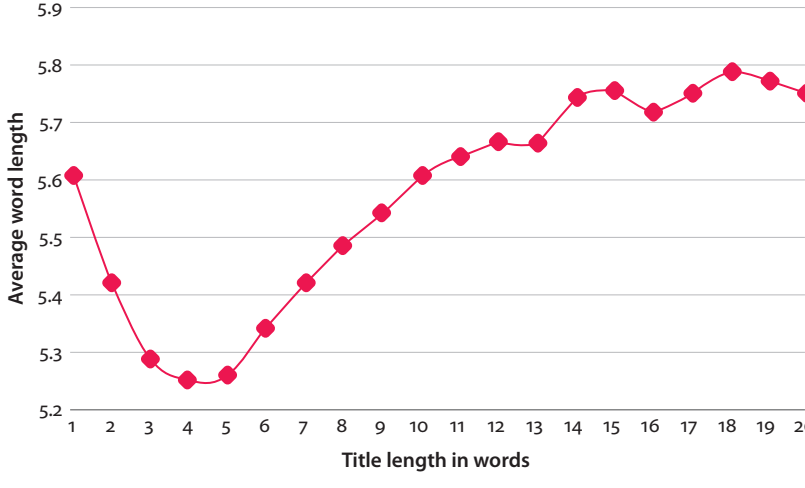


Figure 3. Average word length in a title vs. title length in the bibliographic corpus

structure of the Polish language and of other flectional languages. Three- and four-word titles are almost obligatorily expressed by two or three meaningful lexemes and a function word which determines the relation between them. This is illustrated by the structure of titles such as *Podręcznik dla ośmioklasistów* (Textbook for Eighth Graders), *Droga do Emaus* (The Road to Emmaus), or *Ułan i dziewczyna* (The Lancer and the Girl). With longer phrases, the number of function words no longer deforms the result, because the number of lexical segments also increases. It can also be expected that in analytical languages using articles this anomaly will appear in quite frequent 3- and 4-word titles (e.g., *Histoire de la terre* "History of the Earth"). However, a histogram prepared according to the same principles in the case of agglutinative languages would certainly look different.

As in previous cases, the corpus of titles was compared with the corpus of the general language. Figures 3 and 4 display a seemingly similar pattern: a sharp decline in average word length followed by an increase. However, differences are noticeable between both corpora. The average word length in titles is lower than the similar parameter in the general language. As we have already said, this is probably due to the fact that titles should be simple to read, thus composed of shorter words.

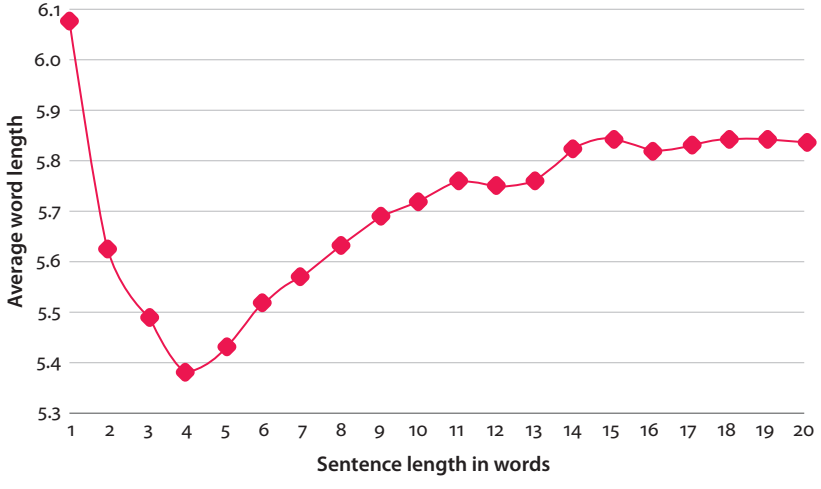


Figure 4. Average word length in a sentence vs. sentence length in the general language (National Corpus of Polish)

The analysis of the relationship between the length of the titles, as well as sentences in the corpus of the general language, and the average length of the words they contain might suggest that Menzerath-Altmann's law would be applicable here. This law states that 'the increase of the size of a linguistic construct results in a

decrease of the size of its constituents'. However, such a relation is valid only for the immediate constituents – and in the case of a sentence such a constituent is a clause, not a word. The same applies to titles which additionally do not have a clear syntactic status. It should come as no surprise, therefore, that the curves in Figures 3 and 4 do not show any law-like tendency, at least from the point of view of quantitative linguistics.

Significant results were obtained in the study of the POS distribution in the corpus of titles as opposed to the general language. The numbers in Table 2 indicate significant differences between the Polish language used in normal communication and the specific language of the title. The language of titles is generally highly nominalized (the proportion of nouns in the title corpus is 57.15%, and in the general language 43.45%), while the proportion of verbs in the bibliographic corpus is significantly lower. Clear differences are also visible in the pronoun group (0.43% vs 1.97%). This is due to the fact that titles are most often impersonal phrases (e.g., *Basics of cellular biology*), while in the general language, the frequency of pronouns increases with the presence of dialogue. While the above differences are due to purely genological specificities, the very high frequency of numbers is due to the fact that titles (and subtitles) often contain indications of edition numbers, parts or dates.

Table 2. Distribution of the main POS in the bibliographic corpus (titles) and in the general language

POS	Bibliographic corpus	General language
Noun	57.15%	43.45%
Verb	3.00%	15.24%
Adj.	11.21%	10.75%
Adv.	1.12%	3.91%
Pron.	0.43%	1.97%
Prep	9.14%	10.90%
Conj.	4.97%	3.96%
Num.	5.44%	0,62%

5. Results: Statistical distributions

The estimation of statistical distributions of spectral lists of vocabulary in the bibliographic and reference corpus also gave interesting results. To compare bibliographical data with data representing the general language we will use the grouped frequency distribution or the frequency spectrum $V(m,N)$ with $m \geq 1$, which is defined as the number of types with frequency m in a sample of N tokens. Formally:

$$V(m, N) = \sum_{i=1}^{V(N)} I\{f(i, N) = m\},$$

where $V(N) = \Sigma_m V(m, N)$ and the indicator function $I\{x\}$ is equal to 1, if expression x is true, and zero otherwise. There is a natural connection between rank-frequency distribution and frequency spectrum:

$$V(m, N) = \sum_i I\{f(i, N) \geq m\}, - \sum_i I\{f(i, N \geq m + 1)\}$$

We considered as potentially relevant statistical distributions typically used in research on lexical data (Baayen 2001). After some preliminary tests it appeared that the best fit of word frequencies was obtained for the generalized Zipf distribution in the case of the general language and the Sichel model for the corpus of titles. The rank frequency distribution described by the Zipf-Mandelbrot law is of the form:

$$f(i, N) = \frac{K}{(i + b)^a},$$

where $a > 1$ and $b \geq 1$ are parameters and K is the normalizing constant (Mandelbrot 1962).

The Sichel model uses the generalized inverse Gauss-Poisson distribution as a description of word probabilities (Sichel 1975). The probability density function for the generalized inverse Gauss-Poisson distribution has the following form:

$$g(x) = Mx^{a-1} \exp(-\frac{x}{c} - \frac{b^2 c}{4x}).$$

The normalizing constant M is of the form $M = \frac{(2/bc)^{a+1}}{K_{a+1}(b)}$, where K_a denotes the modified Bessel function of the second kind of order a. The maximum likelihood estimators of the generalized inverse Gauss-Poisson distribution parameters are described in Sichel (1982).

After some preliminary tests it appeared that the best fit of word frequencies was obtained for the generalized Zipf distribution in the case of the general language and the Sichel model for the corpus of titles. The results of the estimation are presented in Table 3, where ZM and GIGP denote the Zipf-Mandelbrot distribution

and the generalized inverse Gauss-Poisson distribution respectively. They prove that bibliographical data (titles) differ statistically from the general language (if one assumes that it is represented by the National Corpus).[7]

Table 3. Statistical distributions of word spectra in the general language and in the bibliographical corpus

	Distribution	**Par. *a***	**Par. *b***	**Par. *c***
Bibliographical corpus (titles)	GIGP	–0.5277	0.00113	0.0014
Reference corpus (general language)	ZM	0.5877	0.00288	–

As a comparison, Table 3a presents the values of the parameters of the fitted Zipf-Mandelbrot distribution for the bibliographical corpus data and the generalized inverse Gauss-Poisson distribution for the general language data.

