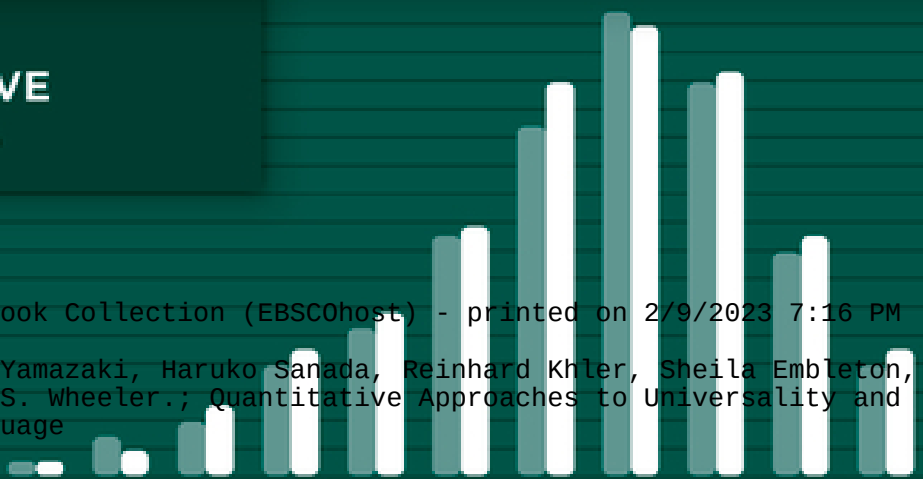
DE GRUYTER
MOUTON

QUANTITATIVE APPROACHES TO UNIVERSALITY AND INDIVIDUALITY IN LANGUAGE

Edited by Makoto Yamazaki, Haruko Sanada, Reinhard Köhler, Sheila Embleton, Relja Vulanović and Eric S. Wheeler

QUANTITATIVE LINGUISTICS



Quantitative Approaches to Universality and Individuality in Language

Quantitative Linguistics

Edited by
Reinhard Köhler and George Mikros

Volume 75

Quantitative Approaches to Universality and Individuality in Language

Edited by
Makoto Yamazaki, Haruko Sanada, Reinhard Köhler,
Sheila Embleton, Relja Vulanović and Eric S. Wheeler

DE GRUYTER
MOUTON

ISBN 978-3-11-062808-1
e-ISBN (PDF) 978-3-11-076356-0
e-ISBN (EPUB) 978-3-11-076363-8
ISSN 0179-3616

Library of Congress Control Number: 2022940833

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Editors' Foreword

In terms of the number of languages dealt with, the study of languages can be divided into two categories: those that analyze a single language and identify the characteristics of that language, and those that deal with several languages and identify the phenomena and tendencies that they share.

Neither of the two methodologies is irrelevant. An analysis in one language can be applied to other languages and lead to the discovery of general tendencies, and sometimes, a general trend is applied to individual languages to observe its applicability. Occasionally, exceptional phenomena may be found, from which a refinement of the law may be derived. The famous Zipf's law, initially proposed based on observations in a few languages, has since been confirmed as a general law in many languages. It is fair to say that the analysis and interpretation in individual languages are not just a question of the language itself; they always need to be examined to see whether they are also valid on a universal level.

This book includes 16 peer-reviewed articles that use quantitative methods to conduct these methodologies in language research. Recently, most data-based linguistic investigations have adopted quantitative methods. It is important to emphasize that the papers in this book do not merely take a quantitative approach; they also offer qualitative considerations, such as interpretations based on linguistic theory, which are well-balanced.

In total, 34 authors from 13 countries contributed papers to the book: Austria, Canada, China, Czech Republic, France, Greece, Italy, Japan, Poland, Qatar, Russia, Spain, and Slovakia.

The languages covered in the book are Catalan, Czech, Chinese, Italian, Mambila (dialects of Nigeria and Cameroon), Japanese, Polish, Romanian, Russian, Spanish, Slovenian, and Ukrainian. Additionally, there are papers on the observation of linguistic laws for many languages; thus, the number of languages appearing in the book is much larger.

Moreover, in terms of content, the book covers various topics: phonology, vocabulary, grammar, style, spoken and written languages, and classical and modern languages. The general classification is as follows. Some authors are classified into several topics: those dealing with universal language laws, such as Menzerath-Altmann Law and Zipf's Law (Radek Čech, Barbora Benesova, and Ján Mačutek; Xinying Chen, Kim Gerdes, Sylvain Kahane, and Marine Courtin; Antoni Hernández-Fernández, Juan María Garrido, Bartolomé Luque, and Iván González Torre; Tereza Motalova; Kateřina Pelegrinová; Haruko Sanada; Yawen Wang and Haitao Liu); those attempting to estimate or automatically classify data (Michele A. Cortelazzo, Franco M. T. Gatti, George K. Mikros, and Arjuna Tuzzi; Yoshifumi Kawasaki; Tatiana A. Litvinova and Olga A. Litvinova; Adam Pawłowski and

https://doi.org/10.1515/9783110763560-202

Tomasz Walkowiak; Gen Tsuchiyama); those dealing with lexical distribution in one or more texts (Jiří Milička, Václav Cvrček, and David Lukeš; Makoto Yamazaki); those discussing analytical methods (Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler); and those comparing several languages to review rules or to find trends (Xinying Chen, Kim Gerdes, Sylvain Kahane, and Marine Courtin; Ján Mačutek and Emmerich Kelih; Yawen Wang and Haitao Liu).

It is a great pleasure for the editors to publish this book as part of the Quantitative Linguistics series. We thank all the authors for their valuable contributions, despite the COVID-19 pandemic. The preparation of this book was supported by the directors of the International Quantitative Linguistics Association (IQLA) and by the National Institute for Japanese Language and Linguistics (Japan) and its affiliated organization, the Centre for Corpus Development.

The Editors
Makoto Yamazaki
Haruko Sanada
Reinhard Köhler
Sheila Embleton
Relja Vulanović
Eric S. Wheeler

Contents

Radek Čech, Barbora Benešová, Ján Mačutek

Why does negation of the predicate shorten a clause?

Abstract: According to the Menzerath-Altmann law, the mean word length is greater in shorter clauses than in longer ones. In Czech, negation is mostly realized by adding the prefix ne- to the beginning of the word, which makes the word longer (and, consequently, it also increases the mean word length in the clause). Therefore, we predict that clauses in which the predicate is in the affirmative form are longer than ones with the negative predicate. We test the hypothesis on a sample of 59 pairs of affirmative and negative forms of the same verb from the Prague Dependency Treebank 3.0.

Keywords: clause length, negation, Menzerath-Altmann law

1 Introduction

Language is a complex system composed of many units of different kinds which interact with each other. The complexity of the system seems to be the cause of the difficulty connected with describing the system and with explaining its properties. However, some properties of many different complex systems are results of relatively simple "mechanisms" (Barabási & Albert 1999, Newman 2010) which have a decisive impact on the system behaviour. Capturing some of the mechanisms in the form of an empirically testable law (or a hypothesis, at least) enables a verification of its validity and opens a way towards an explanation of system properties.

In this paper, we analyse the impact of the Menzerath-Altmann law (Cramer 2005, MAL hereafter), which expresses a very general mechanism controlling a relation between the sizes of language units (for details see Section 2), on certain grammar characteristics. According to the MAL, there is a systematic relation between lengths of language units belonging to the neighbouring

Acknowledgement: J. Mačutek was supported by the grant VEGA 2/0096/21.

Radek Čech, University of Ostrava, e-mail: cechradek@gmail.com
Barbora Benešová, University of Ostrava, e-mail: benesovaba@seznam.cz
Ján Mačutek, Mathematical Institute, Slovak Academy of Sciences and Constantine the Philosopher University in Nitra, e-mail: jmacutek@yahoo.com

https://doi.org/10.1515/9783110763560-001

levels in the language unit hierarchy – in a simplified way, the longer the "higher" unit, the shorter the mean length of the "lower" unit. It means that a change in length of one unit should cause a change in length of the unit from the neighbouring levels (e.g. if words are made longer, clauses are expected to become shorter).

We focus on negation in Czech which is (not exclusively, but in the vast majority of cases) realized by adding the prefix *ne-* to the beginning of a word, e.g.

(1) *Marie přišla,*
[Mary came]

(2) *Marie* ***ne****přišla.*
[Mary did not come]

We will consider only this realization of the negation in our paper.

This prefixation means that the word becomes one syllable (and also one morpheme, but we measure word length in syllables in this paper) longer, which, in accordance with the MAL, should generally make the clauses with negated forms of the word shorter than ones containing the same word without the negative prefix. It must be emphasized that the MAL is of a stochastic character and the law represents a general tendency. Thus, some instances which do not follow the law are admissible – as an example, see clauses (1) and (2) which have the same length (counted by the number of words) despite the fact that the mean word lengths in syllables (and also in morphemes) differ. In other words, the validity of the stochastic law is manifested on a large sample and the existence of some counterexamples does not mean a violation of the law (as is the case with a deterministic law).

The aim of the study is to test the following hypothesis based on the MAL: Clauses with the negative form of the predicate contain (on average) fewer words than clauses with the affirmative form of the predicate.

The paper is organized as follows. Section 2 provides a brief description of the MAL, focusing on its realization at the level of clauses and words. The methodology we applied and the language material which was used are presented in Section 3. Section 4 brings the results, together with their interpretation for some verbs which do not behave according to the hypothesis. Finally, Section 5 concludes the paper with a short discussion.

2 Menzerath-Altmann law

The MAL was formulated for the first time by the German linguist Paul Menzerath as a relation between the duration of sounds in a syllable and length of the syllable in which the sounds occur (Menzerath 1928 – longer syllables consist of relatively shorter sounds), and later as a relation between word length and length of syllables in the word in a dictionary (Menzerath 1954 – the more syllables a word contains, the shorter its syllables are on average).

It is, however, valid much more generally, presumably for all immediate neighbours in the language unit hierarchy (such as e.g. phoneme – morpheme/syllable – word – clause – sentence), see examples provided by Altmann (1980) and by Cramer (2005). Mačutek et al. (2019) suggested a very general formulation of the law as follows: *The mean size of constituents is a function of the size of the construct*, where a construct is a higher language unit (e.g. word) composed of constituents, i.e. lower-level units (e.g. syllables). The usual mathematical expression of this general form of the MAL is

$$y(x) = ax^b e^{-cx}, \tag{3}$$

with *y(x)* being the mean size of constituents if the size of the construct is *x*; *a*, *b*, and *c* are parameters. Very often, a more simple formula is sufficient, namely

$$y(x) = ax^b, \tag{4}$$

with parameter *b* attaining a negative value. It is a special case of formula (3) for $c = 0$. The function (4) is decreasing, and the MAL can be reworded as *the longer the construct, the shorter its constituents* (which is true also for the original Menzerath's observations).

The MAL is valid, among others, also for the relation between sentence length (measured in the number of its clauses) and clause length (measured in words), see Köhler (1982) for English, Heups (1983) for German, and Teupenhayn and Altmann (1984) for eight languages (German, English, French, Swedish, Hungarian, Slovak, Czech, and Indonesian). For all texts under study, the mean clause length decreases with the increasing sentence length, i.e. formula (4) can be used to model the relation.

The MAL can be interpreted in a context of both Zipf's principle of least effort (Zipf 1949) and the synergetic linguistic theory (Köhler 2005). Specifically, the MAL expresses the mechanism which controls proportions of lengths of units of different linguistic levels. These proportions can be seen as a result of the speaker's and hearer's communication requirements which are determined

by a strategy to achieve a communication goal(s) with the least effort. In other words, the co-existence of these requirements leads to a dynamic equilibrium among lengths of linguistic units.

Specifically, negation makes the predicate verb one syllable longer. The equilibrium is thus shifted and the speaker is "forced" – by the least effort requirements – to "find" it by making shorter the clause in which the negative predicate occurs. On the contrary, the use of affirmative predicate verbs allows for longer clauses. Thus, if the hypothesis is not falsified, we can state that there is a systematic relation between negation of the predicate and length of the clause, the fact which has not been observed yet, to our knowledge.

3 Methodology and language material

A clause is considered as the construct that consists of units of a lower degree that constitute it, i.e. words. Within the MAL context, the word is traditionally regarded as the closest lower unit to the clause (see Section 2), however, some approaches have accepted the syntactic phrase as the neighbouring level (cf. Mačutek et al. 2017). Although there are many definitions of the clause, they do not usually differ essentially and they share some crucial features. For instance, according to Crystal (2003), the clause is "a unit of grammatical organization smaller than the sentence, but larger than phrases, words or morphemes". Similarly, in the Prague Dependency Treebank 3.0 (Bejček et al. 2013, PDT 3.0 hereafter), which is used for this analysis (see below), clauses are defined as "grammatical units out of which complex sentences are built. A clause typically corresponds to a single proposition expressed by a finite verb and all its arguments and modifiers (unless they constitute clauses of their own)" (Mikulová et al. 2013). This annotation of clause is used in our analysis. As for the lower unit, i.e. the word, we determine it in accordance with the annotation of the PDT 3.0, which means that a word is identified as a sequence of letters between spaces.

In this study, only predicates were chosen for the analysis, mainly because the predicate constitutes a root of a clause structure, according to dependency grammar formalism (Mel'čuk 1988, Hudson 2010, Osborne 2019), and, consequently, it has a decisive impact on the clause structure, including its length. Further, we decide to compare average lengths of clauses containing different forms (affirmative vs. negative) of the same verb. Specifically, clauses containing either affirmative or negative form of certain verbs were chosen, creating pairs that enabled measuring and testing the differences in size of clauses containing the same verb, but either with or without negation. Fifty-nine pairs of

verbs were tested altogether, with the minimal number of occurrences of each form of the verb in the PDT 3.0 being twenty. We distinguish among different word forms (for verbs in Czech they can differ for different grammatical categories such as person, gender, number, tense, etc.), i.e. we did not lemmatize the sentences.

4 Results

The mean lengths of the clauses which contain the affirmative and negative forms were enumerated and compared for each of the 59 verbs from our sample (see Section 3). The significance of the differences between them were statistically tested by the Mann-Whitney-Wilcoxon test (the Student t-test cannot be applied because the data are not distributed normally, as was shown by the Shapiro-Wilk test). Results are presented in Tab. 1, where AFF is the affirmative verb form in Czech, ENG its English equivalent,[1] f(AFF) is the frequency of the affirmative form in the PDT 3.0, NEG is the negative verb form in Czech, f(NEG) the frequency of the negative form in the PDT 3.0, CL(AFF) and CL(NEG) are the mean lengths of clauses which contain the affirmative and the negative verb form, respectively, and p is the p-value of the test. Verbs which do not behave according to our hypothesis (i.e. the mean clause length is higher for their negative forms) are highlighted in bold.

For 48 out of 59 verbs, the hypothesis from Section 1 is corroborated, as the clauses with the affirmative forms of these verbs are on average longer (the p-value of the Mann-Whitney-Wilcoxon test is below 0.05 for 24 verbs, below 0.01 for 15 verbs).

The remaining 11 verbs contradict the hypothesis. However, this behaviour can be explained for the majority of them. For instance, the only case when the clauses containing the negated predicate are significantly longer than clauses with the affirmative predicate is represented by pair of verbs *myslím* [I think] and *nemyslím* [I do not think]. Formally, the word *myslím* [I think] is a finite verb which has the function of the predicate. However, a closer observation of particular clauses reveals that the word *myslím* has two different grammatical functions in the sample. In many sentences it is used as an adverb or a particle which expresses uncertainty of the statement (actually, it is a parenthesis). Grepl and Nekula (2017) provide as an example e.g. the sentence *Bude myslím pršet* (which,

1 Naturally, the English translations of the verbs depend on the context. We present here the most obvious “dictionary translations”.

Tab. 1: Clause length for affirmative and negative forms of verbs.

AFF	ENG	f(AFF)	NEG	f(NEG)	CL(AFF)	CL(NEG).	p
jsou	they are	2641	nejsou	302	9.18	8.05	< 0.01
bude	he/she/it will	2199	nebude	309	10.08	8.74	< 0.01
byl	he was	1785	nebyl	202	9.54	8.50	< 0.01
má	he/she/it has	1783	nemá	272	9.45	8.08	< 0.01
bylo	it was	1390	nebylo	185	9.13	8.62	0.12
jsem	I am	1264	nejsem	41	7.07	4.95	0.12
jsme	we are	1250	nejsme	34	8.00	6.74	0.17
byla	she was	1200	nebyla	134	10.04	8.02	< 0.01
může	he/she/it can	1013	nemůže	203	10.15	9.09	< 0.01
budou	they will	896	nebudou	127	10.56	8.63	< 0.01
měl	he had/should	810	neměl	112	10.72	8.47	< 0.01
mají	they have	753	nemají	144	8.94	7.99	0.11
musí	he/she/it/they must	701	nemusí	120	9.12	8.81	0.51
byly	they were	679	nebyly	77	10.47	8.43	< 0.01
jde	**he/she/it goes**	**623**	**nejde**	**119**	**8.34**	**8.76**	**0.10**
měla	she had/should	563	neměla	65	11.30	8.57	0.01
mohou	they can	466	nemohou	88	11.09	8.57	< 0.01
měli	they (masc. anim.) had/ should	373	neměli	57	10.12	7.93	< 0.01
patří	he/she/it/they belong(s)	366	nepatří	24	10.97	6.79	< 0.01
byli	**they (masc. anim.) were**	**352**	**nebyli**	**29**	**9.22**	**9.34**	**0.96**
mělo	it had/should	300	nemělo	54	11.08	11.00	0.71
mohl	he could	289	nemohl	70	10.27	8.80	0.02
chce	he/she/it wants	287	nechce	54	9.69	7.87	0.02
měly	they had/should	268	neměly	29	11.91	9.93	0.06
znamená	he/she/it means	230	neznamená	32	7.28	4.63	0.06
platí	he/she/it pays/is true	219	neplatí	41	8.20	6.44	0.07
máme	**we have**	**212**	**nemáme**	**51**	**7.19**	**7.47**	**0.80**
stojí	he/she/it stands/costs	210	nestojí	21	9.23	6.33	< 0.01
podařilo	it succeeded	194	nepodařilo	44	12.55	10.89	0.02
došlo	it came	186	nedošlo	39	9.81	8.00	0.03
mohli	they (masc. anim.) could	184	nemohli	35	10.18	7.69	< 0.01
mohla	she could	180	nemohla	20	11.31	9.60	0.14
mám	I have	168	nemám	45	6.13	6.02	0.64
můžeme	we can	155	nemůžeme	50	8.99	6.82	0.02
chtějí	they want	152	nechtějí	31	8.78	7.45	0.19
budeme	we will	139	nebudeme	24	7.94	7.63	0.84
dá	he/she/it will give	139	nedá	63	7.77	7.19	0.43
chtěl	he wanted	131	nechtěl	35	8.37	8.06	0.87
existuje	he/she/it exists	122	neexistuje	63	7.75	6.24	0.17
myslím	**I think**	**122**	**nemyslím**	**28**	**1.81**	**2.82**	**<0.01**
stalo	he/she/it became	117	nestalo	24	8.58	4.79	< 0.01

Tab. 1 (continued)

AFF	ENG	f(AFF)	NEG	f(NEG)	CL(AFF)	CL(NEG).	p
šlo	**it went**	**116**	**nešlo**	**30**	**8.12**	**9.10**	**0.14**
chceme	we want	84	nechceme	25	7.35	5.56	0.12
chtěli	they (masc. anim.) wanted	84	nechtěli	25	8.19	6.24	0.06
ví	**he/she/it knows**	**82**	**neví**	**62**	**4.41**	**4.69**	**0.63**
vidí	he/she/it/they see(s)	67	nevidí	26	9.28	7.81	0.31
hrozí	he/she/it/they threaten(s)	61	nehrozí	33	8.43	6.48	0.04
víme	**we know**	**61**	**nevíme**	**27**	**3.18**	**3.56**	**0.77**
dokáže	he/she/it can do	54	nedokáže	20	9.09	8.60	0.67
mění	he/she/it/they change(s)	53	nemění	33	9.45	7.48	0.03
vím	I know	52	nevím	75	2.50	2.27	0.04
mohu	I can	47	nemohu	33	7.40	6.06	0.35
brání	**he/she/it/they defend(s)**	**44**	**nebrání**	**22**	**8.55**	**9.14**	**0.78**
chci	I want	40	nechci	34	5.30	4.76	0.53
zbývá	**he/she/it remains**	**38**	**nezbývá**	**24**	**7.55**	**9.71**	**0.13**
vědí	**they know**	**37**	**nevědí**	**22**	**5.32**	**5.59**	**0.47**
věděl	**he knew**	**34**	**nevěděl**	**20**	**4.59**	**5.10**	**0.48**
zná	he/she/it knows	29	nezná	22	7.31	6.36	0.63
souhlasí	he/she/it/they agree(s)	24	nesouhlasí	31	8.42	6.65	0.53

according to the PDT 3.0 syntactic formalism, consists of two clauses, *Bude pršet* and *myslím*) with the literal translation *I think it will rain*, but its meaning is *Probably it will rain* or *Supposedly it will rain* etc. In these sentences, *myslím* is annotated as a one-word clause in the PDT 3.0. Consequently, the mean length of clauses with the affirmative predicate decreases substantially.

Further, the biggest difference between clause lengths among verbs which contradict our hypothesis is observed for *zbývá* [he/she/it remains] and *nezbývá* [he/she/it does not remain], one can find. Here, for 21 out of 23 instances of the negative form, the verb occurs in the syntactic phrase *nezbývá než* [there is no other way but to . . .]. However, this syntactic structure (i.e., verb + *než*) is not attested in the affirmative form in the sample. Obviously, *nezbývá než* cannot be considered a "pure" negation of affirmative form *zbývá*.

A brief examination of all verbs which contradict the hypothesis suggests that considering only the presence or absence of the prefix *ne-* can be too rough a criterion to extract the affirmative and negative forms of verbs from a corpus, and that both semantics and phraseology play an important role. A finer-grained approach which would take them (and possibly other factors) into account can shed more light on the problem and thus enhance results achieved in this pilot study.

5 Conclusion

Our results indicate that the MAL captures a "mechanism" which has indeed a very strong impact on properties of language units. A slight increase in the mean word length (we remind reader that we focus solely on the predicate and its negation, which makes it one syllable or one morpheme longer) is reflected in shorter clauses in 48 out of 59 verbs (i.e. in more than 80% of them). For one half of them (24 out of 48), the difference is statistically significant if the significance level is set to be 0.05. The remaining 11 verbs display the opposite tendency (i.e. negative clauses are longer), but this fact can be explained by different functions which the affirmative and negative form of those verbs have in sentences. This is true especially for the verb *myslím* [I think] (see Section 4), the only one for which the negative clauses are even significantly longer.

The paper opens also several other problems which can be solved only after larger corpora from several languages are analyzed. First, verbs seem to have a special position in clauses, one could say that they are more important than other words (see e.g. Čech et al. 2011). It would be interesting to check whether negation of e.g. adjectives (which is in Czech realized mostly in the same way, i.e. by adding the prefix *ne-*) has the same effect. Second, negation of the predicate is realized by different means in different languages. In Czech, a one-syllable prefix is added, while e.g. in English or in French two one-syllable words are often needed (e.g. *I understand* vs. *I do not understand*, or *je comprends* vs. *je ne comprends pas*). The question is whether the "reaction" of clause length is "stronger" in such cases.

References

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

Barabási, Albert-László & Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439). 509–512.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. „Prague Dependency Treebank 3.0." http://ufal.mff.cuni.cz/pdt3.0/ (accessed 18 October 2021)

Cramer, Irene M. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, 659–688. Berlin/New York: de Gruyter.

Crystal, David. 2003. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Čech, Radek, Ján Mačutek & Zdeněk Žabokrtský. 2011. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and its Applications* 390(20). 3614–3623.

Fowler, Henry W., Francis G. Fowler & Della Thompson. 1995. *The Concise Oxford Dictionary of Current English*, 9th edn. Oxford: Clarendon Press.

Grepl, Miroslav & Marek Nekula. 2017. Postojová částice. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [New encyclopedic dictionary of Czech]. https://www.czechency.org/slovnik/POSTOJOVÁČÁSTICE (accessed 18 October 2021)

Heups, Gabriela. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler & Joachim Boy (eds.), *Glottometrika* 5, 113–133. Bochum: Brockmeyer.

Hudson, Richard. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Köhler, Reinhard. 1982. Das Menzeratsche Gesetz auf Satzebene. In Werner Lehfeldt & Udo Strauss (eds.), *Glottometrika* 4, 103–113. Bochum: Brockmeyer.

Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, 760–774. Berlin/New York: de Gruyter.

Mačutek, Ján, Jan Chromý & Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics* 26(1). 66–80.

Mačutek, Ján, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni & Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, *Pisa, Italy, 2017*, 100-107. Linköping: Linköping University Electronic Press.

Meľčuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Albany (NY): State University of New York Press.

Menzerath, Paul. 1928. Über einige phonetische Probleme. In *Actes du premier Congrès international de linguistes*, 104–105. Leiden: Sijthoff.

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

Mikulová, Marie, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková & Zdeněk Žabokrtský. 2013. „From PDT 2.0 to PDT 3.0 (Modifications and Complements).“ https://ufal.mff.cuni.cz/pdt3.0/doc/tr54.pdf (accessed 18 October 2021)

Newman Mark E. J. 2010. *Networks. An introduction*. Oxford: Oxford University Press.

Osborne, Timothy. 2019. *A dependency Grammar of English: An Introduction and Beyond*. Amsterdam/Philadelphia: John Benjamins.

Teupenhayn, Regina & Gabriel Altmann. 1984. Clause length and Menzerath's law. In Joachim Boy & Reinhard Köhler (eds.), *Glottometrika* 6, 127–138. Bochum: Brockmeyer.

Zipf, George K. 1949. *Human Behavior and the Principle of the Least Effort. An Introduction to Human ecology*. Cambridge (MA): Addison-Wesley.

Xinying Chen, Kim Gerdes, Sylvain Kahane, Marine Courtin

The co-effect of Menzerath-Altmann law and heavy constituent shift in natural languages

Abstract: The present paper tries to link the Menzerath-Altmann law (MAL) to the Heavy Constituent Shift (HCS) phenomenon and discuss their co-effect in human natural languages. We deduce a hypothesis based on MAL and HCS and then try to empirically verify it by investigating multiple language data from Surface-Syntactic Universal Dependencies (SUD). Our results show that the hypothesis is valid across the complete set of typologically diverse languages and the co-effect of MAL and HCS appears to be a very regular universal.

Keywords: Menzerath-Altmann law, heavy constituent shift, surface-syntactic universal dependencies, co-effect, natural languages

1 Introduction

Menzerath's law, also known as the Menzerath-Altmann Law (MAL), predicts that the increase of the size of a linguistic construct results in a decrease of the average size of its components (Altmann & Schwibbe 1989; Hřebíček 1995; Cramer 2005a). Despite the success of the research on MAL, it seems that this powerful law is still not strongly connected to other traditional linguistic discussions that go beyond a mere application of the definitions to various linguistic units. In this study, we aim to address this point by linking MAL to the Heavy Constituent Shift (HCS) phenomenon and discuss their co-effect in human natural languages.

The article is organized as follows. Section 2 starts with a reminder of some studies on MAL. HCS is introduced in Section 3 and our co-effect hypothesis is presented in Section 4. Section 5describes the methodology of this study and the

Acknowledgment: This work is supported by the National Social Science Fund of China (18CYY031).

Xinying Chen, Xi'an Jiaotong University, e-mail: chenxinying@mail.xjtu.edu.cn
Kim Gerdes, University Paris Saclay, e-mail: kim@gerdes.fr
Sylvain Kahane, Université Paris Nanterre, e-mail: sylvain@kahane.fr
Marine Courtin, University Sorbonne Nouvelle, e-mail: rinema56@gmail.com

https://doi.org/10.1515/9783110763560-002

data resource that the study is based on. Section 6 summarizes the results and the conclusion is provided in Section 7.

2 Menzerath-Altmann law

The Menzerath-Altmann law is one of the most discussed linguistic laws; the majority of related studies are focused on verifying this law in certain linguistic constructs with different texts and languages as well as trying to interpret the parameters (Altmann 1980 & 2014; Gustison et al. 2016; Cramer 2005b; Mikros & Milička 2014), for example, examining whether longer words (in the number of syllables) have shorter syllables (in the number of graphemes for phonemes), or if longer clauses (in the number of words) have shorter words (in the number of syllables) in different human languages (Menzerath 1954; Kelih 2010). A small minority of studies discuss the language features that might influence the results of MAL, such as registers (Hou et al. 2020; Xu & He 2020). Meanwhile, MAL has started to transcend quantitative linguistics and is also gaining attention from other disciplines, such as biology (Li 2012; Ferrer-I-Cancho & Forns 2009).

3 Heavy constituent shift

HCS (Ross 1967; Stallings et al. 1998) is a well-known phenomenon of syntax. Based on the concept of "heavy constituents" that are composed of more words (and syllables) than "light constituents", it states that heavier constituents tend to be shifted to the end of the clause. Here is the example 5.56 from Ross (1967: 306):

(1) a. I'll ***give*** some **to my good friend from Akron**.
b. I'll ***give*** **to my good friend from Akron** some.

In this example, the constituent 'to my good friend from Akron' has six words and it is heavier than the constituent 'some' which only has one word. Therefore, it should be shifted to the end of the sentence, as in the further examples from the GUM (The Georgetown University Multilayer Corpus) English treebank of Universal Dependencies (Zeldes 2017):[1]

1 These examples have been collected with the following grew-match request: pattern {X – [obl | advmod] -> B; X – [obj] -> C; X ≪ B; B ≪ C} without {X -> D; X ≪ D}.

(2) a. [. . .] I might capture them and ***learn*** from them **the secrets which the moon had brought upon the night**. (fiction_moon-9)
b. [. . .] the bartender will ***recount*** for the customer **the definition of the santorum neologism**. (interview_coktail-15)
c. [. . .] a scenery made of sand and rocks which ***have*** vaguely **the shape of a castle**. (voyage_guadeloupe_17)
d. [. . .] the adjustments and calculations ***take*** into account **the weighted nature of the data**. (academic_discrimination-51)
e. [. . .] the only candidate who ***embodies*** both physically and philosophically **the growing diversity of the commonwealth**. (interview_libertarian-11)

This commonly observed language phenomenon has been noted by several linguists before Ross (1967). Here are three citations by French linguists from the 18th and 19th centuries, given in Kahane (2020):

> The [complements[2]] must be as close as possible to the governing word, which would not be the case if one were to put the longest [complement] first, which would move the shortest one too far away. (Buffier 1709: 313)

> When several complements fall on the same word, it is necessary to put the shortest one first after the completed word; then the shortest of those that remain and so on until the longest of all, which must be the last. It is important for the clarity of the expression, *cujus summa laus perspicuitas*,[3] to move what serves as the complement as little as possible away from a word. However, when several complements contribute to the determination of the same term, they cannot all follow it immediately; and all that remains is to bring the one that we are forced to keep away from it as close as possible to it: this is what we do by putting first the one which is the shortest, and keeping the longest for the end. (Beauzée 1765: 7)

> When several complements fall on the same word, give the most concise form to the one immediately following the complete word and, as you go along, give the complements a more developed and extensive expression. (Weil 1844: 97)

Note that HCS has first been observed for SVO languages such as French and English, where the complements are produced after the verb that governs them. The term heavy constituent shift has been coined by Ross in the framework of transformational grammar, with the idea that heavy constituents were shifted from some

2 In the French tradition, complement means argument constituents as well as modifier constituents depending on the verb. We will keep this sense in the paper.
3 'whose highest praise is clarity' (a variation of the famous quote from Quintilian's The Orator's Education stating that the oratory's "basic virtue is clarity").

initial position to the final place. For Buffier and Beauzée, light complements must simply be produced before heavy complements. Weil introduced an additional idea: If you want to produce two complements in a given order, make the second one heavier than the first one. In other words, it is not because a complement is heavy that you put it in the second place, it is because it is in the second place that you make it heavier (and, again, there is absolutely no shift in this framing of the phenomenon).

There are still debates concerning the definition of 'heavy'. Although the theoretical discussion is valuable, for the empirical data analysis in this study, we take the operational definition of 'heavy', namely, having more words.

4 Co-effect hypothesis: Combining the heavy constituent shift and the Menzerath-Altmann law

Both HCS and MAL are associated with the constituent size. This suggests that HCS and MAL interfere with each other. We can deduce a hypothesis based on these two premises.

To be more specific, we investigate and compare the size of different constituents in two types of clauses that have either one or two complements to the right of the word X:[4]

(3) ~XAB (the word X has two complements A and B to its right, and A precedes B)

(4) ~XC (the word X has only one complement C to its right)

We will focus on words X, when it is the verbal head of a clause. a, b, c corresponds to the size (the number of words) of the constituents A, B, and C.

First, according to MAL, we can expect that the average size of two complements (case 1.) is smaller than the size of the unique element (case 2.):

(5) $(a + b) / 2 < c$

4 Note that this simplified definition allows any number of dependents to the left of X and does not take into account the presence and size of any elements to the left of X which might be part of the projection of X. We will see in Section 6 that taking into account possible elements to the left does not significantly alter the results.

And then, according to HCS, we also expect that B is heavier than A:

(6) $a < b$

When we combine I & II, we can get that:

(7) $a = (a + a) / 2 < (a + b) / 2 < c \Rightarrow a < c$

We thus presume that we should observe "$a < c$" in empirical data, and if our hypothesis is validated, this language phenomenon can be seen as the co-effect of MAL and HCS in human natural languages.

5 Methodology

Constituents, from the viewpoint of dependency syntax, are projections of a node in the dependency tree, that is the node and all the nodes that it dominates.

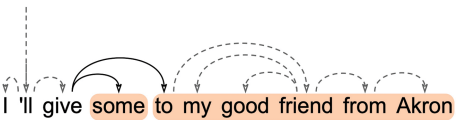


Fig. 1: The dependency tree of the sentence 'I'll give some to my good friend from Akron.

As we can see in Fig. 1, there are two dependencies (bold lines) that fall on the right side of the verb ***give***. Each branch heads one constituent. These two constituents are the two complements of ***give***. In our study, the size of a constituent will be determined by the number of words it contains. The two complements of ***give*** on its right, have respectively size one and size six.

What we are investigating in this paper are two types of clauses, namely, ~XAB and ~XC. In which, ~ represents left branches of the tree. It should be noted that here we do not take into consideration the size of possible left tree branches. For instance, the following three sentences would all be considered as ~XAB clauses:

(8) a. I definitely give some to my good friend from Akron.
(including two left tree branches)
b. I give some to my good friend from Akron.
(including one left tree branches)
c. Give some to my good friend from Akron.
(including zero left tree branches)

And all the following three sentences would be considered as ~XC type clauses:

d. I probably did the job. (including two left tree branches)
e. I did the job. (including one left tree branch)
f. Do the job. (including zero left tree branches)

Furthermore, we strictly limited the numbers of the right tree branches to either one or two. For instance, the following clauses would not be considered in our analysis:

g. I tried. (including zero right tree branches)
h. I told her the truth eventually. (including three right tree branches)

To test our hypothesis, we chose the Surface-Syntactic version (SUD 2.7, Gerdes et al. 2018 & 2019) of the Universal Dependencies treebank set (Nivre et al. 2016). The dataset includes 183 treebanks in 104 languages from various typological groups, with a majority of Indo-European languages. For some languages, several treebanks have been developed. In this pilot study, we are more interested in the general picture, and we combine all the treebanks of a language into one collective treebank. Therefore, we take global measures across all trees of each language.

After clearly defining all the conditions, we first filter out ~XAB type and ~XC type clauses from each dependency treebank we study. We only look at X that are verbs and A, B, C that are subjects or complements. More specifically, we only look at the complements with the dependency tag 'subj', 'comp', 'mod', or 'udep' ('udep' is an underdetermined relation that subsumes both 'comp' and 'mod').[5] For each clause we collect, we compute the size of the constituents A, B or C and store them as *a*, *b*, or *c*, and then we calculate the

5 Of course we also take into account all possible extensions of these tags, such as 'comp:obj', 'compl:obl', 'comp:aux', etc.

mean value of all *a*, *b*, and *c* on the whole treebank. By comparing the mean value of *a* and *c*, we can either accept or reject our hypothesis.

For the numbers of ~XAB and ~XC clauses in each language, see Tab. 1 in the Appendix.

6 Results

We filter out languages with very sparse data that have less than 20 measures of *a* or *c*. This reduces the number of languages to 80. Our results in Fig. 2 show that all languages appear above the diagonal. It reflects that our hypothesis "$a < c$" is verified across these typologically different languages.

The colors and shapes in Fig. 2 roughly represent language groups.[6]

- Indo-European languages: triangles
  - Indo-European-Romance: brown
  - Indo-European-Baltoslavic: purple
  - Indo-European-Germanic, including the English Creole *Naija*: olive
  - Other Indo-European: blue
- Sino-Austronesian: green stars
- Agglutinating languages: red plus signs
- Other languages: black squares

For this article, the actual values of *a*, *b* and *c* in each language are presented in Tab. 1 in the Appendix.

6 We provide an interactive interface on https://typometrics.elizia.net/ allowing for an easy visual exploration of the data as in Fig. 2. The raw data can be found on https://github.com/typometrics/datapreparation.

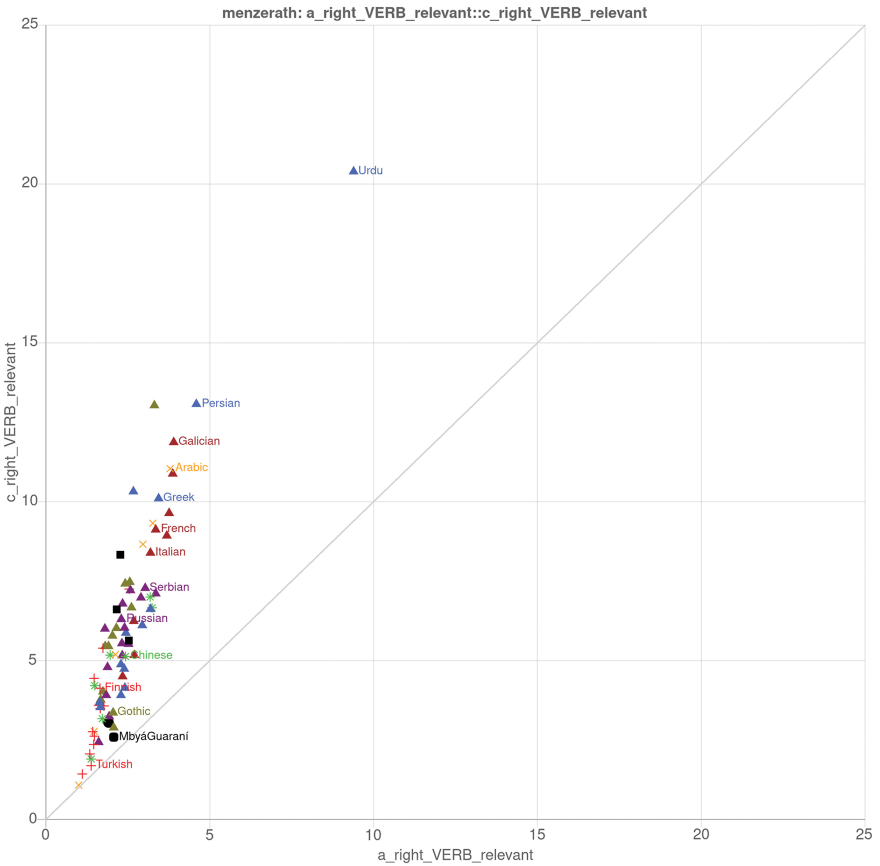


Fig. 2: The average size *c* of C constituents is bigger than the average size *a* of A constituents across the 80 languages of SUD 2.7 where we have at least 20 occurrences of corresponding structures.

7 Conclusion

Our results show that our hypothesis is valid across the complete set of typologically diverse languages that are present in SUD treebanks. The co-effect of Menzerath-Altmann Law and Heavy Constituent Shift appears to be a very regular universal.

Our pilot study shows that by making use of the recently available coherently annotated multilingual SUD, we can bridge MAL with traditional linguistic discussions such as the HCS, and therefore expand the scope of studies on MAL.

Meanwhile, there are still various details to be investigated in the future. For example, we need to explore what happens to the left of the governor, in particular for verb-final languages. It might also be worthwhile to verify the measures for all kinds of clauses, not only the clauses that have a verbal head.

Note also that the data resource for languages is unevenly distributed. Some languages, such as German, English, Czech, Arabic, etc., have large treebanks, while treebanks of some languages have very limited sizes. We still have to evaluate in the future how much the sample size would affect the results. Also, even for the same language, treebanks annotated by different teams can vary from each other. We have to consider the effect of fusing treebanks in the future. Last but not least, we can gradually ease the control factors, reduce the constraints for selecting samples, to test the boundary conditions of the co-effect phenomenon.

Appendix

Tab. 1: Values of *a*, *b*, *c* and the numbers of selected clauses in each language.

Language	a	b	c	Number of ~XAB trees	Numbers of ~XC trees
Afrikaans	3.31	13.37	13.03	211	1281
Akkadian	1.9	4.4	1.71	10	276
Akuntsu	0	0	1.2	0	10
Albanian	2.53	6.16	5.98	19	51
Amharic	1.0	1.08	1.08	49	326
AncientGreek	2.41	5.2	4.14	8725	25192
Apurinã	1.08	2.67	1.61	12	54
Arabic	3.79	13.43	11.04	27627	25993
Armenian	3.19	8.91	6.62	182	2274
Assyrian	1.14	1.43	3.46	7	26
Bambara	2.53	9.27	5.63	168	1122
Basque	1.91	3.4	3.05	551	3997
Belarusian	2.33	5.07	5.16	4060	14177
Bhojpuri	7.1	9.9	9.95	10	74
Breton	2.29	4.44	3.91	228	480
Bulgarian	2.52	5.56	5.52	2491	9637
Buryat	1.0	12.33	5.2	3	81
Cantonese	1.96	5.21	5.17	85	647
Catalan	3.87	10.7	10.88	9561	22072
Chinese	2.43	5.14	5.13	564	21956
Chukot	1.11	1.94	1.43	54	254

Tab. 1 (continued)

Language	a	b	c	Number of ~XAB trees	Numbers of ~XC trees
ClassicalChinese	1.38	2.34	1.9	1895	27462
Coptic	2.12	7.84	5.19	1634	2163
Croatian	2.9	7.22	6.98	2290	9913
Czech	2.58	7.14	7.21	32884	97789
Danish	2.15	6.44	6.02	2456	4502
Dutch	2.42	7.19	7.42	3039	7525
English	2.61	6.3	6.67	13502	36424
Erzya	1.43	2.88	2.76	291	972
Estonian	1.74	4.81	5.39	9123	17000
Faroese	1.74	6.18	4.02	1144	1892
Finnish	1.65	3.99	4.13	10145	22200
French	3.35	9.74	9.12	21348	47025
Gaelic	2.44	8.51	5.87	2160	1247
Galician	3.9	12.0	11.87	2542	7803
German	2.56	7.27	7.47	30612	36399
Gothic	2.05	4.77	3.36	1598	3739
Greek	3.44	10.36	10.1	1348	3094
Hebrew	3.26	10.29	9.32	3527	6222
Hindi	5.71	6.71	16.93	17	4741
HindiEnglish	2.29	4.94	4.88	250	932
Hungarian	2.54	8.82	7.25	386	1334
Icelandic	1.81	6.55	5.45	23643	35949
Indonesian	3.18	7.66	7.0	2801	9876
Irish	2.94	6.94	6.11	2654	1706
Italian	3.19	9.18	8.39	12833	34151
Japanese	4.29	1.88	2.74	17	61
Karelian	1.77	3.74	3.57	69	138
Kazakh	0	0	1.37	0	19
Khunsari	0	0	3.4	0	5
Komi	1.66	4.59	3.49	80	371
Komi-Permyak	1.1	2.0	3.04	10	53
Korean	0	0	2.59	0	233
Kurmanji	1.64	7.07	3.69	28	304
Latin	2.7	7.19	5.17	14020	43318
Latvian	2.32	6.01	5.54	3130	15899
Lithuanian	3.35	7.33	7.11	709	5271
Livvi	1.61	4.63	3.59	38	82
Maltese	2.96	7.9	8.66	873	3547
Manx	2.39	7.05	4.74	223	88
Marathi	2.8	1.6	3.1	5	20
MbyáGuaraní	2.07	3.07	2.59	43	300
Moksha	1.46	2.17	2.36	24	88
Mundurukú	0	0	2.11	0	19

Tab. 1 (continued)

Language	a	b	c	Number of ~XAB trees	Numbers of ~XC trees
Naija	1.67	4.53	3.77	10948	35429
Nayini	0	0	5.0	0	3
NorthSami	1.48	2.61	2.62	822	1669
Norwegian	2.03	6.1	5.78	14499	29070
Old Turkish	1.0	16.0	5.0	1	1
OldChurchSlavonic	1.61	3.67	2.43	1666	4045
OldEastSlavic	1.93	3.61	3.24	4560	8697
OldFrench	2.34	5.64	4.5	3952	11790
Persian	4.59	11.33	13.07	228	9922
Polish	1.88	4.84	4.79	10347	27611
Portuguese	3.69	9.56	8.93	10412	26994
Romanian	2.68	6.68	6.24	18678	45215
Russian	2.3	6.23	6.3	19485	72153
Sanskrit	1.66	3.18	3.53	204	1001
Serbian	3.03	7.22	7.28	1314	4870
SkoltSami	1.36	2.09	3.34	11	50
Slovak	1.84	4.01	3.91	1687	6982
Slovenian	1.8	5.98	6.0	2215	9008
Soi	0	0	6.0	0	1
South Levantine Arabic	1.47	3.4	2.77	30	56
Spanish	3.76	10.3	9.64	19650	43411
Swedish	1.91	6.01	5.45	5196	8793
SwedishSign	2.07	3.52	2.89	27	88
SwissGerman	1.57	5.29	5.93	7	30
Tagalog	1.72	3.1	3.17	72	69
Tamil	1.0	1.0	1.08	1	39
Telugu	0	0	1.32	0	25
Thai	3.24	6.94	6.66	781	2326
Tupinambá	8.0	5.0	0	1	0
Turkish	1.38	2.79	1.69	101	1561
TurkishGerman	1.47	4.96	4.44	212	606
Ukrainian	2.4	6.08	6.03	1946	5985
UpperSorbian	2.34	6.11	6.79	122	244
Urdu	9.39	7.32	20.39	31	1737
Uyghur	1.34	2.54	2.06	50	102
Vietnamese	1.49	4.45	4.21	1226	3946
Warlpiri	1.0	1.0	1.31	2	16
Welsh	2.67	9.84	10.32	798	656
Wolof	2.16	7.14	6.61	881	3890
Yoruba	2.27	9.66	8.33	160	376

References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn (ed.), *Glottometrika*, 2. 1–10.