Table 3a. Statistical distributions of word spectra in the general language and in bibliographical corpus

	Distribution	**Par. a**	**Par. *b***	**Par. *c***
Bibliographical corpus (titles)	ZM	0.5330	0.00133	–
Reference corpus (general language)	GIGP	–0.5995	0.00006	0.0047

This may come as a surprise as the visual inspection of the fits presented in Figures 5 and 6 suggests that the observed and the expected data match almost perfectly. However, human perception is usually misleading in the case of big amounts of data (cf. Grotjahn & Altmann 1993). When the chi-squared tests are applied and threshold values are respected, the opposite is proven: both of the fits fail and the advanced hypotheses of similarity should be rejected. This means that both corpora have different lexicostatistical characteristics. Indirectly one should conclude that publication titles, when assembled in a relatively large corpus, do not have the properties of 'normal' language as it is used for communication by humans.

On the other hand, it is known that the Pearson chi-square goodness-of-fit test rejects all null hypotheses if the sample size is sufficiently large and for this reason it creates a problem with correct interpretation of the test prediction. In Mačutek & Wimmer (2013), it was suggested that it is possible to solve this problem by taking into account not only significance level, but also the so-called test resistance. One

7. Calculations have been performed with the *ZipfR* package (Evert & Baroni 2007; http://zipfr.r-forge.r-project.org/).

of the simplest possible tests of this type is based on C, the discrepancy coefficient defined as:

$$C = \frac{\chi^2}{N},$$

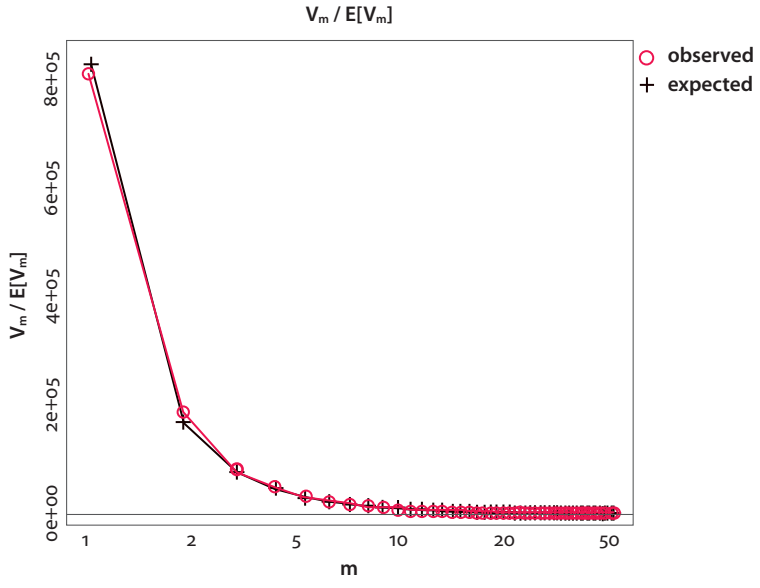
where χ^2 is the value of the Pearson test statistics for the data considered and N is the size of the sample.

When the differences between theoretical and empirical relative frequencies are fixed, the chi-square statistic increases linearly with the sample size. For this reason, Cressie & Read (1984) propose the use of discrepancy to evaluate the quality of fit. For the linguistic data, as a rule of thumb, a fit is considered to be satisfactory if $C < 0.02$. Several more advanced possible solutions of this problem have been mentioned and discussed in Mačutek & Wimmer (2013). In Table 4 we present the values of the chi-square statistics and the corresponding values of the discrepancy coefficient C, for the best fit distributions presented in Table 3 and for the alternative fit from Table 3a.

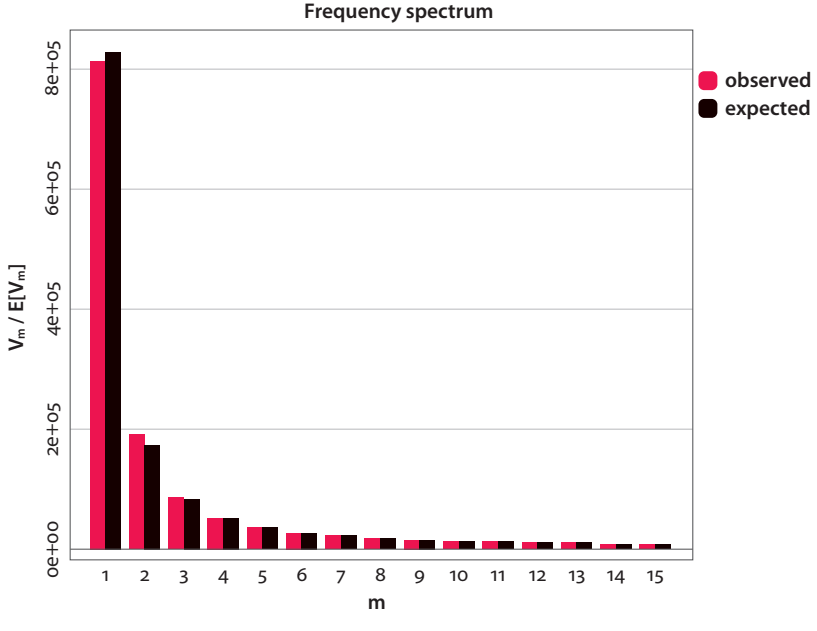
Table 4. The result of the Pearson goodness-of-fit test and corresponding discrepancy coefficient for the distributions of word spectra in the general language and in the bibliographical corpus

	Distribution	χ^2	C
Bibliographical corpus	GIGP	108.733	0.000029
Reference corpus	GIGP	5007.673	0.000021
Bibliographical corpus	ZM	1402.910	0.000369
Reference corpus	ZM	3514.091	0.000015

The values of coefficient C are consistent with results presented in Figures 5–6. The fit for the bibliographical corpus is better than for the general language data, and the value of C suggests that the distribution from the classes usually considered in the literature gives a reasonable fit. Of course, it seems to be interesting to examine the distributions from the broader classes of possible distributions and check the obtained fit using characteristics other than the discrepancy coefficient. However, this would require using the raw data instead of frequency spectrum data on which the presented analysis is based.

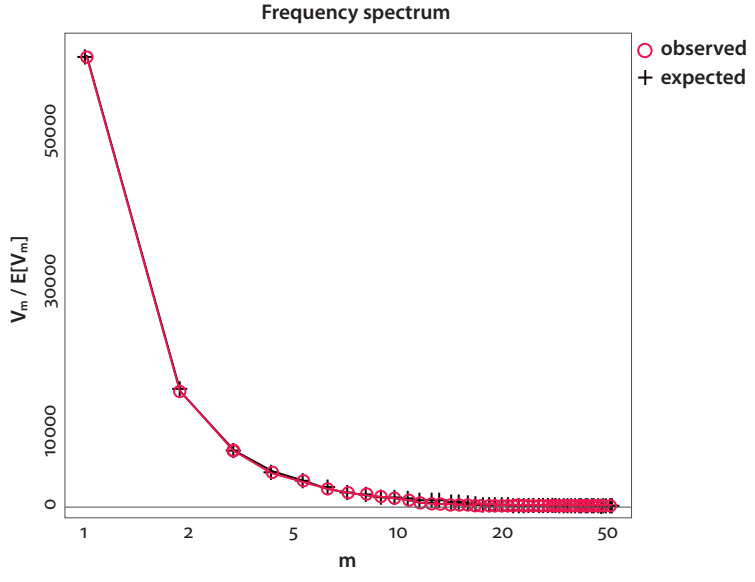


a.

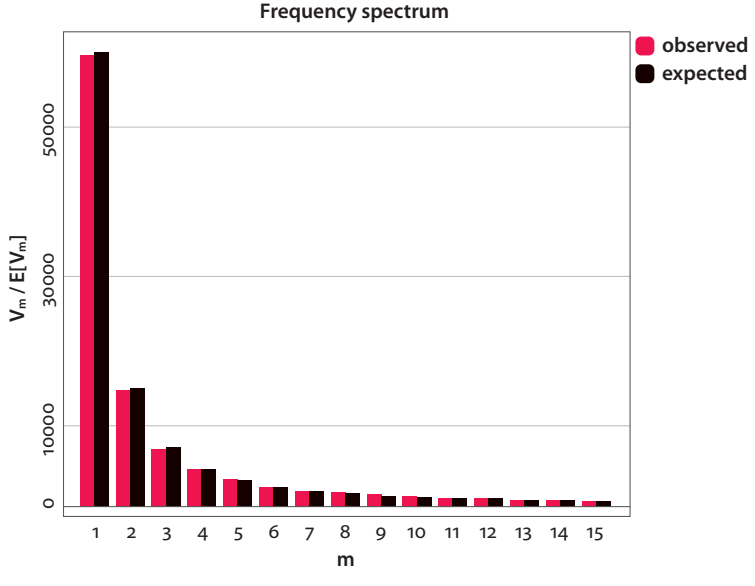


b.

Figure 5. Left panel: The frequency spectrum for the general language data (circles), the Zipf-Mandelbrot fit (solid line and crosses). Right panel: Bar plot for first 15 spectrum elements



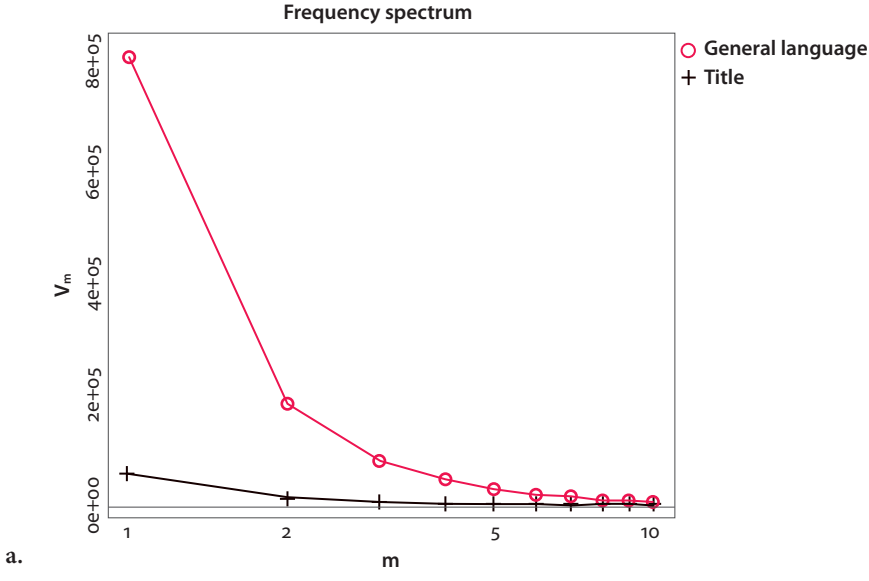
a.



b.

Figure 6. Left panel: The frequency spectrum for the bibliographical corpus (circles), the generalized-Gauss-Poisson fit (solid line and crosses). Right panel: Bar plot for first 15 spectrum elements

The above conclusion, which actually confirms the hypothesis put forward at the beginning, is supported by the histogram of the frequency of words in both corpora (Figure 7), which shows significant differences.



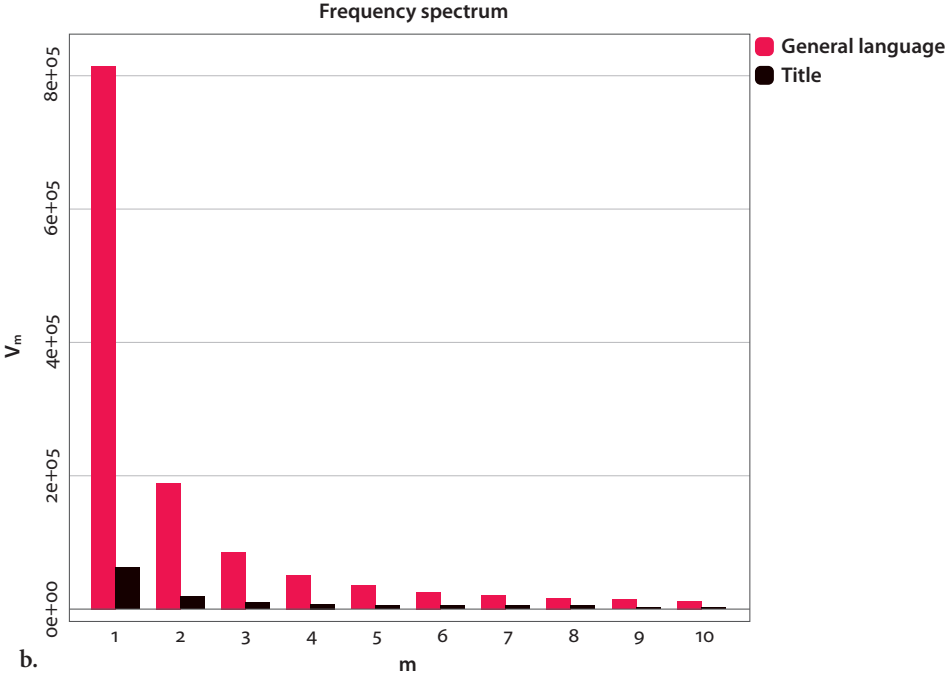


Figure 7. Left panel: The frequency spectrum for the general language data (circles) and the title data (crosses). Right panel: Bar plot for first 15 spectrum elements

6. Conclusions

The aim of this study was to present the quantitative characteristics of the bibliographic corpus and to compare it with the reference corpus of the general language. The tests carried out show that the 'language of titles', i.e., a large corpus composed of book titles extracted from the Polish National Library databases, differs in many respects from the language used in normal communication. It was shown that basic quantitative parameters of both corpora are different (Table 1), and that there are striking differences in the distribution of frequency of parts of speech (Table 2). The same tendency was observed when histograms of title and sentence lengths frequencies were compared (Figures 1 and 2): both corpora turned out to have different quantitative characteristics. We also analyzed the relationship between the average word length in titles and the length of a title (Figures 3 and 4). Both corpora were different and no law-like tendency (e.g., resembling Menzerath-Altmann's law) was observed.

Interesting results were obtained while estimating word spectra distributions of both corpora. It turned out that the generalized Zipf-Mandelbrot distribution and the Generalized Inverse Gauss-Poisson distribution give the best fitting levels. The evaluation of the best fit of both distributions proved to be a problem, as the most common chi-square goodness-of-fit test is designed so that with large corpora the result is always negative. Therefore, in order to evaluate the goodness of fit, resistance and discrepancy tests were used (Table 4). As shown in Figure 7, statistical distributions of word spectra in the bibliographical corpus (titles) and in the general language do not follow the same pattern.

The final conclusion that emerges from the analysis of the results confirms the general principle of statistical linguistics which states that texts created in natural, spontaneous communication differ significantly from texts prepared by researchers as artificial collections, according to arbitrary criteria, such as, for example, the identical context of use, function, or genre of the text. One of the differences between natural texts and 'artificial' text corpora is that the latter does not necessarily follow the statistical laws of language, determined in the context of natural communication. These laws are not just mathematical formulas that the researcher 'matches to a line', that is, to an empirical histogram of some relationship of observed variables. Rather, they reflect, using the formalized language of mathematics, the possibilities and limitations of the human brain, which in a peculiar but effective way optimizes processing of perceived stimuli, aiming at the best adaptation of the human being to its information environment.[8]

8. One of the basic rules of human information processing is the principle of least effort. To a large extent, it shapes the form of language (length of units, speech sound sequences, etc.), but the question of whether it works similarly in other areas of language, for example, in semantics or axiology of the world image, remains open.

A general remark that can be indirectly inferred from the analyses carried out indicates that bibliographies are these days becoming the subject of linguistic and cultural studies, where they are treated as fully fledged text corpora containing valuable, reliable, and easily quantifiable data on culture, society, science, technology, etc. Admittedly, bibliographic corpora are also an emerging object of analysis in quantitative linguistics. Although it is difficult to say whether these new types of data will become a favourite topic of lexicostatistical research, they will certainly be the preferred target of text-mining algorithms – in this case, quantitative research will play an important, albeit auxiliary, role.

Funding

This research was supported by the Polish National Science Center (NCN) under grant 2016/23/B/HS2/01323 "Methods and tools of corpus linguistics in the research of a bibliography of Polish book publications from 1997 to 2017".

The NLP tools used in this study were developed by the CLARIN-PL consortium.

Sources

BN Data: http://data.bn.org.pl/
CLARIN-PL infrastructure: http://clarin-pl.eu
NKJP: http://www.nkjp.uni.lodz.pl/
WCRFT2 morphosyntactic tagger: http://ws.clarin-pl.eu/tager.shtml
ZipfR package: http://zipfr.r-forge.r-project.org/

References

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-010-0844-0

CHBB. 1999–2019. *The Cambridge history of the book in Britain*. Volumes 1–7. Cambridge: Cambridge University Press.