Altmann, Gabriel. 2014. Bibliography: Menzerath's law. *Glottotheory* 5(1). 121–123.

Altmann, Gabriel & Michael H. Schwibbe. 1989. *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim/Zürich/New York: Olms.

Beauzée, Nicolas. 1767. *Grammaire générale ou Exposition raisonnée des éléments nécessaires du langage, pour servir de fondement à l'étude de toutes les langues* [General grammar or Rational exposition of necessary elements to serve as the foundation for the study of all languages], vol. 2: Syntax. Paris: Barbou.

Buffier, Claude. 1709. *Grammaire françoise sur un plan nouveau* [French grammar on a new plan]. Paris: Le Clerc- Brunet-Leconte & Montalant.

Cramer, Irene M. 2005a. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, 659–688. Berlin/New York: De Gruyter.

Cramer, Irene M. 2005b. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics* 12(1). 41–52.

Ferrer-I-Cancho Ramon & Núria Forns. 2009. The self-organization of genomes. *Complexity* 15(5). 34–36.

Gustison, Morgan L., Stuart Semple, Ramon Ferrer-i-Cancho, & Thore J. Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19). E2750–E2758.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*, *November 1*, *Brussels*, *Belguim*.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *Treebanks and Linguistic Theories (TLT 2019)*, *Syntaxfest, August 28–29*, *Paris, France*.

Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics* 6(1). 17.

Hou, Renkui, Chu-Ren Huang, Kathleen Ahrens & Yat-Mei Sophia Lee. 2020. Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. *Digital Scholarship in the Humanities* 35(1). 54–66.

Hřebíček, Luděk. 1995. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Trier: Wissenschaftlicher Verlag Trier.

Kahane, Sylvain. 2020. How dependency syntax found its modern form in the French Encyclopedia: from Buffier (1709) to Beauzée (1765). In Nicolas Mazziotta & András Imrényi (eds.), *Chapters of Dependency Grammar: A Historical Survey from Antiquity to Tesnière*, 85–132. Amsterdam/Philadelphia: John Benjamins.

Kelih, Emmerich. 2010. Parameter interpretation of Menzerath law: evidence from Serbian. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives*, 71–79. Wien: Praesens.

Li, Wentian. 2012. Menzerath's law at the gene-exon level in the human genome. *Complexity* 17(4). 49–53.

Mačutek, Jan, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni & Joakim Nivre (eds.): *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 100–107. Linköping Electronic Conference Proceedings.

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 Fengxiang Fan, Emmerich Kelih, Reinhard Köhler, Ján Mačutek & Eric S. Wheeler (eds.), *Empirical Approaches to Text and Language Analysis*, 180–189. Lüdenscheid: RAM-VERLAG.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016), Portorož, Slovenia*, *May 23–28*, 1659–1666.

Ross, John Robert. 1967. *Constraints on Variables in Syntax*. Cambridge: Massachusetts Institute of Technology dissertation.

Stallings, Lynne M., Maryellen C. MacDonald & Padraig G. O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language* 39(3). 392–417.

Weil, Henri. 1844. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes* [About word order in ancient languages in comparison to modern languages]. Paris: Crapelet.

Xu, Lirong & Lianzhen He. 2020. Is the Menzerath-Altmann law specific to certain languages in certain registers?. *Journal of Quantitative Linguistics* 27(3). 187–203.

Zeldes, Amir. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation* 51(3). 581–612.

Michele A. Cortelazzo, Franco M. T. Gatti, George K. Mikros, Arjuna Tuzzi

Does the century matter? Machine learning methods to attribute historical periods in an Italian literary corpus

Abstract: This study aims to analyse an Italian literary corpus from a diachronic perspective using machine learning methods. With reference to a basis of texts written between the 16th and the 21st century, the aim is to apply a well-known robust machine learning (ML) algorithm (Random Forest – RF) in order to see how the texts are classified in four different partitions, representing periodizations theorized by four Italian literature scholars. The corpus we employed for training the ML algorithm includes 420 Italian texts: 100 texts from the 16th century, 27 from the 17th, 57 from the 18th, 100 from the 19th, 100 from the 20th, and 36 from the 21st. In order to vectorize the texts, we used the Author's Multilevel N-gram Profile (AMNP) (Mikros and Perifanos, 2013; Cortelazzo, Mikros, and Tuzzi, 2018), a document representation method that takes into account a diverse set of linguistic features (i.e., ngrams of increasing length – unigrams, bigrams, trigrams – and ngrams of increasing level – character, word). Each text was split into text chunks of 2000 words in length, and then it was transformed into AMNP vectors.

The results of this research have shown an impressive accuracy in classification with the Random Forest algorithm since the precision in the four periodizations reached a minimum value of 89% in the partition-based Migliorini's theories and a maximum value of 97% in the partition based on Cella's ones.

Looking at the misclassification cases, particularly in Migliorini's training, it's interesting to notice that when Random Forest makes a mistake in classifying text chunks into a century, its error is usually of +/– 1 century.

Keywords: machine learning, classification, diachronic corpora, Italian literature, random forest, Author's Multilevel Ngram profile

Michele A. Cortelazzo, Università degli Studi di Padova, e-mail: cortmic@unipd.it
Franco M. T. Gatti, Università degli Studi di Padova, e-mail: franco.gatti.1@phd.unipd.it
George K. Mikros, Hamad Bin Khalifa University, e-mail: gmikros@hbku.edu.qa
Arjuna Tuzzi, Università degli Studi di Padova, e-mail: arjuna.tuzzi@unipd.it

https://doi.org/10.1515/9783110763560-003

1 Introduction

The history of the Italian language doesn't have a uniform periodization that can be considered shared by all scholars. Several models have been proposed, most of them based on external criteria (i.e., historical phenomena). The most important models are the following four:

1. A partition by century (Migliorini 1960; Marazzini 1994). With a few exceptions, in this case, the periodization relies on factors external to the evolution of the language. In these manuals, each chapter is dedicated to one century of the history of the Italian language (the exceptions concern the period of origins, up to the 14th century, and the most recent period, from the second half of the 19th century);
2. A partition based on internal criteria, i.e., on the evolution of the language itself, particularly at the morphosyntactic level. Durante (1981) distinguishes five periods: the High Middle Ages, when first appeared texts in the Vulgar Italians; the Middle Ages, when Tuscan took on particular relevance among other Vulgar Italians; the 16th and the 17th centuries, when modern aspects of the Italian language emerged; the 18th and the 19th centuries (the period of the Europeanization and then of national unification); the 20th century;
3. A partition dividing the history of Italian into three periods, delimited by two fundamental events: the codification of Italian in the 16th century and the political unification of Italy (Cella 2015);
4. A partition that acknowledges great relevance to the modern and contemporary periods. Antonelli (2018) distinguishes an "Ancient Italian" (from its origins to 1750 circa; we will consider 1763 as the final year), a "Modern Italian" (from 1764 to 1945), a "Contemporary Italian" (from 1946, the year of the first Italian election by universal suffrage, which led the Country to a transition from monarchy to republic, to date).

We are not aware that a verification has ever been carried out on which of these periodizations appears more justified by the evolution of the vocabulary. And this is the purpose of this research: to verify, with the aid of computational stylistics methods, whether or not these periodizations mirror the evolution of the Italian language, as it emerges from the lexical profiles of literary writings selected for this research, and to compare them in terms of the classification results.

2 Corpus

This work was carried out on a corpus of 420 prose texts (novels, short stories, treatises, epistolaries) produced starting from the 16th Century up to the 21st. We didn't select texts from previous periods because only from the 16th century has a codified language model been recognized in Italian culture and shared in all geographic areas of today's Italian territory. Before then, texts from different geographic origins showed a strong diatopic differentiation, which, at least at this point of the research, could have been a disturbing element for the aims of this research.

Three corpora already collected for other purposes have been the basis for the constitution of this 420 prose texts corpus, and two out of three among these pre-existing corpora were already a source for studies on the Italian language. From the 16th to the Mid-20th century, the source for the period was the BIZ, Biblioteca Italiana Zanichelli (Stoppelli 2010), the latest version of an electronic collection of one thousand texts from the Italian written tradition, mainly literary. The source for the second half of the 20th century was De Mauro (2007). He collected a corpus of 100 novels among winners or those selected for the final stage of the Strega Prize, one of the most prestigious Italian literary prizes. Finally, a corpus prepared for stylistic research related to the works signed by Elena Ferrante (Tuzzi & Cortelazzo 2018), also called PIC (Padova Italian Corpus, Savoy 2018), was the source for the most recent years.

The most consistent part of the corpus was taken from the BIZ. In order to build the overall corpus, only those works that were best suited for our research were selected from it: only works in prose (or mainly in prose) were included, and works in dialect were excluded. Moreover, in order to have a good balance within the corpus, some of the works of those authors excessively represented in the corpus have been eliminated. Among the other two corpora, we selected those works that we subjectively considered more relevant for their respective periods. The aim was to collect 100 texts for the best represented centuries (16th, 19th, 20th centuries) and to include the highest number of available texts for the other centuries (17th and 18th centuries; for the 17th century, poorly represented in the BIZ, the corpus has been integrated with some works specifically digitized for this research). For the first part of the 21st century, we selected 36 texts to have an overall corpus of 420 works.

As the following schemes show clearly, the corpus is equally balanced (considering the number of historically available texts for each partition) for each, and every one of the hypotheses of periodization examined.

Distribution of texts within the corpus:

Scheme 1 [Migliorini]

Time period	Texts (N)	%
XVI Century	100	23.8
XVII Century	27	6.4
XVIII Century	57	13.6
XIX Century	100	23.8
XX Century	100	23.8
XXI Century	36	8.6

Scheme 2 [Durante]

Time period	Texts (N)	%
XVI–XVII Century	127	30.2
XVIII–XIX Century	157	37.4
XX–XXI Century	136	32.4

Scheme 3 [Cella]

Time period	Texts (N)	%
XVI Century – 1860	207	49.3
1861 – XXI Century	213	50.7

Scheme 4 [Antonelli]

Time period	Texts (N)	%
before 1763	173	41.2
from 1764 to 1945	155	36.9
from 1946 to XXI Century	92	21.9

3 Methods

In order to investigate which one of the four schemes achieves a satisfying confirmation with our computational stylistics methods, we transformed our corpus according to a document representation model named AMNP: Author's Multilevel N-gram Profile (Mikros & Perifanos 2013 and 2015). The basic idea of

this document representation method is to create vectors for each text which are composite and represent different linguistic levels using different linguistic features (e.g., bigrams, trigrams, etc.). More specifically, we can create a multilevel ngram profile by extracting the N most frequent ngrams from the text in increasing steps of ngram size and at the same time increasing also the linguistic base of the ngram moving from characters to words. With this method, we create a composite vector of character bigrams, trigrams, and word unigrams and bigrams, which defines the quantitative profile of our text. One of the advantages of this document representation method is that we create vectors that correspond to a wide variety of linguistic levels and can simultaneously capture diverse linguistic encodings, e.g., prefixes, suffixes, semantic components, phrasal elements, etc. The resulting ngram profiles train a ML supervised classification algorithm (in this research, Random Forest or RF). Through the training phase, each classification category (in this research, the time period) is associated with a complex quantitative ngram pattern. Using this methodology, each text can be attributed to its predicted time period. To do so, the first step was to set a corpus configuration for each one of the schemes above: four corpora identical in their texts and labeled following Migliorini's, Durante's, Cella's, and Antonelli's theories, respectively. Then, four separate analyses using the AMNP document representation method were carried out, one for each corpus configuration.

After encoding the corpus in UTF-8 format, we split it into 28,096 chunks of 2000 words in length each and computed four lists: potentially, the first list contains the 500 most frequent character bigrams in the corpus, the second list contains the 500 most frequent character trigrams, the third includes the 500 most frequent word unigrams (words), and the fourth the 500 most frequent word bigrams. These four lists are meant to capture different language levels, and they are pivotal to this multilevel analysis: by combining them, we obtain a single matrix with 2.000 columns. Then, we can exploit this matrix to draw up linguistic profiles by means of which we can identify the classes we designed in each configuration of the corpus.

The classification algorithm we selected for training using the AMNP vectors was the 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) 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 concerning 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 is 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 many advantages, especially when we are using biased and unbalanced data, as very frequently is the case with readability studies. The most prominent ones (Kho 2018) are listed below:

- Parallelizable: They are parallelizable, meaning that we can split the process into multiple machines to run. This results in faster computation time. Boosted models are sequential in contrast and would take longer to compute.
- Great with high dimensionality: Random forests are great with high dimensional data since we work 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 to Outliers and Non-linear Data: Random forest handles outliers by essentially binning them. It is also indifferent to non-linear features.
- Handles 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 a random forest, we are averaging the variance and have a low bias and moderate variance model.

As a classification target, we set the time periods classes as they were defined by the four different theoretical partitions we employ in this study. 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 5-fold cross-validation, i.e., the Random Forest will do its training on 5 folders containing 1/5 of the total amount of text chunks each (28,096/5 chunks, i.e., circa 5,619 chunks each). Therefore, the final result on the Random Forest run is the arithmetic mean between the results of each folder. This process has been done once for each corpus configuration (Migliorini's, Durante's, Cella's, and Antonelli's) in order to compare the results of Random Forest classification on them.

4 Results

For **Migliorini**'s configuration, we had the 420 texts corpus (N) split into 6 classes:

Partition	Time period	No. novels	No. text chunks
A	1500–1599 (**16th** Century)	100	7,279
B	1600–1699 (**17th** Century)	27	2,163
C	1700–1799 (**18th** Century)	57	1,830
D	1800–1899 (**19th** Century)	100	7,406
E	1900–1999 (**20th** Century)	100	6,192
F	2000–2019 (**21st** Century)	36	3,226

The classification results of **Random Forest** training on **Migliorini**'s configuration is reported in Tab. 1, as a percentage of correct classification of the chunks:

The best accuracy in Migliorini's configuration was reached with 59 mtry,[1] and its value is 0.89.

Aside from the tables containing the percentages of classification for each of the corpus's four configurations, we also calculated the Precision, Recall, and F1 Score. While the accuracy reported for each table is the total accuracy of the classification task, Precision, Recall, and F1 Score have to be calculated for each class.

The Precision, Recall and F1 Score values for each one of the 6 classes outlined by Migliorini are reported in Tab. 2.

1 Mtry is a specific training parameter in the Random Forest algorithm and it defines the number of variables available for splitting at each node of a decision tree. For extensive discussions

Tab. 1: Random Forest training on Migliorini's partition.

Migliorini		Reference					
		A	**B**	**C**	**D**	**E**	**F**
Prediction	A	99.8%	26.8%	14.0%	0.9%	0.1%	0.0%
	B	0.0%	70.7%	1.0%	0.0%	0.0%	0.0%
	C	0.0%	0.3%	67.2%	0.1%		0.0%
	D	0.2%	2.1%	17.8%	96.5%	16.8%	0.8%
	E	0.0%	0.0%	0.1%	2.5%	78.7%	11.5%
	F	0.0%	0.0%	0.0%	0.0%	4.5%	87.7%

Tab. 2: Precision, Recall and F1 Score for Migliorini's partition.

Migliorini			
Class	**Precision**	**Recall**	**F1 Score**
A	0.89	1.0	0.94
B	0.99	0.71	0.83
C	0.99	0.67	0.80
D	0.83	0.97	0.89
E	0.90	0.79	0.84
F	0.91	0.88	0.89

For **Durante**'s configuration, we had the 420 texts corpus (N) split into 3 classes as below:

Partition	Time period	No. novels	No. text chunks
A	1500–1699	127	9,442
B	1700–1899	157	9,236
C	1900–2019	136	9,418

The classification results of **Random Forest** training on **Durante**'s configuration are reported in Tab. 3, as a percentage of the correct classification of the chunks:

The best accuracy in Durante's configuration was reached with 59 mtry, and its value is 0.94.

about the influence of mtry in Random Forest applications see Cutler *et al.* (2007) and Strobl *et al.* (2008).

Tab. 3: Random Forest training on Durante's partition.

Durante		Reference		
		A	B	C
Prediction	A	99.4%	3.1%	0.0%
	B	0.6%	93.4%	9.5%
	C	0.0%	3.5%	90.4%

The Precision, Recall, and F1 Score values for each one of the 3 classes outlined by Durante are reported in Tab. 4.

Tab. 4: Precision, Recall, and F1 Score for Durante's partition.

Durante			
Class	**Precision**	**Recall**	**F1 Score**
A	0.97	0.99	0.98
B	0.90	0.93	0.92
C	0.96	0.9	0.93

For **Cella**'s configuration, we had the 420 texts corpus (N) split into 2 classes as below:

Partition	Time period	No. novels	No. text chunks
A	1500–1860	207	14,117
B	1861–2019	213	13,979

The classification results of **Random Forest** training on **Cella**'s configuration are reported in Tab. 5, as a percentage of the correct classification of the chunks:

Tab. 5: Random Forest training on Cella's partition.

Cella		Reference	
		A	B
Prediction	A	96.3%	1.9%
	B	3.7%	98.1%

The best accuracy in Cella's configuration was reached with 59 mtry, and its value is 0.97.

The Precision, Recall, and F1 Score values for the 2 classes outlined by Cella are reported in Tab. 6.

Tab. 6: Precision, Recall, and F1 Score for Cella's partition.

Cella			
Class	**Precision**	**Recall**	**F1 Score**
A	0.98	0.96	0.97
B	0.96	0.98	0.97

For **Antonelli**'s configuration we had the 420 texts corpus (N) split in 3 classes as below:

Partition	Time period	No. novels	No. text chunks
A	1500–1763	173	10,767
B	1764–1945	155	9,845
C	1946–2019	92	7,484

The classification results of **Random Forest** training on **Antonelli**'s configuration is reported in Tab. 7, as a percentage of the correct classification of the chunks:

Tab. 7: Random Forest training on Antonelli's partition.

Antonelli		Reference		
		A	**B**	**C**
Prediction	A	98.7%	3.2%	0.0%
	B	1.3%	94.5%	8.0%
	C	0.0%	2.3%	92.0%

The best accuracy in Antonelli's configuration was reached with 59 mtry, and its value is 0.95.

The Precision, Recall and F1 Score values for each one of the 3 classes outlined by Antonelli is reported in Tab. 8.

Tab. 8: Precision, Recall, and F1 Score for Antonelli's partition.

Antonelli			
Class	**Precision**	**Recall**	**F1 Score**
A	0.97	0.99	0.98
B	0.93	0.95	0.94
C	0.97	0.92	0.94

5 Discussion and conclusions

Looking at the overall accuracy results (Migliorini: 0.89, Durante: 0.94, Cella: 0.97, Antonelli: 0.95), we can say that Random Forest achieves high performances in each one of the configurations representing the four scholars' theories. Migliorini has the lowest accuracy value among the four configurations (0.89) and the highest number of classes (6), while the highest accuracy value has been recorded by Cella (0.97), which features the lowest number of classes (2).

Some of the most interesting things to discuss are related to misclassification cases. Looking at the above mentioned tables, we can say that, generally, when misclassification occurs, it's between adjacent periods. This can be seen quite clearly in Migliorini's classifications since we have more classes to observe. Looking at class E (representing the 20th Century), we see the 78.7% of the chunks were correctly classified, and the remaining 21.3% were split up between 19th (class D) and 21st (class F) centuries; no chunk belonging to the 20th century has been classified in the 16th, 17th or 18th century. This may be a good starting point in favour of saying that supervised machine learning methods (in this case, Random Forest) may be useful not just to classify texts for a specific periodization, but they may also be used to detect which particular features of each period better distinguish that period from the others. Moreover, while periods distant in time are rarely misclassified between themselves, this process may be used to find which language characteristics were abandoned at a certain point and which ones arose in a specific period. The only relevant case of misclassification involving periods not adjacent is class C (18th century) of Migliorini's classification.

References

Antonelli, Giuseppe. 2018. *Museo della lingua italiana*. Milano: Mondadori.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45 (1). 5–32.

Cella, Roberta. 2015. *Storia dell'italiano*. Bologna: Il Mulino.

Cortelazzo, Michele A., Georgios K. Mikros & Arjuna Tuzzi. 2018. Profiling Elena Ferrante: a Look Beyond Novels. In: Domenica Fioredistella Iezzi, Livia Celardo & Michelangelo Misuraca (eds.), *JADT 2018. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. 165–173. Roma: UniversItalia.

Culter D. Richard, Thomas C. Edwards Jr., Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson & Joshua J. Lawler. 2007. Random Forests for classification in Ecology. *Ecology* 88. 2783–2792.

De Mauro, Tullio (ed.). 2007. *Primo tesoro della lingua letteraria italiana del Novecento*. Torino: UTET.

Durante, Marcello. 1981. *Dal latino all'italiano moderno. Saggio di storia linguistica e culturale*. Bologna: Zanichelli.

Kho, Julia. 2018. "Why random forest is my favorite machine learning model." *Towards Data Science*. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

Koehrsen, Will. 2018. "An Implementation and Explanation of the Random Forest in Python." https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Marazzini, Claudio. 1994. *La lingua italiana. Profilo storico*. Bologna: Il Mulino.

Migliorini, Bruno. 1960. *Storia della lingua italiana*. Firenze: Sansoni.

Mikros, Georgios K. & Kostas Perifanos. 2013. Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In Eduard Hovy, Vita Markman, Craig Martell H. & David Uthus (eds.), Papers from the *2013 AAAI Spring Symposium "Analyzing Microtext"*, *25–27 March 2013*, *Stanford, California. Palo Alto*, *California*: AAAI Press. 17–23.

Mikros, Georgios K. & Kostas Perifanos. 2015. Gender Identification in Modern Greek Tweets. In: Arjuna Tuzzi, Martina Benešová & Ján Mačutek. (eds.), *Recent Contributions to Quantitative Linguistics* (Quantitative Linguistics 70). 75–88. Berlin: De Gruyter.

Savoy, Jacques. 2018. Elena Ferrante unmasked. In: Arjuna Tuzzi & Michele A. Cortelazzo, (Eds.), *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova*, *7 September 2017*. 129–139. Padova: Padova University Press.

Stoppelli Pasquale (ed.). 2010. *Biblioteca italiana Zanichelli. DVD-ROM per Windows per la ricerca in testi, biografie, trame e concordanze della letteratura italiana*. Bologna: Zanichelli.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9. 307.

Tuzzi, Arjuna, & Michele A. Cortelazzo. 2018. What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer. *Digital Scholarship in the Humanities* 33 (3). 685–702.

Sheila Embleton, Dorin Uritescu, Eric S. Wheeler

Too much of a good thing

Abstract: "More data" seems to be a desirable part of any research effort. However, from our experiences in studying the geography of dialects, we illustrate some of the pitfalls of having too much data. Our conclusion is that the idea of more data needs to be balanced with appropriate selection and insight.

Keywords: quantitative study of language variation, quantitative methods, big data, Mambila languages, Chinese languages, multidimensional scaling

1 Introduction

"More is better", it is said, and especially when it is more data. It seems reasonable that the more data we have about a subject, the more understanding we can gain. So-called "Big Data" is a popular concept just now and there is much truth to this claim. Quantitative studies have always relied on having enough data to make a meaningful model. From early counts of words (e.g. Zipf 1949; Herdan 1956) to modern studies over digitized collections of millions of tokens, having enough data seems essential.

It is no less the case in our work (*see Background*) on language variation vs. geographic distribution. It is essential to have enough geographic locations, and enough linguistic items at any location to be able to assess how much the linguistic variation correlates with geographic factors such as distance, travel time, and even the scale of distance (cf. distances across China vs. distances across a small set of communities in Africa).

However, in at least one of our studies (*see Mambila*), we have found that too much detail hides the patterns we expect to find. In particular, the data set from the Mambiloid languages found along the Nigeria-Cameroon border

Acknowledgement: We would like to most sincerely thank Dr. Bruce Connell (Glendon College, York University) and Dr. Robert Sanders (Komatsu University) for their active discussion, their helpful advice, and for the use of their data. Our wonderful colleague Dr. Uritescu died April 15, 2020. May he rest in peace.

Sheila Embleton, York University, e-mail: embleton@yorku.ca
Dorin Uritescu, York University
Eric S. Wheeler, York University, e-mail: eric.wheeler@sympatico.ca

https://doi.org/10.1515/9783110763560-004

region (Connell 2000, 2001, 2006) shows every location with some differences to every other location – some minor, others more basic, but in general all different. Hence, the multidimensional scaling (MDS) pictures that we use to show linguistic similarities and differences (see Wheeler 2005) are, at best, uninformative.

A more meaningful approach groups linguistic variants according to their status as cognates. The labour and expertise required to do this over the more than 800 data sets is prohibitive unless we can use an automated process – and the process we investigate is one of eliminating some differences (such as tone variation, vowel differences, etc.; we try various combinations) until we have a sequence of consonants that will match likely cognates. By hiding detail, we bring out patterns that are of interest. Less information does more.

Having identified this challenge in one case, we realize that there have been other cases where we got better results with less data (or more precisely, with smaller subsets of more homogeneous data). In our Romanian studies (*see Romanian*), grouping linguistic variation according to grammatical types has proven to give better correlations between language and geography. Using meta-data labels on data items greatly enhances our ability to search for and identify such groupings.

In our Chinese studies (*see Chinese*), we find the "whole" MDS picture overwhelming, but the ability to look at a few dialects at a time (using the interactive data viewer) allows us to see relationships more clearly, and more focussed on some point(s) of interest.

Our theoretical conclusion (*see Theory*), then, is that the principle of "More is better" needs to be counter-balanced by a principle that says "Less is better when less is more informative". And automatic processes, applied axiomatically, do not necessarily produce the best results; the researcher must (creatively) find or select the model that provides the most insightful results.

2 Background

In various projects, the authors have tried to make data on language variants available digitally, with convenient access to the data, selection of data by user defined criteria, and visualization of the data directly or in some processed format. In particular, we have looked at the use of multidimensional scaling (MDS) as a way of visualizing the relative differences among (read "the distance between") language variants. While the work has examined English (Embleton and Wheeler 1997a) and Finnish (Embleton and Wheeler 1997b), our most extensive project (RODA, Romanian Online Dialect Atlas) is on a conservative area of

north-west Romania, reflecting the eastern branch of the Romance languages (Embleton, Uritescu and Wheeler 2002, 2003, 2004, 2007a,b, 2008a,b,c, 2009).

One aspect of this work has been the comparison of geographic distance to linguistic distance, using MDS to visualize the distances (Embleton, Uritescu and Wheeler 2012, 2015, 2016). That has led to the view that geography (however defined, for geography may be a stand-in for something more abstract, such as communication connectedness) accounts for a major proportion of language variation, but clearly geography is not all of the story.

In carrying this work further, we have considered the effect of the scale of geography: what happens when the geography covers large distances (as in China) or relatively small distances (as in the Mambila languages). And so, we have collaborations with others (see below), which are works in progress; they lead us to challenges with how MDS visualizes their data sets.

3 Mambila

In the border region of Nigeria and Cameroon, there are a number of communities using different (but related) language variants. They have been studied by Connell (p.c.) who has provided us with a set of digitized data, from 41 locations in 17 regions, with 881 glosses based on the Swadesh list of basic vocabulary (Swadesh 1952, 1955) and other similar lists. The transcriptions are detailed, including notation of tone and differences in both consonants and vowels, even for words that might be phonemically "the same". Most items are different in some way (see Fig. 1).

mother	mma	mí	mɛ̄́	me	mī	mí	màmá	kwìjɛ̀	mē	mī́	ki	matʃini
father	da	tá	tɛ̀	tɛ̀	tər	tɛlɛ	tàtā	gbāmbīɛ̀	te	tər	tei	tətʃini

Fig. 1: A small sample of Mambila data, showing that some items are very similar but not identical. The last pair is probably a collocation, with the second element being the 3rd person singular possessive pronoun "his/her", again obscuring the potential similarities with other data items.

As a result, a simple geographic map showing a row of data (an "interpretive" map for one gloss), from all the locations, could have about as many variants as there are locations (see Fig. 2). The same holds for most other glosses, and for the MDS picture that takes all the glosses together.

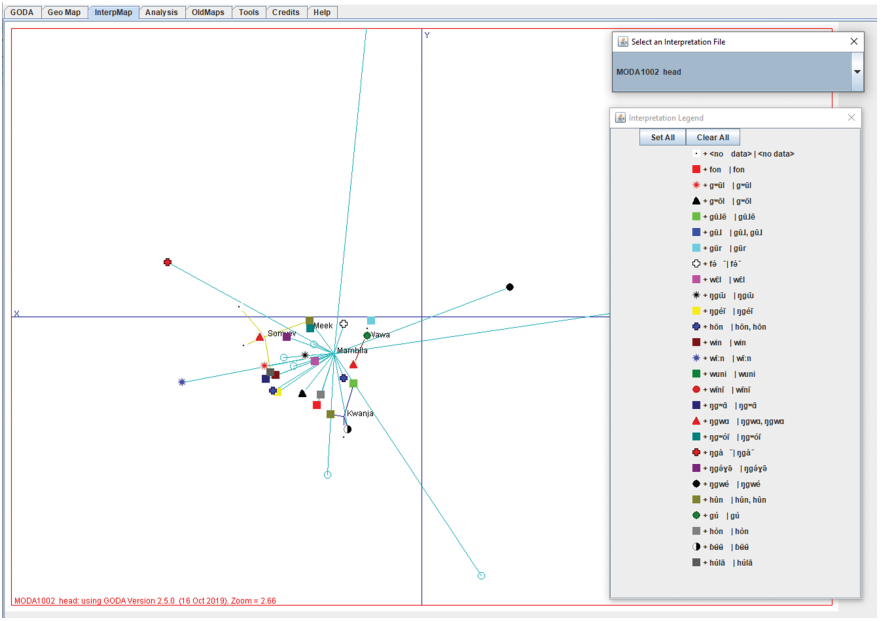


Fig. 2: An interpretive map and legend for the gloss "head", showing almost as many legend items as there are geographic locations.

Many of the legend items are (probable) cognates and could be combined as a single item. However, the work of doing that manually is prohibitive. We discuss a possible solution below.

4 MDS maps

When every point is different from every other point, the result in an MDS picture is an n-dimensional ball; when such a ball is then projected to 2-D or 3-D space, the result is not very informative. In the case of our 881 data rows (called "plain" data), many of the points projected to the same place, and the rest were spread out not very far along a line (with two outliers). In short, the MDS map was not helpful. See Fig. 3.

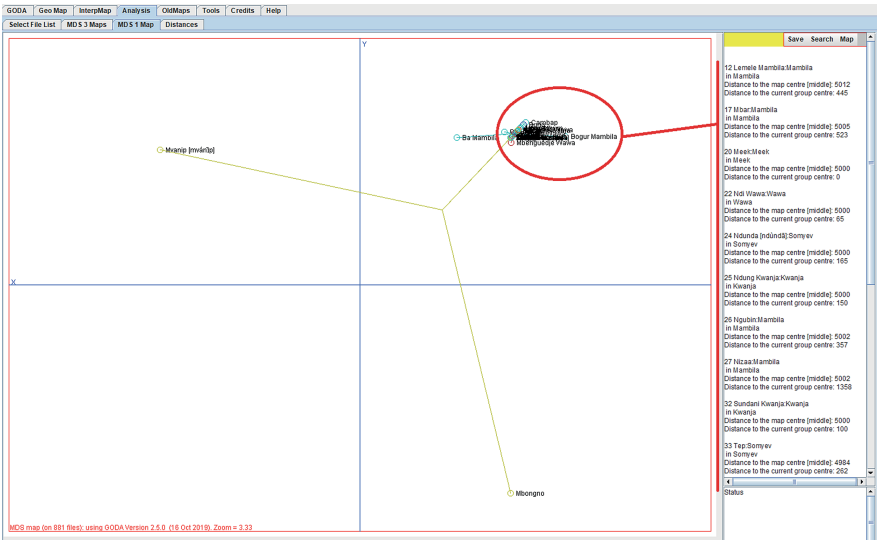


Fig. 3: MDS map on the plain data, showing many points at the same location (in the red circle, with two outliers). The side box shows part of the list of locations that all project to the same place on the MDS picture.

5 Automatic grouping of data

One solution to this problem is to combine data items that are similar. For example, in the "plain" case, ŋgʷā, ŋgwɑ, ŋgʷóī, ŋgá̠, ŋgə́ɣə̄, and ŋgwé (set 1) are each different, as different as ŋgʷā and wīnī, and even though there are obvious similarities to all the items in set 1. Data points could be combined to reflect cognates (which is probably the case here) but overall that requires the manual assessment by a knowledgeable person, over 881 maps, each with up to 30 or more legend items; for us, that is a prohibitive amount of work.

But, if the legend item were translated to a form that, say, ignored vowels, tones and other diacritics, and that grouped consonants by place of articulation or type, it would be possible to automatically group set 1 as "Ng" in contrast to "wN" (for "wīnī" and related forms). Fig. 4 is the legend for "head", so translated.

While the translation may not always group cognates, it does seem to produce a reasonable approximation, and in any case, it does group forms that look similar. With this translation, the MDS maps become more informative. Here is an MDS picture created over 881 glosses using the translated data (see Fig. 5). Many locations now show a linguistic position that is distinctive for that location.

<LEGEND>	MODA Tr1002		head
<L>	1	<no data>	| <no data>
<L>	2	bN	| fon
<L>	3	gr	| gʷūl, gʷōl, gūlē, gūl, gūl, gūr
<L>	4	b	| fə́¯, ɓūū¯
<L>	5	wr	| wɛ̄l
<L>	6	Ng	| ŋgɯī, ŋgéī, ŋgʷɑ̄, ŋgʷóī, ŋgá¯
<L>	7	hN	| hōn, hōn, hûn, hòn, hûn
<L>	8	wN	| win, wī:n, wuni, wīnī¯
<L>	9	Ngw	| ŋgwɑ, ŋgwɑ, ŋgwé
<L>	10	Ngg	| ŋgə́ɣə¯
<L>	11	g	| gú
<L>	12	hr	| húlā

Fig. 4: Legend for "head" using automatic translation of the forms listed after the vertical bar. Compare this with Fig. 2.

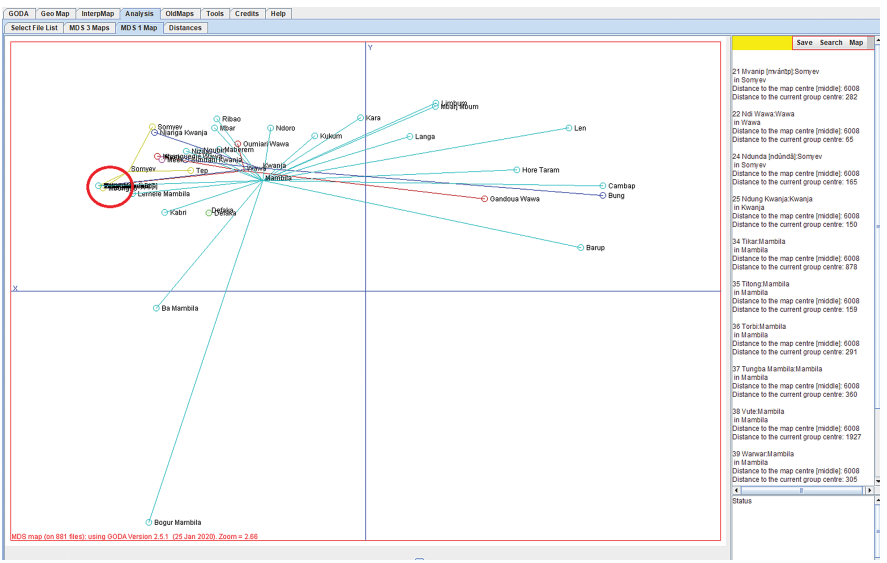


Fig. 5: MDS on 881 glosses with the data translated as in Fig. 4, showing many locations that are distinct (even though there are still several that are not separated, in the red circle).

Thus, the lesson we draw from this situation is that too much detail obscures our view of the differences in the Mambila language setting, and that omitting some of the detail helps us see patterns that are really there. Of course, we are not really destroying data, but rather adding classification information. Still, it is

a case of looking at less data at a time, and doing that for reasons that come from outside the data (namely, the recognition of the existence of cognates, or some technical equivalent).

6 Romanian

In retrospect, that is what we did with our Romanian data. When we looked at "all" the data we had on a large set, we generated an MDS picture that collected most of the sub-regions in one big area, and only clearly separated two of the areas (Fig. 6). The two southern areas (South-West and South-East) are distant from the main group, as expected; two northern areas (Oaș-North and Oaș-Core) which are known to be "distinctive" ended up in the main group, along with all the central sub-regions.

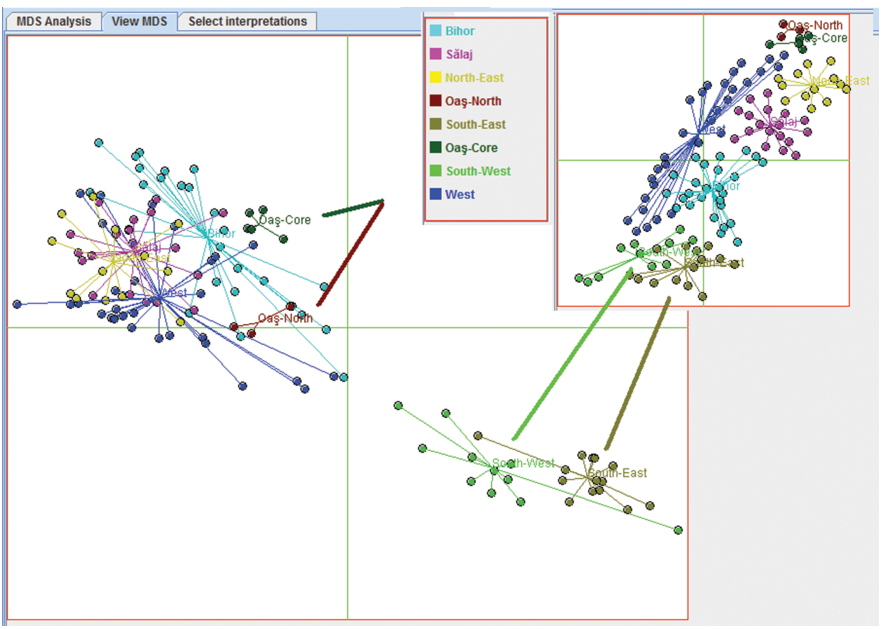


Fig. 6: MDS picture based on "all" the data (with the geographic map in the insert). The two southern sub-regions are separate, but the others are all in one area, including the two northern areas.

In contrast, when a knowledgeable analyst selected a subset of the data that was known to reflect dialect differences, the MDS picture was clearer (see Fig. 7). Not

only were the southern and northern areas clearly set apart, but the central areas were also spread out, although the similarities in these regions were also captured by their overlaps. Using less data, it was possible to see patterns that otherwise would have been hidden (perhaps because weighting all the data equally caused some items to counter-balance other items).

It is reasonable to argue that the selection of data should not be so arbitrary as to create a pattern. We divided our data by function: phonetic, lexical, and morphological. In doing so, we got MDS pictures like those in Fig. 7, although the different functions each gave a somewhat different picture of the dialect situation (Embleton, Uritescu and Wheeler 2016).

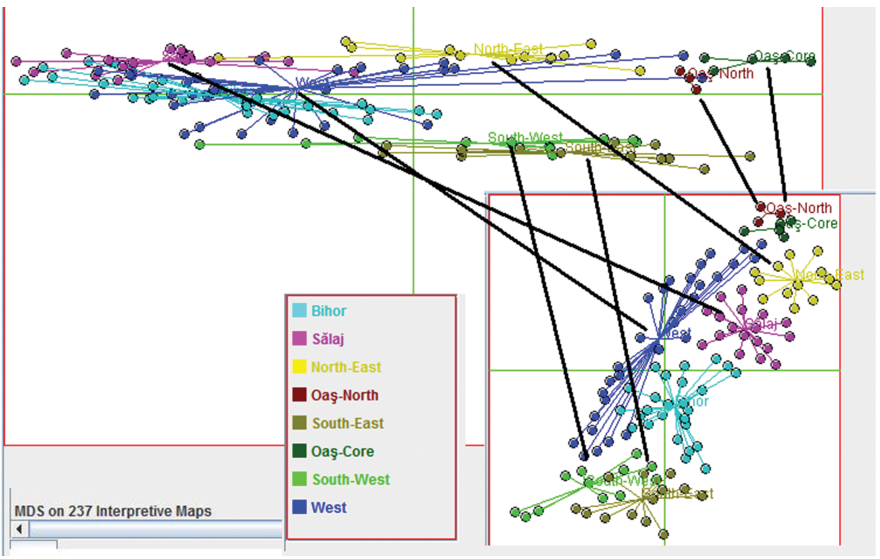


Fig. 7: MDS on a selection of the data (with the geographic map in the inset, and with lines connecting corresponding regions in the two maps). The result of the selection is a better distribution of the sub-regions.

Our Romanian data and the Romanian Online Data Atlas (RODA) has been enhanced to include "meta data" (RODA2; see Embleton, Uritescu and Wheeler 2015, Uritescu 2018). Using metadata, it is possible to select subsets of the full data in a way that is meaningful to the analyst (for example, one could select the forms that are used only by "old people" or that have a certain syntactic construction). Again, selective use of the data allows us to see more.

7 Chinese

Our experience with Chinese data (collaboration with Robert Sanders p.c.) highlights yet another aspect of being selective with data. If we look at all the data we have for Chinese (34 regions, multiple locations and several language varieties, some of which are orally different languages), we see a complex picture (Fig. 8):

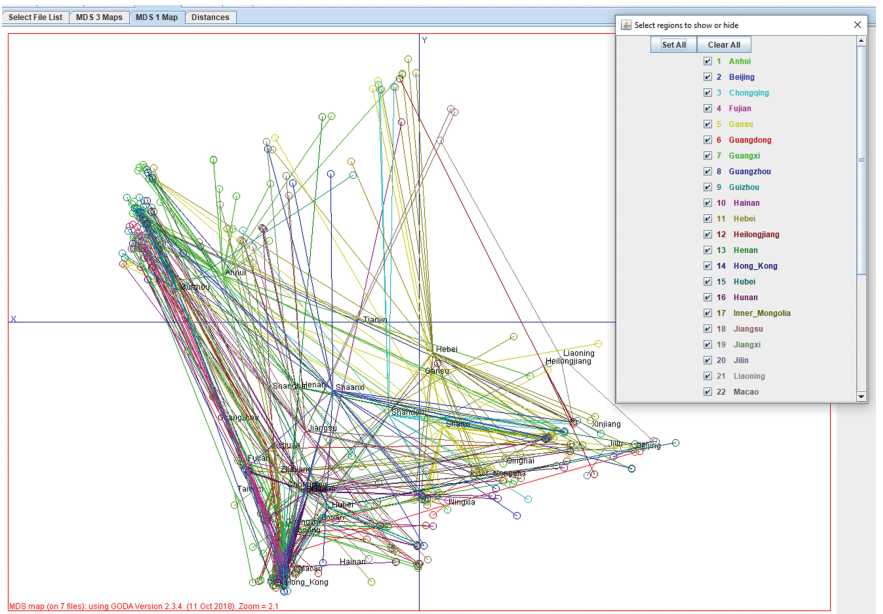


Fig. 8: MDS picture of a set of Chinese data, with 34 regions showing at once.

However, with our software, we have the possibility of seeing only some subsets at a time (see Fig. 9). And so, we can identify the relationships between two or three regions clearly, in a way that is less obvious when we try to see all the data at once. The lesson is obvious, but it is worth noting: less data (at one time) can be more informative.

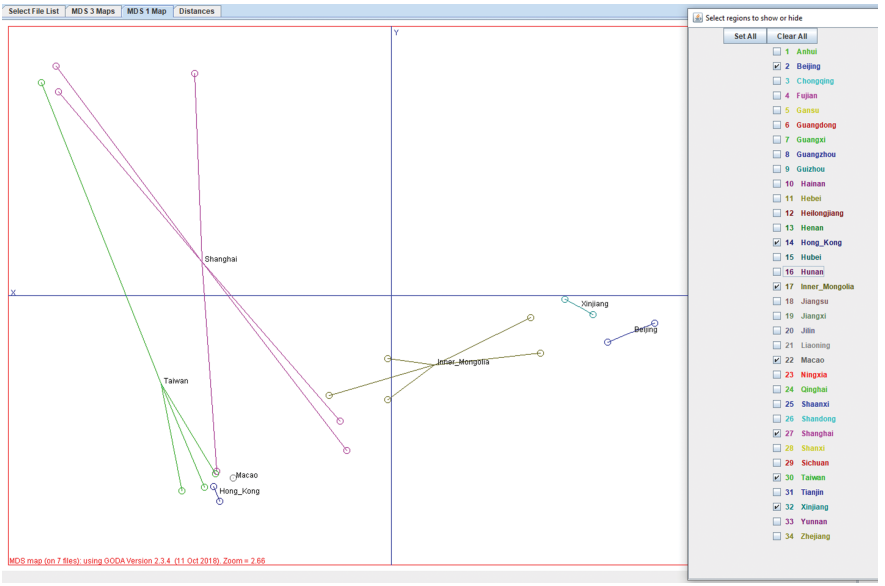


Fig. 9: MDS picture of some Chinese data showing only 7 regions (Taiwan, Hong Kong, Macau, Shanghai, Inner Mongolia, Beijing, and Xinjiang). The regions are each well separated linguistically, and sometimes diverse within a region, as with Shanghai and Taiwan. With less data showing than in Fig. 8, the relationships are more obvious.

8 Theory

Our theoretical conclusion is that the principle of "More is better" needs to be counter-balanced by a principle that says "Less is better when less is more informative". The choice of what data to select, and how to select it must rest with an informed (or inspired) analyst. One cannot automatically get meaningful interpretations of data by simply looking at lots of data. Doing that gives too much opportunity for obscuring what is a significant pattern, by averaging data so that the differences are hidden, or by flooding the user with so much pattern that any particular pattern is lost in the general.

Automatic processes, applied axiomatically, do not necessarily produce the best results; the researcher must (creatively) find or select the model that provides the most insightful results. To be sure that Big Data is effective, it pays to be selective.

References

Connell, Bruce. p.c. Data for Mambila languages. unpublished.

Connell, Bruce. 2000. The integrity of Mambiloid. In Herbert E. Wolf & Orin D. Gensler (eds.), *Proceedings of the 2nd World Congress of African Linguistics*, *Leipzig*, 197–213. Cologne: Rüdiger Köppe Verlag.

Connell, Bruce. 2001. An Introduction to the Mambiloid Languages. In Ngessimo M. Mutaka & Sammy B. Chumbow (eds.), *Research Mate in African linguistics. Focus on Cameroon: A Fieldworker's Tool for Deciphering the Stories Cameroonian Languages Have to Tell; In Honor of Professor Larry M. Hyman*, 79–92. Cologne: Rüdiger Köppe Verlag.

Connell, Bruce. 2006. Mambila. In Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics*, *Second Edition*, vol. 7, 473–475. Oxford: Elsevier.

Embleton, Sheila & Eric S. Wheeler. 1997a. Multidimensional Scaling and the SED Data, In Wolfgang Viereck & Heinrich Ramisch (eds.) *The Computer Developed Linguistic Atlas of England 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.

Embleton, Sheila & Eric S. Wheeler. 2000. Computerized dialect atlas of Finnish: Dealing with ambiguity. *Journal of Quantitative Linguistics* 7. 227–231.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2002, 2007a. "*Online Romanian 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. 2003. "*Romanian online dialect atlas*". International Colloquium of IQLA – International Quantitative Linguistics Association, University of Georgia, Athens, Georgia, May 28, 2003.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2004. Romanian online dialect atlas. 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.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2007a. Data capture and presentation in the Romanian online dialect atlas. In Wladyslaw Cichocki, Wendy Burnett & Louise Beaulieu (eds.), Papers from Methods XII. 12th International Conference on Methods in Dialectology. *Linguistica Atlantica* 27–28. 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* (Quantitative Linguistics 62), 87–96. Berlin/New York: Mouton de Gruyter.

Embleton, Sheila, Dorin Uritescu and Eric S. Wheeler. 2008a. *Digitalized Dialect Studies: North-Western Romanian*. Bucharest: Romanian Academy Press.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2008b. Defining User Access to the Romanian Online Dialect Atlas. *Dialectologia et Geolinguistica* 16. 27–33.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2008c. Identifying dialect regions: Specific features vs. overall measures using the Romanian online dialect atlas and multidimensional scaling. Methods XIII Conference, August 2008, Leeds, UK. In Barry Heselwood & Clive Upton (eds.) 2009. *Proceedings of Methods XIII. Papers from the Thirteenth International Conference on Methods in Dialectology*, 2008, 79–90. Frankfurt am Main: Peter Lang.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2009. Data management and linguistic analysis: MDS applied to RODA. Presented to the Trier Symposium on Quantitative Linguistics, Trier, Germany, December 2007. In Reinhard Köhler (ed.). 2009. *Studies in Quantitative Linguistics* 5. 10–16. Lüdenscheid: RAM-Verlag.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2012. Effective comparisons of geographic and linguistic distances. In Gabriel Altmann, Peter Grzybek, Sven Naumann & Relja Vulanović (eds.). *Synergetic Linguistics. Text and Language as Dynamic Systems*. 225–232. Vienna: Praesens Verlag.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2015. Analyzing dialect variation with metadata: The new format of the Romanian online dialect atlas – Crisana, *VIIIth Congress of the International Society for Dialectology and Geolinguistics, 14–18 September, Eastern Mediterranean University, Famagusta, North Cyprus*.

Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2016. An expanded quantitative study of linguistic vs geographic distance using Romanian dialect data. In Lu Wang, Reinhard Köhler & Arjuna Tuzzi (eds.), 2018. *Structure, Function and Process in Texts*, *Proceedings of Qualico 2016*, 25–33. Lüdenscheid: RAM-Verlag.

Herdan, Gustav. 1956. *Language as Choice and Chance*. Groningen, Netherlands: P. Noordhoff Ltd.

Sanders, Robert (p.c.) Data collected by him, his colleagues and associates.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452–463.

Swadesh, Morris. 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*. 21. 121–137.

Uritescu, Dorin. 2018. Pe marginea noii versiuni informatizate a Atlasului lingvistic al Crișanei / In connection with the new online version of the Linguistic Atlas of Crișana /. In Veronica Ana Vlasin, Dumitru Loșonți, Nicolae Mocanu (eds.), *2018. Lucrările celui de-al XVII-lea Simpozion Internațional de Dialectologie (Cluj-Napoca, 8–9 septembrie 2016)*, 327–342. Cluj-Napoca: Argonaut şi Scriptor.

Uritescu, Dorin, *et al*. RODA 2 = Dorin Uritescu (coord.), Sheila Embleton & Eric S. Wheeler, *Romanian Online Dialect Atlas*, second edition. http://uritescu.ca/RodaAnalysis/home/index (temporary link).

Wheeler, Eric S. 2005. Multidimensional scaling for linguistics. In Reinhard Koehler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*. 548–553. Berlin: Walter de Gruyter.

Zipf, G. K. 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. facsimile 1965. New York: Hafner.

Antoni Hernández-Fernández, Juan María Garrido,
Bartolo Luque, Iván González Torre

Linguistic laws in Catalan

Abstract: In this work, we explore and review linguistic laws, in the case of Catalan, going from *prelinguistic* to higher linguistic levels and addressing both speech and writing. We show evidence supporting the theory that linguistic laws are universal patterns in human language that are more robust in the oral corpus than in writing. This reinforces the "physical hypothesis," which argues that linguistic laws could have a physiological and biophysical origin, and they are reflected in written texts as a consequence of speech symbolization. However, future work is necessary to increase empirical evidence by deeply analyzing other language corpora, propose new cognitive and physical models that clarify the mathematical formulation of some statistical patterns, find more evidence to explain the relationship between prelinguistic and higher linguistic levels, and understand the results reported up to date from a global interdisciplinary perspective of language theory.

Keywords: Catalan, linguistic laws, physical hypothesis, zipf's law, brevity law, menzerath-altmann's law, theory of language

1 Introduction

Linguistic laws are statistical patterns observed in linguistic elements that can be mathematically formulated and quantitatively estimated using real data (Köhler 2005). The discussion of their origin is still open, although it has lately

Acknowledgment: I.G.T. and A.H.-F. were supported by the project PRO2020-S03 (RCO03080 Lingüística Quantitativa) and PRO2021-S03HERNANDEZ by Institut d'Estudis Catalans. A.H-F. was also supported by the grant TIN2017-89244-R (MACDA) (Ministerio de Economía, Industria y Competitividad, Gobierno de España).

Antoni Hernández-Fernández, Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia, Institut d'Estudis Catalans, e-mail: antonio.hernandez@upc.edu
Juan María Garrido, Universidad Nacional de Educación a Distancia, e-mail: jmgarrido@flog.uned.es
Bartolo Luque, Universidad Politécnica de Madrid, e-mail: bartolome.luque@upm.es
Iván González Torre, Universidad del País Vasco y Universidad Politécnica de Madrid, e-mail: ivan.gonzalez.torre@upm.es

https://doi.org/10.1515/9783110763560-005

been argued that they could come from phenomena related to biophysical dynamics (Torre et al. 2019, Hernández-Fernández et al. 2019), due to the characteristic of self-organized critical systems (SOC) (Bak 1996) observed in both speech (Luque et al. 2015) and literary texts (Gromov & Migrina 2017). In this paper we review traditional linguistic laws (Altmann & Gerlach 2016), as well as some other more recently proposed ones (Torre et al. 2019) – summarized in Tab. 1, in the case of Catalan, analyzing statistical patterns both in speech and transcriptions that range from *prelinguistic* levels to the highest linguistic units.

Catalan is a Romance language spoken in the Western Mediterranean by more than ten million people, with other small communities of speakers spread around the world. Catalan is the official language in Andorra and co-official in Spain, specifically in Catalonia, the Valencian community – where the local linguistic variety is called *Valencian* – and the Balearic Islands. Catalan is a language with a medium phonological richness that shows the average values regarding words' entropy (with an entropy rate of 5.84, with languages mean around 5.97, see Bentz et al. (2017)) and morphological complexity (also in terms of unigram word complexity, being ranked in the 202nd position among 520 languages, see Bentz et al. (2016)).

The phoneme inventory of Catalan is slightly different depending on the dialect considered. In the case of the Central variety, the one spoken in Barcelona, which is the one considered here, it includes eight vowels and twenty consonants (Recasens 1993). One outstanding phonetic feature of Catalan is the number of processes that can modify the acoustic realisation of their elements, e.g., vowels /a/ and /e/ are pronounced as a schwa (/ə/) when they appear in unstressed position in the Central variety – except in some particular cases as in compounds nouns; and vowels in contact are often merged into a single long vowel. There are processes that also affect consonants, e.g., some of them (such as /r/) may disappear in word-final position while others may be affected when appearing in the end-of-syllable position through the effect of voicing (if the consonant is not voiced) and devoicing (if the consonant is voiced) (Recasens 1993). These kinds of acoustic processes take place in many languages but they are overlooked when linguistic laws are explored in written corpus instead of oral units.

Research on linguistic laws in speech faces the challenge of defining accurate units of study. Speech units are not isolated elements, as they constitute part of a wider phonological system and obey principles related to contextual dependencies shaped by the dynamics of language acquisition, comprehension, production and processing (Schwartz et al. 2015). However, the broad definition of speech units is still a hot topic in linguistics, engineering, physics, and phonetics (Toman et al. 2006, Cárdenas et al. 2016), particularly in reference to the

phenomena of coarticulation (Menzerath & Lacerda 1933). As units of study, we here consider phonemes, words, and breath groups – the set of utterances articulated in the course of a single exhalation, abbreviated as BG (Tsao & Weismer 1997) –, although we are aware that they are mere idealizations of their phonic realities.

In this paper, we review quantitative linguistic laws in Catalan at several linguistic levels in speech and transcriptions and also explore them at *prephonemic* ranges. Thresholds methodology does not require any previous speech segmentation (Luque et al. 2015) – it, therefore, avoids the controversy of coarticulation – and has been successfully applied for reporting linguistic laws in prelinguistic ranges where there is no cognitive influence (Torre et al. 2017, Torre et al. 2020). We address the links between orality and symbolic expression of language (writing), and between the physical manifestation of language complexity – represented by linguistic law – at *prephonemic* ranges and whether they could be emerging and shaping the language in higher scales. Furthermore, our fundamental objective is to establish bridges with the community of Quantitative Linguistics to achieve a broad cross-disciplinary scientific approach to the study of languages, taking here the Catalan language as a model and a case study.

2 Resources and methods

2.1 Linguistic level: Phonemes, words, and BG

We use the Glissando corpus to review linguistic laws in orality and writing in Catalan. Glissando includes more than 12 hours of speech, totaling over 30,000 phonemes, 80,000 words, and 20,000 BG, recorded under optimal acoustic conditions, orthographically transcribed, phonetically aligned, and annotated with prosodic information (Garrido et al. 2013). The results of exploring this aligned database were extracted from Hernández-Fernández et al. (2019) and summarized in Fig. 1, illustrating six linguistic laws – Zipf's law, Herdan Heaps' law, Size-rank law, Brevity law, and Menzerath-Altmann's law – studied in physical units and their symbolic expressions. We also considered the law of lognormality (Torre et al. 2019), which was not possible to completely analyze in Glissando Corpus due to the insufficient accuracy in the alignment. However, we assume it to be a universal pattern that we had not been able to report in that case due to technical issues (Hernández-Fernández et al. 2019), so we consider that law for the discussion.

Tab. 1: Main linguistic laws (Torre et al. 2019, Hernández-Fernández et al. 2019, Torre et al. 2017).

Linguistic Law	Formula	Details
Zipf	$f(r) \sim r^{-\alpha}$	f: frequency r: rank α: parameter
Herdan-Heaps	$V \sim L^{\beta}$	L: text size or time elapsed V: vocabulary β: parameter
Guttenberg-Richter	$P(E) = E^{-\phi} F(E/E_{\varepsilon})$	E/E_{ε}: dimensionless energy ϕ: parameter F: scaling function
Brevity	$f \sim \exp(-\lambda \ell),\ \lambda > 0$	f: frequency ℓ: size λ: parameter
Size-rank	$\ell \sim \theta \log(r),\ \theta = \frac{\alpha}{\lambda}$	ℓ: size r: rank θ: parameter
Menzerath-Altmann	$y(n) = an^{b}\exp(-cn)$	n: size of the whole y: size of the parts a, b, c: parameters
Lognormality law	$P(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$	t: time duration σ, μ: parameters

In this work, Zipf's law (Fig. 1a) was studied only in written transcripts both for words and phonemes, where in the latter case, it takes the form of Yule's law. Meanwhile, Herdan-Heaps' law (Fig. 1b), Brevity law for the case of words (Fig. 1d), and Menzerath-Altmann's law in the BG-words version (Fig. 1f) were studied in symbolic and physical units. Size-rank law (Fig. 1c) can also be studied in both versions but in Hernández-Fernández et al. (2019) was only explored in physical units. Finally Brevity law for phonemes and Menzerath-Altmann's law in the BG-words version (Fig. 1f) can only be reported in physical quantities.

2.2 Prelinguistic ranges

The study in prelinguistic ranges uses Catalan audio signal from KALAKA-2 TV broadcast speech database (Rodríguez-Fuentes et al. 2012), originally designed

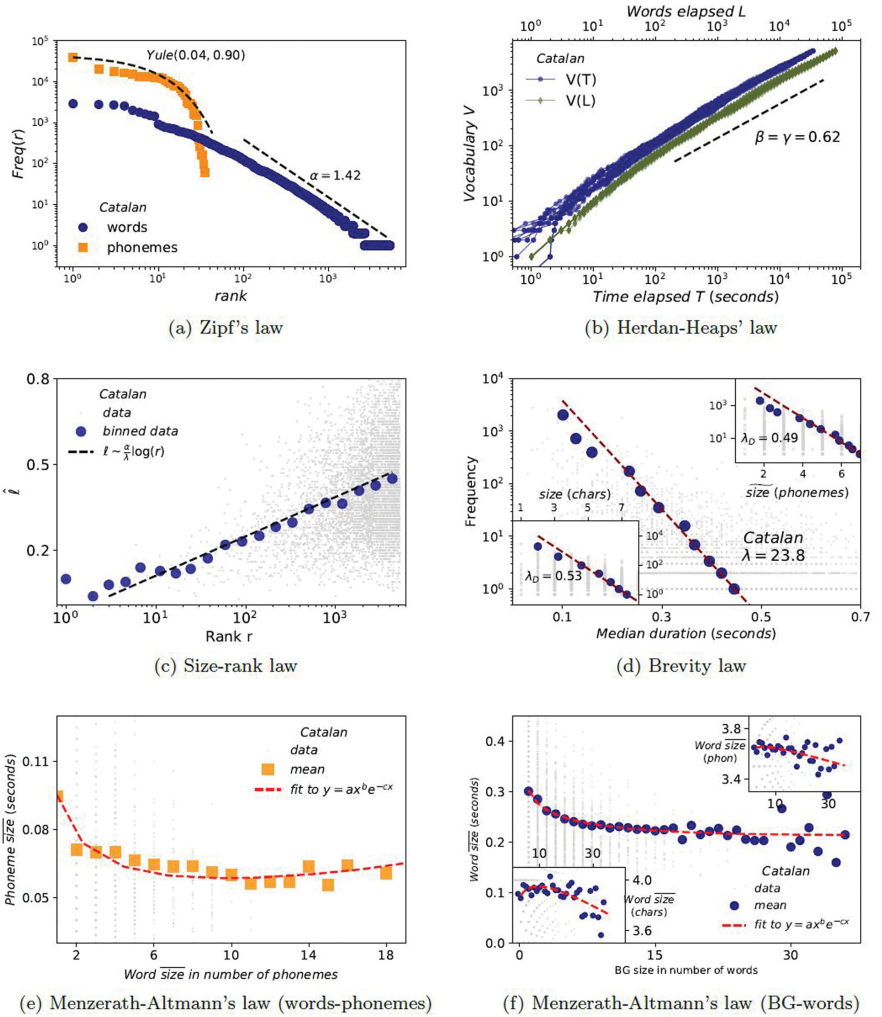


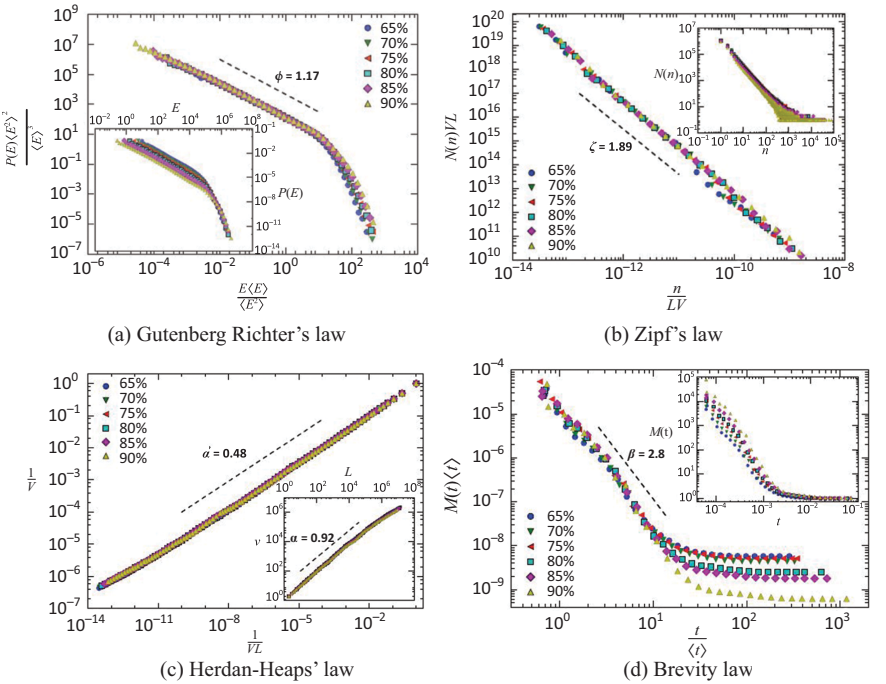
Fig. 1: Main linguistic laws in higher linguist levels: speech and texts (Hernández-Fernández et al. 2019).

for language recognition evaluation purposes, containing 4 hours of both planned and spontaneous speech throughout diverse environmental conditions. The audios are processed making use of *threshold methodology*, which maps any signal or time series into a sequence of voice events separated by silence. This methodology does not require the knowledge of the underneath corpus (Luque et al. 2015), and segments the signal into events that do not necessarily correspond to any linguistic unit. Those events are then used to explore four linguistic laws –

Guttenberg-Richter's law, Zipf's law, Herdan-Heaps' law, and Brevity law – at levels beyond the cognitive level. We here recover the results from Torre et al. (2017) and summarized them in Fig. 2, to discuss later what relationship there is between statistical patterns observed in microscales and dynamics taking place in higher cognitive levels that shape the expression of language. The main panel of Fig. 2a shows a log-log plot of the collapsed and threshold-independent energy release distribution, also called the Guttenberg-Richter law. The main panel of Fig. 2b represents, in a logarithmic scale and after appropriate compressing, the Zipf's law in the probability-frequency formulation. In the main panels of Figs. 2c and 2d, threshold-independent and log-log plots of Herdan-Heaps' and Brevity law are also shown. All the slopes of the free scale regimes are depicted inside each panel.

3 Results and discussion

Zipf's classic works revealed the existence of optimization pressures that affect human language particularly pointing out two of them: the force of *unification*, as well as the force of *diversification*. Those opposing forces respectively address the interests of the speaker and the hearer and, in effective communication, they reach a balance (Zipf 1935, Zipf 1949) which has been shown to be close to a phase transition between maximizing the information transfer and saving the cost of signal use (Ferrer-i-Cancho 2005b). Since then, quantitative methods for studying the language have also been applied to the explorations of animal acoustic communication (Kershenbaum et al. 2016, Gustison et al. 2016), psycholinguistics theories, and cognitive science, among others. However, a rigorous mathematical approach to linguistic laws is not always fulfilled (Fedzechkina & Jaeger 2020). Here we use the Catalan case study to address what we consider to be the two most important questions still open: (1) even though the empirical evidence of linguistic laws has been accumulated during decades, orality has been many times overlooked and the differences, similarities, and relationships between linguistic laws in speech and writing are still unknown; and (2), do linguistic laws really have a cognitive-physical origin, or are they being shaped by deeper biophysical or physiological processes?



(a) Gutenberg Richter's law (b) Zipf's law

(c) Herdan-Heaps' law (d) Brevity law

Fig. 2: Main linguistic laws in prelinguistic levels. These language laws were recovered by shredding the voice in 'speech events' several orders of magnitude below the phonological level (Torre et al. 2017).

3.1 Linguistic laws in orality and written Catalan

Despite the fact that orality and non-written language are the forerunners of writing, many linguistic and cognitive science theories are based on a foundation of symbolic representation typical of writing. We here confront some paradigms of symbolic language, discussing the results obtained when studying linguistic laws with the physical magnitudes of speech and with the more traditional symbolic aspects of writing.

3.1.1 Symbolic laws

Results for Zipf's law in ranks formulation, $f(r) \sim r^{-\alpha}$, showed in Catalan a scaling exponent of $\alpha \sim 1.42$ in agreement with those previously reported, pointing out the

robustness of this classical linguistic law even in speech transcription. The phoneme frequency distribution was fitted to a Yule distribution, as reported in the literature, but more experimental evidence and a theoretical framework are needed to justify this claim. Zipf's law is intrinsically formulated in symbolic units as it measures the frequency of linguistic elements given their written representation, so homography prevails over homophony in semantic approaches (Hernández-Fernández et al. 2016). The speech variability suggests whether a formulation closer to orality would be possible; in this way, if a word is expressed with different acoustic elements, it could be more accurate to count the frequencies of each acoustic element instead of the idealized symbolic one. This, in turn, would open other choices such as considering units of study based on physical quantities instead of written idealizations. Among other questions, it would be interesting to address whether the variability of Zipf's law (Ferrer-i-Cancho 2005b, Baixeries et al. 2013) could be reported using non-symbolic oral (physical) units.

3.1.2 Linguistic laws: Speech and text

In previous works, we showed how linguistic laws have a better fit when considering physical rather than symbolic quantities (Torre et al. 2019, Hernández-Fernández et al. 2019). This is statistically appreciated in the better performance of goodness of fit for the physical version of Menzerath-Altmann's law versus the symbolic one (summarized in Tab. 2). BGs are physical units defined as the set of utterances articulated in the course of a single exhalation, which may or may not coincide with grammatically well-constructed sentences. We encourage the use of this methodology and units as they consider the physiological reality of speech rather than normative grammar

Tab. 2: Sample caption Menzerath-Altmann word-BG adjustment parameters for Catalan (Hernández-Fernández et al. 2019).

	Menzerath-Altmann fit parameters			
	a	***b***	***c***	R^2
BG-Words (time duration)	0.33	−0.15	−0.003	0.47
BG-Words (characters)	3.7	0.04	0.004	0.23

Nevertheless, for the case of Herdan-Heaps' law, Torre et al. (2019) showed that symbolic and physical versions of the law have the same scaling exponents, finding the first strong mathematical relationship between linguistic laws in orality and written corpora. This recent result was also reported when extending the research to the case of Catalan corpus (Hernández-Fernández et al. 2019). We aim to find similar relationships for other linguistic laws that have been explored in speech and in transcriptions. This would help us to better understand the dynamics that are conservative between both forms of language expression and to detect which ones are missed due to to the simplification that takes place when going from speech to the transcription or symbolization.

Zipf's law of abbreviation, or the universal tendency of more frequent words to be shorter (Zipf 1935, Ferrer-i-Cancho 2016), has been verified in a wide range of human languages (Kanwal et al. 2017). We mathematically formulated this law with optimal compression principles of information theory that takes the form of an exponential relationship between the frequency of appearance of words and their length. This formulation adjusts very satisfactorily to real data of Catalan – Fig. 1d – and other languages (Torre et al. 2019, Hernández-Fernández et al. 2019), particularly for physical units of language, which reinforces the hypothesis that advocates considering (acoustic) distinctiveness as a key factor in the listener's processing (Meylan & Griffiths 2017). It should even go further because the distinctiveness of linguistic elements of speech, in addition to duration, includes other physical magnitudes of speech, such as energy or acoustic frequency. The discussion of the precise mathematical formulation of brevity law is still open and recent researchers have reported, and tested out with English written corpora, a mathematical model that justifies a power-law-decay for the brevity-frequency phenomenon (Corral & Serra 2020). However, there are slight differences in the underneath methodology that could explain the different models (exponential versus power-law) proposed. This recent approach also indicates that the origin of Zipf's law could be fully linguistic as it depends crucially on the length of the words and to be connected with Brevity law (Corral & Serra 2020), as already pointed out when studying size-rank law (Torre et al. 2019). Further research and more empirical studies in several languages are still needed to clarify the proper mathematical formulation for brevity law or if the type of unit considered and the modality –orality or writing– is of paramount importance to explain the differences.

In previous work, the so-called size-rank law (Fig. 1c) was only shown in physical quantities and the exponent of its mathematical formulation was for the first time connected with Zipf's law and Brevity law (Torre et al. 2019, Hernández-Fernández et al. 2019). It would be entirely plausible to investigate this law in symbolic terms and more effort is needed in this direction to compare speech and text. Regarding the word-length distribution –or more generally,

the type-length distribution where *type* refers to any linguistic unit – Torre et al. (2019) reported a lognormality law that holds for all linguistic levels when measuring the length in physical units. This is consistent with some previous analyses reported for speech but different from a recent work, where the researchers propose a gamma distribution for characterizing the word frequency-length distribution in written corpus (Corral & Serra 2020). Indeed, we found that phonemes and BG are best fitted by lognormal distribution in speech, whereas for the case of words, beta and gamma distributions have akin statistical likelihood (Torre et al. 2019). Anyway, the differences when fitting the best probability distribution may be explained by the great variability of speech – where each linguistic realization potentially has a different acoustic duration – in contrast with the fixed length of written words.

3.2 Linguistic laws in prelinguistic scales

The total number of alveoli in human lunges is estimated to amount between 274 and 790 million (Ochs 2004), so speech is based on a myriad micro-avalanches of expelled air that are later modulated to produce sounds. The energy E released during voice events in speech is a direct measure of the vocal fold response function under air pressure perturbations and this energy shows a scale-free distribution. This result has been interpreted as a Guttenberg-Richter law for the voice, which could indicate that air exhalations share dynamics typical of self-organized critical systems (Bak 1996, Torre et al. 2020). The exponent associated with scale-free energy release distribution in Catalan ($\varphi = 1.17$) is compatible with results obtained in 16 languages of different linguistic families reinforcing the hypothesis of a physiological phenomenon behind it (Torre et al. 2017). Note that the duration of these energy release events ranges from less than 1 ms to the order of seconds, therefore, ranging from time duration shorter than phonemes to longer than words. The study of energy in speech production and its relationship with the underlying linguistic laws remains a crucial problem to be explored in quantitative linguistics. The study of Zipf's law –Fig. 2b – and Herdan-Heaps' law –Fig. 2c– in prelinguistic level also holds the same relationship between their scaling exponents as broadly reported on higher linguistic levels –Figs. 1a and 1b. Remarkably, the power law recently reported for brevity law in written corpora and discussed above (Corral & Serra 2020), is analogous to the fit found in prelinguistic levels (Torre et al. 2017), in contrast with the exponential law listed in Tab. 1 that was recovered for Catalan, Spanish and English in speech (Hernández-Fernández et al. 2019). Analyzing the continuity of Brevity law across scales and linguistic levels is another research to be done. Hence, the prelinguistic level would be pointing out

that there is continuity between a strictly biophysical (non-linguistic) level and language, explaining therefore the origin of linguistic laws. That is, if linguistic laws are scaling laws, then we should also recover them at their origin in the production of speech, as is the case. So, analogously to what happens in language, the most frequent speech events in this prelinguistic level potentially require a lower amount of energy to be produced, which is in harmony with the law of brevity and the compression principle (Ferrer-i-Cancho et al. 2013), as well as with the principle of least effort (Zipf 1949).

4 Conclusions

Quantitative Linguistics analyzes structures of language, its properties and interrelations with other communication systems to discover the laws underlying the human language phenomena, through the use of quantitative techniques (Köhler 2005). Many universal principles have been defined and explored previously in our field, e.g., the compression principle or the form-meaning mappings optimization (Ferrer-i-Cancho 2016, Kanwal et al. 2017); however, most approaches seem to forget that language is in its origin an acoustic communication system rather than symbolic.

First of all, it is necessary to accumulate empirical evidence straight from speech instead of writing to avoid the biases produced by the segmentation of the study units and the influence of writing technology. Besides, linguistic laws in Catalan and other languages are more robust in orality than in writing, enlighting us about the importance of recovering orality as a source of key information on quantitative linguistics, as the pioneering works of Menzerath and Lacerda (1933) or Zipf (1935, 1949) already did. In any case, the mathematical relationship between orality and writing must be analyzed to have a big picture of the dynamics of language.

Secondly, the presence of linguistic laws not only in human language but also in other communication systems (Heesen et al. 2019) forces us to explore their possible biophysical origin. The traits shared with other primates would justify the appearance of universal principles of communication, such as the law of brevity, which would enter into competition with other principles of communication, such as the priority that the message reaches the receiver (Ferrer-i-Cancho & Hernández-Fernández 2013). The physical hypothesis suggests that physiological constrictions of the brain and human vocal apparatus are shaping the emergence of linguistic laws and would justify *i)* its appearance in prelinguistic levels and in other species that do not have writing and *ii)* that linguistic laws

are better satisfied when studying speech than written texts. If this is true, as has been seen for the linguistic laws of Catalan at the prelinguistic (Torre et al. 2017) and at the linguistic levels (Hernández-Fernández et al. 2019), there are then two possibilities: a) that the linguistic laws have the same mathematical form in the different scales of study, or b) that their mathematical formulation changes depending on the scale. In this sense, a) Zipf's and Herdan-Heaps' laws show continuity in their empirical evidence in Catalan, and b) a power law fits the law of brevity at the prelinguistic level, but an exponential law was founded on the linguistic level, although other authors claim to have found this potential law for the law of brevity in written texts (Corral & Serra 2020), thus adding evidence in favor of a).

Finally, based on the evidence of multiplicative processes present at many anatomical levels in the brain (Buzsáki & Mizuseki 2014) and the evidence of SOC in neuronal activity (de Arcangelis et al. 2006), one could speculate on the existence of *another continuity* between the mechanisms of neuronal control of speech and oral production. Hence, there could be a leap of self-organized criticality phenomena emerging through different scales, including neuronal level, alveolar control, and language production, that should be the subject of future interdisciplinary research.

References

Altmann, Eduardo G. & Martin Gerlach. 2016. Statistical laws in linguistics. In Mirko Degli Esposti, Eduardo G. Altmann & François Pachet (eds.), *Creativity and Universality in Language*, 7–26. (Lecture Notes in Morphogenesis). Cham: Springer.

de Arcangelis, Lucilla, Carla Perrone-Capano & Hans J. Herrmann. 2006. Self-organized criticality model for brain plasticity. *Pysical Review Letters* 9(2). 028107.

Baixeries, Jaume, Brita Elvevåg & Ramon Ferrer-i-Cancho. 2013. The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE* 8 (3). e53227.

Bak, Per. 1996. *How Nature Works: The Science of Self-Organized Criticality*. New York: Springer.

Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw & Ramon Ferrer-i-Cancho. 2017. The entropy of words – Learnability and expressivity across more than 1000 Languages. *Entropy* 19 (6). https://www.mdpi.com/1099-4300/19/6/275.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig & Tanja Samardžić. 2016. A comparison between morphological complexity neasures: Typological data vs. language corpora. In The COLING 2016 Organizing Committee *(ed.), Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), Osaka, Japan*, 142–153. https://www.aclweb.org/anthology/W16-4117

Buzsáki, György & Mizuseki Kenji. 2014. The log-dynamic brain: How skewed distributions affect network operations. *Nature Reviews Neuroscience* 15 (4). 264–278.

Ferrer-i-Cancho, Ramon. 2005a. Zipf's law from a communicative phase transition. *European Physical Journal B* 47(3). 449–457.
Ferrer-i-Cancho, Ramon. 2005b. The variation of Zipf's law in human language. *European Physical Journal B* 44(2). 249–257.
Ferrer-i-Cancho, Ramon. 2016. Compression and the origins of Zipf's law for word frequencies. *Complexity* 21. 409–411.
Ferrer-i-Cancho, Ramon & Antoni Hernández-Fernández. 2013. The failure of the law of brevity in two new world primates. Statistical caveats. *Glottotheory International Journal of Theoretical Linguistics* 4(1). 45–55.
Ferrer-i-Cancho, Ramon, Antoni Hernández-Fernández, David Lusseau, Govindasamy Agoramoorthy, Minna J. Hsu & Stuart Semple. 2013. Compression as a universal principle of animal behavior. *Cognitive Science* 37 (8). 1565–1578.
Cárdenas, Juan Pablo, Iván González, Gerardo Vidal & Miguel Angel Fuentes. 2016. Does network complexity help organize Babel's library? *Physica A: Statistical Mechanics and its Applications* 447. 188–198.
Corral, Álvaro & Isabel Serra. 2020. The brevity law as a scaling law, and a possible origin of Zipf's law for word frequencies. *Entropy* 22 (2). 224. https://www.mdpi.com/1099-4300/22/2/224.
Fedzechkina, Masha & T. Florian Jaeger. 2020. Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition* 196. 104115. http://www.sciencedirect.com/science/article/pii/S0010027719302896
Garrido, Juan Maria, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme de la Mota, Cesar Gonzalez, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, Yesika Laplaza, Francisco Vizcaíno, Eva Estebas, Mercedes Cabrera & Antonio Bonafonte. 2013. Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation* 47(4). 945–971.
Gromov, Vasilii A. & Anastasia M. Migrina. 2017. A language as a self-organized critical system. *Complexity* 2017. 9212538.
Gustison, Morgan L., Stuart Semple, Ramon Ferrer-i-Cancho & Thore J. Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences* 113(19). E2750–E2758.
Heesen, Raphaela & Catherine Hobaiter, Ramon Ferrer-i-Cancho & Stuart Semple. 2019. Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B: Biological Sciences* 286 (1896). 20182900.
Hernández-Fernández, Antoni, Bernardino Casas Fernández, Ramon Ferrer-i-Cancho & Jaume Baixeries i Juvillà. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. *Lecture Notes in Computer Science 9918*. 19–29
Hernández-Fernández, Antoni, Iván González Torre, Juan-María Garrido & Lucas Lacasa. 2019. Linguistic laws in speech: The case of Catalan and Spanish. *Entropy* 21(12). 1153.
Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson & Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition* 165 (2017). 45–52.
Kershenbaum, A. et al. 2016. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews* 91(1). 13–52.

Köhler, Reinhard. 2005. Sources of information (Informationsquellen). In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski, (eds.), *Quantitative Linguistik / Quantitative Linguistics*. Berlin: De Gruyter, 2005, 1003–1014.

Luque, Jordi, Bartolo Luque & Lucas Lacasa. 2015. Scaling and universality in the human voice. *Journal of The Royal Society Interface* 12(105). 20141344.

Menzerath, Paul & Armando de Lacerda. 1933. Koartikulation, Steuerung und Lautabgrenzung: eine experimentelle Untersuchung. Berlín: F. Du¨mmler.

Meylan, Stephan C. & Thomas L. Griffiths. 2017. Word forms – not just their lengths- are optimized for efficient communication. *CoRR* abs/1703.01694. http://arxiv.org/abs/1703.01694

Ochs, Matthias, Jens R. Nyengaard, Anja Jung, Lars Knudsen, Marion Voigt, Thorsten Wahlers, Joachim Richter & Hans Jørgen G. Gundersen. 2004. The number of alveoli in the human lung. *American Journal of Respiratory and Critical Care Medicine* 169(1). 120–124.

Recasens, Daniel. 1993. *Fonètica i fonologia*. Barcelona: Enciclopèdia Catalana.

Rodríguez-Fuentes, Luis J., Mikel Penagarikano, Amparo Varona, Mireia Díez & Germán Bordel. 2012. KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 99–105.

Schwartz, Jean-Luc, Clément Moulin-Frier & Pierre-Yves Oudeyer. 2015. On the cognitive nature of speech sound systems. *Journal of Phonetics* 53. 1–4.

Toman, Michal, Roman Tesar & Karel Jezek. 2006. Influence of word normalization on text classification. *Proceedings of InSciT* 4. 354–358.

Torre, Iván G., Bartolo Luque, Lucas Lacasa, Jordi Luque & Antoni Hernández-Fernández. 2017. Emergence of linguistic laws in human voice. *Scientific Reports* 7. 43862.

Torre, Iván G., Bartolo Luque, Lucas Lacasa, Christopher Kello & Antoni Hernández-Fernández. 2019. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science* 6 (8). 191023.

Torre, Iván G., Oriol Artime, Antoni Hernández-Fernández & Bartolo Luque. 2020. ¿Es el habla una señal crítica auto-organizada? *Inter Disciplina* 8 (20).113–128.

Tsao, Ying-Chiao & Gary Weismer. 1997. Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research* 40 (4). 858–866.

Zipf, George Kingsley. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press.

Yoshifumi Kawasaki

Dating and geolocation of medieval and modern Spanish notarial documents using distributed representation

Abstract: This paper proposes a method to probabilistically date and geolocate medieval and modern Spanish notarial documents. Our model is inspired by distributed representation of words. We construct a neural network for learning spatio-temporal similarity among words so that the words that appear in documents written within a chrono-geographically close area have similar embedding vectors. The spatio-temporal similarity is learned in a framework of multi-task learning to acquire more suitable representation than when learned independently. One of the advantages of our model is its ability to detect the most contributing words to estimation. The degree of contribution corresponds to vector norm. We observed that most of the words with largest norm turn out to be those closely related with certain chrono-geographical area, well-known to Hispanic Philology. Our proposed model constitutes the first step toward developing a quantitative method of dating and geolocation accompanied by empirical evidence.

Keywords: dating, geolocation, distributed representation

1 Introduction

Dating and geolocation are the tasks of assigning an estimated date and place of issue to the documents, respectively. Philologists working on ancient texts need to develop an algorithm to identify the provenance of undated texts and to verify their authenticity. The first step in tackling this challenging problem is to establish a reliable methodology based on dated documents and then to evaluate its efficiency.

To this end, we introduce a novel method to acquire word embeddings that capture both chronological and geographical occurrence patterns. Word embeddings are a state-of-the-art technique widely utilized in the NLP community

Acknowledgment: This work was supported by JSPS KAKENHI Grant Number 18K12361.

Yoshifumi Kawasaki, The University of Tokyo, e-mail: ykawasaki@g.ecc.u-tokyo.ac.jp

https://doi.org/10.1515/9783110763560-006

to represent words as a fixed-length dense vector capturing some linguistic relationship among words: semantic, morphological, syntactic, and/or stylistic (Mikolov, Yih, & Zweig, 2013).

The contributions of this paper are threefold: (1) We propose a novel architecture that acquires spatio-temporal word vectors; (2) We demonstrate that our algorithm achieved high predictability with a low margin of error; (3) Our model enables linguistic explanations based on a quantitative contribution of each word to classification.

The rest of the paper is structured as follows. In Section 2, we briefly discuss related studies. Section 3 introduces our method of spatio-temporal joint embedding. The experimental results, accompanied by an in-depth analysis, are presented in Section 4. Finally, we conclude the paper with a brief summary and discuss the future work in Section 5.

2 Related research

From a linguistic perspective, dating has been commonly conceived as a classification task (Kumar, Lease, & Baldridge, 2011; Tilahun, Feuerverger, & Gervers, 2012). Time is discretized with a given granularity (year, decade, etc.), and the built multi-class classifier determines the most probable timepoint. Geolocation has also been addressed in the same framework, where the partitioned square grids or the administrative units (cities, states, etc.) are utilized as categories (Rahimi, Cohn, & Baldwin, 2017).

In the preceding studies, time and space are dealt with separately. We are the first to handle the two closely related dimensions in a unified fashion. Following a general approach, we treated dating and geolocation as text classification tasks.

3 Methods

We propose a novel technique to capture a spatio-temporal occurrence pattern of words in a single vector using multi-task learning (Goldberg, 2017). We construct a neural network for learning spatio-temporal similarity among words in such a way that the words appearing in documents of close origin have similar vector representation.

Let d denote the document in the corpus. The document d is a sequence of words $\{w_1, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_{|d|}\}$, where w_i is the target word and $\{w_{i-m}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+m}\}$ are its context words. The window-size $m \in \mathbb{N}$ is a hyperparameter. For each word $w \in V$, $\boldsymbol{v}_w$ denotes its word vector, where V is the vocabulary. The learning is done in such a way that the joint probability $P(t, l|w)$ is maximized. To be more concrete, the mean vector of the target word and its context words $\boldsymbol{h}_{w_i} = \frac{1}{2m+1} \sum_{-m \le j \le m} \boldsymbol{v}_{w_{i+j}}$ predict both the date $t \in T$ and place $l \in L$ assigned to d. Here, T and L are the sets of timepoints and locations, respectively. This joint embedding is expected to capture spatio-temporal information with only a single architecture (Fig. 1, left). The total number of parameters to be learned amounts to $|V|D + |\boldsymbol{W}_t| + |\boldsymbol{W}_l|$, where D is the embedding size, $\boldsymbol{W}_t$ the output weight matrix for dating, and $\boldsymbol{W}_l$ that for geolocation.