Cressie, Noel & Timothy R. C. Read. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 46(3). 440–464. https://www.jstor.org/stable/2345686. https://doi.org/10.1111/j.2517-6161.1984.tb01318.x

Evert, Stefan & Marco Baroni. 2007. ZipfR: Word frequency distributions in R. In Sophia Ananiadou (ed.), *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, 29–32. Prague: Association for Computational Linguistics. http://www.stefan-evert.de/PUB/EvertBaroni2007.pdf (9 September, 2020.) https://doi.org/10.3115/1557769.1557780

Febvre, Lucien & Henri-Jean Martin. 1958. *L'apparition du livre*. Paris: Albin Michel.

Green, Jonathan, Frank McIntyre & Paul Needham. 2011. The shape of incunable survival and statistical estimation of lost editions. *The Papers of the Bibliographical Society of America* 105(2). 141–175. https://doi.org/10.1086/680773

Grotjahn, Rüdiger & Gabriel Altmann. 1993. Modelling the distribution of word length: Some methodological problems. In Reinhard Köhler & Burghard B. Rieger (eds.), *Contributions to quantitative linguistics*, 141–153. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-011-1769-2_9

Lahti, Leo, Jani Marjanen, Hege Roivainen & Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloguing & Classification Quarterly* 57(1). 5–23. https://doi.org/10.1080/01639374.2018.1543747

Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20(3). 227–240. https://doi.org/10.1080/09296174.2013.799912

Mandelbrot, Benoît. 1962. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jacobson (ed.), *Structure of Language and its Mathematical Aspects* (Proceedings of Symposia in Applied Mathematics 12), 190–219. Providence, RI: AMS.

Schwetschke, Gustav. 1850. *Codex nundinarius Germaniae literatae bisecularis. Teil: 1564 – 1765*. Halle: G. Schwetschke's Verlags-Handlung und Buchdruckerei. https://reader.digitale-sammlungen.de//resolve/display/bsb11199701.html (9 September, 2020.)

Schwetschke, Gustav. 1877. *Codex nundinarius Germaniae literatae bisecularis. Teil: Forts. 1766 bis 1846*. Halle: G. Schwetschke's Verlags-Handlung und Buchdruckerei. https://digital.slub-dresden.de/werkansicht/dlf/102071/1/0/ (9 September, 2020.)

Sichel, Herbert S. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association* 70. 542–547. https://doi.org/10.2307/2285930

Sichel, Herbert S. 1982. Asymptotic efficiency of the three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika* 69. 467–472. https://doi.org/10.2307/2335423

Tolonen, Mikko, Jani Marjanen, Hege Roivainen & Leo Lahti. 2019a. Quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52(1). 57–78. https://doi.org/10.1080/01615440.2018.1526657

Tolonen, Mikko, Jani Marjanen, Hege Roivainen & Leo Lahti. 2019b. Scaling up bibliographic data science. In Costanza Navarretta, Manex Agirrezabal & Bente Maegaard (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, 450–456. Copenhagen: University of Copenhagen. https://cst.dk/DHN2019Pro/papers/40_2019DHNBDS.pdf (9 September, 2020.)

Analysis of English text genre classification based on dependency types

Yaqin Wang
Guangdong University of Foreign Studies

The present study aims to explore whether dependency type can be used as a distinctive text vector for classifying English genres. Three classification methods, namely principal component analysis, hierarchical clustering, and random forest were employed to investigate the clustering effect. Results show that dependency type is an effective measure in distinguishing text genres, especially between spoken genre and written genre.

Keywords: dependency type, genre classification, principal component analysis, hierarchical clustering, random forest, spoken English, written English

1. Introduction

Previous studies in the field of text categorization have focused on the comparison of different algorithms which aims for better classification results (Baharudin et al. 2010; Zaghloul et al. 2013; Zhang et al. 2015). When it comes to specific methods, on the one hand, the most common approaches employed are lexical features, such as common word frequency (Kessler et al. 1997; Stamatatos et al. 2000a, 2000b; Feldman et al. 2009). On the other hand, syntactic information is also important in studying text genre classification (Baayen et al. 1996; Stamatatos et al. 2000a). Typical paradigms are passives, nominalizations, and counts of the frequency of various syntactic categories (e.g., part-of-speech tags) (Biber 1993, 1995; Hou et al. 2014; Hou & Jiang 2014). However, to the best of our knowledge, explicit academic attention seems to be insufficient to the internal syntactic relationship in the field of text classification.

Dependency grammar describes syntactic relationships among words. Dependency distance, the linear distance between the governor and the dependent, can be used as a useful measurement for investigating hidden language universals (Liu 2008; Futrell et al. 2015; Liu et al. 2017). Dependency direction, the word

https://doi.org/10.1075/cilt.356.17wan

order of dependents and governors, has been a key measure of interest in the field of cross-language linguistic work (Hiranuma 1999; Eppler 2005; Liu 2010; Jiang & Liu 2015). Regarding genre classification, Liu et al. (2009b) noted that there exist some differences of dependency direction between conversation and written genres of Chinese. From a statistical point of view, Wang & Liu's (2017) study suggested that genre affects dependency distance and dependency direction significantly; however, the effect is small.

Apart from these two metrics, another important feature of dependency relation is dependency type. The definition of a dependency relation is proposed as: a binary and asymmetrical relation between two linguistic units (Tesnière 1959; Hudson 1990; Liu et al. 2009a). Dependency type describes the grammatical relations of words in a sentence; thus, it can well capture the syntactic relationship between words. Can this feature be a useful metric for text classification and genre judgement?

Several studies using syntactic annotations under the framework of dependency grammar have been conducted in two typical classification tasks, i.e., text genre classification and authorship verification (Hollingsworth 2012; Gao & Feng 2011; Rygl 2014). Hollingsworth (2012) used grammatical dependency relations for identifying authorship and found that syntactic features outperform lexical features. Gao & Feng (2011) employed ten dependency relations with distinctive distribution between oral and written Chinese. These studies suggest that applying dependency relations to text classification is feasible and effective. Nevertheless, little consideration has been given to the role that the dependency type plays in classifying English genres and texts.

The present study, therefore, explores whether dependency type can be used as a distinctive text vector for classifying English genres. It may shed new light on the quantitative analysis of genre analysis and the features of dependency types as well. More importantly, it may broaden the applied range of text clustering and text classification.

Three main research objectives are as follows:

- What is the distribution of dependency types in different genres of English?
- As a kind of text vector, can dependency type be a useful feature in text clustering?
- If so, do all dependency types play the same role in classifying texts? (If not, why?)

For the second question, the present study uses two clustering methods, principal component analysis and hierarchical clustering, to investigate the performance of dependency type in text clustering and classification. The random forest method was then selected to answer the third question, that is, to determine the different degrees of importance of dependency relations.

The study is organized as follows: Sections 2 and 3 display the treebank establishment and three methods conducted. Section 4 interprets the results, followed by the conclusion and suggestions for further study in Section 5.

2. Treebank establishment

The current research used the BNC (Burnard 2000) as the data source. The BNC (British National Corpus) is a corpus of about 100 million words of contemporary spoken and written British English. The written texts of the BNC are composed of two sorts of text, 'imaginative' and 'informative' texts. The 'informative' texts include eight domains, namely 'applied science', 'arts', 'belief', 'commerce', 'leisure', 'natural science', 'social science', and 'world affairs'. Due to limited time and space, the present study did not choose all kinds of informative text. Instead, texts were randomly collected from all eight domains as the representative of 'informative' genre. Another genre, 'written to be spoken', which is composed of scripted television materials and play scripts (Burnard 2000), written with the aim for speaking in public, was also included in the study. This genre may share characteristics both with oral discourse and written texts. Spoken texts were also chosen as the data source. Thus, there are four genres, i.e., 'spoken', 'informative', 'imaginative', and 'written to spoken', in the current research. In each genre, 10 texts, each consisting of around 2000 tokens, were randomly selected from 4 genres, 40 texts in total. Prior examination on the text was carried out.

After the data extraction, the Stanford parser (version 3.7.0) (de Marneffe & Manning 2008; Nivre et al. 2016) was employed to output universal dependency among the words in an input sentence (two terms, i.e., 'universal relation' and 'dependency type', are used interchangeably in the current research). Dependencies were printed in the CoNLL 2007 format (Nivre et al. 2007) as the following table shows. An English dependency treebank was then established.