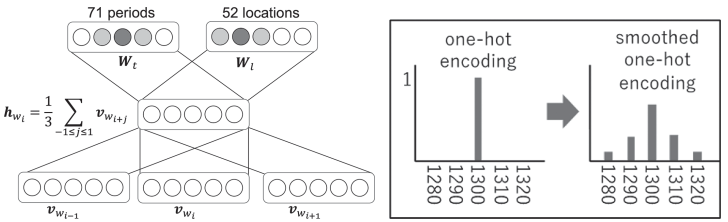


Fig. 1: Illustration of architecture for joint spatio-temporal embedding (left). Here, the window size is set to $m = 1$. Schematization of how the one-hot encoding is transformed into a smoothed one with a center point of the period 1300 (right).

We decompose the aforementioned joint probability $P(t, l|w)$ into the product of two probabilities under the conditional independence assumption (3.1). Here, $\boldsymbol{v}_t$ and $\boldsymbol{v}_l$ are the weight vectors corresponding to the timepoint t and location l, respectively.

$$P(t, l|w) = P(t|w)P(l|w) = \frac{\exp(\boldsymbol{h}_w \cdot \boldsymbol{v}_t)}{\sum_{t' \in T} \exp(\boldsymbol{h}_w \cdot \boldsymbol{v}_{t'})} \frac{\exp(\boldsymbol{h}_w \cdot \boldsymbol{v}_l)}{\sum_{l' \in L} \exp(\boldsymbol{h}_w \cdot \boldsymbol{v}_{l'})} \tag{3.1}$$

Another novelty of our methodology lies in relaxing the commonly used one-hot vector encoding using kernel function $K(x, x'; \sigma) = \exp\left(-\left\|x - x'\right\|^2 / \sigma^2\right)$. Here, $\left\|x - x'\right\|$ represents the difference between the point x and the center point x' and corresponds to the temporal difference and geographical distance in dating and geolocation, respectively. With $\sigma \to 0$, the smoothed vector coincides with the original one-hot vector. By way of illustration, the one-hot vector $(0, 0, 1, 0, 0)$ can be converted into a smoothed one $(0.1, 0.2, 0.4, 0.2, 0.1)$ with a low value assigned

to the adjacent points in proportion to the distance from the center point (Fig. 1, right). Let $\boldsymbol{q} = (q_1, \ldots, q_M)$ denote the M-dimensional one-hot vector, where M represents the number of categories. Our method updates the i-th element q_i guaranteeing that the sum of the smoothed vector is equal to 1:

$$q_i \leftarrow \frac{K(x_i, x'; \sigma)}{\sum_x K(x, x'; \sigma)} \tag{3.2}$$

This transformation allows for reflecting prior knowledge on temporal and geographical contiguity, leading to a smoother distribution.

From the aforementioned explanation, we define the loss function to be minimized as follows:

$$Loss = -\sum_i^n \frac{\alpha_i}{\sum_i \alpha_i} \left(\sum_{t \in T} k_{it} \log P(t|w_i) + \sum_{l \in L} k_{il} \log P(l|w_i) \right) \tag{3.3}$$

Here, $k_{it} \in [0,1]$ and $k_{il} \in [0,1]$ are the smoothed answer label of the i-th token for dating and geolocation, respectively. Here, n is the length of concatenation of all the training documents and $\alpha_i = 1/N_{(t,l)}(w_i)$ is the weight given to w_i, computed as the inverse of the overall token counts $N_{(t,l)}(w_i)$ at the timepoint t and the location l where w_i is seen. This weighting is aimed at leveling the unequal distribution of words across the corpus. The importance of a word is discounted when the word comes from the timepoint and place where the number of tokens is high.

Finally, let us describe the categorization procedure. The document vector $\boldsymbol{d}$ is constructed as the mean of the word embeddings seen in the document: $\boldsymbol{d} = \frac{1}{|d|} \sum_{w \in d} \boldsymbol{v}_w$. Obviously, the word vectors with high norms contribute more. We only take into consideration words seen in the document, ignoring those that do not appear in the training data. We decided to simplify document representation in the detriment of syntactic information conveyed by word order. Once $\boldsymbol{d}$ is computed, the document vector is converted into two probability distributions, one for dating and another for geolocation, using softmax functions with the same respective output weight matrices $\boldsymbol{W}_t$ and $\boldsymbol{W}_l$ obtained during the training phase. The document is then attributed to the timepoint and location that presents the highest probability.

4 Experiments

4.1 Dataset

The experiment was conducted using the medieval and modern Spanish documents corpus *CODEA+ 2015 (Corpus de Documentos Españoles Anteriores a 1800)*.[1] This corpus contains nearly 2500 documents of a legislative, administrative, and legal nature, composed mostly in present-day Spain from the 12th to 18th century.

From the 2156 documents bearing date and place of issue, a randomly chosen 80% were used as training data for learning distributed representation, 10% as a validation set for hyperparameter tuning, and the remaining 10% as a test set for dating and geolocation. We used a critical version of texts (*presentación crítica*), where the spelling, as well as unification/separation of words, were modernized based on unified criteria. The text was lowercased, and the non-alphabetical characters were deleted. Furthermore, we manually eliminated explicit information referring to the date and place of issue. The words with a frequency of less than 10 in the training data were converted into a single token UNK. Special tokens BOS and EOS were added at the beginning and end of the text, respectively. We did not apply tokenization. Overall, the vocabulary size $|V|$ was 6.5K, and the number of tokens was 1.1M.

The time spanning from 1090 to 1809 was discretized into 71 periods with an interval of a decade: $T \equiv \{1090, 1100, \ldots, 1800\}$. By way of illustration, the period $t = 1090$ comprises the years from 1090 to 1099. Thus, the date of the documents was converted into the corresponding period. The set of locations L consists of the 50 present-day Spanish provinces.

4.2 Analysis of trained word vectors

The results presented below come from the following hyperparameter setting: embedding size $D = 50$, window size $m = 2$, kernel bandwidths $(\sigma_t, \sigma_l) = (20, 50)$. We fixed the mini-batch size to 1000. The training was stopped after 20 epochs.

1 I am indebted to Prof. Pedro Sánchez-Prieto Borja of the University of Alcalá, Spain, for providing me with digitized text data of the corpus, without which this study would have been impossible.

The left panel of Fig. 2 visualizes in two dimensions the acquired word embeddings via *t*-SNE (Van Der Maaten & Hinton, 2008). The words plotted closely are similar in their pattern of spatio-temporal occurrence. The vertical axis roughly represents chronology, and the horizontal one indicates geography and topics. We can observe older forms at the bottom and modern ones at the top. More specifically, at the middle-lower center, we find Medieval Latin words, at the lower-center old Astur-Leonese forms, and at the lower-right old Navarro-Aragonese terms.

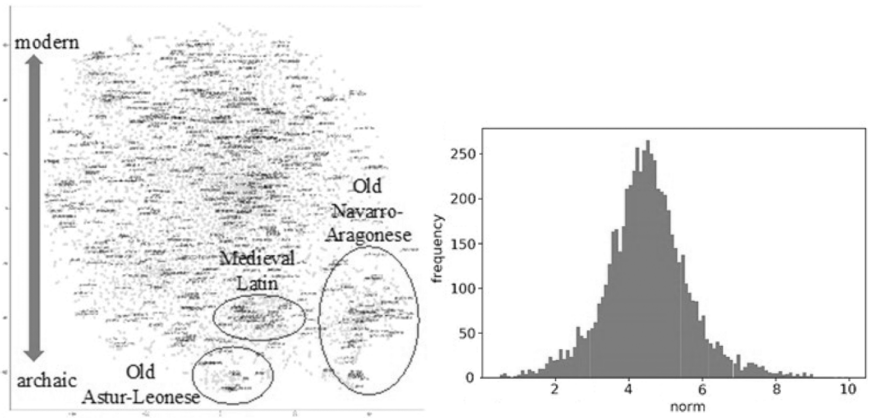


Fig. 2: *t*-SNE visualization of word embeddings (left) and their norm distribution (right).

The norm distribution is shown on the right panel of Fig. 2. The majority of words with the lowest norm are, as was expected, the terms readily found anytime and anywhere, hence deprived of spatio-temporal traits: *de*, *en*, *por*, *a*, *los, que, con*, *la*, etc. Conversely, the words with the highest norm are mostly toponyms, proper nouns, and locutions indicative of specific chrono-geographical origin: *polbos, trujiello, santorcat, gordon, tableros, fregoria, fito, dites,* etc.

Figure 3 depicts the results of dating and geolocation of *possedir* "to possess". Interestingly enough, they are in perfect agreement with their old Aragonese origins. The darkness of the shaded areas on the map corresponds to the posterior probability: the darker the shading, the higher the probability. The nearest neighbors of *possedir* include *ont, ditas,* and *nozer*, all of which are congruent with philological insights (Zamora Vicente, 1967).

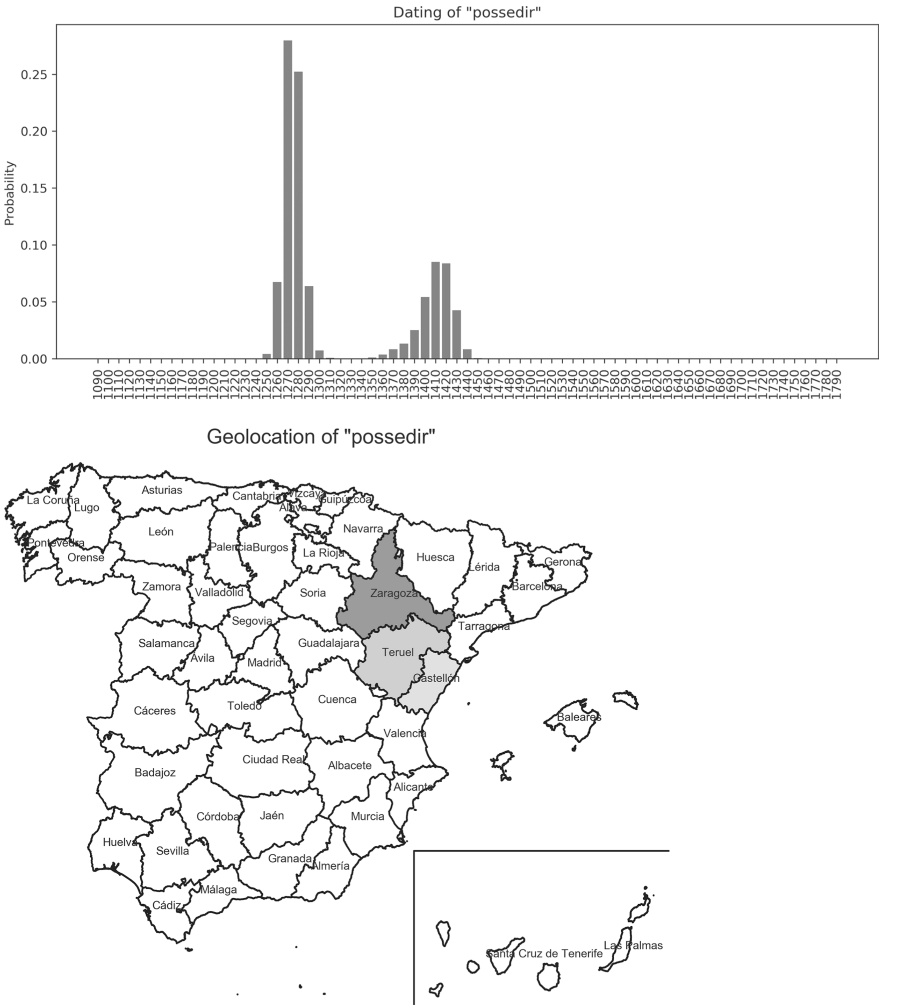


Fig. 3: Outcome of dating of *possedir* "to possess" (top). Illustration of the results of geolocation (bottom).

4.3 Classification results

The performance of dating was measured by how far the true period was from the estimated one. For instance, the document incorrectly classified into the preceding or the following period presented an error of 10 years. The mean

absolute error (MAE) was 24 years. The left panel of Fig. 4 plots estimated dates against the true ones. Most of the dots lie along the diagonal line. The performance of geolocation was evaluated using the geographical distance between the correct and estimated provinces. The error distribution is shown on the right of Fig. 4. We attained an MAE of 104 km with an overall accuracy of 56%. It is only fair to note that, against our expectations, the separately learned temporal and spatial embeddings, each with half the dimensionality of the joint ones, yielded comparable or occasionally higher predictability, which implies that the multi-task learning might have brought about underfitting in each task, and that time and space should be considered discretely.

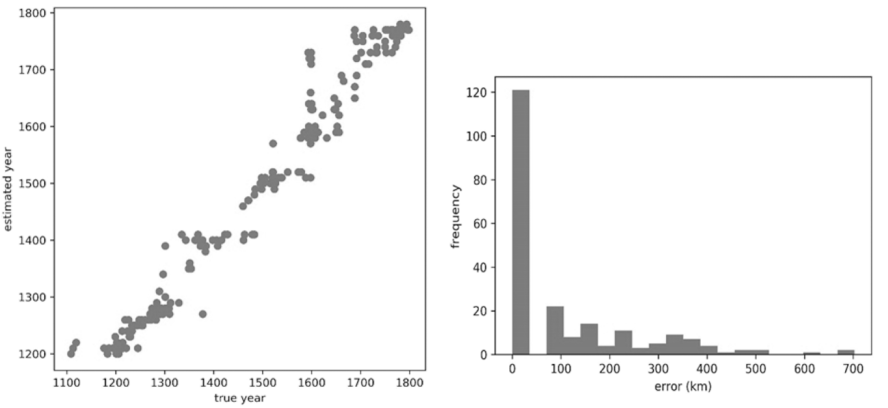


Fig. 4: Illustration of the results of dating, where the estimated dates are plotted against the true ones (left). Error distribution of geolocation (right).

A sensitivity analysis revealed that the larger window size generally led to performance improvement for $m \in \{0, 1, 2, 3\}$; that, expectedly, the larger embedding size resulted in better fitting for $D \in \{25, 50, 100\}$; that $\sigma_t = 20$ yielded the best results for $\sigma_t \in \{0, 10, 20, 30\}$, which reflects relative temporal continuity; that the small bandwidth $\sigma_l = 50$ was the most effective for $\sigma_l \in \{0, 50, 100, 200\}$, which suggests that each location has its own peculiarities not shared with its neighboring areas; and that the token weighting in the optimization was beneficial.

We conclude the section with our examination of a concrete case of estimation. Figure 5 shows the results of the dating of the Doc. 301 (AD 1276, Palencia). Dating was mostly accurate with an error of 6 years. The place of issue, Palencia, was correctly identified, followed in the ranking by the neighboring provinces.

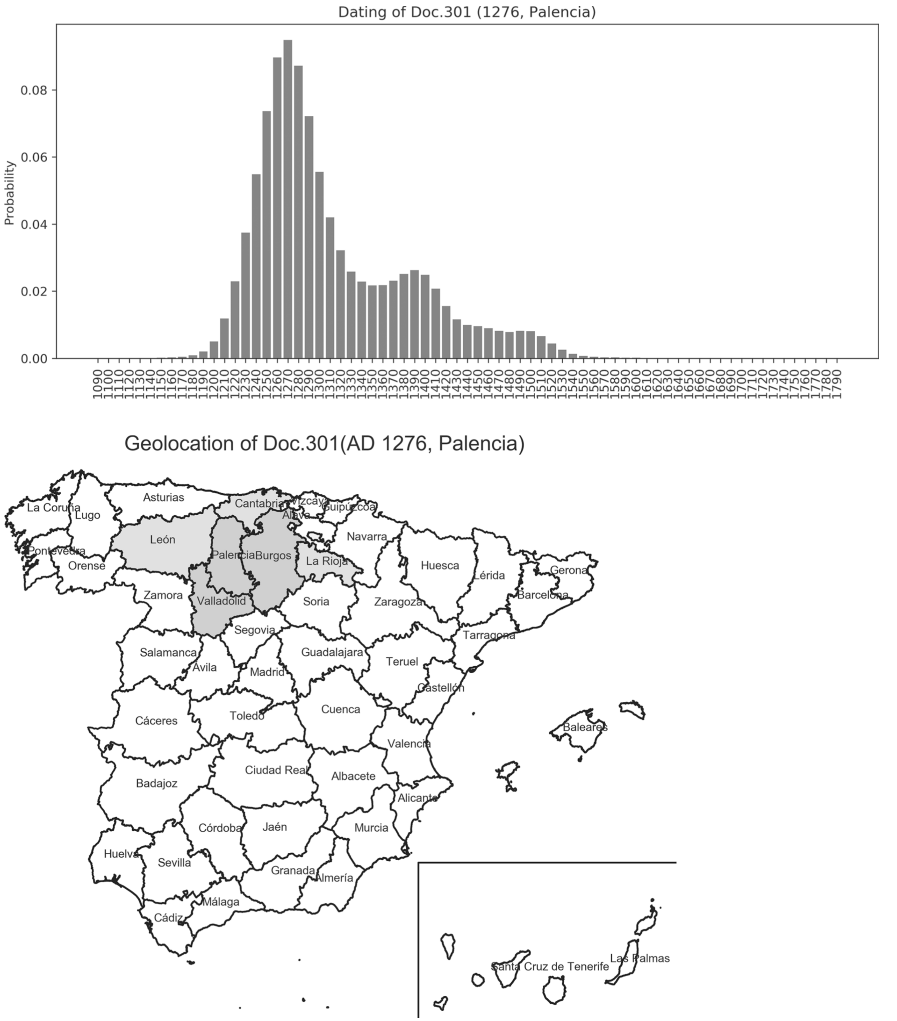


Fig. 5: Results of dating (top) and geolocation (bottom) of the Doc. 301 (AD 1276, Palencia).

The fairly smooth distribution arising from the relaxed one-hot encoding helped to interpret the classification results. Among the words with the highest norm were *cumo*, *cabillo*, *part*, and *viren*, all pointing to the old mid-western origin of the text (Menéndez Pidal, 1999).

5 Conclusions and future work

This paper presented an algorithm to date and geolocate undated documents, based on the novel spatio-temporal joint embedding technique. Our experiment demonstrated that estimations are feasible with an allowable margin of error and that the outcome agrees with philological insights. An interesting extension is to explore the possibility of leveraging character-level embeddings to be learned using a recurrent neural network. In so doing, we could exploit orthographic variation as well as morphological commonality, which are not addressed in this study. Another line of research is to detect word-level anomalies in the text by discovering words that would exhibit extremely low cosine similarity with the overall document vector. The identified words are possibly archaisms, modernisms, dialectal variants, or posterior corrections.

References

Goldberg, Yoav. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies 10*(1). 1–311.

Grupo de Investigación Textos para la Historia del Español (GITHE). n.d. *CODEA+ 2015 (Corpus de documentos españoles anteriores a 1800)*. http://www.corpuscodea.es/ (accessed 9 April 2020)

Kumar, Abhimanu, Matthew Lease & Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, 2069–2072. New York: Association for Computing Machinery.

Menéndez Pidal, Ramón. 1999 [1918]. *Manual de gramática histórica española*, vigésima tercera edición. Madrid: Espasa-Calpe.

Mikolov, Tomas, Wen Tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Association for Computational Linguistics.

Rahimi, Afshin, Trevor Cohn & Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 209–216. Association for Computational Linguistics.

Tilahun, Gelila, Andrey Feuerverger & Michael Gervers. 2012. Dating medieval English charters. *Annals of Applied Statistics* 6(4). 1615–1640.

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

Zamora Vicente, Alonso. 1967 [1960]. *Dialectología española*, segunda edición muy aumentada. Madrid: Gredos.

Tatiana A. Litvinova, Olga A. Litvinova

Cross-modal authorship attribution in Russian texts

Abstract: The problem of determining the author of an anonymous text from a set of possible authors based on a comparison of stylometric features ("authorship attribution") has been studied extensively, but it remains extremely difficult to address in a forensic scenario which is often characterized by a limited number of short texts of known authorship that do not match with the questioned document in terms of the domain (topic, genre, mode – written/oral, etc.). Cross-domain authorship attribution is a difficult and understudied task, cross-modal authorship attribution being the least examined. We performed experiments in cross-modal authorship attribution on a Russian text corpus with a small number of written and oral texts per author and of different genres, modelling real-world forensic scenario. We tested several types of context-independent features and applied principal component analysis (PCA) and cluster analysis to find similarities in data. Results of one-way ANOVA for text PCA coordinates and categorical variables ("factors") "Author", "Mode" and "Mode_Author" show that most of the context-independent features change with mode shift but in an author-dependent way. Morphological features first introduced in this paper discriminated authors despite the mode, although classification models of the combined feature set showed the highest accuracy. Discourse features showed the main effect for mode. Our research contributes to the understanding of the level of stability of idiolectal features during a mode change and highlights the extremely complex and intricate nature of intra- and interidiolectal variations as well as the necessity to include spoken data in authorship attribution studies.

Keywords: authorship attribution, idiolect, idiolectal variation, Russian language

Acknowledgment: The work is supported by the grant of Russian Science Foundation No 18-78-10081 "Modelling of the idiolect of a modern Russian speaker in the context of the problem of authorship attribution".

Tatiana A. Litvinova, Olga A. Litvinova, Voronezh State Pedagogical University, Corpus Idiolectology Lab, e-mail: centr_rus_yaz@mail.ru

https://doi.org/10.1515/9783110763560-007

1 Introduction

The problem of authorship attribution (AA), i.e. determining the author of an anonymous text, is of great importance for national security because of the rapid growth of internet communication and the need to determine the author of malicious content. Authorship attribution is a hot research topic with a lot of papers reporting high accuracies of predictive models. Despite this fact, there is a range of problems yet to be addressed. Most of the papers (especially research in computer science and literature studies) focus on the tasks of AA far from the viewpoint of a typical forensic scenario. Namely, most of the studies deal with 1) a large number of possible authors; 2) a large volume of texts or/and a large number of texts per author; 3) match of training and test texts in the genre, topic, length, etc.; 4) features and models which are difficult to interpret (e.g., character n-grams) (Kestemont et al. 2019; Neal et al. 2018).

A cross-domain scenario, i.e. the one where texts with known authorship and disputed texts come from different domains (genres, topics, modes, etc.), though very common in the real world, remains difficult for classifiers (Kestemont, M. et al. 2019). Meanwhile, some aspects of cross-domain AA, e.g., a cross-modal type where texts come from different modes (oral/written) are surprisingly understudied, in part due to the difficulties in obtaining appropriate corpora (Goldstein-Stewart et al. 2009). Overall, spoken data are rarely used in AA, which is quite surprising, – given the development of speech-to-text technologies and lack of training data typical of forensic scenario, spoken data are at least worthy of research attention. Meanwhile, as some studies show (Litvinova et al. 2018), even in a tight (somewhat unrealistic) topic-genre-controlled scenario, a mode shift causes dramatic changes in the linguistic styles of the speakers (texts were produced by the same individuals). Moreover, as was shown in Litvinova et al. (2018), a shift in mode affects a broader range of stylometric features (idiolectal markers) than does a change in topic and text type. When texts with known authorship differ from a questioned document not only in mode but also in the other factors of intraidiolectal variation (which is a typical real-world scenario), will we still be able to attribute the disputed text? Which features are least affected by the factors of intraindividual variation?

The aim of this paper is to assess the level of stability of the so-called context-independent stylometric features in oral and written texts by the same authors and their discriminative ability in a cross-modal real-world-like scenario (short texts, small number of suspects, mismatch of training and test texts in the topic, genre, etc.) (Litvinova and Gromova, 2020). We tested several types of content-independent features. Some of the features are already widely used in the AA domain, while other markers have been presented for the first time. A special focus

of our paper is on searching for the latent structure of the data in terms of the effect of different factors of idiolectal variation on text similarity. Therefore our research contributes to the understanding of a complex picture of the intra- and interindividual variation.

2 Cross-modal authorship attribution

Studies in authorship attribution in the cross-modal scenario are sparse. Research by Aragón (2016) and Litvinova et al. (2018) is dedicated to the assessment of the level of stability of idiolectal features under mode shift. Aragón (2016) investigated oral and written picture descriptions over a range of hand-crafted lexical, morphosyntactic and discursive features, and revealed that the morphosyntactic variables (such as the use of subordinated relative pronouns, modal verbs, the position of thematic adjunct) showed the highest level of consistency over mode change, while discursive and pragmatic features were highly variable. Litvinova et al. (2018) examined the level of stability of a range of context-independent stylometric features under mode, topic and text-type shift and showed that the proportion of periods, certain conjunctions and the overall proportion of conjunctions (means of "making boundaries," according to the authors) were stable, while lexical complexity features were mode-dependent.

However, neither of the studies deals with the task of AA, unlike that by Goldstein-Stewart and colleagues (2009), which reported the results of several experiments on AA, including cross-modal settings. Using a set of topical words, function words and features extracted by means of the LIWC software,[1] researchers reported the lowest accuracy of the model for this scenario in comparison to cross-genre and cross-topic settings. Thus their results are in line with the findings reported in Litvinova et al. 2018. Overall, spoken data are surprisingly rarely used in AA.

All of the above-mentioned studies were implemented on texts, although different in mode, topic and register, produced in similar conditions during a short period of time by a highly homogeneous pool of authors (university students). On the one hand, this allows researchers to control the effect of the different factors on linguistic production (as far as it is possible), but, on the other hand, this "sanitized" version of the AA task simplifies a complex picture of intra- and interidiolectal variation. Besides, this research failed to take into account the intercorrelations of linguistic variables.

1 https://liwc.wpengine.com/.

3 Materials and methods

3.1 Corpus description

For this particular study, we used texts included in the "RusIdiolect" Corpus, which is being developed at the Corpus Idiolectology Lab (Litvinova 2022).[2] RusIdiolect Corpus contains both multiple texts by the same authors and metadata describing the authors (gender, age, for some – personality test scores, etc.) and a communication situation where the texts were produced (experimental vs. natural, genres, etc.), thus providing users with information on factors of intra- and interidiolectal variation.

For this particular study, we were seeking to balance the total volume of texts from the same authors and the number of written and oral texts, while keeping different levels of similarity in terms of genre and communication situation they were produced in. The mean length of the texts is rather small (391 words, SD=145) to make the task as close to the real-world scenario as possible. All the texts are in Russian, and all the authors are native speakers of Russian. The corpus description is presented in Tab. 1.

Tab. 1: Characteristics of the authors and texts.

A pensioner, a former university teacher of English, female, 80 (Author_C)	**An editor for a university publishing house, female, 36 (Author_B)**	**A housewife (an English teaching degree), female, 30 (Author_A)**
Written	*Written*	*Written*
Text 1 – Picture description (as part of an experiment) (189)	Text 1 – Essay "What does a city of my dream look like?" (as part of an experiment) (379)	Text 1 – Picture description (as part of an experiment) (319)
Text 1 – Picture description (as part of an experiment) (206)	Text 2 – Review for the film recently seen by the author (as part of an experiment) (398)	Text 2 – Essay "What annoys you most of all?" (as part of an experiment) (272)
Text 3-5 – Parts of fiction books written by the respondent 1 to 5 years ago (372, 465, 270)	Text 3 – Picture description (as part of an experiment) (387)	Text 3 – Description of the previous day in chronological order (as part of an experiment) (320)

2 The corpus is freely available at https://rusidiolect.rusprofilinglab.ru/.

Tab. 1 (continued)

A pensioner, a former university teacher of English, female, 80 (Author_C)	**An editor for a university publishing house, female, 36 (Author_B)**	**A housewife (an English teaching degree), female, 30 (Author_A)**
	Text 4 – Part of the article in the sports journal (346)	Text 4 – Description of the previous day in non-chronological order (as part of an experiment) (417)
Total written: *1 502*	**Total written: 1 510**	**Total written: 1 332**
Oral	*Oral*	*Oral*
Text 1 – Part 1 of a spontaneous everyday conversation, 664	Text 1– Essay "Imagine you wake up in somebody else's body" (as part of an experiment) (460)	Text 1 – Picture description (as part of an experiment) (290)
Text 1 – Part 2 of a spontaneous everyday conversation, 609	Text 2 – Another picture description (as part of an experiment) (369)	Text 2 – Description of a day (as part of an experiment) (277)
Text 1 – Part 3 of a spontaneous everyday conversation, 503	Text 3 – Essay "Your attitude to WWW" (deceptive opinion) (as part of an experiment) (718)	
Total oral: 1776	**Total oral: 1 547**	**Total oral: 567**
Professor of History, male, 47 (Author_E)	**Professor of Philology, male, 47 (Author_F)**	**Barman, blogger, male, 33 (Author_D)**
Written	*Written*	*Written*
Text 1 – Post on social media (504)	Text 1 – Review of a performance (557)	Texts 1–6 – Posts (251, 233, 150, 495, 210, 173)
Text 2 – Part of a research paper (426)	Text 2–4 – Parts of fiction books written by the respondent 1 to 5 years ago (532, 207, 743)	
Text 3 – Forum posts (565)		
Total written: 1 495	**Total written: 2 039**	**Total written: 1 512**
Oral	*Oral*	*Oral*
Text 1-2 – Videolecture (326, 339)	Text 1-2 – Videolecture (479, 505)	Text 1-3 – Vlogs (430, 362, 247)

Tab. 1 (continued)

Professor of History, male, 47 (Author_E)	Professor of Philology, male, 47 (Author_F)	Barman, blogger, male, 33 (Author_D)
Text 3-4 – Vlogs (340, 501)		
Total oral: 1 506	**Total oral: 984**	**Total oral: 1 039**

Oral texts were orthographically transcribed by hand (each text was transcribed independently by two professional linguists), punctuations marks were added based on punctuation rules and intonation of the speaker.

3.2 Methods

Since our aim is not to construct the classifier to attribute texts in terms of their authorship with the highest possible accuracy but to reveal the effect of mode change on the level of stability of idiolectal features and patterns of interaction of different contextual (**Mode, Author**) and textual variables, we focus mainly on exploratory methods capable of finding a latent structure in data.

As exploratory methods, we use the principal component analysis (PCA) followed by cluster analysis. PCA has been widely used in AA studies since the seminal study by Binongo et al. (1999). E.g., PCA was performed on 50 most frequent words with relation to different factors of interidiolectal variation (age, genre, education) in Baayen et al. 2002, with no clear authorial, but genre and educational structures revealed (factors being analyzed separately). We aimed to reveal the effect of combined factor **Mode_Author** as well as **Mode** and **Author** separately on the level of similarity of texts. We also use a technique that combines the advantages of PCA and clustering – Hierarchical Clustering on Principal Components (HCPC) (Husson et al. 2010a) as implemented in FactoMineR, a package dedicated to exploratory data analysis in R (Husson et al. 2010b). We also used FactoShiny[3] and FactoExtra (Kassambara 2017) packages for the visualization of PCA and HCPC results.

So, we aim to obtain the typology of texts based on the text variables. We took texts by 5 authors as active individuals as well as all text variables except for *text length* which was taken as supplementary. Author A was taken as a

3 http://factominer.free.fr/graphs/factoshiny.html.

supplementary individual. **Mode, Author, Mode_Author** were considered as supplementary qualitative variables (factors). All the textual variables are standardized to attribute the same influence to each linguistic unit.

For the classification experiments, we used R package Stylo (Eder et al. 2016). We performed two series of experiments: firstly, we used oral texts by pair of authors for training and their written texts for testing; secondly, we performed pairwise leave-one-out cross-validation based on all the texts by two speakers.

3.3 Feature description

We tested several sets of features that are considered context-independent. Only the features which could be extracted both from transcribed oral and written texts were chosen. Another principle for feature selection is their corpus independence. While a lot of AA papers make use of corpus-dependent features like *n*-most frequent words of the corpus or apply feature selection techniques like tf-idf (which is often useful in terms of accuracy of the predictive models), we believe that in order to make more generalized conclusions, more corpus-independent features should be employed.

According to Oakes (2019: 103), PCA is normally used with from 10 to over 100 variables, so we attempted to follow this rule while constructing our feature set. In a study of this kind, there should ideally be at least a 4:1 ratio between subjects and variables (Ibid). However, there are a lot of datasets (e.g., genome data, but language data as well) where the number of variables is many times higher than that of individuals. In this case, "diagonalising the scalar products matrix instead of the correlation matrix as the FactoMineR package does" (Husson et al. 2010b: 28) is recommended. Below we describe our feature sets in more detail.

3.3.1 Function words and punctuation marks (FW_P)

This type of features is widely used in AA studies and considered context-independent, although it was shown (Mikros and Argiri 2007; Goldstein-Stewart et al. 2009) that at least some function words are domain-dependent. Moreover, the composition of this feature set differs from study to study and little attention is paid to the behavior of certain units from this list as well as to their linguistic nature. Some studies (see Goldstein-Stewart et al. 2009, for example) use lists of stop words for a given language, but the problem of sparsity inevitably arises in the case of short texts. To overcome these problems, we have compiled several

lists of the most frequent function words in Russian. Using RuTenTen,[4] one of the largest corpora of modern Russian, and the Average Reduced Frequency as a criterion, we compiled two lists – 1000 most frequent word forms and 1000 most frequent lemmas. After that, content words were deleted manually, resulting in two lists – **1000Word_forms** (334 units) and **1000Lemmas** (176 units). Next, we deleted from these lists all the words except for prepositions, particles and conjunctions, resulting in **1000Word_forms_Strict** (72) and **1000Lemmas_Strict** (62) lists. We also applied the same procedure for the first **100 word forms** (resulting in two lists with 91 units for a non-strict list and 41 for a strict list) and **100 lemmas** (71 units and 36 units, respectively). We also compiled the list of punctuation marks (PM, 10) since, as it was shown by Kulig and colleagues (2017), they behave statistically as the most frequent words (which are mostly function words). Thus, ***FW_P*** feature set consists of the percentages of the most frequent function word forms (91), 10 PM and the total percentage of PM as features (Allpuntc). In order to avoid sparsity, we set a frequency threshold: features with zero values in more than 10% of the texts were removed. The resulting feature set compiles 13 features (9 FW, 3 PM, Allpuntc).

3.3.2 Lexical sophistication (Lex_Soph)

This type of feature describes the concept of vocabulary sophistication in terms of its frequency in a given language. They use the information on the frequency of words outside the analyzed texts. We developed **11** features: percentage of word forms in texts from 1000Word_forms; 1000Word_forms_Strict; 100Word_forms; 100Word_forms_Strict; lemmas from 100Lemmas; 100Lemmas_Strict; 1000Lemmas; 1000Lemmas_Strict; 9 MFW (from FW_P); Allpuntc; 9 MFW + Allpuntc. Lemmatization was performed using TreeTagger.[5]

3.3.3 Lexical richness and diversity (Lex_Rich)

The most popular feature of this type is the type-token ratio (TTR), however, it is dependent on text length. For this study, we aimed to use features that are considered length-independent (as far as it is possible for this kind of feature set). Researchers do not seem to agree on the problem of the effect of text length on such

4 https://www.sketchengine.eu/rutenten-russian-corpus/.
5 https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

features (Fergadiotis et al. 2015). Based on a thorough literature review, we have chosen the Moving-Average Type-Token Ratio (MATTR) and calculated this index separately for words and lemmas. This was done with the MATTR software (version 2.0 for Windows)[6] with a window size = 50 words. We also calculated the differences between these two variables, which was shown to be effective for author and genre discrimination (Čech and Kubát 2018). We also have chosen several indicators calculated using QUITA software with minimal dependence on the length (Entropy, Av_Tokens_Length, R1, RRmc, Λ, Adj_Mod, WritersView, Clength_R, Token_Length_Max) (for more details, see Kubát et al. 2014). In addition, we decided to include TTR on all the words and on content lemmas only calculated on English translations of original texts performed using Google translate service. to retrieve additional information on the lexical diversity of the texts. We also used two indices related to the log frequency of features from the CELEX database. Although these indices are conceptually related to lexical sophistication, we included them in this group since they were also performed on English translations. We also added the indices reflecting concreteness, imageability, familiarity and meaningfulness of content words. The last 6 variables were also calculated on the translated texts using the Coh-Metrix software (McNamara et al. 2014).

3.3.4 Morphological features (Morph)

We performed POS-tagging using TreeTagger and then calculated 10 length-independent features of lexical richness from ***Lex_Rich*** set but this time not on word forms and lemmas but on POS tags. We also calculated the following indices: Lexical Density (ratio of content to function POS), F-measure (Heylighen and Dewaele 2002), Activity (ratio of verbs to the sum of verbs and adjectives), Quality (ratio of adjectives to nouns), Details (ratio of adjectives and adverbs to nouns and verbs), Deicticity (ratio of pronouns to the sum of adjectives and nouns), Prep_Noun (ratio of prepositions to nouns). We also calculated Idea Density (Brown et al. 2008) on translated texts using CPIDR software.[7]

6 http://ai1.ai.uga.edu/caspr/#CPIDR.
7 http://ai1.ai.uga.edu/caspr/#CPIDR.

3.3.5 Discourse features (Disc)

Since there are no tools for the extraction of the discourse feature for Russian, we translated text into English and then processed them using Coh-Metrix. We calculated indices for referential cohesion (noun overlap, stem overlap, content word overlap, etc.), verb overlap, statistics on connectives overall and on certain categories, temporal cohesion, etc. (28 features overall; for a detailed description of the indices we refer the reader to McNamara et al. 2014).

3.3.6 Readability (Read)

To calculate this type of feature, we used the R package quanteda (Benoit et al. 2018). We only selected features which do not depend on the syllable detection since this function works reliably only for English texts. Also, we included mean sentence length (MSL) in this feature set. Coh-Metrix readability measures calculated on translated texts (Flesch Reading Ease, Flesch–Kincaid Grade Level, L2 Readability index) were added as well. Overall, 14 features compile this set.

3.3.7 Total (Total_set)

Overall, 104 features were used.

4 Results

4.1 Finding patterns in data

First of all, we checked all the results of PCA against the reference values (the 0.95-uantile of the distribution of the percentages obtained by simulating data tables of equivalent size on the basis of a normal distribution, see Husson et al. 2010b: 28). All the obtained values were higher than the reference ones, therefore the variability explained by PCA is significant. The analysis of the graphs does not detect any outliers. No missing values were detected. The number of dimensions used for follow-up clustering was chosen based on a scree test performed using R package Factoextra.

R package FactoMineR provides the results of one-way ANOVA for PCA coordinates of individuals and each categorical variable as well the values of the

chi-square test for cluster partitioning and categorical variables. The confidence ellipses were constructed to assess the separability of classes (Fig. 1). We summarized the results of the analysis in Tab. 2. Due to the lack of space, we provide the results only for the first two dimensions for each feature set.

Tab. 2: Effect of categorical variables on feature sets.

Feature set	Categorical variable related to dimensions (ANOVA for PCA coordinates of individuals and categorical variable p-value)		Categorial variable related to cluster partitioning (chi-square test p-value)
	1st Dim	2nd Dim	
FW_P	**Mode_Author** (0.002), **Mode** (0.005)	**Mode_Author** (0.03)	**Author** (0.03)
Lex_Soph	None	**Mode_Author** (<0.0001), **Mode** (0.008), **Author** (0.03)	None
Lex_Rich	**Mode_Author** (0.0008), **Mode** (0.003), **Author** (0.05)	**Author** (0.01), **Mode_Author** (0.02)	**Mode_Author** (0.0002), **Mode** (0.001), **Author** (0.003)
Morpho	**Mode_Author** (0.003), **Author** (0.006)	**Mode_Author** (0.001), **Author** (0.03)	**Author** (0.001), **Mode_Author** (0.002)
Read	**Mode_Author** (<0.0001), **Mode** (0.007), **Author** (0.01)	None	**Mode_Author** (0.0002), **Author** (0.007), **Mode** (0.03)
Disc	**Mode** (<0.00000001), **Mode_Author** (<0.0000001)	None	**Mode** (<0.0000001), **Mode_Author** (p<0.0000001)
Total_set	**Mode_Author** (<0.00001), **Mode** (0.00004)	**Mode_Author** (0.00002), **Mode** (0.001), **Author** (0.04)	**Mode_Author** (0.0002), **Mode** (0.001)

Confidence ellipses (Fig. 1) showed clear separation on **Mode** for ***Lex_Rich*** and ***Disc***, slight overlap for all the rest sets, except for ***Morpho***, which showed the smallest overlap for **Author** and the largest overlap for **Mode**. Clustering with ***Lex_Rich*** revealed a clear "author+mode" structure. All the sets showed the main effect for ***Mode_Author*** for Dim1 except for ***Disc*** features which exhibited the main effect for **Mode**. ***Mode_Author*** was first in the rank of factors explaining for

distances between variables on Dim2 for all feature sets except for ***Lex_Rich***, which showed the main effect for **Author**.

It is interesting to comment on the coordinates of texts by supplementary individual F_30 (Fig. 1). The coordinates of the texts of this individual were predicted using the PCA parameters obtained on the texts written by active individuals.

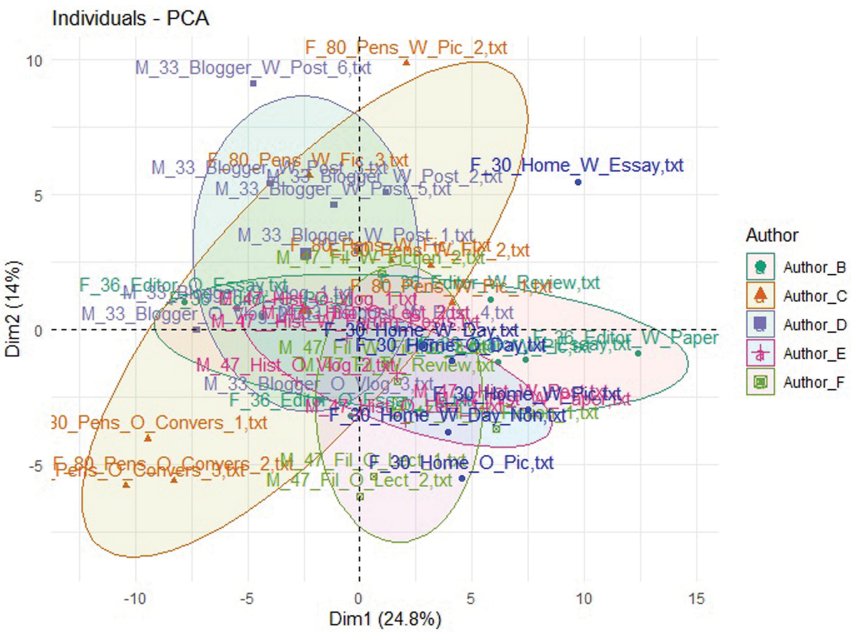


Fig. 1: PCA with **Total_set** and supplementary individual.

This female author was hired to conduct the task of writing/speaking via social media and sent all the texts 24 hours after receiving the assignment. She was the only author without texts produced outside the experiment. The small distance between oral and written day descriptions could be explained by the fact that oral text was produced immediately after writing on the same topic or even based on it (the transcriber commented on these texts saying that they felt 'as if she was reading'). The picture description, which was assessed by transcribers as "natural", is closer to oral texts by the other authors. An essay written by this author contained borrowings from the Internet. Although the plagiarism was detected and removed, this fact could have affected the structure of the remaining text or paraphrasing could be used. This could explain the position of this text far from the other texts by this author.

4.2 Supervised approach

Despite a small number of texts, we performed classification experiments in authorship attribution to assess the discriminative ability of different sets of features. As the problem of closed-set attribution in a real-world forensic scenario usually involves 2 (rarely 3) possible authors (Litvinova and Gromova, 2020), we performed pairwise classification. For each of the 6*5/2=15 pairs of the authors, the authorship of each written text was estimated on the basis of the remaining oral texts of the pair of authors. The model based on ***Total_set*** has the highest averaged accuracy (**67.63%)** and the lowest SD (15.85). ***Morpho*** has averaged accuracy of **63.17**% with high SD (30), which means that these features are extremely useful in separating certain authors but quite useless to separate the others.

We also performed pairwise leave-one-out cross-validation using Cosine Delta. An increase in the number of texts for training naturally leads to that in the accuracy rate, although the discrimination ability of feature sets is still author-dependent as the analysis of the confusion matrix suggests. ***Morpho*** performed better than the other sets in terms of the averaged accuracy – **73.24 %** (SD=16.58), while the averaged ***Total_set*** accuracy in this scenario equals **90.41** % (SD = 5.9). Of course, these observations should be taken with caution due to the very limited number of texts in this experiment, but the proposed feature set (***Morpho***) is obviously in need of further investigation on a broader corpus in terms of its discriminative ability in the cross-modal scenario – both separately and in combination with other features.