Table 1. CoNLL 2007 format of output dependency

1	What	WP	WP	0	*root*
2	is	VBZ	VBZ	1	*cop*
3	art	NN	NN	1	*nsubj*

As Table 1 shows, the first column represents the position number of each token (including the punctuation); the middle two columns are PoS tags of corresponding tokens; the next column is the position number of each word's governor. The last column is the specific dependency type. 39 dependency types are found in these

four genres.[1] Among these types, two types, i.e., *dep, root*, are worth mentioning. *Dep* arises when the parser cannot accurately determine the relation type. It may be due to a parser error or an ungrammatical construction. This type is kept in the study since it may signify different stylistic features. Some genres may have more ungrammatical sentences than do others. The same reason holds for keeping the type *root* in the analysis. *Root* refers to the head of a sentence, rather than a specific relation between words. It can be used as a useful parameter in measuring the distribution of dependency types of different genres.

Frequencies of all dependency types of 40 texts were then counted from the output and used as text vectors for further analysis.

3. Methods

3.1 Principal component analysis

For the dataset which consists of 39 variables, techniques which reduce dimensions while still retaining much information, such as principal component analysis (PCA) (Hotelling 1933), are needed. As a statistical procedure, PCA transforms the observations of dependency types into a set of values of linearly uncorrelated principal components. These linear principal components successively have maximum variance for the data. Based on this method, one can have a glance at the relationship between texts.

3.2 Hierarchical cluster analysis

Hierarchical clustering analysis (HCA) traces structure in a matrix of distances. Agglomerative clustering, one of the hierarchical clustering techniques, was implemented in the study. It is a bottom-up approach which starts with single points and then agglomerates into groups, and then larger groups, and so on. Euclidean distance was employed to measure the distance between text vectors. Like PCA, this method can also present the relationship between the genre and dependency type. The result of this method is shown in a tree of clusters, a so-called dendrogram. The dendrogram can explicitly show the clusters of texts. Then the cophenetic correlation coefficient (CPCC) was calculated for the validation of the clustering result. CPCC assesses how faithfully a dendrogram preserves the pairwise distances

1. English universal dependencies are listed in the website: http://universaldependencies.org/en/dep/all.html. However, not all relations in the website are found in the parsing result. Please check Appendix A for specific information of dependency types occurring in the present study.

between the original unmodeled data points. Therefore, it can indicate how well the clustering result represents the original data.

3.3 Random Forest

It is necessary to determine whether each dependency type is important for text classification. Random forest (Breiman 2001) is an ensemble learning method for classification. It generates many classification trees and each tree "votes" for the most popular class. The forest then chooses the classification having the most votes. It can be used to rank the importance of variables in a classification problem.

Hou & Jiang (2014) employed this method to study the degrees of importance of word PoS tags. Likewise, the present study selected this method to determine the number of important dependency types in genre classification.

The above three text mining methods and relevant figures were all implemented using the R project.

4. Results and discussion

Before we proceed to the discussion of the clustering result of the texts, let us first have a glance at the general distribution of universal relations of four genres.

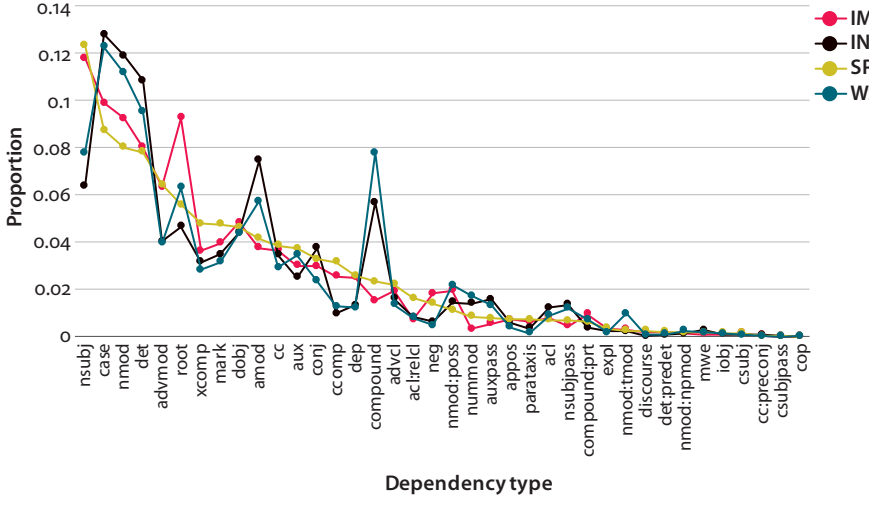


Figure 1. The distribution of universal relations in four genres

(In this and the following figures and tables, 'informative genre' is referred to as IN, 'imaginative' is IM, 'written to be spoken' is WS, 'spoken' is SP.)

Just from eyeballing the graph, one can notice that differences across genres do exist. 'Imaginative' genre and 'spoken' genre have more proportions of *ccomp, nsubj, root*, etc., than do 'informative' and 'written to be spoken' genre. In contrast, 'informative' and 'written to be spoken' genres have more frequencies of *amod* and *nmod* than do spoken texts.[2] This distribution is consistent with stylistic transformation from spoken texts to written texts. For instance, as the author mentioned before, *dep* is retained for investigating stylistic features. Spoken texts have higher proportions of *dep* than do the other three genres, since spoken texts' peculiar features, i.e., ungrammatical structures, repetition, fillers, etc., provide challenges for the parser. The definition of *ccomp* is as follows: the clausal complement of a verb or adjective is a dependent clause which is a core argument. The frequency of *ccomp* accounts for 3.2% in 'spoken' texts' dependency type distribution, while it occupies 0.9% in 'informative' texts. This indicates that spoken genre tends to have more clausal complements (finite or non-finite) than do written texts. Detailed analysis of the distribution of dependency types merits further study.

Since there exist variations in dependency types across genres, the question remains of whether text vectors characterized by types can classify genres. Are all these types useful in genre categorization? The following section tries to answer these questions.

4.1 PCA

Figure 2 shows the clustering result based on the first two principal components. It is generally consistent with the pre-specified structure of the data; spoken English (the narrow circle lying on the left part of the graph) is isolated from the 'informative' and 'written to be spoken' genre. Although belonging to the written texts, the 'imaginative' genre (the second circle on the left) is closer to the 'spoken' genre compared with the other two written genres.

According to Baayen (2008), there is a rule of thumb that principal components which account for 5% variance are important. The important principal components (which at least have 5% variance) are listed in Table 2.

Table 2. The importance of principal components

	PC1	PC2	PC3	PC4	PC5
Standard deviation	3.75	2.20	1.87	1.43	1.29
Proportion of Variance	0.37	0.13	0.09	0.05	0.04
Cumulative Proportion	0.37	0.50	0.59	0.64	0.69

2. The proportions of dependency types across genres are displayed in Appendix B.

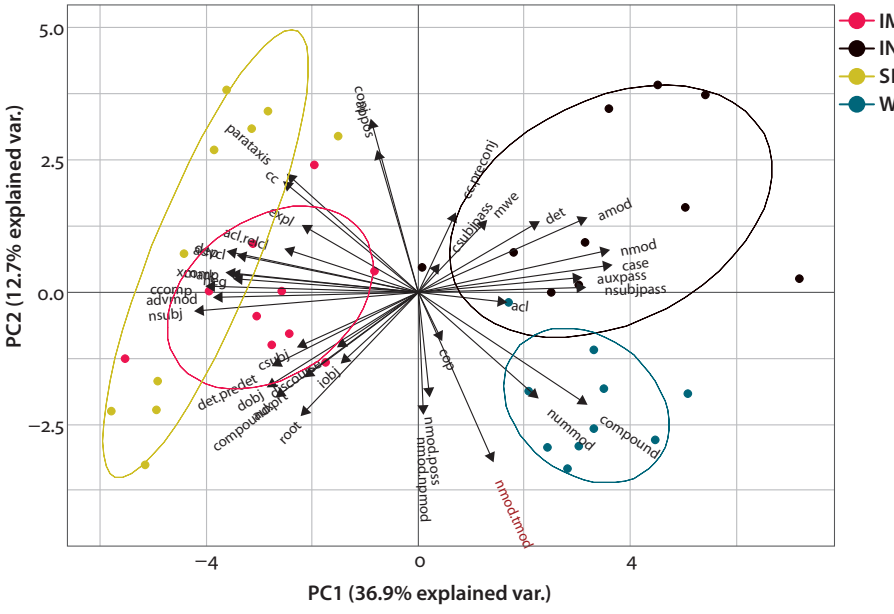


Figure 2. The result of PCA in terms of the first two principal components

The first four principal components account for about 64% of the whole variance. The cumulative proportion of the first two components achieves 50%, which means that Figure 2 can interpret the relationship of texts well. The above results show that dependency types can be used as a useful measurement to classify different genres.