5 Conclusions

There are several lessons we have learned from this study. Firstly, determining the authors even in a closed-set (the simplest in comparison with open-set one) scenario with a limited number of texts per author different in genre and mode using stylometric features and techniques is an extremely difficult task. Secondly, the stylometric features we tested are dependent on the combined factor ***Mode_Author*****,** i.e., these features are affected by mode change but in an author-dependent way. Discourse features were the only feature set that represented the main effect for mode. A new type of features related to morphological diversity of the texts introduced in this paper turned out to be the least mode-dependent and allowed us to distinguish certain authors with high accuracy. Their effectiveness and dependencies on the other factors of idiolectal variation (first of all genre) should be examined more thoroughly in future studies, including the cross-modal scenario.

Thirdly, a real picture of intra- and interidiolectal variation is very complex and mode is an important factor of intraidiolectal variation. Researchers in AA should not limit themselves to the analysis of written texts taking into account a huge number of oral texts (vlogs, etc.) that appeared lately on the Internet as well as the development of the speech-to-text services.

References

Aragón, Garazi J. 2016. *An analysis of authorship attribution: Identifying linguistic variables in oral and written discourse*. Madrid: Universidad complutense Madrid. Master's Thesis.

Baayen, Harald, Hans van Halteren, Anneke Neijt & Fiona Tweedie. 2002. An experiment in authorship attribution. In Annie Morin & Pascale Sébillot (eds.), *JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles, Saint-Malo, France, 2022*, 29–37. St. Malo: Université de Rennes.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller & Akitaka Matsuo. 2018. Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30). https://doi.org/10.21105/joss.00774

Binongo, Jose & Marc A. Smith. 1999. The application of principal component analysis to stylometry. *Literary and Linguist Computing* 14. 445–466.

Brown, Cati, Tony Snodgrass, Susan J. Kemper, Ruth Herman & Michael A. Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods* 40. 540–545.

Čech, Radek & Miroslav Kubát. 2018. Morphological richness of text. In Masako Fidler & Václav Cvrček (eds.), *Taming the corpus. From inflection and lexis to interpretation*, 63–77. New York: Springer International Publishing.

Eder, Maciej, Jan Rybicki & Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *R Journal* 8(1). 107–121.

Fergadiotis, Gerasimos, Heather Harris Wright & Samuel B. Green. 2015. Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech Language, and Hearing Research* 58. 840–852.

Goldstein-Stewart, Jade, Ransom Winder & Roberta Sabin. 2009. Person Identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009). Athens, Greece, 2009*, 336–344. Athens: Association for Computational Linguistics.

Heylighen, Francis & Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 7. 293–340.

Husson, Francois, Julie Josse & Jerome Pagès. 2010a. Principal component methods – hierarchical clustering – partitional clustering: why would we need to choose for visualizing data. http://factominer.free.fr/more/HCPC_husson_josse.pdf

Husson, Francois, Le Sebastien & Jerome Pagès. 2010b. *Exploratory Multivariate Analysis by Example Using R*. Boca Raton: Chapman & Hall/CRC Press.

Kassambara, Alboukadel. 2017. *Practical Guide To Principal Component Methods in R (Multivariate Analysis Book 2)*. CreateSpace Independent Publishing Platform.

Kestemont, Mike, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast & Benno Stein. 2019. Overview of the cross-domain authorship attribution task at PAN 2019. In *Notebook Papers of CLEF 2019 Labs and Workshop*s, Lugano, Switzerland, 2019, 1–15. Lugano: CEUR-WS.org.

Kubát, Miroslav, Vladimír Matlach & Radek Čech. 2014. *QUITA – Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Kulig, Andrze, Jaroslaw Kwapien, Tomasz Stanisz & Stanislaw Drozdz. 2017. In narrative texts punctuation marks obey the same statistics as words. *Information Science* 375. 98–113.

Lê, Sebastien, Julie Josse & Francois Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25. 1–18.

Litvinova, Tatiana & Anastasiya Gromova. 2020. *Компьютерные технологии в судебной автороведческой экспертизе: проблемы и перспективы использования* [Current problems of forensic authorship analysis and the possibility of their solution with the use of computer methods: problems and prospects]. *Science Journal of VolSU. Linguistics* 19. https://doi.org/10.15688/jvolsu2.2020.1.7

Litvinova, Tatiana, Olga Litvinova & Pavel Seredin. 2018. Assessing the level of stability of idiolectal features across modes, topics and time of text production. In *Proceedings of FRUCT 2018*, *Bologna Italy*, *2015*, 223–230. Helsinki: FRUCT Oy.

Litvinova, T. 2021. RusIdiolect: A New Resource for Authorship Studies. In Tatiana Antipova (eds) Comprehensible Science. ICCS 2020. Lecture Notes in Networks and Systems, vol 186. P. 14–23 Springer, Cham. https://doi.org/10.1007/978-3-030-66093-2_2

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy & Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Mikros, George K. & Eleni K Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, *Amsterdam, Netherlands*, *2007*, 1–7. Amsterdam: CEUR.org.

Neal, Tempestt J., Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, Damon L. Woodard. 2018. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)* 50. 1–36.

Oakes, Michael P. 2019. Statistics for Corpus Linguistics. Edinburgh University Press. 272 pp.

Ján Mačutek, Emmerich Kelih

Free or not so free? On stress position in Russian, Slovene, and Ukrainian

Abstract: Stress position in three languages with free stress is scrutinized. Although there are no deterministic rules for stress in such languages, some statistical tendencies are clearly observable. Simple mathematical models for the mean stress position and for the relative mean stress position are suggested. As a "by-product," we show that word length distributions differ among different parts of speech.

Keywords: free stress, stress position, word length, Slavic languages

1 Introduction

This contribution provides some quantitative insight into the stress position in three Slavic languages (Russian, Slovene, and Ukrainian). Stress in these languages can be placed on any syllable and is therefore often described as free or unpredictable (Hyman 1977). While such a statement is true in the sense that there are no deterministic rules for stress position, the stress position displays some statistical tendencies.

In the first step, we summarize some results from older quantitative stress studies in Russian (cf. the overview by Kempgen 1995: 32–35, who re-analyzed data by Moiseev 1976). The data indicate that stress in Russian occurs predominantly in the second half of words and that it gravitates towards the middle of the word. Thus, based on these findings, a tentative assumption about a general "tendency towards the centre" can be made. In the second step, we will present new results from two other Slavic languages with free stress, namely Slovene and Ukrainian. In particular, we will present data on the stress position (also distinguishing different parts of speech). Some (preliminary) mathematical models, which predict the stress position as a function of word length, will be developed.

Acknowledgment: J. Mačutek was supported by the grant VEGA 2/0096/21.

Ján Mačutek, Mathematical Institute, Slovak Academy of Sciences and Constantine the Philosopher University in Nitra, e-mail: jmacutek@yahoo.com
Emmerich Kelih, University of Vienna, e-mail: emmerich.kelih@univie.ac.at

https://doi.org/10.1515/9783110763560-008

Finally, our findings allow us to formulate some preliminary rules for free stress (for the time being valid for the analyzed Slavic languages): it is free, but not entirely random (the distribution of stressed syllables is not uniform, but certain medium positions are preferred), and it is predictable (albeit not deterministically, but only stochastically).

2 Linguistic stress

Stress is an important suprasegmental feature of sounds, syllables, words, multiword expressions, and sentences. Hence there are multiple possibilities of a quantitative analysis of stress and stress position, a domain which is notoriously less researched in quantitative linguistics. Therefore, our contribution is focused on word accent in Slavic languages only, where already some quantitative studies and data are available. Word stress is usually related to a stressed syllable, being characterised by pitch change, greater duration, and a greater intensity of a vowel (Clark & Yallop 1995: 328ff).

In phonetics, different kinds of accents are distinguished, e.g., a dynamic one, usually named as stress, and a "musical" pitch accent. Another distinction concerns the position of the accent within words, which can be either fixed or free. Fixed stress is positionally restricted, as, for instance, in Czech, Latvian, Finnish, Hungarian, etc. A more fine-grained distinction can be made when one distinguishes whether stress is, e.g., on the first (initial) syllable, the last (final) syllable, or the penultimate syllable. Hyman (1977) and Gordon (2002) provide empirical data on stress position in the languages of the world. Based on Hyman's (1977) analysis of over 300 languages, it appears that approximately 37% of them have initial stress, 31% final stress, 25% penultimate stress, and the remaining few have either peninitial or antepenultimate stress. According to Hyman (1977), only these five stress systems are attested in world fixed stress languages.

In languages with free stress (like e.g., Russian, Bulgarian, Spanish, Italian, English, etc.), the freedom of stress is utilized for marking differences in the meaning (e.g., in German *umf**a**hren* 'to drive round' vs. ***u**mfahren* 'to knock someone over') or for marking different parts of speech (e.g., in English the noun *pr**o**gress* vs. the verb *progr**e**ss*). Free-stress languages seem to be less present, since according to the analysis by Hyman (1977), 138 out of 444 languages (31%) are of this type. Stress in these languages can occur practically anywhere; however, from a morphological perspective, particular rules and tendencies of placement of the stress can be obtained as, for instance, a tendency towards stress placement on the morphological root or on prefixes and suffixes.

From a statistical point of view, one has to ask about stochastic tendencies of the stress position, i.e., are there some preferred patterns to be found in the stress placement? It seems unlikely that the stress placement in so-called free stress languages is indeed absolutely random[1] (in the sense of the uniform distribution, i.e., with each syllable being equally probable to carry stress).

3 Stress in Slavic languages

Slavic languages form an important group of the Indo-European language family. Members of this group comprise both fixed stress and free stress languages. Regarding stress, there is an interesting internal differentiation within Slavic languages (cf. Krüger 2009). On the one hand, all contemporary West Slavic languages have fixed stress (Polish has penultimate stress, while Czech, Slovak, and Sorbian have initial stress). On the other hand, within the South Slavic languages, only Macedonian is characterized to have a fixed stress system (antepenultimate; cf. however, Kempgen 2008, who observes a switch to the initial stress placement). All other South Slavic languages (Bosnian, Bulgarian, Croatian, Serbian, and Slovene) have free stress. The same holds for all contemporary East Slavic languages (Belarusian, Russian, and Ukrainian).

As far it is known from historical and comparative linguistics, Proto-Slavic was a language with free stress (cf. Baerman 1999). Thus, seen from a historical perspective, the introduction of fixed stress systems seems to be an innovation among Slavic languages.

4 Language material and data

4.1 Russian

The Russian stress system has been analysed in much detail (cf. Lehfeldt 2012), also by means of quantitative analyses. Regarding its quantitative properties, in

1 In some languages, there are several linguistics factors which seem to determine the stress placement, even if the language is considered to have "free" stress position. Therefore, languages with a predictable free stress are sometimes distinguished. Phonotactic restrictions (syllable weight, long vs. short vowels in case of quantity-sensitive languages), parts of speech, conjugation and declination patterns, intonation, and pragmatic-contextual realizations of speech are, among others, discussed as relevant factors.

particular, Kempgen (1995: 32–35) has to be mentioned due to his summary of the most important empirical observations of the accent position. In Russian, as can be expected for a free stress language, practically any syllable of the word can be stressed, but every single word form is characterized by one particular stress position (there are, of course, a few examples where a variation in stress can be observed). However, it has been recognized in the past by many researchers, as reported by Kempgen (1995: 32), that there is an obvious tendency towards the centre, i.e., towards the syllable in the middle of a particular word.

This observation has been further specified in two aspects. First, stress occurs but rarely in the peripheries of words (i.e. in the initial and in the final position), which is a natural consequence of the preference towards the centre. Next, the second half of the word is clearly preferred. What this means in the empirical dimension can be seen in Tab. 1, where data on stress positions in 49,483 Russian word forms taken from an orthoepic Russian dictionary are presented (cf. Moiseev 1976 for the original counts and further details).

Tab. 1: Stress position in Russian (WL – word length, SP – stress position) according to Kempgen (1995: 34), based on Moiseev (1976).

SP	1	2	3	4	5	6	7	8
WL								
2	3834	5183						
3	3138	8224	6189					
4	447	5241	6139	1440				
5	50	340	3685	1970	324			
6	2	59	332	1592	388	55		
7		1	24	199	405	66	6	
8			1	7	50	63	3	
9					3	7	7	
10					1	1	2	4
11								1

A closer look at the data reveals a remarkable mechanism in stress placement in Russian. As already pointed out by Kempgen (1995: 33), one can easily observe that the peak of every single frequency distribution systematically moves with the increasing word length from the left edge to the right edge of the word form. This relation between word length and stress position can be, according to Kempgen (1995: 33), represented by a simple linear model, where the average position of stress can be predicted based on word length – the longer the word form is, the more to the right (counted from the beginning of word forms) stress

tends to be placed. In any case, the data give strong evidence that the stress position in Russian is a statistical mass phenomenon, which is predictable based on word form length.

4.2 Slovene

Slovene is a representative of South Slavic languages. For Slavic linguistics, Slovene is of particular interest because, in the contemporary language, both a pitch accent and a stress accent can be observed. This fact reflects a rather complex situation in the various dialects of the language. Moreover, contrary to Russian, Slovene is also quantity-sensitive, but the length of vowels is inherently connected with stress (cf. Priestly 1993: 390).

According to our knowledge, Slovene has never been analysed statistically with respect to stress position. We analysed the stress position in Slovene on the basic vocabulary from a Slovene-German learner's dictionary (cf. Kelih & Vučajnk 2018). Moreover, we are in a position to distinguish different parts of speech. Our data basis is as follows: 1522 nouns, 273 adjectives, and 528 verbs. Lemmas are taken as basic units (i.e., in the case of nouns, word forms in nominative singular; adjectives in the short form of masculine nominative singular, as commonly given in dictionaries of Slovene; and verbs as infinitives). The lemma analysis thus excludes (and this is important to note) the possible change of the stress position in other cases, e.g., Nom. Sg. *m**o**st* ('bridge') is considered, but Gen. Sg. *most**u*** is not. One further characteristic of free stress in Slovene is a remarkable tendency towards variation. There are many word forms where an alternate stress positioning would be allowed, e.g., *z**i**dati* and *zid**a**ti* ('to build', 'to construct'), both common in the Slovene Standard language. However, for the sake of simplicity, for our analysis, we took the accent position, which is mentioned in the most important dictionary of contemporary Slovene in the first position (cf. SSKJ 2014 at fran.si). Frequencies of stress positions in Slovene can be found in Tab. 2.

It is obvious that stress tends to occur in the middle of words and that it mostly avoids word peripheries. However, as opposed to Russian, there is no clear preference for the second half of the word. In shorter nouns and adjectives (those consisting of two and three syllables), stress on the first half of words prevails, while in longer ones, the opposite is true. Verbs are an extra category, as the infinitive of Slovene verbs ends in the suffix *-ti* (there are a few exceptions ending in the suffix *-či*), which never carries stress. Regardless of verb length, the penultimate position (i.e., the placement of stress on the last possible syllable) is the most frequent.

Tab. 2: Stress position in Slovene (SP – stress position, WL – word length).

	nouns					adjectives				verbs				
SP	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**5**
WL														
1	171					34								
2	377	166				78	26			42				
3	95	333	55			15	67	15		97	145			
4	10	86	156	6		1	7	20	4	1	86	101		
5		9	23	26				3	3			21	22	
6				7	2								1	2

4.3 Ukrainian

Ukrainian belongs, alongside Russian, to the subgroup of Eastern Slavic languages (it is, therefore, interesting to check for similarities and differences between these two languages). The situation concerning quantitative-oriented investigations on stress position in Ukrainian is similar to that of Slovene – we are not aware of any published results (Łukaszewicz & Mołczanow 2018: 259 even write that Ukrainian "remains a *terra incognita* in phonological literature").

The Ukrainian data we present in Tab. 3 were created from the corpus of Ukrainian works of fiction (see http://www.mova.info/cfq1.aspx?fdid=hproz2018). We chose the same approach as for Slovene, i.e., we considered nouns, adjectives, and verbs separately, and we worked with word lemmas (nominative singular for nouns, masculine nominative singular for adjectives, infinitive for verbs). First, 1500 most frequent lemmas from each of the three parts of speech were taken from the corpus. Some of them were then disregarded, because either they were not included in the Ukrainian-Slovak dictionary by Popeľ (1960), which was used to determine stress positions, or stress position was ambiguous (the dictionary offers two possibilities for some words, without an indication which one is preferred, occurs more often, etc.). Thus, we analysed a sample consisting of 1221 nouns, 1122 adjectives, and 1369 verbs.

Once again, syllables in the centre of a word carry stress more often than those at word peripheries. As far as the influence of word length is concerned, the Ukrainian data display a tendency similar to those from Slovene – in two-syllable words, the beginning of the word is preferred; three-syllable ones do not provide a clear picture; while in longer words, the second half is stressed more often.

Tab. 3: Stress position in Ukrainian (SP – stress position, WL – word length).

	nouns					adjectives					verbs				
SP	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
WL															
1	208					7					1				
2	380	201				179	90				52	27			
3	56	225	62			61	426	44			128	310	35		
4		30	39	3		5	85	170	2		60	203	291	4	
5			10	3	2			22	28		5	63	117	45	
6				2					1	2		5	13	6	2
7													1	1	

5 Results

5.1 Stress position

For all three languages under study, it is obvious (see Tab. 4, which also contains a column with Ukrainian nouns and adjectives, i.e. verbs were excluded, the reason to be explained later in this section) that the mean stress position moves to the right as word length increases – mathematically speaking, if $MSP(x)$ is the mean stress position in x-syllabic words, $MSP(x)$ is an increasing function of the variable x. In addition, we also introduce the notion of the relative mean stress position in x-syllabic words,

$$RMSP(x) = \frac{MSP(x) - 1}{x - 1}$$

The relative mean stress position can be interpreted as the "centre of gravity" with respect to stress. For example, the stress in a three-syllabic word can be on either of the first, second, or third syllable. One can imagine it as a line segment with endpoints 1 and 3, i.e., as a line segment of length two. The formula above normalizes the segment so that its endpoints are 0 and 1, with possible positions of stress adjusted accordingly (i.e., in the case of a three-syllable word, possible stress positions are 0, 0.5, and 1, which correspond to the first, second, and third syllable, respectively). The values of the relative mean stress positions can be found also in Tab. 4. We note that in Tab. 4, we disregarded the single word of length 11 in Russian and two words of length 7 in Slovene, as we considered these samples too small to be used in our analyses.

Tab. 4: Dependence of mean stress position (MSP) and relative mean stress position (RMSP) in Russian, Slovene, and Ukrainian on word length (WL).

	Russian		Slovene		Ukrainian		Ukrainian (without verbs)	
	MSP	**RMSP**	**MSP**	**RMSP**	**MSP**	**RMSP**	**MSP**	**RMSP**
WL								
2	1.57	0.57	1.28	0.28	1.34	0.34	1.34	0.34
3	2.21	0.61	1.83	0.42	1.92	0.46	1.99	0.50
4	2.65	0.55	2.58	0.53	2.51	0.50	2.64	0.55
5	3.34	0.59	3.39	0.60	3.02	0.51	3.45	0.64
6	4.02	0.60	4.33	0.66	3.39	0.48	4.40	0.68
7	4.75	0.63						
8	5.48	0.64						

The mean stress position can be modelled by the linear function

$$MSP(x) = ax + b$$

which achieves an excellent fit for all three languages ($a = 0.65$, $b = 0.20$, $R^2 = 0.99$ for Russian; $a = 0.77$, $b = -0.38$, $R^2 = 0.99$ for Slovene; and $a = 0.52$, $b = 0.36$, $R^2 = 0.99$ for Ukrainian with all words taken into account).

The relative mean stress positions are increasing only for Slovene. However, for Russian, only three values are problematic (the sequence of relative mean stress positions is increasing for word lengths from 4 to 9). First, there is a decrease at the very end, but the sample contains only eight words of length 10 (see Tab. 1), hence the robustness of the value is questionable, and it can easily change if more words are added. Second, an increase can be observed at the beginning. We remind that the Russian sample differs from the other two in two aspects: (a) it is comprised of all word forms, not only lemmas, (b) there are no restrictions with respect to parts of speech, while only nouns, adjectives, and verbs were chosen for Slovene and Ukrainian (where frequently used items were analysed). For the time being, we do not know anything about the behaviour of stress in particular parts of speech in Russian, but as most synsemantic words are short, they are represented in higher proportion among shorter words. In addition, the stress in some Russian words is movable, e.g., *ruk**a*** ('hand') in Nom. Sg., but *r**u**ku* in Acc. Sg, *r**u**ki* in Nom. Pl. Kempgen (1995: 34) claims that a movable accent occurs much more often in very frequent words, but according to the famous Zipf's law of brevity (cf. Zipf 1949, or recently Casas et al. 2019), these words tend to be short. Consequently, some fluctuations in stress positions can be

caused by movable stress, and possibly also by different stress patterns in synsemantic words (which is an open question).

The only irregularity in Ukrainian can be observed at the end for words of length 6, and it disappears if verbs are disregarded. As opposed to Slovene, reflexive verbs in Ukrainian are written, according to the orthography of the language, together with the reflexive suffix *-sya*, which never carries stress,[2] e.g., бо**я**тися [boy**a**tisya] in Ukrainian vs. ***ba**ti se* (both: 'to be afraid') in Slovene. Thus, reflexive verbs are "artificially" made longer in Ukrainian, and stress can never fall on their last syllables. This fact could – albeit speculatively – explain the decrease in the relative mean stress position.

If the Ukrainian words are re-analysed with verbs omitted (the data can be found in Tab. 4), and if we consider only Russian words of length from 4 to 9, the fit of the linear model to the mean stress positions remains excellent, and parameter values in the model differ less among the three languages ($a = 0.67$, $b = -0.01$, $R^2 = 0.99$ for Russian; $a = 0.76$, $b = 0.23$, $R^2 = 0.99$ for Ukrainian). The relative mean stress positions in the re-analysed data can be modelled by the function

$$RMSP(x) = 1 - cx^d$$

with $c = 0.83$, $d = -0.41$, $R^2 = 0.99$ for Russian; $c = 1.15$, $d = -0.66$, $R^2 = 0.99$ for Slovene; and $c = 1.03$, $d = -0.64$, $R^2 = 0.95$ for Ukrainian. The limit of the function is 1 for increasing x, which means that it respects the upper theoretical limit of the relative mean stress position. The values of the parameters c and d are strongly correlated (with the Pearson correlation coefficient –0.95), which indicates that a model with only one parameter (which would be easier to interpret) might be sufficient.

The overall relative mean stress position evaluated from all data, which are presented in Tabs. 1–3, is 0.58 for Russian, 0.41 for Slovene, and 0.44 for Ukrainian.[3] The relative mean stress with the value of 0.5 would mean that, on average, exactly the middle of a word is stressed; one can see that in Russian, the second half of words is preferred, with the opposite tendency in Slovene and Ukrainian. It remains an open question whether this difference is caused by different sample methodologies (all word forms vs. word lemmas; basic

2 The same orthographic principle is followed in Russian.

3 If only words with length from 4 to 9 are considered in Russian, and if verbs are omitted in Ukrainian, the overall relative mean stress positions do not change much – they attain values 0.57 for Russian and 0.45 for Ukrainian.

vocabulary vs. the most frequent words vs. no such restrictions; all parts of speech vs. nouns, adjectives, and verbs).

5.2 Word length in different parts of speech

Our research brings also an interesting "by-product" – for Slovene and Ukrainian. It is possible to test whether word length distribution differs among nouns, adjectives, and verbs.[4] Data are presented in Tab. 5. We remind that the data represent the length of word types (as they were taken from dictionaries) as opposed to tokens (which is a more usual approach in word length studies, cf. Grzybek 2006).

Tab. 5: Word length distribution for nouns, adjectives, and verbs in Slovene and Ukrainian.

	word length	1	2	3	4	5	6	7
Slovene	nouns	171	543	483	258	58	9	
	adjectives	34	104	97	32	6	0	
	verbs	0	42	242	198	43	6	
Ukrainian	nouns	208	581	343	72	25	2	0
	adjectives	7	269	531	262	56	3	0
	verbs	1	79	473	558	230	26	2

The chi-square test reveals that if word length distributions of all three parts of speech are tested, the null hypothesis (which says that the proportions do not differ among parts of speech) is rejected for both languages (p-value less than 0.01 in both cases). However, the difference between nouns and adjectives (i.e., without verbs) remains significant for Ukrainian, but not for Slovene (with a p-value equal to 0.11)

6 Conclusion

It is a well-known fact that even in languages with fixed stress, not all words "behave" as they should according to their typological affiliation. According to

4 We do not have this possibility for Russian, as data which we took from Kempgen (1995) (originally from Moiseev 1976) do not distinguish among parts of speech.

Hyman (1977: 56), for instance, in languages that are supposed to have final stress, the final syllable indeed carries stress only in 90% of the word forms. This observation leads to the question of where, in fact, to draw the line between languages with fixed and free stress. The behaviour of stress is, in principle, stochastic.

A similar problem appears, as we have shown, in languages with free stress. The freedom of stress position is by no means unlimited, but particular patterns and tendencies can be observed. Thus, one has to agree with Hyman (1977: 56), who asks: "How do we decide what is 'dominant'?" There is no trivial answer, but simple counting already provides remarkable hints for further discussion.

The data and analyses presented above give us a possibility to summarize and generalize tendencies in stress positions up to a certain extent (we are aware of the fact that the analysis of three languages from one language family is of limited relevance). There seems to be quite a systematic trend in languages with free stress – the stress prefers positions in the middle of words and moves towards the end of the word with the increasing word length. This trend must be verified in several typologically diverse languages.

Our investigations also suggest several directions for future research. First, data from this paper were obtained by two different approaches (all word forms in Russian vs. word lemmas in Slovene and Ukrainian). Both approaches are relevant, but they are not directly comparable. All three languages under study are fusional, and different cases are marked by suffixes which make word forms longer. Moreover, there is also the phenomenon of the moving stress, which has an impact on stress positions in word forms, but cannot be taken into account if we restrict ourselves to lemmas. Second, stress position can be studied not only in dictionaries (i.e., on word types), but also in texts on word tokens (where proportions of shorter words will be higher than in dictionaries). Finally, it seems that word length behaviour (we remind that word length is one of the main factors which determine stress position) differs among different parts of speech, which is a topic that deserves systematic research, again on the levels of both types and tokens.

References

Baerman, Matthew. 1999. *The Evolution of Fixed Stress in Slavic*. München: Lincom.

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.

Cubberley, Paul V. 1980. The Suprasegmental Features in Slavonic Phonetic Typology. (Bibliotheca Slavonica 20). Amsterdam: Hakkert.

Clark, John & Colin Yallop. 1995. *An Introduction to Phonetics and Phonology*. Oxford: Blackwell.

Gordon, Matthew. 2002. A factorial typology of quantity-insensitive stress. *Natural Language & Linguistic Theory* 20(3). 491–552.

Grzybek, Peter. 2006. History and methodology of word length studies. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, 15–90. Dordrecht: Springer.

Hyman, Larry M. 1977. On the nature of linguistic stress. In Larry M. Hyman (ed.), *Studies in Stress and Accent. Southern California Occasional Papers in Linguistics 4*, 37–82. Los Angeles: University of Southern California.

Kelih, Emmerich & Tatjana Vučajnk. 2018. Slovensko-nemški tematski slovar: osnovno in razširjeno besedišče. 4500 gesel, frazemov in stavčnih primerov. Grund- und Aufbauwortschatz Slowenisch-Deutsch. 4500 Lemmata, Phrasen und Satzbeispiele. Klagenfurt/Celovec-Ljubljana/Laibach-Wien/Dunaj: Hermagoras/Mohorjeva.

Kempgen, Sebastian. 1995. *Russische Sprachstatistik*. München: Otto Sagner.

Kempgen, Sebastian. 2008. Das Makedonische – auf dem Weg zur Anfangsbetonung? In Gabriel Altmann, Iryna Zadorozhna & Yuliya Matskulyak (eds.), *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th Anniversary of Professor V.V. Levickij*, 311–318. Chernivtsi: Knihi-XXI.

Krüger, Kersten. 2009. Freier Akzent (Flexion). In Tilman Berger, Karl Gutschmidt, Sebastian Kempgen & Peter Kosta (eds.), *The Slavic Languages. An International Handbook of Their History, Their Structure and Their Investigation. Volume 1*, 86–100. Berlin/New York: de Gruyter.

Lehfeldt, Werner. 2012. Akzent und Betonung im Russischen, 2., verbesserte und erweiterte Auflage. München: Sagner.

Łukaszewicz, Beata & Janina Mołczanow. 2018. Leftward and rightward stress iteration in Ukrainian. In Bartłomiej Czaplicki, Beata Łukaszewicz & Monika Opalińska (eds.), *Phonology, Fieldwork and Generalizations*, 259–275. Berlin: Peter Lang.

Moiseev, Aleksandr. 1976. Mesto slovesnogo udarenija v sovremennom russkom literaturnom jazyke. *Studia Rossica Posnaniensia* 7. 77–87.

Popeľ, Ivan. 1960. *Ukrajinsko-slovenský slovník* [Ukrainian-Slovak dictionary]. Bratislava: Slovenské pedagogické nakladateľstvo.

Priestly, Tom M. S. 1993. Slovene. In Bernard Comrie & Greville G. Corbett (eds.), *The Slavonic Languages*, 388–451. London: Routledge.

SSKJ. 2014. *Slovar Slovenskega Knjižnega Jezika. Prva knjiga A-Pa & Druga knjiga Pe-Ž* [Dictionary of Slovene literary language. The first volume A-Pa & The second volume Pe-Ž], 2nd edn. Ljubljana: Cankarjeva založba.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge (MA): Addison-Wesley.

Jiří Milička, Václav Cvrček, David Lukeš

Unpacking lexical intertextuality: Vocabulary shared among texts

Abstract: This paper focuses on lexical intertextuality, namely the three following intertextual properties: 1) the number of word-types shared by two texts; 2) the number of word-types shared by all texts in a collection; 3) the number of word-types shared by equal-sized segments of a collection. We have observed that the relation between the number of texts and the number of shared types follows a power law; similar behavior can be seen if text borders are disregarded and the corpus is artificially divided into equal-sized segments. The number of shared types is proportional to the size of these sequences. We developed baseline models for the number of shared types, i.e. models predicting the number of types shared by texts if all tokens were randomly shuffled and evenly spread among texts. The comparison between the empirical data and the baseline model can be used for contrastive purposes, to compare the number of shared types in corpora of different languages.

Keywords: lexicon, intertextuality, corpus, random model, number of types

1 Introduction

It is a deep-seated conviction of mainstream quantitative linguistics that quantitative regularities (or laws) are to be derived only from within texts as naturally occurring and homogeneous units of language (see Köhler – Altmann 2005). Supposedly, linguistic laws applicable to isolated texts are blurred or distorted (cf. Strauss – Grzybek – Altmann 2007, Čech – Kosek – Mačutek – Navrátilová 2020) when observed on text aggregates (e.g. in a language corpus).

In contrast to this, we believe that insisting on the conceptualization of language as a sum of individual texts would be a significant oversimplification. It has been pointed out several times in qualitative studies that intertextual links are an integral part of the textual ecosystem (e.g. Teubert 2005) and play an important role in facilitating communication. Quantitative examples of these

Jiří Milička, Charles University, e-mail: milicka@centrum.cz
Václav Cvrček, Charles University, e-mail: vaclav.cvrcek@ff.cuni.cz
David Lukeš, Charles University, e-mail: david.lukes@ff.cuni.cz

https://doi.org/10.1515/9783110763560-009

links can be found on various levels, e.g. the frequency of a word in one text can be estimated from its frequencies in other texts (which enables comprehension in situations when the word is missing or cannot be recognized). We would thus argue that texts are not independent entities, but rather an interconnected ecosystem which influences even basic linguistic characteristics such as frequencies of units.

In the present paper, we focus on lexical intertextuality in relation to the variability of texts. Specifically, we study the three following intertextual properties:

1. the number of word-types shared by two texts;
2. the number of word-types shared by all texts in a collection;
3. the number of word-types shared by equal-sized segments of a collection.

In all three cases, relations between the number of texts and number of shared types are investigated and the emerging laws of lexical intertextuality are explored. Moreover, empirical values are measured, and theoretically motivated baseline models are derived. These quantitative models are meant to provide a baseline for comparison with authentic linguistic data: they predict the expected number of types shared by all texts in a hypothetical corpus whose tokens have been randomly shuffled among the texts. The randomization component of these models distorts the structure of typical real-world intertextual relations (in real texts, tokens of a given type are typically more likely to clump together, as opposed to evenly spreading out across the corpus), which provides us with a backdrop against which this structure can be studied.

2 Data

For empirical testing, we used SYN2015 (Křen et al. 2015), a 100M word representative corpus of contemporary written Czech. The corpus consists of texts from 1990 onwards (with most of them published in the timespan 2010–2014). As for genre and register composition, it comprises texts from three major domains: fiction, non-fiction and newspapers, and thus covers a large spectrum of non-private printed texts. For details about the corpus composition, see Křen et al. 2016.

3 Comparing two texts

We begin by examining a special case of our question: how many shared word types can we expect in two texts randomly chosen from a corpus? In real life, the shared vocabulary of a set of texts depends on many non-trivial factors

(their topics, authors, genres or registers, grammatical constraints, etc.). Quantifying any single one of them is challenging enough, let alone integrating all of them into a comprehensive model which would predict the expected number of shared types in response to these factors. In order to evaluate the observed number of shared types we can establish as a reference point a theoretical baseline value for the average number of shared types in a hypothetical scenario where the factors are levelled out. This is achieved by randomizing the tokens appearing in both texts, which makes all tokens equally likely to occur at any position within the two texts. As we will see below, this yields a much simpler model which only depends on text lengths and type frequencies.

4 Baseline model

Let us consider a corpus consisting of two texts, the first one comprising the set of word types A and being a tokens long, the second one comprising the set of word types B and being b tokens long. The word type T has absolute frequency t in the corpus.

If we randomly shuffle the corpus tokens, we can derive the probability that type T is represented only in text A and not in text B, based solely on the combinatorial properties of the system and ignoring the tendency of tokens of the same type to clump together:

$$p(T \in A \wedge T \notin B) = \frac{\binom{a}{t}}{\binom{a+b}{t}} \tag{1}$$

Hence the probability that type T is only in text A, or only in text B:

$$p(T \notin A \vee T \notin B) = \frac{\binom{b}{t} + \binom{a}{t}}{\binom{a+b}{t}} \tag{2}$$

The corpus consists of only two texts. Thus the probability that type T is *both* in text A and text B:

$$p(T \in A \wedge T \in B) = p'(T \notin A \vee T \notin B) = 1 - \frac{\binom{b}{t} + \binom{a}{t}}{\binom{a+b}{t}} \tag{3}$$

To get C, the average count of types that are represented in both randomized texts, we need to sum up the probability $p(T \in A \wedge T \in B)$ for all types in the text ($A \cup B$ being the set of all types in the corpus).

$$C_{A,B} = \sum_{T \in A \cup B} 1 - \frac{\binom{b}{t} + \binom{a}{t}}{\binom{a+b}{t}} \tag{4}$$

5 Practical implementation

In order to avoid large numbers, let us transform the formula into a recurrent form. First let us do the following substitution:

$$C_{A,B} = \sum_{T \in A \cup B} 1 - \frac{\binom{b}{t} + \binom{a}{t}}{\binom{a+b}{t}} = \sum_{T \in A \cup B} 1 - Z_{t,a} - Z_{t,b} \tag{5}$$

where

$$Z_{t,a} = \frac{\binom{a}{t}}{\binom{a+b}{t}} \qquad Z_{t,b} = \frac{\binom{b}{t}}{\binom{a+b}{t}} \tag{6}$$

Now we can proceed to express the two functions in their recurrent form:

$$Z_{1,a} = \frac{a}{a+b} \qquad Z_{1,b} = \frac{b}{a+b}$$

$$Z_{t,a} = Z_{t-1,a}\frac{a-t+1}{a+b-t+1} \qquad Z_{t,b} = Z_{t-1,b}\frac{b-t+1}{a+b-t+1} \tag{7}$$

6 Empirical data

As a sanity check, we compared the results of the formula with physically randomized texts. Two texts were randomly chosen from the corpus (the length of the first one being 12,639 tokens, and the second one 1050 tokens). The two texts had 203 types in common. After 1M randomizations (i.e. shuffling the tokens of both texts) of the two texts, the average number of shared types was 338.308, which is in accordance with the model that predicts the value of 338.305 shared types. As expected, it is also higher than the 203 types that the texts originally had in common, before shuffling.

The average ratio between the number of shared types in two texts randomly chosen from the corpus and the model value that was counted from the two text's word type frequency distribution is 0.5211 (500,000 trials). The lower the ratio, the higher the tendency of tokens of the same type to clump together, compared to the randomized model. In our corpus, no single pair of texts overcame the ratio of 1, which means that in real-life texts, the types tend to clump together and their distribution is uneven (see the distribution of ratios in Fig. 1). But in principle the ratio can exceed 1 for two very similar or even identical texts.

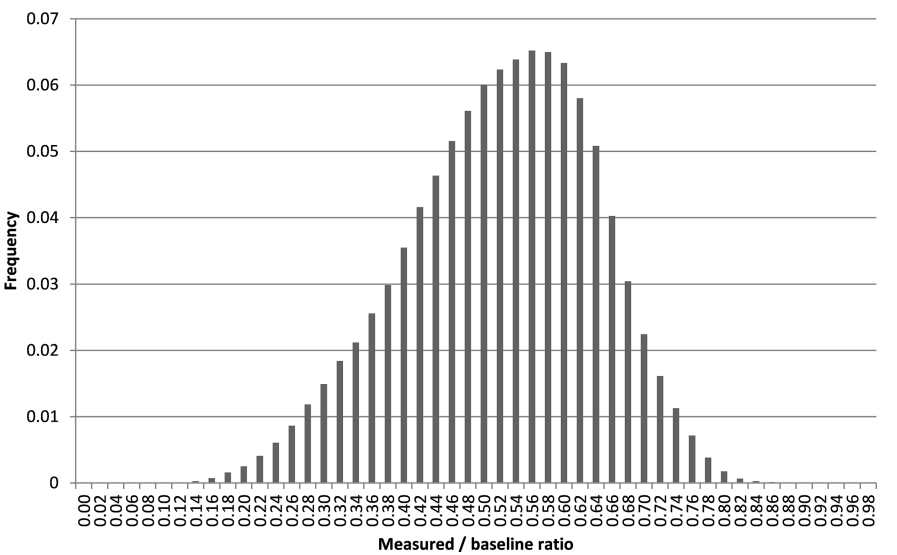


Fig. 1: Distribution of ratios between the actual number of shared types in two randomly chosen texts, and the model upper bound on shared types for those two texts (mean = 0.5211 on 500,000 trials).

7 Comparing multiple texts

Let us now consider a randomized corpus consisting of g texts, each of them having a set X_q of types so that the set of all word types in the corpus is G, with $G = \cup_q^g X_q$. Each of the texts has n_q tokens, the total sum being $N = \sum_q^g n_q$. Following formula 1, we can calculate the probability that type T with overall frequency t in the corpus exclusively occurs in a text q (or in a set of types X_q) as:

$$p\left(T \in X_q \wedge T \notin G - X_q\right) = \frac{\binom{n_q}{t}}{\binom{N}{t}} \tag{8}$$

A key concept for deriving the general model is the probability of the given type T occurring exclusively in the union of a subset of texts. The sum of these probabilities for every possible subset of texts of a given cardinality will be denoted by a preceding lower index, e.g. the sum of the probabilities of T occurring exclusively in the union of any three texts in the corpus is ${}_3h$, calculated as follows:

$${}_3^g h = \frac{\binom{n_1+n_2+n_3}{t} + \binom{n_1+n_2+n_4}{t} + \ldots + \binom{n_{g-2}+n_{g-1}+n_g}{t}}{\binom{N}{t}} \tag{9}$$

Note that formula 2 is a special case of formula 9 for two texts, so $p(T \notin X_1 \vee T \notin X_2) = {}_1^2 h$. In accordance with formula 3, the expression $1 - {}_1^2 h$ denotes the probability that the given type is in both texts of a two-text corpus. In order to generalize formula 2 for a three-text corpus, we must keep in mind that some of the combinations are covered by multiple ${}_2^3 h$ terms and their intersections thus must be subtracted $\left(\text{e.g. both } \binom{n_1+n_2}{t} \text{ and } \binom{n_1+n_3}{t} \text{ cover all the ways to assign all } t \text{ tokens to } X_1, \text{ i.e. } \binom{n_1}{t}\right)$:

$$p(T \notin X_1 \vee T \notin X_2 \vee T \notin X_3) =$$
$$= \frac{\binom{n_1+n_2}{t} + \binom{n_1+n_3}{t} - \binom{n_1}{t} + \binom{n_2+n_3}{t} - \binom{n_2}{t} - \binom{n_3}{t}}{\binom{N}{t}} = {}^{3}_{2}h - {}^{3}_{1}h \tag{10}$$

If we further generalize the formula to four texts, the intersection to subtract becomes ${}^{4}_{2}h - {}^{4}_{1}h$, because as above, ${}^{4}_{2}h$ actually covers some of the combinations twice and they should be subtracted only once:

$$p(T \notin X_1 \vee T \notin X_2 \vee T \notin X_3 \vee T \notin X_4) = {}^{4}_{3}h - \left({}^{4}_{2}h - {}^{4}_{1}h\right) \tag{11}$$

Given this recursion, we can arrive at the general formula for the probability of T occurring in all g texts in the corpus (analogously to formula 3):

$$p\left(\bigwedge\nolimits_{q}^{g} T \in X_q\right) = 1 - \left({}^{g}_{g-1}h - \left({}^{g}_{g-2}h - \left({}^{g}_{g-3}h - \left({}^{g}_{g-4}h - \left(\ldots {}^{g}_{1}h\right)\right)\right)\right)\right) \tag{12}$$

Which can be abbreviated as:

$$p\left(\bigwedge\nolimits_{q}^{g} T \in X_q\right) = 1 - \sum_{q=1}^{g-1}\left((-1)^{(g-q-1)}\,{}^{g}_{q}h\right) \tag{13}$$

As in the previous case, to get the average count of types that are represented in all texts in the corpus, we need to sum up the probability $p\left(\bigwedge_{q=1}^{g} T \in X_q\right)$ for all types in the corpus:

$$C_G = \sum_{T \in G}\left(1 - \sum_{q=1}^{g-1}\left((-1)^{(g-q-1)}\,{}^{g}_{q}h\right)\right) \tag{14}$$

8 Practical implementation

We are not able to make this formula computable for higher numbers of texts, as the number of computations needed scales with the binomial coefficient of the number of texts over half the number of texts, i.e. $\binom{g}{g/2}$. Therefore, we just estimate its value through simulation, by repeatedly shuffling the tokens of the corpus or of the chosen texts randomly.

Note that this problem can be reinterpreted as the **coupon collecting problem** for unequal probabilities without replacement (or non-uniform coupon collecting problem), where the texts represent different coupon types, the tokens represent the coupons and we need to know the probability that after the t^{th} box, our collection of coupons is complete (i.e. that after randomly taking t tokens of type T from the corpus, there will be at least 1 token from each text). Any good solution to this problem will also provide a workable implementation of our baseline model (Wild et al. 2013).

9 Empirical data

We measured both the empirically attested counts of shared types and empirically obtained baseline counts for sets of 2 to 100 randomly chosen texts. For each iteration, 10,000 random samples of the corresponding number of texts (2, 3, . . ., 100) were drawn from the corpus. For each random sample: 1) the number of shared types was determined; 2) the sample was randomized 10 times so that the baseline value could be determined; 3) the empirical value was divided by the baseline value. Subsequently, averages over the 10,000 samples were computed for the values of the number of shared types, the baseline values, and the empirical / baseline ratio. Each iteration thus ultimately yielded one data point for each of the three charts in Figs. 2 and 3.

Figure 2 shows that a power law can be successfully fitted to both the empirical and the baseline relation between the number of texts in the sample and the number of shared types, with the fitting metric being slightly lower for the baseline values, at least for our data. However, the best fit can be obtained for the 1-displaced power function. The empirical data can be fitted with $y = 1624(x-1)^{-1.0731} (R^2 = 0.9997)$, while the baseline relation can be fitted by $y = 3173(x-1)^{-0.9568} (R^2 = 0.992)$. At this stage, we have no explanation for these models, and we admit that additional work is needed to clarify these relationships.

Figure 3 shows the average ratios between the measured empirical value and the baseline model. As can be seen, the ratio follows a "pipe curve" (cf. Cvrček 2014), i.e. the value initially drops, but then it rises slowly. This "pipe curve" can be successfully fitted by the double power function $y = 0.62x^{-2.17} + 0.37x^{0.0367}$, with $R^2 = 0.991$.