4.2 Text clustering

Another text clustering method, the agglomerative hierarchical analysis, was then carried out and the results are displayed as follows.

The dendrogram in Figure 3 and Table 3 show the relationship of clusters. It indicates that texts are split into two major clusters. The left split is composed of the 'informative', 'written to be spoken' texts. The right cluster includes the 'spoken' texts and 'imaginative' texts. Compared with the relationship between the 'informative' and 'written to be spoken' texts, the 'spoken' and 'imaginative' texts

Table 3. The result of agglomerative hierarchical analysis

	IM	IN	SP	WS
1	10	0	10	0
2	0	10	0	10

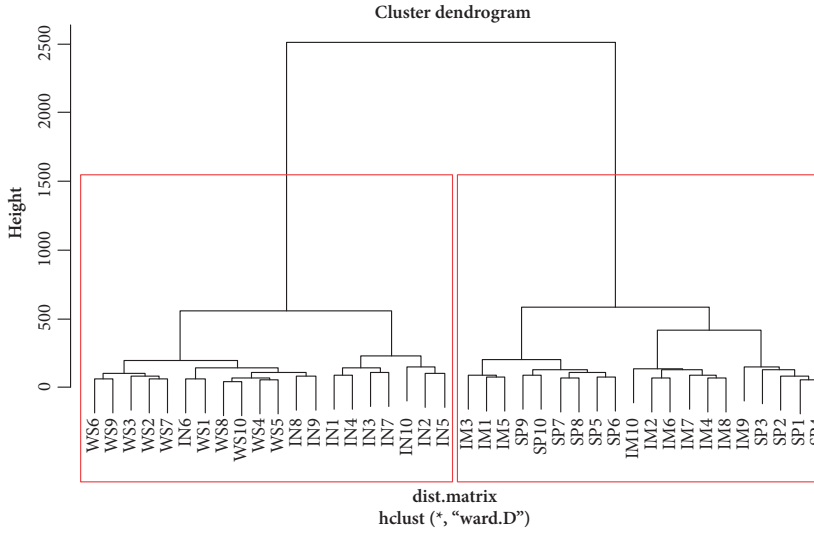


Figure 3. Agglomerative hierarchical analysis of 40 texts

are more mixed in the right split. The correlation between the cophenetic matrix of the dendrogram and the distance matrix of the texts is 0.81. This shows that the clustering result is acceptable.

This result is consistent with what the current chapter found in the last section in terms of the PCA result. It also mirrors the finding regarding the distribution of types. For instance, the 'imaginative' genre and 'spoken' genre have more proportions of *ccomp, nsubj, root*, etc., than do the 'informative' and 'written to be spoken' genre. This shows that although the 'imaginative' genre belongs to written texts, they share more characteristics with spoken texts. This may be accounted for by the large number of conversations in 'imaginative' texts. Another phenomenon worth noting is that the 'written to be spoken' genre, though generally for speaking purposes, is much closer to the informative genre, which belongs to a written genre.

4.3 Random forest

So far, the author has discussed the clustering result using all dependency types as variables. It remains a question whether each dependency type is equally useful in determining genres. If some of them are sufficient for distinguishing genres, there is no need to keep all types. Thus, important universal relations need to be screened out. Random forest was applied to examine the importance of all dependency types.

As Figure 4 shows, dependency types show different levels of importance in the genre classification. This figure shows the values of the mean decrease in Gini

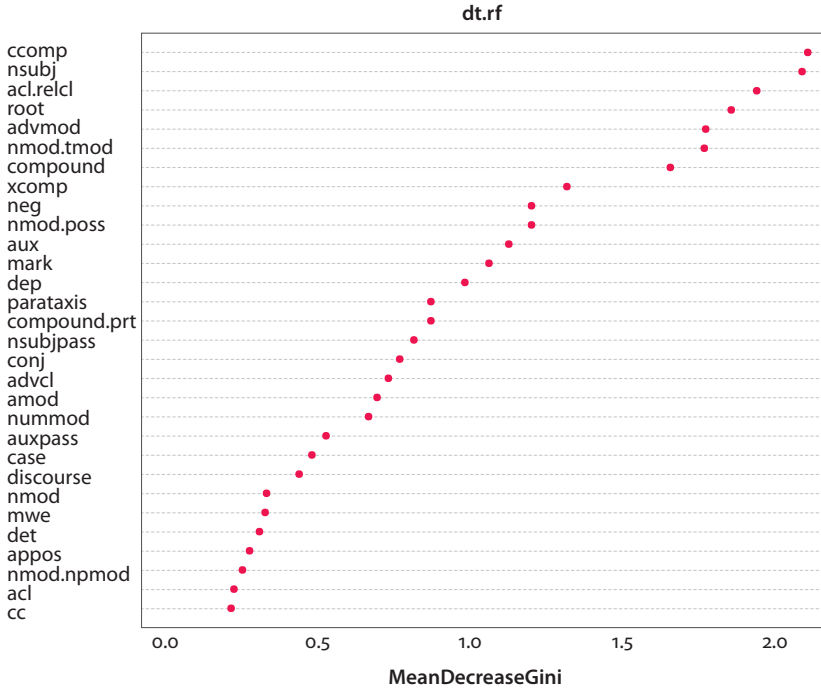


Figure 4. Importance of dependency types

across dependency types which calculate the impurity-level influence of variables to compare their importance (Hou & Jiang 2014). The higher the value of the index, the more important the variable is. *Ccomp, nsubj*, and *acl.relcl* are the top three important types, while *nmod:npmod, acl*, and *cc* are less important ones.

The first ten important dependency types were chosen, namely, *ccomp, nsubj, acl.relcl, root, advmod, nmod:tmod, compound, xcomp, neg, nmod:poss*, and the agglomerative hierarchical clustering was conducted again to see whether fewer text vectors would improve the performance of the clustering result. The correlation between the cophenetic matrix of the dendrogram and the distance matrix of the texts is 0.84. It increases slightly compared with the former result (0.81).

These important dependency types are preferable in text genre classification. Nevertheless, the present study only chose four genres as the data source. Larger numbers of texts and genres need to be explored to validate the clustering feasibility of dependency types in future research.

5. Conclusions

In the current chapter, the distribution of dependency types in four genres was first presented. The results indicated that differences across genres exist due to different genre features.

Several classification methods were employed. The frequencies of dependency types across genres were used as text vectors. PCA and agglomerative hierarchical clustering were first performed and results showed that dependency types can be used as an effective parameter in distinguishing text genres, especially between spoken genre and written genre.

Random forest, which determines the importance of numerous variables, was then performed. The results showed that not all dependency types are important in classifying genres. Ten important types were also tested again to see whether they improve the performance of the clustering result.

The present study verified the importance of dependency types in English genre classification. It may be useful for future applications in text genre categorization and dependency studies. However, this study is but a first foray into the issue and still await further validation using larger numbers of texts from more genres, from more languages. Other lexical or syntactic features can also be proposed together with dependency types to improve the performance of clustering result.

Acknowledgements

The paper has been supported by the MOE Project at Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

Baayen, Harald, Hans Van Halteren & Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3). 121–132. https://doi.org/10.1093/llc/11.3.121

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baharudin, Baharum, Lam H. Lee, Khairullah Khan & Aurangzeb Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology* 1(1). 4–20. https://doi.org/10.4304/jait.1.1.4-20

Biber, Douglas. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics* 19(2). 219–241.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Breiman, Lee. 2001. Random forests. *Machine Learning* 45(1). 5–32. https://doi.org/10.1023/A:1010933404324

Burnard, Lou. 2000. *Reference guide for the British National Corpus (World Edition)*. Oxford: Oxford University Computing Services.

de Marneffe, Marie-Catherine & Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University. https://worksheets.codalab.org/rest/bundles/0x953afe5537074b4b9cd3c57e08e2d865/contents/blob/StanfordDependencies-Manual.pdf

Eppler, Eva M. 2005. The syntax of German-English code-switching. London: University of London dissertation.