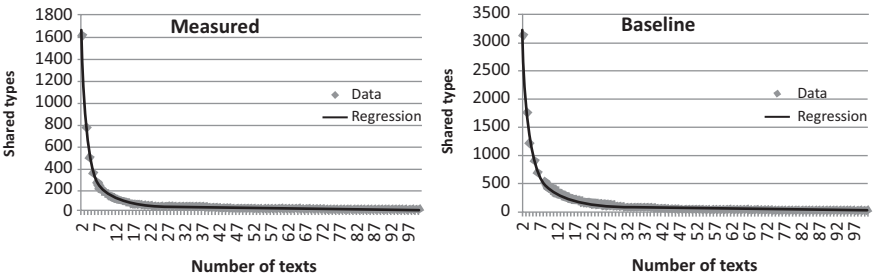


Fig. 2: The number of shared types as a function of the number of texts.

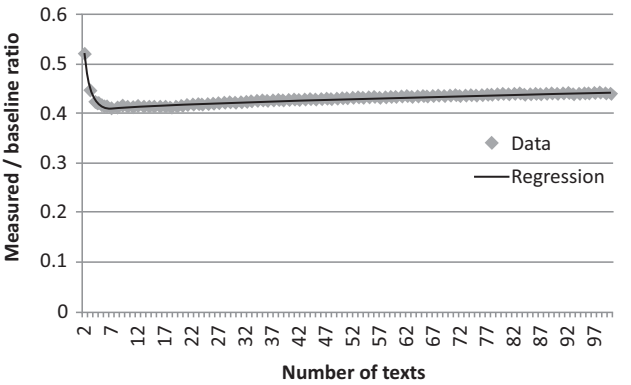


Fig. 3: The relation between the number of texts and the average ratio between the measured value and the baseline model.

10 Comparing equal-sized segments

In many cases, we come across collections of language material which either lack distinctive text borders, or the texts are extremely short (collections of tweets or other social media content, advertisements, political slogans, short messages, internet memes etc.). In such a case, the intersection of types as described in the previous section does not make sense. What we may do instead is to divide the corpus into equally long segments, which makes it possible to apply the same kind of analysis as in the previous section (see Conclusion for suggestions on useful ways of interpreting the results).

11 Baseline model

The division of the corpus into equal-sized segments allows us to use a special case of formula 14 which is computationally much more feasible.

$$C_G = \sum_{T \in G} \left(1 - \sum_{q=1}^{g-1} \frac{(-1)^{(g-q-1)} \binom{g}{q} \binom{nq}{t}}{\binom{ng}{t}} \right) \tag{15}$$

The corpus consists of g segments; each segment comprises n tokens.

12 Practical implementation

In order to avoid large numbers, let us transform the formula into a recurrent form. First, let us do the following substitution of formula 15:

$$C_G = \sum_{T \in G} \left(1 - \sum_{q=1}^{g-1} \left((-1)^{(g-q-1)} Z_{q,t} \right) \right)$$

$$Z_{q,t} = \frac{\binom{g}{q} \binom{nq}{t}}{\binom{ng}{t}} \tag{16}$$

Henceforth:

$$Z_{1,1} = 1 \tag{17}$$

Subsequently, let us initialize $Z_{q,1}$ via the following recurrent formula (which follows from formula 16):

$$Z_{q,1} = Z_{q-1,1} \frac{g - q + 1}{q - 1} \tag{18}$$

Then the full recurrent formula can be used (which also follows from formula 14):

$$Z_{q,t} = Z_{q,t-1} \frac{nq - t + 1}{ng - t + 1} \tag{19}$$

13 Empirical data

As a sanity check, we compared the predictions of the formula with the results of physically splitting up our corpus into equal-sized segments and randomizing them. Two hundred segments, each 2000 tokens in length, were obtained from the corpus; they had 17 types in common. After 500,000 randomizations, the average number of shared types was 29.3308, which is in accordance with the model which predicts a value of 29.3311 shared types. After this initial check, we proceeded to perform a series of measurements for sets of 2 to 100 randomly chosen non-overlapping segments of 10,000 tokens in length, in the same way as described previously, yielding an average empirical value, baseline, and ratio between the two for each iteration.

Figure 4 shows that a power law can be successfully fitted, even in this case, to both the empirical and the baseline relation between the number of texts in the sample and the number of shared types. As in the previous case, the best fit can be obtained for the 1-displaced power function, except that another argument shifting the whole model vertically is needed. The empirical data can be fitted with $y = 625(x-1)^{-0,897} + 36.1$ ($R^2 = 0.996$), and the baseline relation with $y = 1276(x-1)^{-0,7824} + 101.9$ ($R^2 = 0.9995$).

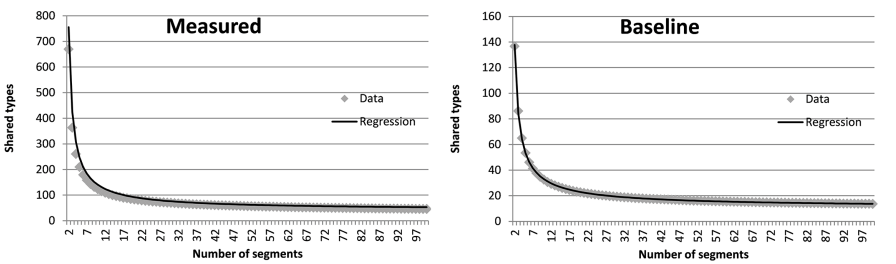


Fig. 4: The number of shared types as a function of the number of segments.

Figure 5 shows the average ratios between the measured empirical value and the baseline model. The ratio trend does not follow a "pipe curve" as in the previous case but decreases steadily, after an initial drop. Both relations involving ratios (in Figs. 3 and 5) can be successfully fitted with the $y = ax^b + cx^d$ function, with the difference that for the natural texts, the d parameter is positive, whereas in the present case of equal-sized segments, it is negative (namely $y = 0.9x^{-3.452} + 0.426x^{-0.050}$; $R^2 = 0.985$).

We also investigated the relation between segment length and the number of shared types: we drew 10,000 random samples of 20 non-overlapping segments

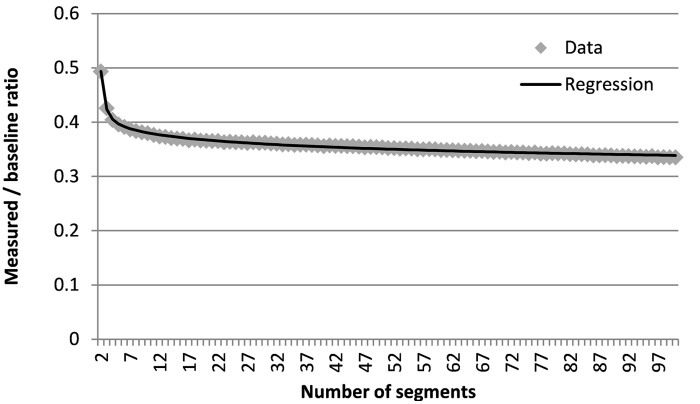


Fig. 5: The relation between the number of segments and the average ratio between the measured value and the baseline model.

of 1000 tokens, counted the average number of shared types, and calculated the average baseline model. The same operation was repeated for progressively longer segments, in 1000 token increments up to 100,000. The results are surprising: if the number of segments is kept constant, then the relation between segment length and the number of shared types is linear (see Fig. 6), the measured empirical value being fitted by $y = 0.00686x + 15.14$ ($R^2 = 0.99994$); the model for the baseline values fits slightly worse, at least in our case and in the present range of values: $y = 0.0286x + 72.2$ ($R^2 = 0.9996$).

As can be seen in Fig. 7, the relation between segment length and the measured / baseline ratio is a decreasing curve which can be fitted by a power function ($y = 1.88x^{-0.179}$; $R = 0.974$).

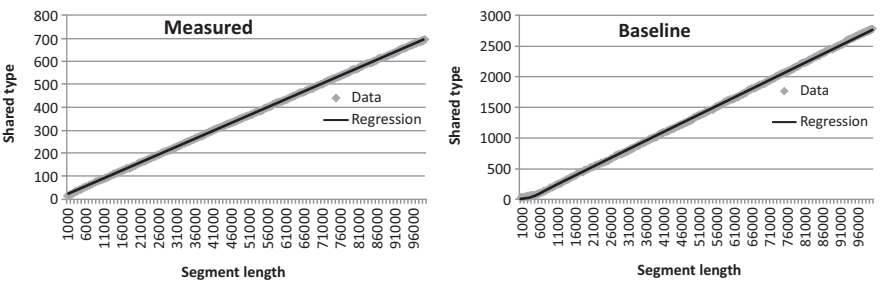


Fig. 6: The number of shared types as a function of segment length.

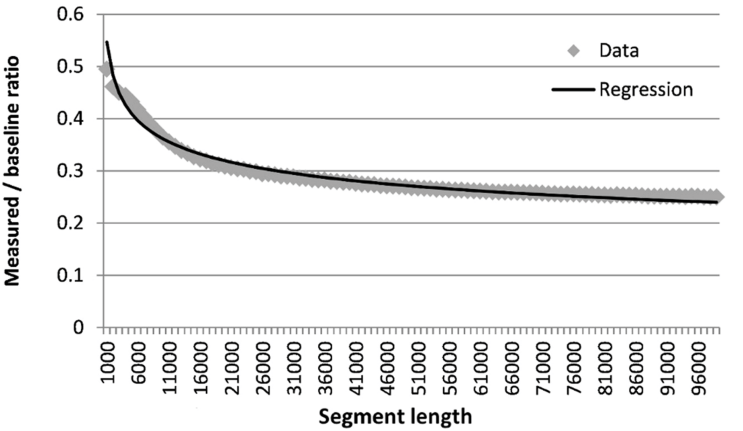


Fig. 7: The relation between the length of segments and the average ratio between the measured value and the baseline model.

14 Conclusion

As we have discovered, the number of shared types in a collection of texts observes lawful behavior. Namely, the relation between the number of texts and the number of shared types follows a power law. Interestingly, similar behavior can be seen if text borders are disregarded and the corpus is artificially divided into equal-sized segments, although the fit is not as good as in the previous case, at least for our data.

Another noteworthy finding of this study is that the number of shared types is proportional to the length of these sequences.

This is not the first time such regularities have been observed on the corpus level, for example Zipf's Law can be successfully fitted even for large and heterogeneous corpora of texts, but these regularities are routinely considered a mere artefact of the same regularities holding true on the level of the text, or that they are superficially similar yet fundamentally different (Montemurro 2002). In contrast, the phenomena we explored in the present study are fundamentally supra-textual and have no meaning below the level of a set of texts. Our findings are thus based on qualities emerging only on the discourse level and demonstrate the fact that relationships among texts are a crucial part of human communication.

Moreover, we have developed baseline models for the number of shared types, i.e. models calculating the expected number of types shared by texts if

all tokens are randomly shuffled among texts. Since there is typically a bias in real texts in favor of tokens of the same type clumping together, the number of word-types shared by either real texts or equal-sized segments (obtained by splitting collections of real texts) is smaller on average than what the baseline model predicts.

The comparison between the empirical data and the baseline model can be used for contrastive purposes to compare the number of shared types in corpora of different languages. We can expect that a corpus harvested from one domain would show a higher relative number of shared types than a heterogeneous corpus sampled from various domains, or that a corpus of text excerpts would be more heterogeneous than a corpus of identical length consisting of whole texts, etc. Exploring these hypotheses is beyond the scope of this short paper; we leave it to future work.

The ratio between the empirical number of shared types and the baseline model can be proposed as a simple discourse homogeneity metric. If the texts in the corpus are too short or the text boundaries are fuzzy, or if we need to compare two corpora with fundamentally different distributions of text lengths, we can approximate this metric by dividing the corpus into equally sized segments. In this case, the baseline model is simpler to implement and compute. Apart from that, as the number of shared types is linearly dependent on segment length, we can compare two or more corpora directly by dividing them into the same number of segments and normalizing the empirical values of shared types by the respective segment lengths.

References

Altmann, Gabriel. 1992. Das Problem der Datenhomogenität. *Glottometrika* 13. 187–298.

Cvrček, Václav. *Kvantitativní určení lexikálního jádra jazyka* [Quantitative delimitation of the core of a language]. *Časopis pro moderní filologii* 96(1). 9–26.

Čech, Radek, Pavel Kosek, Ján Mačutek & Olga Navrátilová. 2020. Proč (někdy) nemíchat texty aneb Text jako výchozí jednotka lingvistické analýzy. *Naše řeč* 103. 24–36.

Köhler, Reinhard & Gabriel Altmann. 2005. Aims and methods of quantitative linguistics. *Problems of Quantitative Linguistics*. 12–42.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Michal Škrabal, Petr Truneček, Pavel Vondřička & Adrian Jan Zasina. 2015. *SYN2015: Reprezentativní korpus psané češtiny* [SYN2015: Representative Corpus of Written Czech]. Ústav Českého národního korpusu FF UK. www.korpus.cz

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka,

Michal Škrabal, Petr Truneček, Pavel Vondřička & Adrian Jan Zasina. 2016. SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 2522–2528.

Montemurro, Marcelo A. & Damián H. Zanette. 2002. New perspectives on Zipf's law in linguistics: From single texts to large corpora. *Glottometrics* 4. 87–99.

Strauss, Udo, Peter Grzybek & Gabriel Altmann. 2007. Word length and word frequency. In Peter Grzybek (ed.), *Contributions to the science of text and language*, 277–294. Dordrecht: Springer.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1). 1–13.

Wild, Marcel, Svante Janson, Stephan Wagner & Dirk Laurie. 2013. Coupon collecting and transversals of hypergraphs. *Discrete Mathematics and Theoretical Computer Science, DMTCS* 15(2). 259–270.

Tereza Motalova

The Menzerath-Altmann law in the syntactic relations of the Chinese language based on Universal Dependencies (UD)

Abstract: The Menzerath-Altmann law (MAL) describes the relationship between two language units – construct and constituent – based on the inverse proportionality of their lengths. The MAL has been applied to various languages and Chinese is no exception. The aim of this study is to apply the MAL to the syntactic dependency structures of Chinese. Its validity is tested on the relations between sentence – phrase – word – (character). We hypothesize that the longer a sentence, the shorter the phrases (measured in words); furthermore, the longer the phrase, the shorter its words (measured in characters). The study yielded the following results: the first hypothesis was corroborated only partially, whereas the second hypothesis was rejected.

Keywords: Chinese language, Menzerath-Altmann law, phrase, syntax, treebank, Universal Dependencies

1 MAL introduction

The Menzerath-Altmann law deals with the relationship between two immediately adjacent language units in a respective hierarchy, with their relationship being based on the inverse proportionality of their lengths. The French linguist Grégoire was the first who discovered this phenomenon in phonology in 1899 (e.g., Kułacka 2010: 257). The German phonetician and psychologist Menzerath consequently performed experiments on the length of syllables (Menzerath 1928) and words (Menzerath 1954) and formulated this phenomenon as follows: “The greater the whole, the smaller its parts” (Menzerath 1954: 101). Several decades later, Altmann proposed a mathematical model for the law and generalized

Acknowledgment: I would like to thank Radek Čech, Xinying Chen and Jiří Milička for thoughtful comments and suggestions during the work on the experiment.
This work was supported by the European Regional Development Fund through the project ‘Sinophone Borderlands – Interaction at the Edges’ (no. CZ.02.1.01/0.0/0.0/16_019/0000791).

Tereza Motalova, Palacký University Olomouc, e-mail: tereza.motalova@upol.cz

https://doi.org/10.1515/9783110763560-010

Menzerath's observation to the following form: "The longer a language construct the shorter its components (constituents)", or even more generally, "The length of the components is a function of the length of language constructs" (Altmann 1980: 124–125). In Altmann's view, construct and constituent represent general concepts which can be replaced by empirical cases (Altmann 1980: 127). Due to the contribution of both researchers, this phenomenon is known as the Menzerath-Altmann law.

Mathematical formulas of the MAL differ in the involvement of different parameters:

Model 1 with two parameters A, b: $y = Ax^b$ (truncated model)
Model 2 with three parameters A, b, c: $y = Ax^b e^{-cx}$ (complete model)

where x is the length of a construct measured in its constituents and y is the average constituent length measured in immediately adjacent lower language units. A, b and c are parameters (Andres et al. 2014: 4). "Parameter a determines the shift on the y-axis and can be understood as the 'starting value' of the fitting curve, while parameter b is responsible for the steepness and 'speed' of the decrease of the curve" (Kelih 2010: 71). "Just as A the parameter b might depend on the language and the language level" (Cramer 2005: 45). Nevertheless, interpretation of the parameters has not been fully developed (Eroglu 2014: 393).

Over the course of discussions on the models, researchers have not reached an agreement on which model fits the data the best. According to Andres (2014), the truncated model seems to be more optimal in comparison with the complete formula due to the accuracy of parameter b. In contrast, the simple formula might not fit the data on lower language units (Andres 2014: 2, 10). In addition, researchers proposed modifications to the originally designed model (e.g., Kułacka & Mačutek 2007, Mačutek & Rovenchak 2011, Milička 2014).

2 MAL in Chinese

The idea of the application of the MAL to Chinese is not entirely new. It has been attracting the attention of researchers for several times. Let us begin by introducing a few selected experiments performed on Chinese with a particular emphasis on the sentence level.

Bohn (2002) tested relations between language units on a Chinese corpus containing reports from the news agency Xinhua. The sentence was measured in

the number of clauses and the clause in the number of words. The words in the constituent position were measured in the number of characters. Bohn defined both the sentence and the clause on the basis of punctuation marks, whereas the word was segmented automatically by means of software. In the case of a sentence–clause pair, the MAL validity was confirmed, whereas in the case of a clause–word pair, the coefficient of determination reached lower values. Bonn also ran an experiment on a single text and the results obtained were the opposite. The coefficient of determination reached a higher value in the case of the clause–word pair (Bohn 2002: 139–143).

Jin and Liu (2017) examined the relationship between sentence and clause lengths (measured in words) in several corpora of different genres. They considered the sentence and the clause to be a segment limited by respective punctuation marks, whereas word segmentation was carried out by the tool of the Chinese Lexical Analysis System ICTCLAS. The experiments revealed that the greatest goodness-of-fit between the MAL and the data was achieved in the case of size-restricted texts (Jin & Liu 2017: 201–202, 218).

Hou et al. (2017) conducted another experiment testing the MAL on the relationship between sentence–clause (word). The segmentation approach was again based on punctuation marks and the ICTCLAS tool. The authors applied the MAL to different registers and they arrived at the conclusion that the relationship defined by the MAL holds true in the case of texts representing the formal written language (Hou et al. 2017: 5–6, 14).

To give another example, Chen (2018) also performed an experiment focusing on the relationship between sentence–clause (word) using punctuation marks for their automatic recognition. Chen used the Lancaster Chinese corpus and, based on the results obtained, concluded that the MAL is valid for these language units (Chen 2018: 2–3).

All the experiments mentioned above demonstrate that the MAL is mostly valid for relations between these language units. This raises a question, however, if the MAL would hold true for relations between language units which we determine not only by the graphical approach but also by using different linguistic criteria. Mačutek et al. (2017) used dependency grammar in order to test the relationship between clause, phrase and word. The clause represented the construct measured in the number of its constituents, i.e., phrases. In the authors' view, the phrase directly depends on the main predicate of a clause and is measured in its neighbouring units – words. The authors applied the MAL to the Prague Dependency Treebank 3.0. It should be mentioned that only main clauses were tested due to a lack of annotation of the predicate in subordinate clauses. The experiment yielded that the MAL is valid in the syntactic dependency structure for the Czech language and the hypothesis – longer clauses

tend to consist of shorter phrases (measured in words) was not rejected (Mačutek et al. 2017: 102–104).

The aim of this paper is the application of the similar approach used by Mačutek et al. (2017) to test the MAL in the syntactic structures of the Chinese language using treebanks released within the Universal Dependencies project (Nivre et al. 2016).

3 Methodology

3.1 Universal dependencies and treebanks

Universal Dependencies (UD) project is an initiative aimed at creating cross-linguistically consistent treebank annotation based on dependency syntax. On the one hand, the annotation contains universal categories applicable across languages, while on the other hand, it also provides extensions specific to a particular language (Universal Dependencies 2020a). In general, "The basic dependency representation forms a tree, where exactly one word is the head of the sentence, dependent on a notional ROOT and all other words are dependent on another word in the sentence" (Universal Dependencies 2020b).

The study examines two UD treebanks, namely Traditional Chinese HK Treebank (Wong et al. 2017) and Parallel Universal Dependency (PUD) treebank (Zeman et al. 2017) for Chinese. The Traditional Chinese HK Treebank consists of 1004 sentences from official proceedings and subtitles. The proceedings were recorded during a meeting of the Legislative Council of the Hong Kong Special Administrative Region held on 12 October 2016. The subtitles were created for three short films shot by students from the School of Creative Media, namely Missing days / 小時光, Tempo in Temple / 廟眾樂樂, What day is today / 今日星期幾. The treebank annotation was performed manually by Herman H. M. Leung and Tak-sum Wong at City University of Hong Kong. Due to a mixture of different genres, we separated this treebank into two parts. The first 650 sentences create the sub-treebank containing only subtitles (hereinafter subtitles), whereas the remaining 354 sentences of the proceedings were analysed separately (hereinafter proceedings). This division was highly motivated by a significant difference between these two genres – the subtitles represent colloquial language, while the proceedings resemble more formal language in its form of long and complex sentences.

The PUD treebank contains 1000 randomly selected sentences from news and Wikipedia. The sentences came mostly from English and were subsequently translated into Chinese by professional translators. In contrast to the Traditional

Chinese HK Treebank, the annotation was first carried out by Google and then converted to UD by the UD team.

3.2 Language units

For the purposes of this experiment, we examined the mutual length relationship between sentence, phrase, word and character.

The sentence belongs to those language units which are easily and reliably recognized on the basis of punctuation marks. The borders of this language unit are formed by the full stop 。, the question mark ? and the exclamation mark !. We consider the sentence to be only a construct measured in the number of its neighbouring units, i.e., phrases.

Our phrase is, in the view of the experiment, identified by any word which directly depends on the main predicate of a sentence, or more precisely, on a root according to the UD. The phrase length equals the sum of all the remaining words dependent on the phrase head including this head itself. This grouping, i.e., a word and its phrase, can be visualized as a subtree hanging from it (Mel'cuk (1988:14). Root, as well as punctuation, are not taken into account. The number of phrases, or more precisely direct dependencies, determines the length of the sentence. It should be pointed out that sentences consisting only of a root were excluded because the number of their direct dependencies relations equals zero.

To make it completely clear, we use the example of a sentence from the proceedings to demonstrate the length of the phrase (Fig. 1). The sentence contains four phrases – the first from the left has the length of one word (但 dàn), the second of six words (程度 chéngdu, 選舉 xuǎnjǔ, 嚴謹 yánjǐn, 這 zhè, 的 de, 次 cì), the third of four words (選舉 xuǎnjǔ, 比 bǐ, 學生會 xuéshenghuì, 大學dàxué) and the last has again one word (更 gèng). The root (不如 bùrú) and the full stop (。) are not counted.

The word segmentation relies entirely on the UD. It should be emphasized that the UD segmentation is based on syntactic words – neither phonological nor orthographic words (Nivre et al. 2016: 1660). Moreover, the Chinese written language does not even use a space between words. Hence, segmentation based on orthography cannot be taken into consideration. For a language such as Chinese, UD utilizes a complex word segmentation algorithm (Universal Dependencies 2020c).

Although the word takes only the role of the constituent of the phrase, it is crucial to determine its constituent. In the view of the MAL, the word should be measured in its immediately adjacent unit. As Grotjahn and Altmann (1993: 141) mentioned, however, defining the concept of the word involves several difficulties to which the "problem of the unit of the measurement" also belongs.

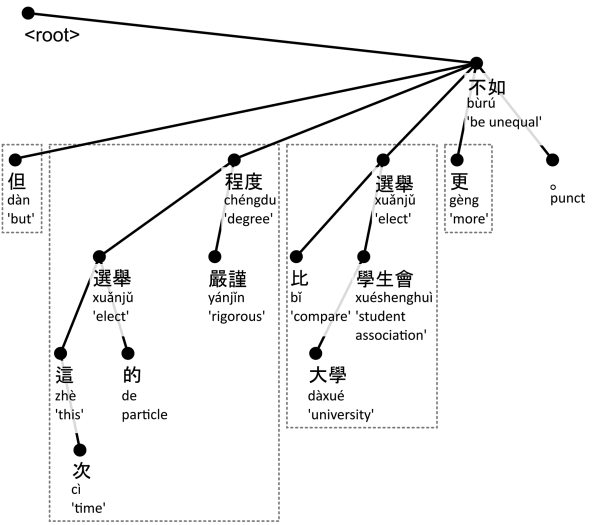


Fig. 1: The syntactic structure of the sentence (ID 848) from Traditional Chinese HK Treebank.
但這次選舉的嚴謹程度比大學學生會選舉更不如。
Dàn zhècì xuǎnjǔ de yánjǐn chéngdu bǐ dàxué xuéshenghuì xuǎnjǔ gèng bùrú.
'However, this election is not as rigorous as the elections of the university student association.'

In the case of Chinese, three language units can be regarded as the word constituents – stroke, component, and character (Chen & Liu: 2016: 10–11). Due to the formation of polysyllabic words, Chinese characters lost their logographic properties. They began to be connected with other characters which resulted in the existence of a certain relationship between them. For this reason, we consider the immediately neighbouring unit of the word to be the character.

The character represents the basic graphic unit of the Chinese writing system that corresponds to one syllable in the spoken language with one exception.[1] All Chinese characters occupy the same graphic square without regard to the number of their strokes and components.

The study primarily examines how the phrase behaves as a construct and constituent in relation to language units that are immediately adjacent to it, namely sentence and word. In terms of the MAL, we assume the following: the longer a sentence, the shorter the phrases (measured in words); furthermore, the longer the phrase, the shorter its words (measured in characters).

1 The exception is the case of er-coloring, for instance the word for painting 画儿 huàr.

4 Results

This section provides the results which we obtained by the application of the MAL to the Traditional Chinese HK Treebank – subtitles and proceedings – and to the Chinese PUD treebank on the levels sentence–phrase and phrase–word. We present the results in the form of tables (Tab. 1 and Tab. 2) containing parameters A, b and, in some cases, parameter c. The reliability of both mathematical models is expressed by the coefficient of determination R^2. We used the statistical software NLREG (Sherrod, 1992–2015) to acquire all the values. The Figs. 2–5 demonstrate the trends of fitting curves which can be observed within the results related to the truncated formula (Model 1) for both language pairs. All the figures were created in RStudio (RStudio Team: 2015). Finally, observations with frequencies $f \leq 5$ were omitted.

Tab. 1: Parameters *A, b, c* and coefficients of determination R^2 for the MAL models applied to the sentence–phrase.

Sample	Model[2]	A	b	c	R^2
Subtitles	Model 1	1.5269	0.1577	–	0.7681
	Model 2	1.5277	0.1617	0.0014	0.7681
Proceedings	Model 1	4.0497	−0.2364	–	0.8055
	Model 2	4.3297	0.1117	0.1240	0.9423
PUD Treebank	Model 1	11.4069	−0.6695	–	0.9888
	Model 2	11.4207	−0.6653	0.0016	0.9888

Generally speaking, the value of the coefficient of determination R^2 fluctuates between zero and one – the higher the value is, the better the respective model fits the data. According to Mačutek and Wimmer (2013: 233), the reliable goodness-of-fit starts with the value $R^2 \geq 0.9$. Some researchers propose, however, lower values – $R^2 \geq 0.7$ (Andres et al. 2012: 15) or $R^2 \geq 0.5$(Andres et al. 2014: 10).

If we look into the details of the results on the level of sentence–phrase (Tab. 1), in the case of the proceedings, the coefficient of determination has a value greater than 0.9 only in connection with the complete formula (Model 2). The PUD treebank shows even better results – both coefficients of determination exceed the value of 0.9. In contrast, the validity of the MAL is not confirmed for the subtitles. As for the phrase, it is too early to conclude that it can be regarded as an appropriate unit of measurement. Let us first address the results which we obtained on the lower level where the phrase was tested in the role of the construct.

2 Model 1: $y = Ax^b$; Model 2: $y = Ax^b e^{-cx}$.

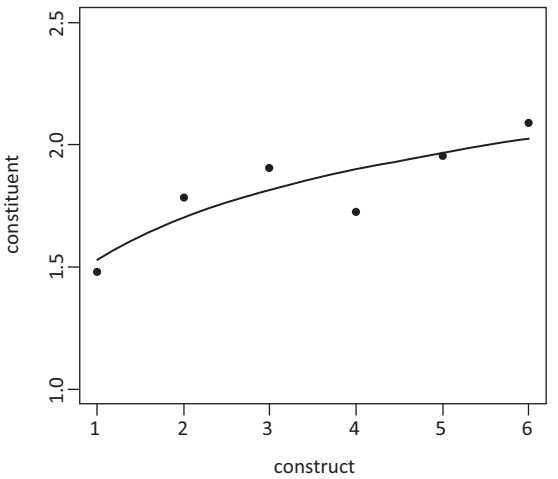


Fig. 2: Relationship between sentence and phrase using the truncated formula (Model 1) – subtitles.

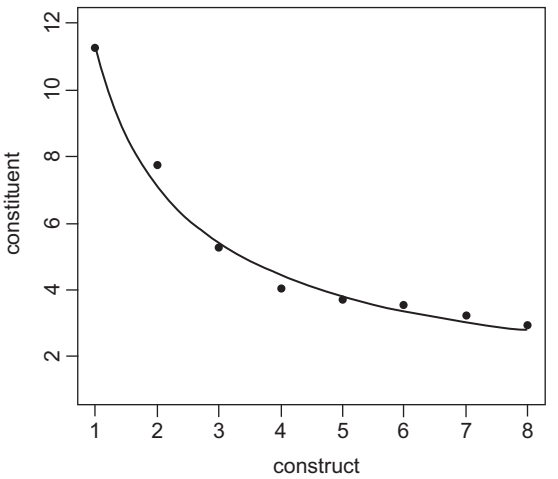


Fig. 3: Relationship between sentence and phrase using the truncated formula (Model 1) – PUD Treebank.

The relationship between the phrase and word (measured in characters) is not in accordance with any of the MAL models (Tab. 2). The fitting curve depicted below tends to be constant or increase rather than decrease (Fig. 4 and Fig. 5). Moreover, the scale of the construct length goes up to 17 words, 33 words and 38

Tab. 2: Parameters *A, b, c* and coefficients of determination R^2 for the MAL models applied to the phrase–word.

Sample	Model[3]	A	b	c	R^2
Subtitles	Model 1	1.3644	0.0239	–	0.2195
	Model 2	1.3663	0.0125	−0.0029	0.2247
Proceedings	Model 1	1.5527	0.0179	–	0.1278
	Model 2	1.5498	0.0228	0.0009	0.1291
PUD Treebank	Model 1	1.7903	−0.0089	–	0.0273
	Model 2	1.6992	0.0724	0.0110	0.4003

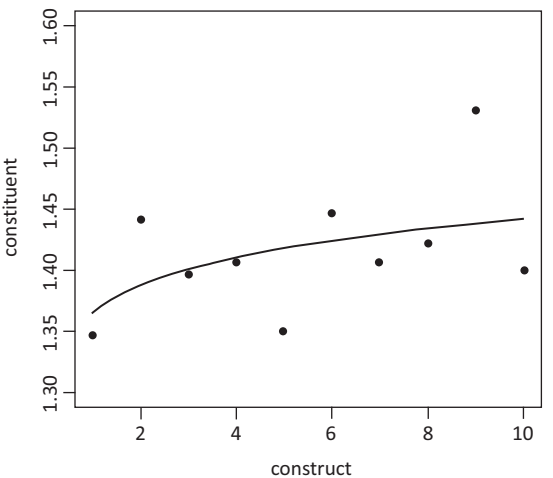


Fig. 4: Relationship between phrase and word using the truncated formula (Model 1) – subtitles.

words per phrase. From the point of view of short-term memory, there should exist a limitation in the length of each language level (Kułacka 2009: 58). The capacity of the working memory has its limits as, for example, mentioned by Miller (1956) in the form of the "magical number seven plus minus two".

3 Model 1: $y = Ax^b$; Model 2: $y = Ax^b e^{-cx}$.

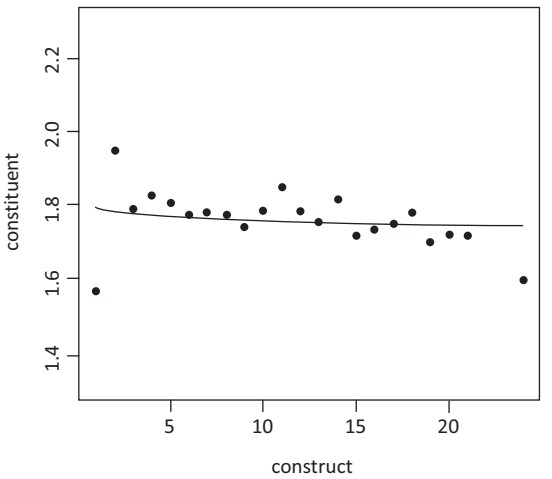


Fig. 5: Relationship between phrase and word using the truncated formula (Model 1) – PUD Treebank.

5 Conclusion and further research

In summary, well-fitting results were achieved in the samples representing the formal style – the results obtained by the application of the MAL to the PUD treebank ($R^2 > 0.9$ for both models) and the proceedings ($R^2 > 0.9$ for Model 2) did not reject our first hypothesis – the longer a sentence, the shorter the phrases (measured in words). The data acquired from the sub-treebank containing the subtitles did not fit any of the MAL models. It appears that the genre has a considerable influence on the results.

We also hypothesized that the longer the phrase, the shorter its words (measured in characters). All the results, however, rejected this hypothesis. The construct lengths indicate that the chosen hierarchy of the language units is incomplete and we perhaps omitted at least one "layer" which is lower than the sentence and higher than the phrase at the same time. For this reason, we propose including a clause between them and running the experiment again. Consequently, a phrase would not be as long in the number of words and not even as deep in the view of the dependency structure.

Inclusion of the clause into our hierarchy does not, however, have to be the only cause of the unsuitability of the MAL. The constituent lengths fluctuate within a very small interval and demonstrate that words consist mostly of one or two characters on average. This fact results in the impossibility of the MAL to

manifest itself within the length relationship between these units. Disagreement with the MAL can be associated with three crucial factors. As Chen and colleagues pointed out in their study, the majority of words are mono-syllables (characters) or bi-syllables (characters) in modern Chinese (Chen et al. 2015: 8). Hence, this length limitation can have an impact on word segmentation and our results. Secondly, word segmentation is based on the UD and its own algorithm, which can also exert an influence to a certain extent. It would be desirable to compare UD word segmentation, for example, with the results of the ICTCLAS tool applied in the above-mentioned studies (cf. MAL in Chinese). Finally, the recent study performed by Chen and Liu (2016: 26) shows that the most appropriate measurement unit of the word for written Chinese is the component. The question arises as to whether or not the component would be a more appropriate measurement unit for UD word segmentation.

This experiment is considered a pilot study in order to test the similar approach applied by Mačutek et al. (2017) to the Czech language. Their experiment proved the MAL validity in the relationship between the main clauses and their phrases (measured in words). Nevertheless, our results showed that we should perform further experiments focusing on the following issues:

1. To include a clause into the hierarchy of language units for Chinese and test it not only as a constituent, but also as a construct;
2. To address the possibility of the existence of a unit which could be in the hierarchy between the phrase and word and subsequently test it as a construct and constituent;
3. To test other measurement units for the syntactic word determined by the UD, for example, the component;
4. To apply this approach not only to UD treebanks but to others if available.

References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. *Glottometrika* 2. 1–10.

Andres, Jan, Martina Benešová, Lubomír Kubáček & Jana Vrbková. 2012. Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics* 19(1). 1–31.

Andres, Jan. 2014. The Moran–Hutchinson formula in terms of Menzerath–Altmann's law and Zipf–Mandelbrot's law. In Gabriel Altmann, Radek Čech, Ján Mačutek & Ludmila Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis. Studies in Quantitative Linguistics* 17, 29–44. Lüdenscheid: RAM-Verlag.

Andres, Jan, Martina Benešová, Martina Chvosteková & Eva Fišerová. 2014. Optimization of Parameters in the Menzerath–Altmann law, II. Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. *Mathematica* 53(2). 5–28.

Bohn, Hartmut. 2002. Untersuchungen zur chinesischen Sprache und Schrift. In Reinhard Köhler (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*, 127–177. Trier: Universitätsbibliothek Trier.

Chen, Heng. 2018. Testing the Menzerath-Altmann law in the sentence level of written Chinese. *Open Access Library Journal* 5(e4747). 1–5. https://doi.org/10.4236/oalib.1104747

Chen, Heng & Haitao Liu. 2016. How to measure word length in spoken and written Chinese. *Journal of Quantitative Linguistics* 23(1). 5–29.

Chen, Heng, Junying Liang & Haitao Liu. 2015. How does word length evolve in written Chinese? *PLoS ONE* 10 (9): e0138567. 1–12. https://doi.org/10.1371/journal.pone.0138567

Cramer, Irene. 2005. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics* 12(1). 41–52.

Eroglu, Sertac. 2014. Menzerath–Altmann law: Statistical mechanical interpretation as applied to a linguistic organization. *Journal of Statistical Physics* 157(2). 392–405.

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: Springer.

Hou, Renkui, Chu-Ren Huang, Hue San Do & Hongchao Liu. 2017. A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics* 24(4). 350–366.

Jin, Huiyuan & Haitao Liu. 2017. How will text size influence the length of its linguistic constituents? *Poznań Studies in Contemporary Linguistics* 53(2). 197–225.

Kelih, Emmerich. 2010. Parameter interpretation of Menzerath law: Evidence from Serbian. In Peter Grzybek & Emmerich Kelih (eds.), Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives, 71–79. Wien: Praesens.

Kułacka, Agnieszka. 2009. The necessity of the Menzerath–Altmann law. A*nglica Wratislabiensia* 47. 55–60.

Kułacka, Agnieszka. 2010. The Coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics* 17(4). 257–268.

Kułacka, Agnieszka & Ján Mačutek. 2007. A discrete formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics* 14(1). 23–32.

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, 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(3). 227–240.

Mačutek, Ján, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni & Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 100–107. Pisa: Linköping University Electronic Press.

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Menzerath, Paul. 1928. *Über einige phonetische Probleme. Actes du premier congre's international de linguistes*. Leiden: Sijthhoff.

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

Milička, Jiří. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics* 21(2). 85–99.

Miller, George. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review* 63(2). 81–97.

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, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666. Portorož: European Language Resources Association.

RStudio Team. 2015. *RStudio: Integrated Development for R. RStudio*, Inc. Boston, MA. http://www.rstudio.com/.

Sherrod, Phillip H. 1992–2015. *NLREG Version 6.6* (Demonstration).

Universal Dependencies. 2020a. "Introduction." Universal Dependencies. https://universaldependencies.org/introduction.html. (accessed 2 March 2020)

Universal Dependencies. 2020b. "Syntax: General Principles." Universal Dependencies. https://universaldependencies.org/u/overview/syntax.html (accessed 2 March 2020)

Universal Dependencies. 2020c. "Tokenization and Word Segmentation." Universal Dependencies. https://universaldependencies.org/u/overview/tokenization.html (accessed 9 March 2020)

Wong, Tak-sum, Kim Gerdes, Herman Leung & John Lee. 2017. Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In Simonetta Montemagni & Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 266–275. Pisa: Linköping University Electronic Press.

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj & Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In Jan Hajič & Dan Zeman (eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–19. Vancouver: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/K17-3001

Adam Pawłowski, Tomasz Walkowiak

Statistical tools, automatic taxonomies, and topic modelling in the study of self-promotional mission and vision texts of Polish universities

Abstract: The websites of higher education institutions fulfil informational and promotional functions. They are addressed to potential stakeholders, in particular candidates for studies and business representatives. This paper deals with the content of texts on the websites of Polish universities in the "Mission", and "Vision" tabs. The corpus of promotional texts provided by universities is analysed using the methods of statistics, topic modelling and automatic taxonomy. In particular, in the empirical part of the paper, we present the quantitative characteristics of the corpus and stylo-statistical measures. We present automatic taxonomies of texts and verify the hypothesis that the taxonomy of texts representing individual universities should reproduce the classification based on the explicit data. We also use the method of topic modelling to reconstruct the main content of the promotional messages of universities.

Keywords: quantitative methods, university mission and vision, topic modelling, automatic taxonomy, academic institutions taxonomy

1 Introduction

Academic institutions strive to communicate their character and attractiveness using strategies based on individual characteristics, combined with the need to adapt to more general rules. The showcase of each university is a website which describes, among other things, its principles, structure and achievements. The website is, of course, a multidimensional message, in which not only text appears, but graphic and audio layers are also present. Without entering into a dispute over which of these media has a stronger effect on the human mind, it can be assumed that in the realm of science and education this is the intellect, which plays a more important role than emotions. And so text, rather than

Adam Pawłowski, University of Wrocław, e-mail: adam.pawlowski@uwr.edu.pl
Tomasz Walkowiak, Wrocław University of Technology, e-mail: tomasz.walkowiak@pwr.edu.pl

https://doi.org/10.1515/9783110763560-011

image, remains the dominant medium of content. The websites of academic institutions are usually very extensive, and not all of their components can be considered relevant with regard to the design of the self-image. For this reason, it has been assumed in this work that **signals intentionally indicating the specificity of a given university are contained in the sub-pages referred to as 'Mission' and 'Vision'**, or in their functional equivalents, such as 'About us', 'About our university', 'Why is it worth studying with us', etc.

2 Goals and hypotheses

In this paper the following research hypotheses were adopted and verified:

1a. The image of a university as expressed in the content of its mission and vision is its **individualising** feature, and thus distinguishes it from other types of universities (e.g., medical universities should differ from military or technical schools).

1b. The image of a higher education institution as expressed in the content of the mission and vision is also its **integrating** feature. It means that to a certain extent it makes a particular university similar to other universities of the same type.

2. It is possible to automatically recognise the key areas of the university's mission and vision as expressed in content words.

There is actually no contradiction between hypotheses 1a and 1b because, according to the adopted assumption, the language of the mission and vision fulfils an integrating function (similarity) in the case of the collections of universities with the same educational and scientific profile, while in the collection of universities with different profiles, a discriminatory function is (or at least should be) more apparent. Building a mission and vision of a university is always a search for balance between typicality and uniqueness.[1] Therefore, if it turns out that universities with different profiles are in the same classification clusters, it would be the result of incorrect constructions of their self-presentation (and a falsification of hypothesis 1a).

1 Looking for deeper, psychological reasons for this attitude, one may reflect that humans always balance on the verge of cognitive conservatism (the need to conform perception of reality to one's knowledge, beliefs and established ideas) and curiosity-driven open-mindedness, which, however, requires additional effort in the communication process.

Hypotheses 1a and 1b will be verified using automatic taxonomy tests, such as stylometry metrics and text similarity analysis based on TF-IDF vectorisation. Their validation is possible due to the fact that the objects investigated are assigned to predetermined types, so it is possible to check the accuracy of automatic mappings based solely on the distribution and frequency of lexemes in texts. Moreover, it will be possible to reveal relationships between universities or types of universities that actually exist but are not explicit. On the other hand, rejection of the hypothesis 1a (lack of discrimination between objects) would imply that the authors of the mission and vision texts were not able to include in them the relevant features specific to the universities they describe. Hypothesis 2 will be verified using the topic modelling method, which allows for the extraction of keywords from a collection of short texts. The resulting sets of topics and keywords will be compared with generally recognised characteristics of university types.

3 Research material

The subject of our research was texts describing the corporate image of all Polish universities. They were of variable length, and appeared in different sections of the websites. Most often, they were located in the "Mission" and "Vision" tabs, but in some cases also in the "About us", "Our university" or "Why is it worth studying with us" tabs. Some universities had several tabs of this kind on their websites, while others had no self-image content at all. For this reason, in Tab. 1, the number of universities and the number of files examined are not equal. The distribution of types of the universities surveyed is presented in Tab. 1.

Tab. 1: Higher education institutions in Poland.

Type of university	Universities	Files
State-owned academic institutions	108	123
State-owned vocational colleges	34	36
Non-public schools	249	226
Together:	391	385

A separate issue is the structure of academic education, which is reflected in the classifications presented below. In Poland, as in most countries of the world, there are state-owned and private universities, here referred to as non-public. In the group of state-owned universities, a distinction is made between

"universal" universities (offering education and research in numerous fields, specialised universities (e.g. technical, military or medical), and vocational schools. The group of non-public schools, on the other hand, is heterogeneous and inconsistent because it includes small and large, weak and prestigious institutions, while the educational profile of many of them is often opaque. The distribution of state-owned academic schools is shown in Tab. 2.