Feldman, Sergey, M. A. Marin, Mari Ostendorf & Maya R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *2009 IEEE International Conference Acoustics, Speech and Signal Processing, 4781–4784*). Taipei: IEEE. https://doi.org/10.1109/ICASSP.2009.4960700

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33). 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gao, Song & Zhiwei Feng. 2011. Research on text clustering based on dependency treebank. *Journal of Chinese Information Processing* 25(3). 59–63.

Hiranuma, So. 1999. Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics* 11. 309–322.

Hollingsworth, Charles. 2012. Using dependency-based annotations for authorship identification. *Text, Speech and Dialogue* 7499. 314–319. https://doi.org/10.1007/978-3-642-32790-2_38

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* 24(6). 417–441. https://doi.org/10.1037/h0071325

Hou, Renkui & Minghu Jiang. 2014. Analysis on Chinese quantitative stylistic features based on text mining. *Digital Scholarship in the Humanities* 31 (2). 357–367. https://doi.org/10.1093/llc/fqu067

Hou, Renkui, Jiang Yang & Minghu Jiang. 2014. A study on Chinese quantitative stylistic features and relation among different styles based on text clustering. *Journal of Quantitative Linguistics* 21(3). 246–280. https://doi.org/10.1080/09296174.2014.911508

Hudson, Richard A. 1990. *English word grammar*. Oxford: Basil Blackwell.

Jiang, Jingyang & Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency treebank. *Language Sciences* 50. 93–104. https://doi.org/10.1016/j.langsci.2015.04.002

Kessler, Brett, Geoffrey Nunberg & Hinrich Schütze. 1997. Automatic detection of text genre. In Philip R. Cohen & Wolfgang Wahlster (ed.), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Stroudsburg, PA: Association for Computational Linguistics.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2). 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 120(6). 1567–1578. https://doi.org/10.1016/j.lingua.2009.10.001

Liu, Haitao, Richard Hudson & Zhiwei Feng. 2009a. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory* 5(2). 161–174. https://doi.org/10.1515/CLLT.2009.007

Liu, Haitao, Yiyi Zhao & Wenwen Li. 2009b. Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics* 45(4). 509–523. https://doi.org/10.2478/v10010-009-0025-3

Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews* 21. 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

Nivre, Joakim, Hall Johan, Kübler Sandra, McDonald Ryan, Nilsson Jens, Riedel Sebastian & Yuret Deniz. 2007. *The CoNLL 2007 shared task on dependency parsing*. In Jason Eisner (ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (915–932)*. Prague: Association for Computational Linguistics.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. *Universal Dependencies v1: A multilingual treebank collection*. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666. Portorož: European Language Resources Association (ELRA).

Rygl, Jan. 2014. Automatic adaptation of author's stylometric features to document types. In *International Conference on Text, Speech, and Dialogue*, 53–61. Cham: Springer. https://doi.org/10.1007/978-3-319-10816-2_7

Stamatatos, Efstathios, Nikos Fakotakis & George Kokkinakis. 2000a. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26(4). 471–495. https://doi.org/10.1162/089120100750105920

Stamatatos, Efstathios, Nikos Fakotakis & George Kokkinakis. 2000b. Text genre detection using common word frequencies. In Martin Kay (ed.) *Proceedings of the 18th conference on Computational Linguistics*, Volume 2, 808–814. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/992730.992763

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.

Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences* 59. 135–147. https://doi.org/10.1016/j.langsci.2016.09.006

Zaghloul, Waleed, Sang M. Lee & Silvana Trimi. 2013. Text classification: Neural networks vs support vector machines. *Industrial Management and Data Systems* 109(5). 708–717. https://doi.org/10.1108/02635570910957669

Zhang, Wen, Xijin Tang & Toshida Yoshida. 2015. TESC: An approach to TExt classification using Semi-supervised Clustering. *Knowledge Based Systems*, 152–160. https://doi.org/10.1016/j.knosys.2014.11.028

Appendix A. English universal relations

acl	clausal modifier of noun
acl:relcl	relative clause modifier
advcl	adverbial clause modifier
advmod	adverbial modifier
amod	adjectival modifier
appos	appositional modifier
aux	auxiliary
auxpass	passive auxiliary
case	case marking
cc	coordinating conjunction
cc:preconj	preconjunct
ccomp	clausal complement
compound	compound
compound:prt	phrasal verb particle
conj	conjunct
cop	copula
csubj	clausal subject
csubjpass	clausal passive subject
dep	unspecified dependency
det	determiner
det:predet	predeterminer
discourse	discourse element
dobj	direct object
expl	expletive
iobj	indirect object
mark	marker
mwe	multi-word expression
neg	negation modifier
nmod	nominal modifier
nmod:npmod	noun phrase as adverbial modifier
nmod:poss	possessive nominal modifier
nmod:tmod	temporal modifier
nsubj	nominal subject
nsubjpass	passive nominal subject
nummod	numeric modifier
parataxis	parataxis
punct	punctuation
root	root
xcomp	open clausal complement

Appendix B. The proportion of dependency type across genres

Type	IM	IN	SP	WS
acl	0.008	0.012	0.007	0.009
acl:relcl	0.007	0.008	0.016	0.008
advcl	0.019	0.016	0.022	0.014
advmod	0.064	0.040	0.065	0.040
amod	0.038	0.075	0.042	0.058
appos	0.007	0.006	0.007	0.004
aux	0.030	0.025	0.038	0.035
auxpass	0.006	0.016	0.008	0.013
case	0.100	0.129	0.088	0.124
cc	0.037	0.035	0.039	0.029
cc:preconj	0.000	0.001	0.000	0.000
ccomp	0.026	0.010	0.032	0.013
compound	0.015	0.057	0.024	0.078
compound:prt	0.010	0.004	0.006	0.007
conj	0.030	0.038	0.033	0.024
cop	0.000	0.000	0.000	0.000
csubj	0.002	0.001	0.001	0.001
csubjpass	0.000	0.000	0.000	0.000
dep	0.025	0.013	0.026	0.012
det	0.081	0.109	0.079	0.096
det:predet	0.002	0.001	0.002	0.001
discourse	0.002	0.000	0.002	0.001
dobj	0.049	0.044	0.047	0.044
expl	0.003	0.003	0.004	0.002
iobj	0.001	0.001	0.001	0.001
mark	0.040	0.035	0.048	0.032
mwe	0.001	0.003	0.002	0.002
neg	0.018	0.006	0.014	0.005
nmod	0.093	0.120	0.081	0.113
nmod:npmod	0.001	0.001	0.002	0.002
nmod:poss	0.020	0.014	0.011	0.022
nmod:tmod	0.003	0.002	0.003	0.010
nsubj	0.119	0.064	0.124	0.079
nsubjpass	0.005	0.014	0.007	0.012
nummod	0.003	0.014	0.008	0.017
parataxis	0.006	0.003	0.007	0.001
root	0.094	0.047	0.056	0.064
xcomp	0.036	0.032	0.048	0.028

In memory of Gabriel Altmann

Eminent linguist, a man with a brilliant mind, and friend

It is with a great sorrow that we received the news of the death of Gabriel Altmann (24 May 1931 – 2 March 2020), our teacher, mentor, colleague, friend, and often a source of scientific inspiration. In the genuine sense of the word, he was the founding father of quantitative linguistics as we know this research field today. Below we present memories of him, as written by several contributors to this volume.

I did some editing work for Gabriel. Although I did not have a chance to work with him face-to-face, I did communicate with him by email over several years. Whatever the project, he was always encouraging about the work, gracious in his thanks, and very congenial overall. Even at such a distance, I feel he was a close colleague, and I am sure all who knew him will miss his unfailing support and insight.

Eric S. Wheeler

I never had the pleasure of meeting Gabriel Altmann in person, but I had been fortunate to establish and maintain scientific contact with him that started about 15 years ago. At that time, I was still a relative novice in the field of quantitative linguistics. Gabriel (this is how I always addressed him in our e-mail correspondence, never calling him Gabi) helped me a lot in the beginning and continued to be very encouraging and supportive over the years. He was always kind to me, fully understanding that I was a mathematician and not a linguist, although I knew something about languages and linguistics. With his support, I was able to accomplish more, and I am forever grateful to him for that.

Relja Vulanovic

I myself hadn't seen Altmann for many many years, perhaps 20 years, so it's hard to know what to write. I think the main thing I would mention would be that he took a keen interest in my very early work in the field just as soon as he heard of it, and he then made sure I was brought into the right loops, like Qualico, IQLA, etc., as well as connecting us to publishers like Brockmeyer, so we could get our early monographs published, and therefore a good solid start to our academic careers.

https://doi.org/10.1075/cilt.356.18

He was always very encouraging, and so provided a good start for young scholars. Also, he had very broad interests, and so nobody felt that their interests or fields were less worthy than anybody else's. So, in short, broadly encouraging and sympathetic, especially to younger scholars. Sheila Embleton

I got in touch with Gabriel at the end of 2017; I remember that we exchanged a couple of e-mails on 30 December. He actually deleted the first e-mail, as he was struggling with some computer issues. As I see it now, I consider it symptomatic of all his career in science – he was always fighting for his cause, trying to get important allies and persuading them to carry out the type of linguistic research he believed in. I recall his style of writing, full of exclamation marks, likeable interferences, and tons of supportive comments; we spent hours discussing the future of linguistics, shaping the contents of *Glottometrics*, doing and correcting research, formatting books. He felt like a bomb, but an extremely productive one – something like a continuous big-bang of ideas for new papers and monographs. And despite being in his late eighties, he did all this on his own, with no institutional backup, just out of pure desire to show new ways for our discipline.

It seems that three years are not a long enough time to get to know a person thoroughly. With a mailbox with more than 1,100 letters connected to the name of Altmann, I feel I may know something. And this something makes me want to push his legacy further. Michal Místecký

For Gabriel Altmann, a wonderful teacher, an eternal inspiration, a true friend.
Alexander Mehler

On March 3, 2020, I heard the lamentable news from my Ph.D. supervisor, Prof. Haitao Liu, that Prof. Gabriel Altmann, the founding father of quantitative linguistics, died on March 2 at the age of nearly 89. Though I got to know Gabriel personally for a relatively short time since July 2019, I was deeply impressed by the diligent, patient and intelligent soul during the period we collaborated.

Long before I got the chance to know him in person; when I first came into the field, I was mesmerized by the beauty and mystery of quantitative linguistics. Gabriel Altmann, who founded quantitative linguistics, was the name printed on the numerous papers and books in the field, remote yet respectable to us students. Later, the name became rather familiar when I started to use the software, Altmann-Fitter, for fitting models frequently. It reminded me of the famous quote by Isaac Newton, "If I have seen further, it is by standing on the shoulders of giants".

I sent the first email to Gabriel on the occasion of the abrupt death of Prof. Fengxiang Fan, my M.A. supervisor, in August 2018. Prof. Liu and I submitted an article to *Glottometrics* introducing Prof. Fan's life and publications. After a few

comments on the manuscript, he started to ask about the details of his old colleague's death, sadly and concernedly in my view, 'What happened to Fan?'

Then he proposed that, if possible, we could edit a special issue based on a collection of Prof. Fan's graduate students' M.A. theses, saying that 'Europe is eager to see what is being done in China'. As the guest editor of the special issue, I started the email correspondence with Gabriel in July last year, which lasted for approximately one month. What a month! Diligently, he replied to the email at lightning speed. Since the correspondent of my emails was the very famous quantitative linguist, I, as a newbie in the field, was nervous and scrupulous over the first couple of emails. He behaved in such an easygoing and humorous way that often reminded me of my own grandfather. It instantly eliminated my self-consciousness when he said in the second email, 'Please, call me simply Gabriel, even my grandchildren do it!'. Over those letters, Gabriel was always ready to offer help and tirelessly gave his advice to young people like us. In one of his emails, Gabriel said that 'if you are ready to write something, analyze the Chinese or English, tell me, I give you a lot of problems.' It would be somehow egoistic to say that it is a pity that I didn't get in touch with him more or earlier. But indeed, I learned a lot from him regarding not only the academic aspect, but also the art of communication in everyday life. He generously gave compliments to encourage me when I asked for his advice and candidly pointed out mistakes whenever he found something wrong, not to mention his wisely humorous personality.

Alas, Gabriel, the wise man is long gone now. He, however, will always be remembered as a wonderful linguist, an inspiration to all of us to go on along the academic path boldly, meticulously, and persistently. Yaqin Wang

I was introduced to Professor Gabriel Altmann by Professor Jadwiga Sambor, who in the 80s and 90s conducted research in the field of quantitative linguistics at the University of Bochum. Although we did not meet personally, we worked together for several years in connection with my research on the application of time series analysis in linguistics. The result was our correspondence (at first traditional, then only by e-mail) and several papers in the *Journal of Quantitative Linguistics*. Indirectly this cooperation allowed me to write my habilitation monograph and foster my scientific career. Prof. Altmann, fully aware of the impoverishment of science in the former Soviet bloc countries after 1989, showed me great help by proposing to apply for the Humboldt Foundation scholarship, which I obtained and benefited from in 2001–2002, while working as a fellow at the University of Trier. I admired his comprehensive linguistic and mathematical knowledge, the accuracy and discipline of reasoning, and finally, the pursuit of respecting the general laws of epistemology, which turn loose considerations about language into scientific argumentation. Adam Pawłowski

Once upon a time, there was a linguist. The linguist lived quite an extraordinary life (judging from his occasional comments in personal conversations and e-mails – just consider his emigration from communist Czechoslovakia and starting a new life in Germany) and had quite extraordinary abilities. He was able to argue (politely, of course) with his thesis supervisor Vladimír Skalička (a famous professor at the Charles University in Prague – let us remind you that we are talking about his study in the 1950s, when the hierarchy at universities was much more pronounced than today, and students were not supposed to enter a free discussion with a teacher, and certainly not to oppose a professor) and convince him that, in that case, the (extraordinary, of course) ideas of the student are much better than the ones of his supervisor. He studied Indonesian and Japanese, languages which are not really common in Central Europe (but why should we expect Gabriel to choose something more ordinary?), and spoke and understood many others. He was an extraordinary student later in his life as well – he attended lectures on statistics, having already a diploma from linguistics (and one could say without an exaggeration that mathematics lost a very talented scholar when he decided to focus on linguistics). He was an extraordinary teacher (in spite of joking that a university would be a perfect institution, if only there were no students), he was able to teach by e-mails long before it became modern, and, as of today, necessary (we are both, in a way, 'products' of his, perhaps even involuntary, 'distance teaching', as our e-mail exchange with him could be labelled nowadays). He was an extraordinary professor emeritus, he came with many interesting ideas and wrote many books and papers after he had retired, his behaviour being exactly opposite to the retiring professor from one of his funny stories (this one can be found in the volume dedicated to the 65th birthday of Reinhard Köhler). We allow ourselves not to speak of his extraordinary scientific career, as such a description would be extraordinarily long.

On a more serious note, Gabriel Altmann was a pleasant, nice man whose company, mainly virtual, but on several occasions also real, we enjoyed very much. He was unselfish (one can only guess how many papers where he refused to be a co-author were written with his help) and modest (he was never shy of making fun of himself). As we come from the same cultural space as he did, we could alternate between scientific topics and things more personal, like anecdotes on his former teachers and colleagues, places we all knew well, etc. It was a pleasure to speak to him and to read his e-mails (which were always made more interesting – or shall we say more extraordinary? – by his witty remarks).

We will miss you, Gabi. Rest in peace. Ján Mačutek and Radek Čech

Writing even a few words for such a brilliant person as Prof. Gabriel Altmann is a challenging task. Due to a series of unfortunate events, I never had the chance to meet Gabriel in person but I had frequent communication with him via emails. I was always impressed by his modesty and his humility. Gabriel was a stellar scientist with contributions to a huge spectrum of scientific fields (from music to linguistics) and definitely one of the most important pioneers in quantitative linguistics. However, at the same time he was a very pleasant and easy to approach person with positive attitude. I was particularly impressed by his continuous engagement in research even in his older years and his enthusiastic commitment to young scientists. For me Gabriel would be always a role model, a scientist that can inspire younger generations and open new paths to our knowledge. George Mikros

Index

Specialists in quantitative linguistics the world over have recourse to a solid and universal methodology. These days, their methods and mathematical models must also respond to new communication phenomena and the flood of data produced daily. While various disciplines (computer science, media science) have different ways of processing this onslaught of information, the linguistic approach is arguably the most relevant and effective. This book includes recent results from many renowned contemporary practitioners in the field. Our target audiences are academics, researchers, graduate students, and others involved in linguistics, digital humanities, and applied mathematics.

ISBN 978 90 272 1010 4
9 789027 210104

JOHN BENJAMINS PUBLISHING COMPANY