Tab. 2: State-owned academic schools in Poland.

Type	Number	Type	Number
Universities	18	Medical Universities	12
Universities of Technology	18	Universities of Fine Arts	19
Economic Universities	5	Military Academies	4
Pedagogical Universities	5	Church Universities	8
Agricultural Universities	6	Clerical Seminars	7
Academies of Physical Education	6	Total:	108

The corpus was created after the collection of research material from the websites of Polish universities (Tab. 3). It consists of 385 files of various lengths and has a volume of 144,079 text words (tokens). The size of many samples clearly differed from the average (374 text words), which had an impact on the overall concentration measures: the coefficient of variation has a value of 0.96, indicating a large diversity in the volume of texts in the corpus.

Tab. 3: Number and length of samples in the corpus.

Number of files in the corpus	385
Total corpus volume (graphic words)	144,079
Average file volume (graphic words)	374.23
Standard deviation	359.73
Coefficient of variation	0.96

4 Theory and methods

In the preliminary stage of the research, some basic stylometric measures were used, namely TTR, logTTR and UBER lexical richness coefficients, K Yule's vocabulary diversity characteristic and Lorenz / Gini's vocabulary concentration coefficient (Tab. 4).

Tab. 4: Basic stylometric measurements used in the study.

Coefficient	Formula	Description
Guiraud (lexical richness)[2]	$log\left(V/\sqrt{N}\right)$	N – no. of words V – no. of different words
logTTR (lexical richness)	$log(V/N)$	N – no. of words V – no. of different words
UBER (lexical richness)[3]	$\frac{log\ (N^2)}{logN\ -\ logV}$	N – no. of words V – no. of different words
Yule's K (diversity measure)[4]	$\frac{\sum_{j=1}^{V}\sum_{i=1}^{j} f_i^2 - N}{N^2} \times 10^4$	f_i – number of occurrences of different words
Lorenz / Gini[5] (concentration coefficient)	$\frac{2\sum_{i=1}^{K} i\hat{f}_i - N}{NV} - 1$	$\hat{f}_i$ – number of occurrences of different words; higher values indicate greater concentration of vocabulary

The use of these coefficients is not only a tribute to the tradition of stylometric research. Tests prove that the univariate measures of lexical richness and concentration are also useful for processing large corpora of applied texts. More advanced analysis of text proximity requires a method of measuring similarity between documents (Piasecki et al. 2018). A common method used for building a vector representation of documents is the bag of words (Harris 1954). The key assumption in this approach is that the sentence can be expressed using an unordered set of frequencies of selected words (Salton & Buckley 1988). A number of task-specific modifications of this method are available. For example, the selected dictionary has typically filtered off the most and least common phrases, as well, stop words may be eliminated (Torkkola 2004). In addition, the number of selected features (words) can be often reduced by transforming the words to their generic form (stemming, lemmatization). The literature commonly suggests weighting the raw counts of word occurrences e.g., in relation to the total length of each document and uniqueness of a given word. Therefore in the experiments performed, we have used the TF-IDF (Salton & Buckley 1988) weighting schema,

2 Guiraud 1960.

3 Cf. Dugast 1979.

4 Cf. Yule 1944.

5 The Gini / Lorenz coefficient in this context is a summary statistic that measures how equitably words are distributed in a text (Wilson 2009).

where raw counts are divided by the maximum frequency in a document and multiplied by the inverted document frequency.

The distance between text vectors could be further analysed by clustering or multi-dimensional scaling. Among different clustering algorithms, the agglomerative method was the one we used (Day & Edelsbrunner 1984). It is based on a "bottom-up" process that starts with the initial set of singleton clusters including one document each. Next, in each step, the two most similar clusters are found and merged into a new one. As a result, a tree-like hierarchy of clusters (also referred to as a dendrogram) is established.

Multidimensional scaling aims to present vectors in a low dimensional space (in our case 2D) in a way that elements which are close in *N*-dimensional space (original vector space) are as close to each other as possible in the low dimensional space. In the experiments carried out, we have used the *t*-distributed Stochastic Neighbour Embedding (t-SNE) method (van der Maaten & Hinton 2008).

Moreover, for the automatic extraction of keywords, the topic modelling technique was used. The aim of such algorithms is to generate collections of words (referred to as 'topics') that are significant and specific to a given corpus on the basis of the word frequency and the scope of their presence in the samples (Blei et al. 2003). Similar to previous methods, it ignores the word order. Among different topic modelling methods, we have chosen the Latent Dirichlet Allocation – LDA (*ibid.*) and Additive Regularization of Topic Models – ARTM (Vorontsov & Potapenko 2014).

5 Results

5.1 Lexicostatistical corpus profile

An introduction to the research proper is the lexico-statistical description of the mission and vision corpus. The basic statistics of lemmatized words[6] in a text give a general idea of its stylistic profile, quantitative relations of its elements and, to a lesser extent, of its content. The values of stylometric coefficients in a corpus do not have an absolute reference scale, so it is difficult to assess *a priori* whether they are high or low. That is why the principle is applied of comparing them to a reference corpus. In our case, we have adopted as a reference system

6 The term 'word' will be understood as a sequence of characters between spaces (graphic word), and in a very few justified cases (co-existence of cumulative and split spelling) – a multiword expression.

the manually annotated part of the National Corpus of the Polish Language.[7] An advantage of this solution is that the compared corpora do not differ in size to a significant degree. Table 5 contains basic stylometric measures calculated on the mission and vision corpus.

Tab. 5: Stylometric description of the mission and vision corpus.

	Mission & Vision corpus	**Polish National Corpus**
Volume (N)	144,079	992,014
Tokens (V)	8,444	54,128
V/N (TTR)	0.059	0.055
Lexical richness		
Guiraud's coefficient	22.25	54.34
log TTR	0.761	0.789
UBER	49.73	65.55
Lexical diversity		
Yule's K	690.86	494.41
Lexical concentration:		
Lorenz / Gini's coefficient	0.847	0.880
Lorenz / Gini's coefficient (nouns)	0.829	0.814

The results of the calculations in Tab. 5 partly confirm the expectations based on intuition which suggests that the profiled corpus of missions and visions will be lexically poorer than the general language and will contain a highly concentrated vocabulary, i.e., a relatively small number of lexemes will have a high text coverage. Guiraud and UBER indicators of lexical richness give a clear signal that the vocabulary of the mission and vision corpus is poorer than that of the general language. This is also confirmed by Yule's K-characteristic, which is roughly defined as an unbiased estimator of the probability of two identical words being drawn at random from the text. Again, the corpus of missions and visions proves to be more homogeneous than the general language. What is surprising, however, is the value of the Lorenz / Gini's diversity index. We expected that the diversity of the corpus of mission and vision would be much smaller, as it contains quite repetitive vocabulary. However, it turned out that, statistically, its lexical diversity does not differ from the average calculated for general language. This is probably due to the presence of a full set of function words in the general language contrasted with a relatively poor set of grammatical structures in the mission and vision corpus. To verify this point, we have calculated the Lorenz / Gini index

7 The volume of the corpus is one million words (cf. http://nkjp.pl/).

exclusively for nouns. This time the intuition and observations based on other indicators was confirmed, although the difference in the values obtained for the corpus of mission and vision and general language was smaller than expected.

5.2 Taxonomies

Two related hypotheses 1a and 1b, are being verified here. According to the first one, artificial intelligence methods that rely solely on vocabulary should prove to be effective in showing that the texts of mission and vision of universities of the same type also display similarity (they should cluster together) but at the same time, they should differ from universities or higher schools having other profiles. On the other hand, texts representing small educational establishments without a distinct identity (mainly small non-public higher schools or colleges), should be considered *a priori* as difficult to classify and devoid of discriminatory features.

The first graph shows the approximate distribution of the universities (represented by the points) in a 2D space. Colours are assigned to individual types of universities. Each text was represented by frequencies of selected words. A predetermined set of full-meaning words, characteristic of a typical academic discourse, was used as a classification criterion.[8] The raw frequencies were weighted using the TF-IDF method (Salton & Buckley 1988). Weighted vectors were mapped on a 2D plane using the t-SNE technique (van der Maaten & Hinton 2008). As it has been often observed, stylometric studies of literary texts give good results when using top frequency function words, while taxonomies of applied texts are more effective when full-meaning words frequencies are used as a criterion.

The taxonomy below (Fig. 1) shows that the centre of the graph is occupied by the non-public schools, which are the most numerous in the corpus (a detailed discussion of their distribution would actually require a separate work). There are, however, clear clusters of academic universities, as well as of medical, military, technical, catholic, and physical education state-owned higher schools. This result is empirically grounded and seems intuitively plausible. On the other hand, against the background of the point cloud, the position of pedagogical universities (overlapping "general" academic universities), as well as economic higher schools, is blurred. This is probably the result of their weak identity. They duplicate many university courses, and they exist partly because of a real demand but partly because of a certain inertia of state establishments which are

8 The list of key words is available at www.repository.clarin-pl.eu/0001.

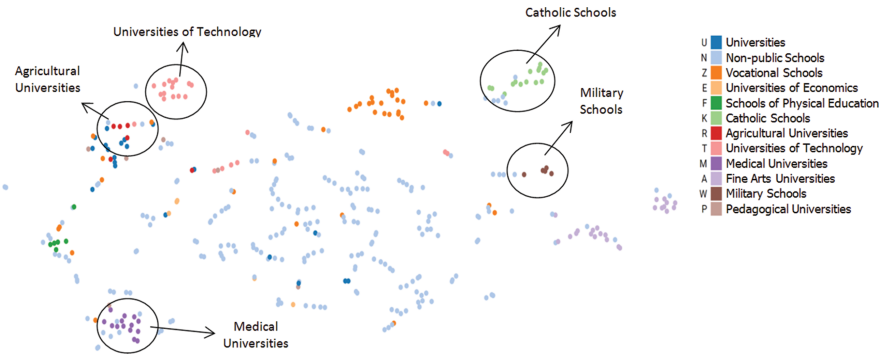


Fig. 1: Distribution of Polish higher education institutions based on the lexical characteristics of their mission and vision texts.

difficult to simply close down or transform. The issue of the low level of clarity of economic universities will be addressed later in the paper, when interpreting the results of the study of the topic modelling.

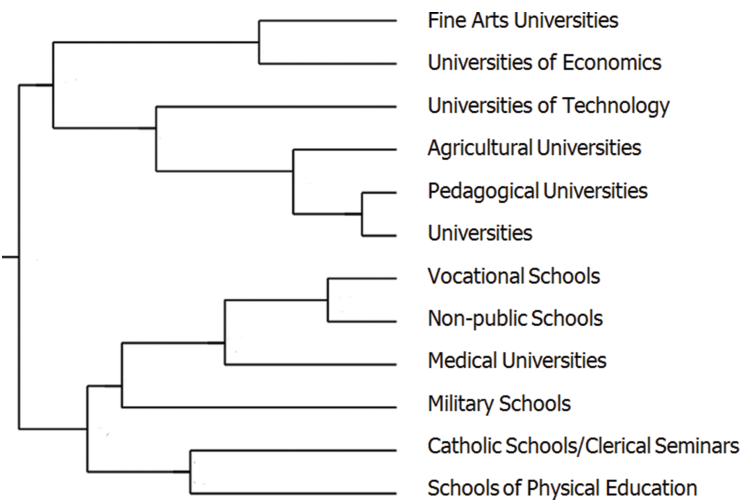


Fig. 2: Distribution of Polish higher education institutions based on the lexical characteristics of their mission and vision texts (merged files).

The graph above (Fig. 2) shows the dendrogram of particular types of schools, prepared using the agglomerative clustering method on a preselected set of full-meaning words (Day & Edelsbrunner 1984). The documents classified were created by merging the mission and vision texts of particular universities or

higher schools of the same type. As can be seen, there are two main clusters, which on the one hand contain great state-owned academic institutions, and on the other hand, private and vocational schools, as well as some highly specialised state-owned higher schools (military, religious and medical). This picture confirms that the language of missions and visions gives a good overview of the profile of universities. One can note that pedagogical schools come very close to academic universities; the same applies to state-owned vocational schools and non-public educational institutions (both types appear close to one another, as they are focused on developing practical skills rather than scientific or academic activities). Very similar results were obtained when particular universities and higher schools were classified and displayed in a 2D space (cf. Fig. 1). It can be jokingly pointed out that the proximity of economic and fine arts universities proves that economics is not so much a craft devised to manage the economy as a quasi magical art that creates reality (cf. the expression 'creative accounting' or 'virtual money').

5.3 Topic modelling

The quantitative analysis of the formal units of language does not give an insight into the content and semantics of the documents examined. A useful method for at least partially overcoming this limitation is the use of topic modelling, which allows the automatic extraction of keywords aggregated in sets (topics). In this case, its use is all the more justified because short texts are being studied, and it is for such small volumes that the method was originally developed and optimised (see Tab. 3).[9]

The goal of applying this method is to discover the topics that would characterise the actual types of universities and the common lexical features of self-image texts of academic institutions. According to hypothesis 2, good results should therefore be obtained with models where the number of relevant (significant) topics corresponds to the number of types of universities or is slightly higher. It can be expected that the particular topics will then match the real types of universities relatively well. In our case, we have distinguished twelve such types in all: academic, technical, economic, pedagogical, agricultural, physical education, medical, artistic, military, and religious universities, vocational schools, and an inconsistent group of non-public higher schools. Therefore, during the

9 Its developers tried to automate the process of extracting topics and keywords from scientific articles and abstracts, which are relatively short in length (cf. Blei et al. 2003).

research, different numbers of topics (from 10 to 25) and two methods of their generation (ARTM and LDA) were tested.

After a series of tests, the result obtained with the LDA method for 20 topics was chosen as the best option. The topics were then evaluated by a group of five specialist respondents who were familiar with the workings of academia. They had to evaluate their semantic consistency (C) and relevance (R) on a scale from 0 to 10. The relevance of a topic is here understood as its compatibility with the actual type of university or higher school. Due to the small number of evaluators, the level of their consistency was not verified. The result obtained reveals the existence of groups of topics referring to five domains, labelled as: 'general academic', 'specific academic', 'non-public', 'educational', and 'vocational schools' (Tab. 6). The group of topics defined as 'general academic' includes lexemes common to all the universities or schools, and not related to any specific type of higher school. Topics defined as 'specific academic' describe those universities that provide education and research in well-defined areas. This applies to higher schools with technical, medical, religious, military, artistic, and physical education profiles. The 'non-public' and 'educational' labels refer to the practical and educational vocabulary that appears in the mission and vision language of mostly private schools. The analysis of the topics generated from the corpus of mission and vision texts of great universities indicates that what makes the 'academic' profile distinct compared to the 'professional' one is the presence of keywords belonging to the broadly understood semantic field of science: so these are such lexeme as 'science', 'scientist', 'research', 'laboratory' etc. The identity of non-public and vocational schools, on the other hand, is determined by the presence of lexemes related to the labour market, employment of graduates, attractiveness of education, personal development, personal success – and thus the fast monetization of education (see other groups or lexemes at the Tab. 6, and Fig. 3).

When analysing this result, one can observe that the hypotheses put forward in the introduction have been only partially verified. Some types of academic schools are actually represented by separate topics, but others seem invisible. A strongly marked identity was observed for art, technical, medical, religious, military, and physical education schools, as well as for professional colleges and some non-public schools. However, there are no distinct topics representing pedagogical, economic and agricultural schools. The same regularity was also observed and described in previous sections.

The issue of the lack of topics indicating the existence of economic and pedagogical schools was partially explained in the previous sections (pedagogical colleges were presented as specific, "reduced copies" of universities). It is worthwhile extending this interpretation even further. Both economics and

Tab. 6: Selection of topics: labels, keywords, evaluation.[10]

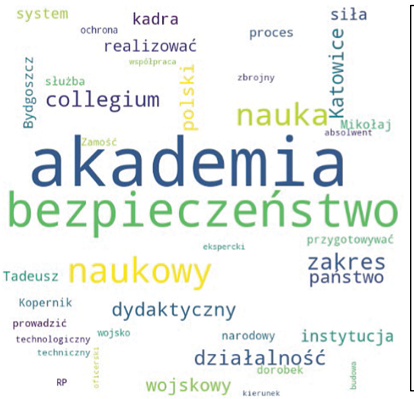
Name	Main keywords (selection)	C	R	Type
UNIVERSITY	scientific, research, tradition	82%	93%	general academic
TECHNOLOGY	university of technology, scientific, engineering	92%	95%	specific academic
MEDICINE	medical, scientific, health, science	87%	97%	specific academic
FINE ARTS	artistic, academy, art, music, culture	86%	97%	specific academic
SECURITY	academy, security, science, military	76%	82%	specific academic
PHYSICAL CULTURE	academic, social, education, physical, culture	35%	67%	specific academic
RELIGION	church, Christian, theological, catholic, seminary, man	91%	98%	specific academic
PROFESSION	school, professional, local, region	43%	72%	vocational college
SCIENCE	scientific, educational, development, system, purpose, cooperation	93%	42%	general academic
BUILDING	laboratory, student, modern, teaching, room, building, library	52%	37%	general academic
STUDIES	faculty, studies, graduate, degree, professional	59%	42%	g. academic / non-public
COLLEGE	school, knowledge, market, life, need, social, work	35%	75%	non-public / educational
HUMAN BEING	man, value, knowledge, skill, learning, ability, shaping	27%	15%	non-public / educational
JOB	work, study, knowledge, graduate, practical	87%	32%	non-public / educational
UNIVERSITY	scientific, research, tradition	82%	93%	general academic

Notations:
C – coherence
R – relevance with regard to the academic segment.

pedagogy studies are relatively inexpensive and are perceived by some candidates as easy. For many poorly talented students, they give the illusion of social advancement through higher education. Therefore, when automatically analysing the language of mission and vision texts, the identity of serious economic and pedagogical universities is blurred among a large number of non-public schools that have a very vague profile but use similar keywords.

Another problem is the very content of the mission and vision of academic institutions of economic profile (there are 5 of them in Poland – see Tab. 2). Basically, the purpose and essence of economics is to manage production and services

10 Labels have been arbitrarily given by the authors of this paper, and not generated automatically by the algorithm. Full list of categories is available at www.repository.clarin-pl.eu/0001.



Selected terms:

academy (akademia)
security (bezpieczeństwo)
science (nauka)
scientific (naukowy)
collegium
military (wojskowy)
didactic (dydaktyczny)
activity (działalność)
staff (kadra)
service (służba)
state (państwo)
scope (zakres)

Fig. 3: Sample caption An example of a topic 'security' in the category 'specific academic' (font size is correlated with word or concept's importance).[11]

in order to earn money and multiply other wealth, leading to increased overall prosperity. However, in no description of the mission of an economic university do the lexemes 'money', 'wealth', 'business', and related words appear. Probably the authors of the mission and vision texts considered them too explicit, referring to supposedly low and unethical incentives (lust for money and power). Instead, there are many high-flying and euphemistic generalisations about progress, innovation or social sensitivity (cf. "an innovative economic university developing creative intellectual potential and educating leaders in response to the challenges of the future", "an independent and socially sensitive university", "education focused on shaping ethical attitudes").

An issue that requires separate comment is the lack of classification topics and clusters that show the clear identity of agricultural universities. This is astonishing because Poland, closing the technological gap that for decades had divided it from the leaders of the European economy, was until recently a typical agricultural country and still remains a great food producer. Agricultural education is therefore popular, well funded and at a very high level. However, it is likely that the language of the missions and visions of agricultural academic schools reflects complexes typical of countries undergoing social and technological modernisation. Rural work, and more generally rurality, is associated in the minds of the descendants of people who have migrated to cities in the last three generations, with primitive living conditions, a lack of culture and technological backwardness, and therefore does not enjoy social prestige. This

11 The other topics are available at www.repository.clarin-pl.eu/0001.

complex of inferior origin is so strongly embedded in contemporary Polish culture that many expressions with negative overtones, connoting, for example, a low level of personal culture, uncouthness, dirt, etc., make use of lexemes traditionally associated with the countryside. It is probably for this reason that agricultural academic schools – some of the most modern in Europe – hide the rural vocabulary under very vague claims that any university could aspire to. On the other hand, there is a lack of literal references to the practice of animal breeding (fodder, porkers, etc.), tillage (harvesting, crops, etc.) or farm facilities (barn, pasture, stables, etc.) – that is, the essence of the village and agriculture. Among the declared areas of academic interest of agricultural schools, we find instead "respect for academic values, the spirit of responsibility for the ideas of humanism, freedom, tolerance, respect for ethical norms, creating attitudes of openness", "seeking and respecting the truth", "openness to new ideas and freedom of expression" etc.

6 Conclusions

In the course of the research presented here, a corpus composed of mission and vision texts of 391 universities in Poland was analysed (Tabs. 1 and 2). It was shown that the corpus of mission and vision is linguistically a specific collection of lexemes, characterized by a poorer vocabulary than that of the general language (Tab. 5). This results from the repeatability of vocabulary related to the academic world, and in particular to research and education. One of the reasons for the linguistic poverty of mission and vision texts is also the lack of diligence on the part of their authors, who often copy the same clichéd formulations from each other. Quantitative analysis of the corpus of mission and vision was carried out using the methods of stylometry (Tabs. 4 and 5), automatic taxonomy and topic modelling. The hypotheses were tested, stating that the image of a university, as expressed in the language of the mission and vision, is on a general level an integrating feature (all universities share similar features appearing in the language they use to describe their missions and visions), but on closer scrutiny it may be also a discriminatory feature (there exist specific subgroups among all the universities or higher schools, using specific terminology). These hypotheses have been verified positively, as specialised universities clustered together both in automatic grouping and in topic modelling analysis. However, it was not possible to indicate the distinct, individual character of economic, pedagogical and agricultural universities. This fact does not, however, undermine the credibility of the obtained result, because it can be explained by cultural (social complex of

rural origin), ethical (negative connotation of lust for money and power) and structural factors (excess of non-public schools with a pedagogical profile, similarity of pedagogical and academic universities).

References

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

Day, William H. E. & Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* 1(1). 7–24.

Dugast, Daniel. 1979. *Vocabulaire et discours*. Genève: Slatkine.

Guiraud, Pierre. 1960. *Problèmes et méthodes de la statistique lexicale*. Paris: Presses Universitaires de France.

Harris, Zellig. 1954. Distributional structure. *Word* 10. 146–162.

Piasecki, Maciej, Tomasz Walkowiak & Maciej Eder. 2018. Open stylometric system WebSty: Integrated language processing, analysis and visualisation. *Computational Methods in Science and Technology* 24(1). 43–58.

Salton, Gerard & Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management* 24(5). 513–523.

Torkkola, Kari. 2004. Discriminative features for text/document classification. *Formal Pattern Analysis&Applications* 6(4). 301–308.

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

Vorontsov, Konstantin & Anna Potapenko. 2014. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In Dmitry I. Ignatov, Mikhail Yu. Khachay, Alexander Panchenko, Natalia Konstantinova & Rostislav E. Yavorsky (eds.), *Analysis of Images, Social Networks and Texts (AIST 2014)*, 29–46. (Communications in Computer and Information Science, vol. 436). Cham: Springer.

Wilson, Andrew. 2009. Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories. *Glottotheory* 2(2). 97–107.

Yule, George Udny. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Kateřina Pelegrinová

Quantitative characteristics of phonological words (stress units)

Abstract: The paper presents a pilot study of certain quantitative characteristics of phonological words in Czech. We use specific set of rules to segment a written text into the phonological words. Since a phonological word is a unit analogous to an orthographic word, we compare the size of the inventory and the rank-frequency distribution of both units.

Keywords: phonological word, orthographic word, size of the inventory, rank-frequency distribution, hapax legomenon

1 Introduction

Word is a frequently discussed concept in linguistics. There are many approaches to how to define word: grammatically, phonologically, lexically, orthographically etc. Whereas many of these (especially orthographic words) have been investigated in detail from the perspective of quantitative linguistics, studies devoted to phonological words are rather exceptional (Best 1998, Těšitelová 1985). Phonological word is a sound unit which is usually defined as a group of syllables with stress (Palková, 2017). In every language, there are rules or rather tendencies of positioning the stress, but these tendencies can be disturbed by other factors: the topic of conversation, the actual mood of the speaker etc. That is why it is difficult to define the boundaries of phonological words precisely. Palková (2004) examined the natural speech signal and sound properties of syllables groupings in Czech and determined the list of unambiguous rules for the segmentation of a text into phonological words.

In this paper, we study phonological words as defined by Palková (2004). Since little is known about them, we begin with the basic quantitative characteristics, such as the size of the inventory and the rank-frequency distribution. We also compare these characteristics with those of orthographic words.

Acknowledgment: This research was supported by the project of Moravian-Silesian Region: "Podpora talentovaných studentů doktorského studia na Ostravské univerzitě III" (07359/2019/RRC).

Kateřina Pelegrinová, Department of the Czech language, Faculty of Arts, University of Ostrava, e-mail: pelegrinovak@gmail.com

https://doi.org/10.1515/9783110763560-012

For the analysis, 16 short stories by two Czech authors were used (eight of them written by Karel Čapek and eight of them by Miloš Čermák). The length of these stories (see Tab. 1) is in the interval from 956 to 1867 orthographic words. Table 1 includes the titles of all the selected short stories and their abbreviations as they are used hereinafter in this paper:

Tab. 1: Titles of analysed texts and their abbreviations and length of the texts *L*.

Karel Čapek's texts	abbreviation	L	Miloš Čermák's texts	abbreviation	L
Ukradený kaktus	**KČ01**	1867	*Taky tě miluju*	**MČ01**	1579
Povídka starého kriminálníka	**KČ02**	1636	*Hádej, kdo ti píše?*	**MČ02**	1705
Zmizení pana Hirsche	**KČ03**	1852	*Neviděli jste Karla*	**MČ03**	956
Histrorie dirigenta Kaliny	**KČ04**	1624	*Láska nebeská*	**MČ04**	1143
Smrt barona Gandary	**KČ05**	1348	*Někdo ji tam má*	**MČ05**	1356
Jehla	**KČ06**	1377	*Štěstí je perspektivní produkt*	**MČ06**	1355
Obyčejná vražda	**KČ07**	1395	*Všechno je jinak*	**MČ07**	1308
Ušní zpověď	**KČ08**	1211	*Prodavač květin*	**MČ08**	1278

2 Methodology

A phonological word (sometimes called a stress-unit) participates in a hierarchy of sound units between the syllable and the intonational phrase. A phonological word is usually defined as a group of syllables with one stressed syllable. According to Palková (2017), however, the phonological word in Czech is determined not only by the stressed syllable but by the whole melody characteristic of this unit. As mentioned above, she derived formal rules to set the boundaries of phonological words for the purposes of text-to-speech synthesis. These rules are based on the study of the natural speech signal. Though phonological words are usually determined by phonological research on speech, as Palková (2004, p. 34) states, "the stress-unit in Czech is relatively easy to define in the written text due to its close connection with the word." In her study, boundaries of the phonological word are determined by the number of syllables in the orthographic word and by the position of this word in a clause-unit. "Clause-unit is defined as text between two punctuation marks in Czech orthography, with only one modification: the mark [,] is automatically added before the conjunction *a*. . . . This solution is possible due to the fact that orthography in

Czech relatively closely follow syntactic dependency relations." (Palková 2004, p. 35). Palková distinguishes three types[1] of position in a clause-unit:

a) Initial position (I) – the phonological word is in the first position within the clause-unit (i.e. after the punctuation marks)
b) Medial position (M) – the phonological word is in any position between the initial and the final positions
c) Final position (F) – the phonological word is in the position immediately before the boundary signal (punctuation) of the clause-unit.

Before describing the rules for segmentation, it is necessary to emphasize that the order of these rules is mandatory. Further, the rules described in this paper differ in some details from the rules of Palková (2004). Specifically, we extend the list of monosyllabic prepositions as they are described in the study of Czech propositions based on the Czech national corpus (Čermák et al., 2009). This list includes monosyllabic prepositions, including those derived from Latin: *na* 'on', *do* 'into', *o* 'about', *za* 'behind', *pro* 'for', *po* 'after', *od* 'from', *u* 'at', *při* 'by', *před* 'before', *bez* 'without', *pod* 'under', *nad* 'above', *přes* 'over', *dle* 'according to', *skrz* 'through', *vstříc* 'towards', *de* 'de', *zpod* 'from under', *in* 'in', *ad* 'ad', *per* 'per', *dík* 'thanks', *vně* 'outside', *vzdor* 'despite', *ob* 'ob', *via* 'via', *ex* 'ex', *stran* 'concerning', *znad* 'from above', *ke* 'to', *ve* 'in', *ze* 'from', *se* 'with'.

In Czech, there is a strong tendency for phonological words to consist of more than just one syllable. Thus, monosyllabic words (hereinafter MW's) tend to join other words and special rules must be applied. Moreover, there is a special group of words among the MW's: the clitics. (Clitics in F position are treated differently than the other MW's.) We adopt the list of clitics from Palková (2004): it includes the monosyllabic variants of the auxiliary verb *být* 'to be': *jsem* 'I am', *jsi* 'you [sg.] are', *je* 'he/she/it is', *jsme* 'we are', *jste* 'you [pl.] are', *jsou* 'they are', *bych* 'I would', *bys* 'you would', *by* 'he/she/it would'; and the short forms of pronouns which have another, full, form: *mi* 'me [dat.]', *ti* 'you [dat.]', *tě* 'you [acc.]', *ho* 'him/it [acc.]', *mu* 'him [acc.]', *ji* 'her' [acc.], *je* 'them' [acc.]; and the reflexive pronouns *se* [acc.] and *si* [dat.].

Below, each of the above mentioned cases is illustrated by an example. In these examples, we mark boundaries of the relevant phonological word by the hash signs; where necessary, boundaries between syllables are marked by a dash.

1 Actually, she distinguishes four types in her study. But the fourth one is needed only for the text–to–speech transformation and is not relevant for the text segmentation into phonological words.

(1) Non-syllabic prepositions (*k* 'to', *s* 'with', *v* 'in', *z* 'from') are joined to the following word:
Přišel *#k* *oknu.#*
come.PTCP.SG.M to window.DAT.SG
'He came to the window.'

(2) Monosyllabic prepositions are joined to the following word:
Přišel *#pod* *okno.#*
come.PTCP.SG.M under window. ACC.SG
'He came under the window.'

(3) Polysyllabic words that stand next to each other are treated as an independent phonological word:
#Přišel# *#pozdě#* *#večer.#*
come.PTCP.SG.M late in_the_evening
'He came late in the evening.'

(4) MW in the F position:
(a) forms an independent phonological word:
Přišel *#včas.#*
come.PTCP.SG.M in_time
'He came in time.'
(b) if MW in this position is clitic, it must be processed as a MW in the M position (see below for the rules):
#Přišel *jsem.#*
come. PTCP.SG.M aux.1. SG
'I came.'

(5) a single MW in the I position is always joined to the following word:
#On *přišel#* *pozdě.*
he come.PTCP.SG.M late
'He came late.'

(6) More MW's in the I position standing next to each other:

(a) Two to four MW's are regarded as a single phonological word:

#Chtěl jsem si dát#
want.PTCP.SG.M aux.1.SG REFL.DAT give.INF
pivo.
beer.ACC.SG
'I wanted to have a beer.'

(b) Five MW's are divided into two phonological words in the ratio 3:2:

#Chtěl jsem si# #tam dát#
want.PTCP.SG.M aux.1.SG REFL.DAT there give.INF
pivo.
beer.ACC.SG
'I wanted to have a beer there.'

(c) Six MW's are divided into two phonological words in the ratio 3:3:

#Chtěl jsem si# #tam dát
want.PTCP.SG.M aux.1.SG REFL.DAT there give.INF
sám# pivo.
alone beer.ACC.SG
'I wanted to have a beer there alone.'

(d) Seven MW's are divided into two phonological words in the ratio 4:3:

#Chtěl jsem si tam# #dát
want.PTCP.SG.M aux.1.SG REFL.DAT there give.INF
sám jen# pivo.
alone just beer.ACC.SG
'I wanted to have just a beer there alone.'

(7) MW's in the M position:

(a) A single MW is joined to the preceding unit:

#Viděl jsem# kočku.
aux.1.SG aux.1.SG cat.ACC.SG
'I saw a cat.'

(b) Two MW's preceded by a unit that contains two to four syllables are joined to the preceding unit:

#Při-šel jsem tam# pozdě.
come.PTCP.SG.M aux.1.SG there late
'I came there late.'

(c) Two MW's preceded by a unit that contains five or more syllables form an independent phonological word:

Ne-za-mý-šle-li *#jsme* *si#* *koupit*
not_intend.PTCP.PL.M aux.1.PL REFL.DAT buy.INF
pivo.
beer.ACC.SG
'We did not intend to buy a beer.'

(d) Three MW's preceded by a unit that contains two or three syllables are joined to the preceding unit:

– *#Chtě-li* *jsme* *si* *dát#*
want. PTCP.PL.M aux.1.PL REFL.DAT give.INF
pivo.
beer.ACC.SG
'We wanted to have a beer.'

(e) Three MW's preceded by a unit that contains four or more syllables form an independent phonological word:

– *Ne-za-mý-šle-li* *#jsme* *si* *dát#*
not_intend.PTCP.PL.M aux.1.PL REFL.DAT give.INF
pivo.
beer.ACC.SG
'We did not intend to have a beer.'

(f) Four MW's preceded by a unit that contains two syllables are joined to the preceding unit:

– *#Chtě-li* *jsme* *si* *tam* *dát#*
want.PTCP.PL.M aux.1.PL REFL.DAT there give.INF
pivo.
beer.ACC.SG
'We wanted to have a beer there.'

(g) Four MW's preceded by a unit that contains three or four syllables: the first two MW's are joined to the preceding unit, the last two form an independent phonological word:

– *#Ne-chtě-li* *jsme* *si#* *#tam*
not_want.PTCP.PL.M aux.1.PL REFL.DAT there
dát# *pivo.*
give.INF beer.ACC.SG
'We did not want to have a beer there.'

(h) Four MW's preceded by a unit that contains five or more syllables form an independent phonological word:
– *Ne-za-mý-šle-li #jsme si tam*
not_intend.PTCP.PL.M aux.1.PL REFL.DAT there
dát# pivo.
give.INF beer.ACC.SG
'We did not intend to have a beer there.'

(i) Five MW's preceded by a unit that contains two or three syllables: the first two MW's are joined to the preceding unit, the last three form an independent phonological word:
– *#Chtě-li jsme si# #tam dát*
want.PTCP.PL.M aux.1.PL REFL.DAT there give.INF
jen# pivo.
just beer.ACC.SG
'We wanted to have just a beer there.'

(j) Five MW's preceded by a unit that contains four syllables form two independent units in the ratio 2:3:
– *Za-mý-šle-li #jsme si# #tam dát*
intend.PTCP.PL.M aux.1.PL REFL.DAT there give.INF
jen# pivo.
just beer.ACC.SG
'We intended to have just a beer there.'

(k) Five MW's preceded by a unit that contains five or more syllables form two independent units in the ratio 3:2:
– *Ne-za-mý-šle-li #jsme si tam#*
not_intend.PTCP.PL.M aux.1.PL REFL.DAT there
#dát jen# pivo.
give.INF just beer.ACC.SG
'We did not intend to have just a beer there.'

(l) Six MW's preceded by a unit that contains two or three syllables: the first two MW's are joined to the preceding unit, the last four MW's form two independent phonological words in the ratio 2:2:
– *#Chtě-li jsme si# #tam dát#*
want.PTCP.PL.M aux.1.PL REFL.DAT there give.INF
#pak jen# pivo.
then just beer.ACC.SG
'We wanted to have then just a beer there.'

(m) Six MW's preceded by a unit that contains four or more syllables form two independent phonological words in the ratio 3:3:
– *Za-mý-šle-li #jsme si tam# #dát*
intend.PTCP.PL.M aux.1.PL REFL.DAT there give.INF
pak jen# pivo.
then just beer.ACC.SG
'We intended to have then just a beer there.'

In this research, we define word as an orthographic unit, i.e. a string of letters delimited by spaces (punctuation was removed before the segmentation). To illustrate this segmentation, the following sentence is be used as an example:

Můj králík rád žere noviny na gauči. 'My rabbit likes to eat newspaper on the sofa.'

This Czech sentence consists of seven orthographic words.

3 Results

3.1 Inventory size

The size of the inventory consists of unique units (i.e. types, as opposed to tokens) that occur in the text at least once.

The inventory size of particular texts is presented in Tab. 2. N_{PW} is the inventory size of phonological words, N_W is the inventory size of words, R_N is the difference between these inventory sizes, i.e. $R_N = N_{PW} - N_W$. A positive value indicates a larger inventory of phonological words than (orthographical) words, and vice versa. In all texts except for one (MČO3) the inventory of phonological words is larger.

Tab. 2: Size of the inventories of phonological words N_{PW} and words N_W, difference of these values R_N.

text	N_{PW}	N_W	R_N	text	N_{PW}	N_W	R_N
KČ01	1053	955	98	MČ01	888	823	65
KČ02	882	772	110	MČ02	954	875	79
KČ03	908	751	157	MČ03	605	611	−6
KČ04	926	835	91	MČ04	718	699	19
KČ05	764	682	82	MČ05	796	737	59

Tab. 2 (continued)

text	N_{PW}	N_W	R_N	text	N_{PW}	N_W	R_N
KČ06	761	685	76	MČ06	772	707	65
KČ07	769	700	69	MČ07	809	774	35
KČ08	688	622	66	MČ08	768	714	54

We assume that there are factors that determine the size of the inventory of phonological words. First, it seems that the length of the text (see Tab. 1) plays an important role. In shorter texts, there is not enough "space" for as many unique combinations of words as in longer texts. This tendency can be seen in Fig. 1.

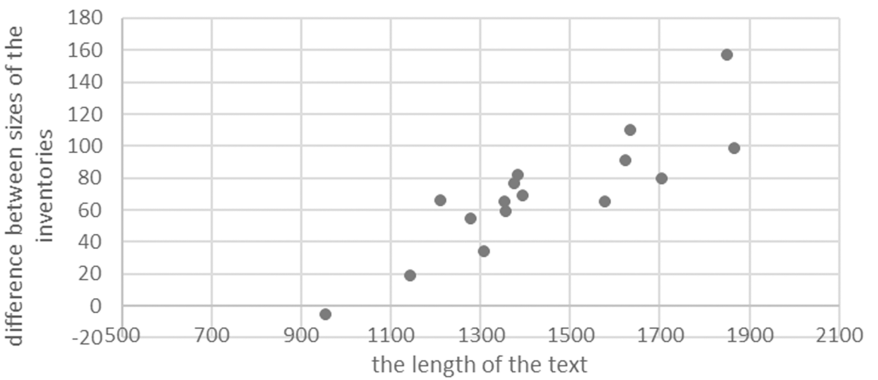


Fig. 1: The relationship between text length in words and the difference of the inventory sizes.

Second, the size of the inventory should be also influenced by the total number of monosyllabic words in the text. According to the segmentation rules, monosyllabic words tend to join another unit rather than forming an independent phonological word. On the other hand, words with more syllables form independent phonological words (unless they are merged with preceding or following MW's). It follows that the difference between inventory sizes will be larger in texts that contain more monosyllabic words: monosyllabic words are joined together or to other words, more combinations and more unique phonological words emerge, and thus the inventory of phonological words increases. Consequently, texts with less monosyllabic words will have similar inventory sizes of phonological words and of orthographic words. In the sample, monosyllabic words were the most frequent ones with respect to the word length, with one

exception, which is the text MČ03. This text is also the shortest text in the sample. We may explain the negative value of the difference between the inventories for the text MČ03 as a consequence of these two factors (i.e. the text length and a the relatively low proportion of MW's).

3.2 Rank-frequency distribution

One way of describing a system is the analysis of its rank-frequency distribution. Rank-frequency distributions of language units (with the exception of phonemes and graphemes, see Grzybek at al. 2009) follow some of the Zipfian distributions (see several contributions in Köhler et al. 2005, and specifically for words e.g. Popescu et al. 2009). We expect phonological words to "copy" the behaviour of words in this respect, i.e. to follow one of the Zipfian distributions. We decided to model the rank-frequency distribution of phonological words by the Zipf-Mandelbrot distribution which is defined as

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

The results of the rank-frequency analysis can be found in Tab. 3.

Tab. 3: Parameters of the Zipf-Mandelbrot distribution applied to model rank-frequency distribution of phonological words.

text	words		phonological words		text	words		phonological words	
	a	b	a	b		a	b	a	b
KČ01	0.7628	2.0932	0.3855	5.4493	**MČ01**	0.7623	1.9310	0.5734	79.4903
KČ02	0.7935	1.7240	0.5085	52.3553	**MČ02**	0.7829	1.9794	0.4857	3.6925
KČ03	0.8567	2.2938	0.5181	3.2764	**MČ03**	0.7009	2.7744	0.3985	2.5697
KČ04	0.7884	1.3923	0.3884	3.3576	**MČ04**	0.7139	2.1184	0.3547	5.1593
KČ05	0.7949	2.4753	0.5439	15.4904	**MČ05**	0.7690	1.8509	0.3887	2.5001
KČ06	0.7896	1.6975	0.4977	50.8786	**MČ06**	0.7794	1.6846	0.5260	37.3073
KČ07	0.7895	2.1185	0.3922	2.9226	**MČ07**	0.7227	2.4787	0.4214	3.0506
KČ08	0.7869	1.8438	0.3918	2.3811	**MČ08**	0.7367	1.6810	0.6030	3.3729

For all texts, the p-value is above 0.99, i.e. the results do not reject our assumption on the level of significance $\alpha = 0.001$.

We also expect that more hapax legomena (units that occur only once in the text) will be observed (in comparison with the distribution of words). This is

implied in the very nature of phonological words: phonological words can consist of more than just one word and thus more unique units can emerge.

In Tab. 4 differences between the numbers of hapax legomena for phonological words and orthographic words are presented.

Tab. 4: Number of hapax legomena of phonological words HL_{PW} and words HL_{W} and the differences between these values R_{HL}.

text	HL_{PW}	HL_{W}	R_{HL}	text	HL_{PW}	HL_{W}	R_{HL}
KČ01	931	706	225	**MČ01**	782	612	170
KČ02	763	554	209	**MČ02**	835	639	196
KČ03	770	512	258	**MČ03**	546	492	54
KČ04	816	637	179	**MČ04**	642	550	92
KČ05	650	487	163	**MČ05**	700	557	143
KČ06	660	500	160	**MČ06**	690	530	160
KČ07	675	519	156	**MČ07**	723	603	120
KČ08	590	453	137	**MČ08**	671	525	146

As expected, the results show the tendency of phonological words to produce more hapax legomena in texts than graphical words. This fact is also connected to the inventory size: in Fig. 2 we can see that texts that contain more hapax legomena of phonological words tend to have a larger inventory of phonological words and consequently the difference between inventories is also larger.

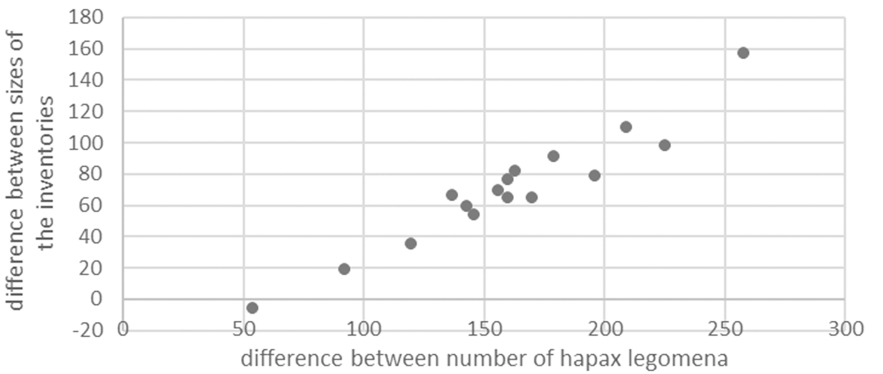


Fig. 2: The relationship between the differences of inventory size and number of hapax legomena.

4 Conclusion

This paper is a pilot study of phonological words in Czech. Phonological words follow the same tendencies and laws that are typical of orthographic words, such as, for example, rank-frequency distribution. The results show that phonological words follow the same Zipfian model but the resulting parameters differ from the parameters for orthographic words. A comparison of inventory sizes of both units shows some differences.

The current research will be followed by an investigation of other properties of phonological words, such as length in terms of number of syllables. Preliminary results indicate that these properties, too, are analogous to those of orthographic words.

References

Best, Karl-Heinz. 1998. Results and perspectives of the Göttingen project on quantitative linguistics. *Journal of Quantitative Linguistics* 5. 155–162.

Best, Karl-Heinz & Otto Rottmann. 2017. *Quantitative Linguistics: An Invitation*. Lüdenscheid: RAM-Verlag.

Čech, Radek, Petr Pajas & Ján Mačutek. 2010. Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics* 17. 291–302.

Čermák, František, Václav Cvrček, Vladimír Petkevič & Tomáš Jelínek. 2009. *Statistiky češtiny* [Statistics of the Czech language]. Praha: Nakladatelství Lidové noviny.

Grzybek, Peter, Emmerich Kelih & Ernst Stadlober. 2009. Slavic letter frequencies: A common discrete model and regular parameter behaviour? In Reinhard Köhler (ed.), *Issues in Quantitative Linguistics*, 17–33. Lüdenscheid: RAM-Verlag.

Köhler, Reinhard, Gabriel Altmann & Rajmund G. Piotrowski (eds.). 2005. Quantitative Linguistics: An International Handbook. Berlin/New York: Walter de Gruyter.

Palková, Zdena. 2004. The set of phonetic rules as a basis for the prosodic component of an automatic TTS synthesis in Czech. In Zdena Palková & Jitka Janíková (eds.), *Phonetica Pagensia* 10, 33–46. Praha: Karolinum.

Palková, Zdena. 2017. Mluvní takt [Stress unit]. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [New encyclopedic dictionary of the Czech language]. https://www.czechency.org/slovnik/MLUVN%C3%8D%20TAKT (accessed 20 March 2020)

Popescu, Ioan Iovitz et al. 2009. *Word Frequency Studies*. Berlin/New York: de Gruyter.

Těšitelová, Marie. 1985. *Kvantitativní charakteristiky současné češtiny* [Quantitative characteristics of the contemporary Czech language]. Praha: Academia.
Wimmer, Gejza, Gabriel Altmann, Luděk Hřebíček, Slavomír Ondrejovič & Soňa Wimmerová. 2003. *Úvod do analýzy textov* [An introduction to the text analysis]. Bratislava: VEDA

Software

Altmann Fitter. 2004. Lüdenscheid: RAM-Verlag.

Haruko Sanada

Explorative study on the Menzerath-Altmann law regarding style, text length, and distributions of data points

Abstract: The present study empirically investigated the Menzerath-Altmann Law for two data sets of newspaper texts which addressed the same topics but have different readerships (i.e. adults and children). Articles in the two data sets have a different style for different readers, and text, sentence, and clause lengths for adults are much longer. However, the relationships among these lengths are similar to each other in the two data sets. From an observation of our data and simulations with imaginary data, the following assumptions can be made: (1) the distribution of data points is systematically related to the distribution of the sum of clause lengths; (2) the distribution of data points seems to be related to averages of clause length if the number of data points is a function of sentence length; and (3) averages of clause length do not directly depend on the text length because the average is determined by a ratio of the number data points and a sum of clause lengths. The number of data points as a function of linguistic properties has not been considered in former studies on the Menzerath-Altmann Law.

Keywords: Menzerath-Altmann law, frequency, newspaper, Japanese, synergetic linguistics

1 Introduction and hypothesis

This study uses texts from Japanese newspapers to conduct an empirical investigation of the Menzerath-Altmann Law (MAL). Altmann generalised the linguistic characteristics described by Menzerath (1954) by stating, 'The longer a language construct the shorter its components (constituents)' (Altmann 1980:1). Since then, several studies have been conducted on the MAL. Most have focused on providing answers regarding (Cramer 2005a) which languages and linguistic levels the MAL would hold true for and which definition of or measurement for linguistic properties would be proper to use for the MAL. In their research, Köhler (1984) and Cramer (2005b) interpreted the parameters of the MAL formula (see below). In the present study, we hypothesised that a

Haruko Sanada, Faculty of Economics, Rissho University, Japan, e-mail: hsanada@ris.ac.jp

https://doi.org/10.1515/9783110763560-013

difference in the type of text addressing the same topic may affect parameters *A* and *b* of the MAL formula.

The present study empirically investigated the MAL in two data sets of newspaper text, not used in former studies, addressing the same topics, but for different readerships (adults and children). Therefore, our hypothesis is as follows: the differences between readerships can be expressed by the parameters of the MAL formula. The present study also observed how individual frequencies of data affect a distribution of DP (data points) as a total of frequencies, sums, or averages of linguistic properties as dependent variables of the MAL. Former studies have not considered this point.

2 Data

We prepared two data sets extracted from newspaper articles. One is extracted from *Mainichi Shinbun* for adults (A), and the other one is from *Mainichi Shogakusei Shinbun* for elementary school children (B). Both have been published daily by the same newspaper publishing company, Mainichi Shinbunsha, since 1872 for (A) and 1936 for (B). The text was downloaded in August 2018 from the publisher's website (https://mainichi.jp/). The publication dates were between 23 July and 5 August, 2018, and articles published 14 days before the date of download are the oldest available on the website. Though the topics of all the articles in the two newspapers do not correspond to each other, 24 articles were extracted from each newspaper addressing similar topics meant for either adults or children, for a total of 48 articles. Articles for children are published two or three days after articles for adults, and are similar to the articles for adults not only in their topics but also in their structures. Therefore, it seems that the articles for children are produced through a revision of the articles for adults. Lengths of morphemes and characters in each article for children were almost half that of those in the articles for adults. Exact numbers are shown later in this paper.

The topics of these articles included society, politics, current events, and meteorological reports. Articles for children were written in the polite style (*desu/masu*), which is often used in school textbooks for children or private letters, while articles for adults were written in the ordinary style (*dearu*). All *kanji* (Chinese script) in the articles for children had margin notes in *hiragana* (phonetic Japanese) to assist with the reading of *kanji*. In Japan, children learn approximately 1,000 *kanji* characters over a span of six years in elementary school. Children can therefore read margin notes if they see a *kanji* character which they do not yet know. Articles for children do not include words using *kanji* characters

except those 1,000, and words comprised of *kanji* characters other than those 1,000 in articles for adults are replaced with other words or expressions in articles for children. In previous studies comparing Japanese texts for children to ones for adults, Yuasa (2006, 2015, 2016) analysed the characteristics, style, and use of *kanji*, and the choice of words or expressions in our data is in line with her studies.

In order to analyse sentences and obtain morphemes from the 48 articles, we employed the Japanese morphological analyser *MeCab* (Graduate Schools of Informatics in Kyoto University et al. 2008), along with the electronic dictionary *UniDic* (National Institute for Japanese Language and Linguistics 2008). The *UniDic* employs morphemes according to the definition of 'short unit', as developed by the National Institute of Japanese Language and Linguistics (National Language Research Institute 1964), and we also employed this definition in the present study. Due to technical reasons related to software, Arabic numerals (1, 2, 3, . . ., 1000, 10000, etc.) were replaced by numbers in *kanji* (一, 二, 三, . . ., 千, 万, etc.). This replacement affects the number of characters, but it can be ignored because the number of characters for the Arabic numerals was very small when compared to the total number of characters.

The present study uses 48 newspaper articles to examine the following: text length measured in sentences (TL(S)), sentence length in clauses (SL(C)), and clause length in morphemes (CL(M)). Furthermore, morpheme length measured by the number of *kanji* characters (ML(Char)) and morpheme length measured in the number of characters when *kanji* is converted to *hiragana* (ML(Rd)) are also examined. An outline of the sizes of linguistic entities for the two data

Tab. 1: Outline of the sizes of linguistic entities for two data sets from newspapers.

(A) Newspaper for adults	Sentence	Clause	Morpheme	Char.	Rd.
Total numbers	311	892	8404	13328	19948
Average	13.0	37.2	350.2	555.3	831.2
Maximum	26	74	687	1026	1632
Minimum	2	4	68	111	199
Variances	36.21	322.72	24958.14	61297.39	140417.81
(B) Newspaper for children	**Sentence**	**Clause**	**Morpheme**	**Char.**	**Rd.**
Total numbers	150	430	4194	6754	9789
Average	6.3	17.9	174.8	281.4	407.9
Maximum	10	28	280	488	677
Minimum	3	8	101	174	238
Variances	3.94	25.41	1706.69	4528.49	10119.53

Abbreviations: Char: Number of characters of morphemes including *kanji* per text. Rd: Number of characters when *kanji* is converted to *hiragana* (phonetic Japanese) per text.

sets is shown in Tab. 1. Maximum quantities, averages, and variances in (A) were greater than ones in (B) on all linguistic levels.

3 Tests of the Menzerath-Altmann law and results

From our previous studies, it can be assumed that the text in these newspapers has a certain homogeneity, because the MAL held true with a group of sentences which were extracted from a newspaper corpus (Sanada: 2016). Therefore, we treated the two data sets like two long texts, which included 24 articles each. The individual articles for children were too short for the MAL.

We investigated four relationships between the two newspapers with the following three linguistic levels:

(1) Text length measured in the number of sentences (TL(S)) as *x* and average sentence length in clauses (SL(C)) as *y*;
(2) Sentence length measured in the number of clauses (SL(C)) as *x* and average clause length in morphemes (CL(M)) as *y*; and
(3) Clause length in morphemes (CL(M)) as *x* and average morpheme length measured in the number of characters with *kanji* (ML(Char)) as *y1*, or average morpheme length measured in the number of characters when *kanji* was converted to *hiragana* (ML(Rd)) as *y2*.

A general formula (1) for the MAL, as shown by Altmann (1980), was employed to obtain a regression curve for four relationships (levels 1 to 3) of two data sets (A and B). Averages of *y* were employed for a regression curve if the number of DP for *y* was greater than 10.

$$y = A\, x^b e^{-c*x} \tag{1}$$

Three of eight relationships (i.e. (2) SL(C) and CL(M) in (A), and (1) TL(S) and SL(C) and (2) SL(C) and CL(M) in (B)), were fitted to a regression curve while the rest were not fitted. The eight relationships and their regression curves are shown in Figs. 1 through 6. Averages of *x* drawn with white circles in the figures were excluded from calculations of the regression curve, because their number of DP was less than 10. There is no DP whose number is smaller than 10 in Fig. 5.

The original data for the relationships between (2) SL(C) as *x* and all averages of CL(M) as *y* in (A) and in (B) are shown in Tabs. 2 and 3, respectively,

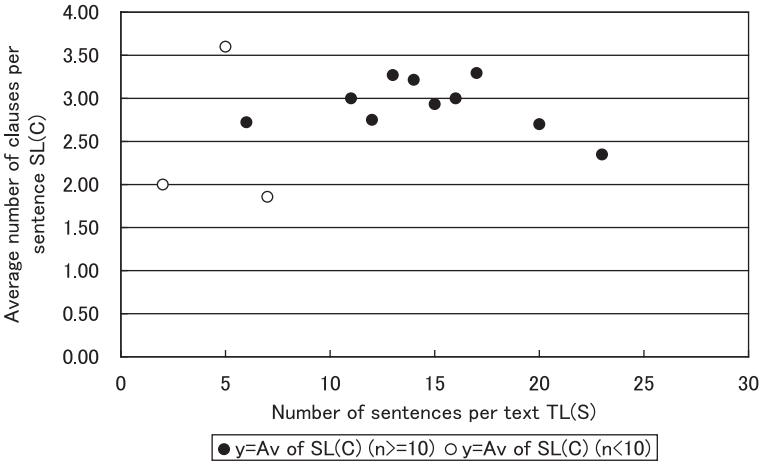


Fig. 1: (1) TL(S) as x and averages of SL(C) as y in (A) the newspaper for adults.

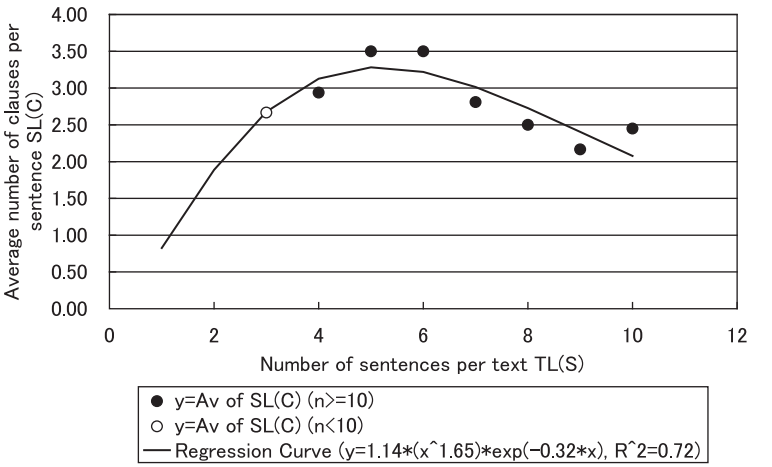


Fig. 2: (1) TL(S) as *x* and averages of SL(C) as *y* in (B) the newspaper for children, with a regression curve $y = 1.14 * (x^{1.65}) * exp^{(-0.32*x)}$ ($R^2 = 0.72$).

which correspond to Figs. 3 and 4. The numbers of DP for *y* and their variances are also listed in Tabs. 2 and 3. Using *Altmann-Fitter* software, we confirmed that the distribution of the number of DP of (2) SL(C) and CL(M) in (B) fits the Hyper-Pascal, Dacey-negative binomial, or the extended positive

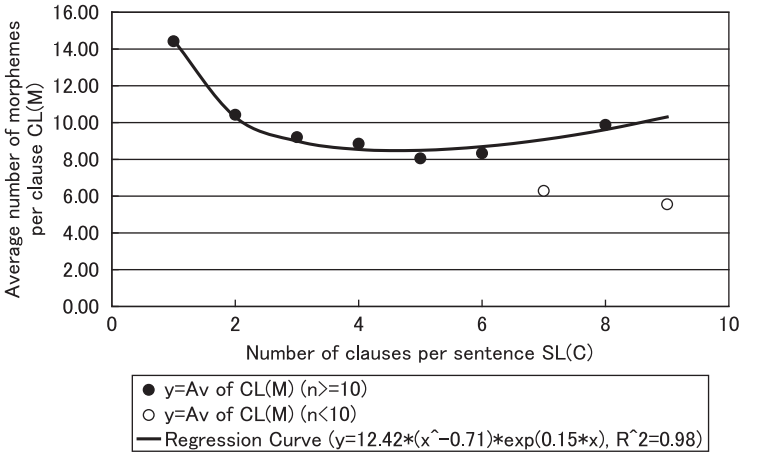


Fig. 3: (2) SL(C) as *x* and averages of CL(M) as *y* in (A) the newspaper for adults.

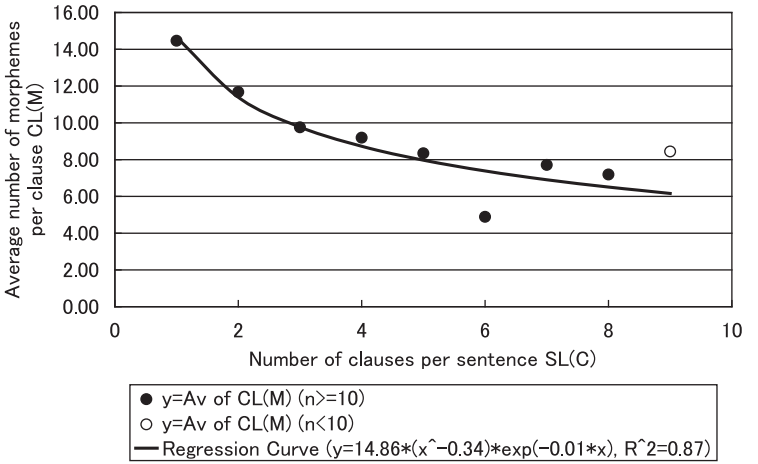


Fig. 4: (2) SL(C) as *x* and averages of CL(M) as *y* in (B) the newspaper for children.

negative binomial distributions,[1] though we could not find a good theoretical distribution for one number in the DP for (2) SL(C) and CL(M) in (A).

1 The positive negative binomial distribution is also known as the zero-truncated negative binomial distribution (Wimmer et al. 1991: 540).

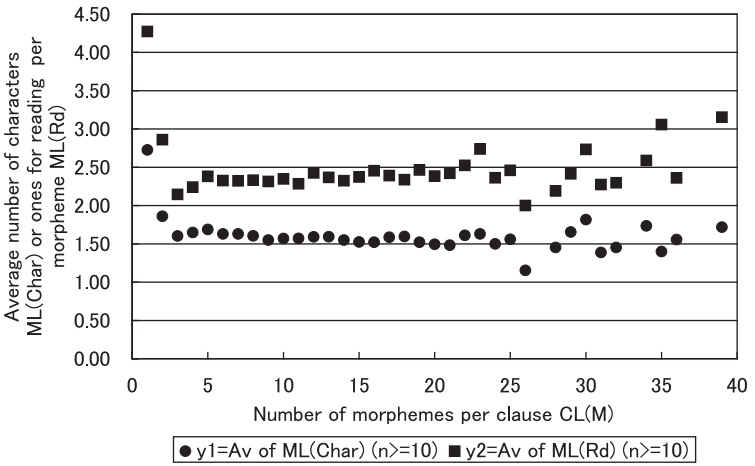


Fig. 5: (3) CL(M) as *x*, averages of ML(Char) as *y1* and averages of ML(Rd) as *y2* in (A) the newspaper for adults.

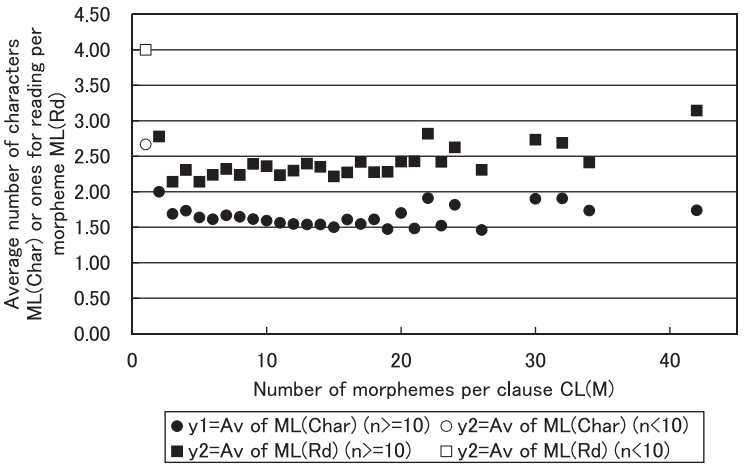


Fig. 6: (3) CL(M) as *x*, averages of ML(Char) as *y1* and averages of ML(Rd) as *y2* in (B) the newspaper for children.

Tab. 2: Relationships for (2) SL(C) and CL(M) in (A) the newspaper for adults.

SL(C) as *x*	1	2	3	4	5	6	7	8	9
Average of CL(M) as *y*	14.42	10.43	9.21	8.85	8.05	8.33	6.29	9.88	5.56
Theoretical values of MAL	14.47	10.3	8.99	8.54	8.49	8.69	9.08	9.62	10.31
Number of DP	59	176	213	192	140	72	7	24	9
Variances of CL(M)	67.63	30.44	30.97	34.15	28.22	25.89	7.06	34.53	2.91

$y = 12.42 * (x^{-0.71}) * exp^{(0.15*x)}$. $R^2 = 0.98$.

Tab. 3: Relationships of (2) SL(C) and CL(M) in (B) newspaper for children.

SL(C) as *x*	1	2	3	4	5	6	7	8	9
Average of CL(M) as *y*	14.46	11.68	9.75	9.19	8.35	4.89	7.71	7.19	8.44
Theoretical values of MAL	14.66	11.39	9.77	8.73	7.97	7.38	6.91	6.51	6.16
Number of DP	26	90	114	88	55	18	14	16	9
Variances of CL(M)	72.48	30.22	30.31	29.18	24.52	8.65	12.63	11.15	16.25

$y = 14.86 * (x^{-0.34}) * exp^{(-0.01*x)}$. $R^2 = 0.87$.

4 Discussion (1): Differences between two data sets with the same topics

We hypothesised that a difference in the type of texts addressing the same topics may affect parameters *A* and *b* of the MAL formula. According to Köhler (1984) and Cramer (2005b), parameter *A* represents a quantity which presumably depends on the language and linguistic level, and parameter *b* represents a shortening tendency and describes the range of structural information.

Observing parameter *b* in Tab. 2 (–0.71) and Tab. 3 (–0.34), we cannot immediately explain why (A) has more complicated text structures than (B). Table 1 shows that maximum quantities, averages, and variances of (A) are greater than ones of (B) on all linguistic levels. However, averages of CL(M) to SL(C) in Tabs. 2 and 3 do not have a large number of differences, and from these numbers it cannot be assumed that (A) has more complicated text structures than (B).

By means of parameter *A* as shown in Tab. 2 (12.42) and Tab. 3 (14.86), we cannot determine if there is a difference between the two newspapers. However, a regression curve displayed in Fig. 2 for a relationship between TL(C) and SL(C) shows a different shape from those in Figs. 3 and 4. In addition, the

scattering of data drawn in Fig. 1 through 6 are similar between the two newspapers on the same linguistic level (i.e. Figs. 1 and 2, Figs. 3 and 4, and Figs. 5 and 6), even though regression curves could not be obtained for some of these six relationships. Therefore, it can be assumed that the shape of a regression curve depends on the linguistic level.

The length of morphemes (ML) is almost stable, which does not depend on any measurement unit (ML(Char) or ML(Rd)). The reason can be assumed that the 'short unit' is employed as the definition, which often consists of two *kanji* characters.

We also made smaller data sets from the 24 articles of (A) and (B) each for the relationships of (2) SL(C) and CL(M) (i.e. two data sets from 12 articles, three data sets from eight articles, four data sets from six articles, and six data sets from four articles). Obtained regression curves show very similar tendencies.

5 Discussion (2): Methodological observations of the Menzerath-Altmann law

From our data, there is no dramatic difference in the relationships between (2) SL(C) and CL(M) though text, sentence, clause, and morpheme lengths are different in the two data sets of (A) and (B). As shown in Tab. 1, maximum quantities, averages, and variances of (A) are greater than the ones of (B) for all linguistic levels. Sentence style and word choice are also different between (A) and (B), even if the topics are similar.

Here a question arises: which linguistic properties affect a curve of the MAL, excluding the language and the linguistic level pointed out by Köhler (1984) and Cramer (2005b)? From our observations, a distribution of the number of DP affects the curve. We describe the characteristics of linguistic properties in the present paper because these are still explorative findings.

If we take the relationship (2) SL(C) and CL(M) as an example, the calculations can be explained as follows: for SL = 1 as x and the number of DP of CL as y is multiplied by each length of CL (1, 2, 3, etc.) and a sum of CL is obtained which corresponds to SL = 1. The sum of CL is divided by a sum of the number of DP of CL and an average length of CL as y is obtained when SL = 1. The process can be regarded as a sum of CL and an average of CL are obtained while the number of DP is weighted by CL (see formulae 2 and 3).

$$\text{Sum of CL} = (\text{CL}_1{}^*\text{DP}_1 + \text{CL}_2{}^*\text{DP}_2 + \ldots \text{CL}_n{}^*\text{DP}_n) \tag{2}$$

and

$$\text{Average of CL} = \text{Sum of CL}/(\text{DP}_1 + \text{DP}_2 + \ldots \text{DP}_n) \quad (3)$$

Table 4 shows calculations used to obtain the averages of CL(M) by SL(C) and their regression formula employing the MAL for (B) the newspaper for children. The numbers of DP by CL and SL, averages of CL, and theoretical values correspond to numbers in Tab. 2. The theoretical values of the MAL in Tab. 4 also correspond to the regression curve in Fig. 4.

Tab. 4: A part of distributions of data points (DP) and sums for (2) CL(M) and SL(C) for (B) the newspaper for children (corresponding to Tab. 3).

SL(C) as *x*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Distribution of DP by SL(C) and CL(M)									
DP for CL(M) = 1				1	1	1			
DP for CL(M) = 2	1	1	1	2	2	1		1	
DP for CL(M) = 3		2	9	8	4	5	1	1	1
. . .									
DP for CL(M) = 42	1								
Total number of DP	26	90	114	88	55	18	14	16	9
DP* CL(M) by SL(C)									
DP * 1 as CL(M)				1	1	1			
DP * 2 as CL(M)	2	2	2	4	4	2		2	
DP * 3 as CL(M)		6	27	24	12	15	3	3	3
. . .									
DP * 42 as CL(M)	42								
Sum of CL by SL	376	1051	1112	809	459	88	108	115	76
Average of CL by SL (Sum of CL/DP)	14.46	11.68	9.75	9.19	8.35	4.89	7.71	7.19	8.44
Theoretical values of the MAL	14.66	11.39	9.77	8.73	7.97	7.38	6.91	6.51	6.16
Variances of CL	72.48	30.22	30.31	29.18	24.52	8.65	12.63	11.15	16.25

The total number of DP and the sum of CL by SL have a certain correlation because the sum of CL is systematically calculated as weighted by each SL. It can be visually confirmed in Fig. 7. It is also confirmed that variances of CL have no

direct correlation with the total number of DP (see Fig. 8). We assumed that a distribution of the number of DP must be a function of SL because averages of CL is a function of SL if a regression curve of the MAL is obtained, and the averages of CL is provided by the number of DP and sums of CL by SL. Theoretical values of DP and the sum of CL are calculated employing the MAL formula (1), and we obtained good results with $R^2 = 0.98$ for both (see Figs. 9 and 10).

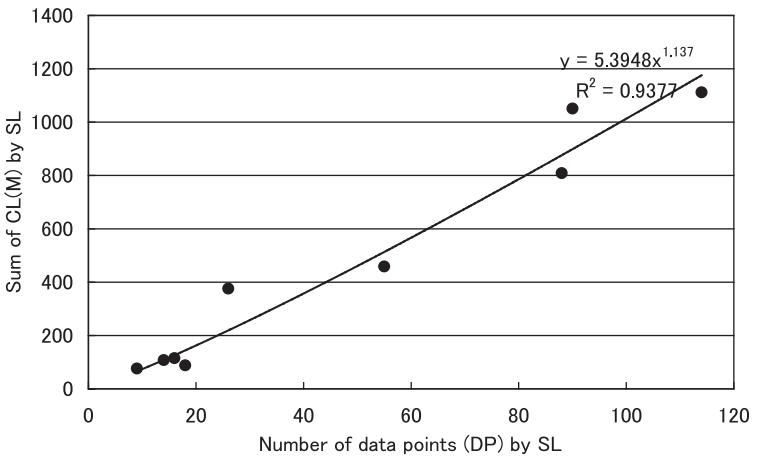


Fig. 7: A relationship between the total number of DP by SL as *x* and sum of CL by SL as *y*.

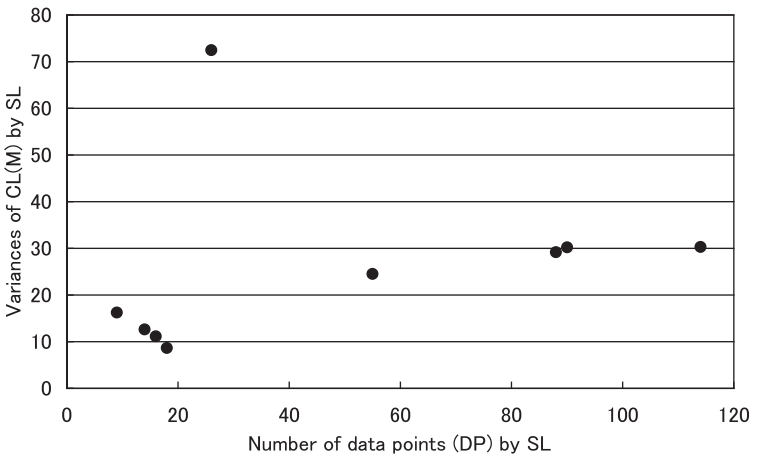


Fig. 8: A relationship between the total number of DP by SL as *x* and variances of CL as *y*.

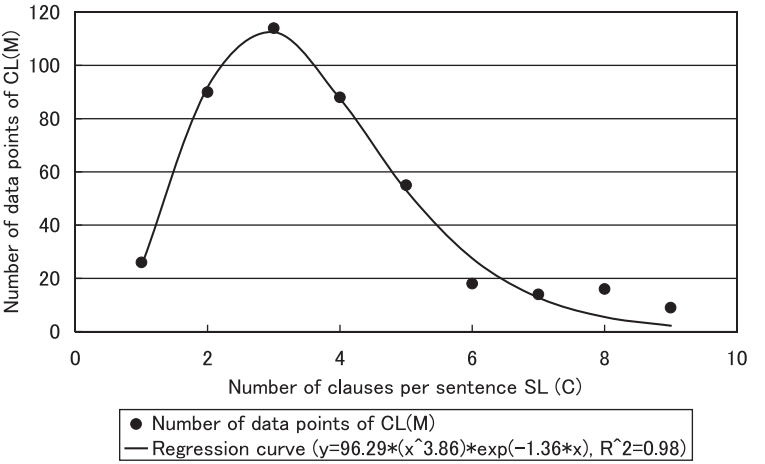


Fig. 9: A relationship between SL(C) as *x* and the number of DP of CL(M) as *y* in (B) the newspaper for children with a MAL regression curve $y = 96.29 * (x^{3.86}) * exp^{(-1.36*x)}$ ($R^2 = 0.98$).

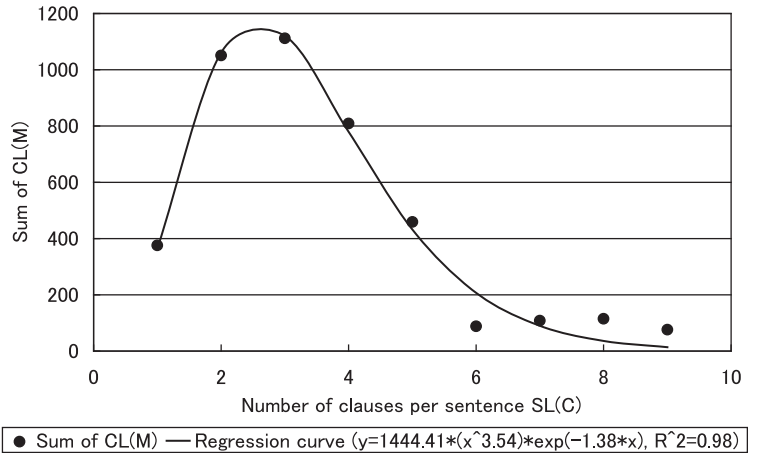


Fig. 10: A relationship between SL(C) as *x* and sum of CL(M) as *y* in (B) the newspaper for children with a MAL regression curve $y = 1444.41 * (x^{3.54}) * exp^{(-1.38*x)}$. ($R^2 = 0.98$).

In order to confirm how individual frequencies of data affect a distribution of DP as a total of frequencies, sums of CL(M), and averages of CL(M), we prepared four tables of imaginary data with SL as *x* and CL as *y* (the items for *x* and *y* are provisional), which are Samples (1) through (4) (see Tabs. 5 through 8). The sums of CL in the tables are obtained from DP weighted by each CL. Sample (1) is a case in which DP are regularly arrayed. DP increases if CL increases, and

Tab. 5: Imaginary data Sample (1).

SL	1	2	3	4	5	6	7	8
CL=1	1	2	3	4	4	3	2	1
CL=2	2	3	4	5	5	4	3	2
CL=3	3	4	5	6	6	5	4	3
CL=4	4	5	6	7	7	6	5	4
. . .								
CL=8	8	9	10	11	11	10	9	8
DP	36	44	52	60	60	52	44	36
Sum	204	240	276	312	312	276	240	204
Average	5.67	5.45	5.31	5.20	5.20	5.31	5.45	5.67

Tab. 6: Imaginary data Sample (2).

SL	1	2	3	4	5	6	7	8
CL=1				10	10			
CL=2			10	10	10	10		
CL=3		10	10	10	10	10	10	
CL=4	10	10	10	10	10	10	10	10
. . .								
CL=8	10	10	10	10	10	10	10	10
DP	50	60	70	80	80	70	60	50
Sum	300	330	350	360	360	350	330	300
Average	6.00	5.50	5.00	4.50	4.50	5.00	5.50	6.00

Tab. 7: Imaginary data Sample (3).

SL	1	2	3	4	5	6	7	8
CL=1	10	10	10	10	10	10	10	10
. . .								
CL=5	10	10	10	10	10	10	10	10
CL=6		10	10	10	10	10	10	
CL=7			10	10	10	10		
CL=8				10	10			
DP	50	60	70	80	80	70	60	50
Sum	150	210	280	360	360	280	210	150
Average	3.00	3.50	4.00	4.50	4.50	4.00	3.50	3.00

Tab. 8: Imaginary data Sample (4).

SL	1	2	3	4	5	6	7	8
CL=1	10	10	10	10	10	10	10	10
. . .								
CL=5	10	10	10	10	10	10	10	10
CL=6	10	10	10	10	10	10	10	
CL=7	10	10	10	10	10	10		
CL=8	10	10	10	10	10			
DP	80	80	80	80	80	70	60	50
Sum	360	360	360	360	360	280	210	150
Average	4.50	4.50	4.50	4.50	4.50	4.00	3.50	3.00

DP increases and decreases again if SL increases. Samples (2), (3), and (4) have the same DP. Sample (2) lacks some data in the smaller CL, Sample (3) lacks some data in the greater CL, and both are symmetrically arrayed. Sample (4) is asymmetrical data. Cells without data are coloured in grey.

DP, sums of CL, and averages of CL for Sample (1) to Sample (4) are drawn in Figs. 11–13. Regression curves of a quadratic polynomial function are also shown in Figs. 11–13 in order to visually confirm the data tendencies. Regression curves of the MAL do not fit these imaginary data.

Compared to distributions of DP in Fig. 11, Fig. 12 shows that distributions of sums of CL have similar tendencies to ones of DP. In Fig. 13, averages of CL for Sample (2) draw a deeper curve than ones for Sample (1), because the average depends on the ratio of DP to the sum of CL. This suggests that in two texts, the same curve of averages can be obtained if the texts have the same ratio of DP to the sum of CL to each other, and that text length does not directly affect the averages of CL. If we compare Samples (2) and (3), it can be observed that the sums of CL and the averages of CL are different even if the data have the same DP because cells lacking data in Tabs. 6 and 7 are different. This caused a difference in the weighted sums of CL in Samples (2) and (3). DP for greater CL affects more than DP for smaller CL regardless of the value of DP. Asymmetry of data in Sample (4) provides a kink in the curve of the average CL, which is shown in Fig. 13.

From these sample data, the following are confirmed: (1) a distribution of DP is systematically related to a distribution of the sum of CL; (2) a distribution of DP seems to be related to averages of CL if DP is a function of SL; and (3) averages of CL do not directly depend on text length, because the average is determined by a ratio of DP and a sum of CL.

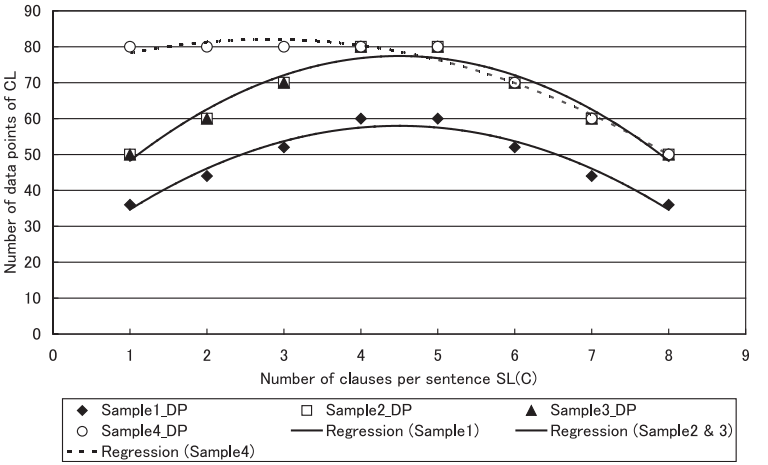


Fig. 11: A relationship between SL as *x* and the number of data points of CL as *y* in Samples (1) to (4).

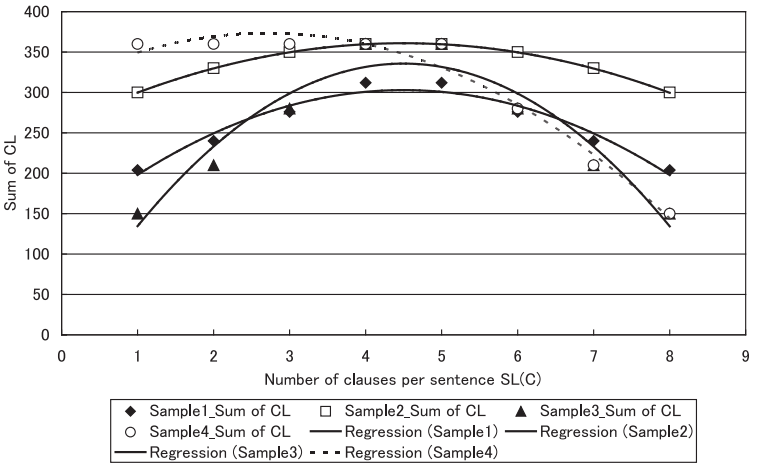


Fig. 12: A relationship between SL as *x* and the sum of CL as *y* in Samples (1) to (4).

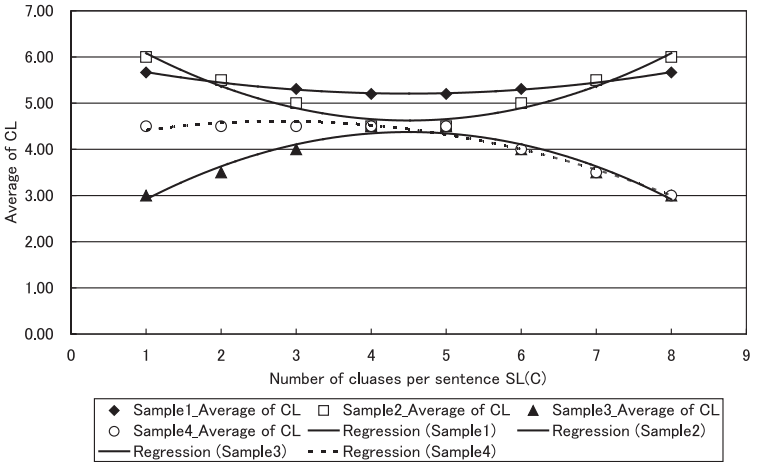


Fig. 13: A relationship between SL as *x* and the averages of CL as *y* in Samples (1) to (4).

6 Conclusions

The present study empirically investigated the MAL for two data sets of newspaper text which addressed the same topics but have different readerships (i.e. adults and children). Articles of the two data sets have a different style for readers, and text, sentence, and clause lengths for adults are much longer. However, relationship tendencies for (1)TL(C) and SL(C), (2) SL(C) and CL(M), (3) CL (M) and ML(Char), and CL(M) and ML(Rd) are similar to each other in the two data sets. The relationship tendencies are individualised according to linguistic levels, as posited by Köhler (1984) and Cramer (2005b).

From an observation of our data and simulations with imaginary data, the following assumptions can be made: (1) a distribution of DP is systematically related to a distribution of the sum of CL; (2) a distribution of DP seems to be related to averages of CL if DP is a function of SL; and (3) averages of CL do not directly depend on the text length because the average is determined by a ratio of DP and a sum of CL. These are still explorative results, and they must be systematically proven in future studies. The number of DP as a function of linguistic properties has not been considered in former studies on the MAL. Therefore, the present study may inform future studies on this law.

References

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

Cramer, Irene M. 2005a. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski, (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch*, 659–688. Berlin/ N.Y.: de Gruyter.

Cramer, Irene M. 2005b. The Parameters of the Altmann-Menzerath Law. *Journal of Quantitative Linguistics 12*(1). 41–52.

Köhler, Reinhard. 1984. Zur Interpretation des Menzerathschen Gesetzes. In Joachim Boy & Reinhard Köhler (eds.), *Glottometrika 6*. 177–183. Bochum: Brockmeyer.

Mainichi Shinbun [Daily newspaper]. Tokyo: Mainichi Shinbusha. (https://mainichi.jp/)

Mainichi Shogakusei Shinbun [Daily newspaper for elementary school children]. Tokyo: Mainichi Shinbusha. (https://mainichi.jp/maisho/)

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

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

Sanada, Haruko. 2016. Menzerath-Altmann Law and the sentence structure. *Journal of Quantitative Linguistics*, *23* (3). 256–277.

Wimmer, Gejza & Gabriel Altmann. 1991. Thesaurus of univariate discrete probability distributions. Essen: Stamm.

Yuasa, Chieko. 2006. *Kodomomuke Bunsho no Joho no Hairetsu: Shogakusei Shinbun wo Taisho ni* [Order of information in the text for children: in the case of newspaper for children]. *Buntairon Kenkyu* [Stylistics], *52*. 41–56.

Yuasa, Chieko. 2015. *Rinji Ichigo no Bunkai wo Tomonau Iikae: Shogakusei Shinbun wo Taisho ni* [A study on paraphrase of temporal combined words targeting schoolchildren]. Saitama *Daigaku Kiyo Kyoyo Gakubu* [Bulletin of Saitama University, Faculty of Liberal Arts], *51*(1). 173–183.

Yuasa, Chieko. 2016. *Shogakusei Shinbun no Juyo Nenreiso ni Yoru Doshi no Kakiwake: Ruijisei to Nan'ido no Renkan* [A choice of verbs adapted for children in the newspaper for children: relationships between the synonymity and the difficulty]. *Goi Kenshu* [Study of vocaburary], *13*. 58–66.

Software and digital dictionaries

Altmann, Gabriel. *Fitter, version 3.1.3.0.* 2013. Lüdenschied: RAM-Verlag.

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. (https://clrd.ninjal.ac.jp/unidic/)

Gen Tsuchiyama

Quantitative analysis of the authorship problem of "The Tale of Genji"

Abstract: This study involves quantitative research on the authorship problem of "The Tale of Genji," which is the most famous Japanese classical literary work, using statistical analysis methods. "The Tale of Genji" was written by Murasaki Shikibu (around 930–1014), who is one of the most famous female novelists from the Heian period (794–1185). It is a full-length tale set at the Heian Imperial Court that depicts the love story of Hikaru Genji, the main character. The tale has been widely read across generations. There is a theory that the final 13 volumes were penned by different authors because of differences in the stage and the main character. Japanese literary research has been unable to resolve this authorship problem, and it is still unclear whether there had been one or more authors for this literary work. This study applies quantitative methods to solve the authorship problem of this tale.

In this study, a word n-gram and word length n-gram were used for analysis in order to resolve the authorship problem, but the result showed no difference in measurement characteristics between the final 13 and the other volumes of "The Tale of Genji." The authorship problem of the literary work has been discussed for a long time, but the quantitative analysis performed by this study shows that there is no evidence supporting the theory that there might have been multiple authors of "The Tale of Genji." It can, therefore, be concluded that "The Tale of Genji" was likely written by a single author.

Keywords: stylometry, authorship attribution, classical Japanese literature

1 Introduction

This study aims to investigate the authorship problem of "The Tale of Genji" using statistical analysis methods. "The Tale of Genji" is a famous work in classical Japanese literature that was written by Murasaki Shikibu (around 930–1014), who was a female novelist from the Heian period. It is a full-length tale that has been widely read across generations. The main setting of this tale is at the Heian

Gen Tsuchiyama, Center for Interdisciplinary AI and Data Science, Ochanomizu University, Japan, e-mail: tsuchiyama.gen@ocha.ac.jp

https://doi.org/10.1515/9783110763560-014

Imperial Court, and it depicts the love story of Hikaru Genji, the main character. "The Tale of Genji" is comprised of 54 volumes; Volume 43 and beyond are about the story of the main character after his death. In particular, the 10 volumes that follow the first 45, in which the lead character becomes Hikaru Genji's child, Kaoru, are called Uji Jujo, leading to the thought that the author of these volumes was different.

"Kacho Yojo," which was written by Kanera Ichijou (1402–1481), states that all volumes with the exception of Uji Jujo were the works of Murasaki Shikibu, while Uji Jujo was written by Dainino Sanmi. In recent times, on the other hand, others have rejected the theory that Uji Jujo was written by a different author. Ikeda (1951) and Ohno (1984) noted that while there are differences in other volumes such as those in prose style, these are not sufficient to prove that the tale was written by a different hand. The author of Uji Jujo is still presumed to be Murasaki Shikibu. In this regard, there is still no resolution on the actual authorship of Uji Jujo. In addition to this, Niomiya Sanjo, which is the three volumes (Vol. 42, 43, 44) that preceded, is inconsistent with the story of Uji Jujo, so there is another authorship problem aside from that of Uji Jujo.

In Japan, with regard to modern literature, themes related to the identification of authors using quantitative methods have been widely researched. On the other hand, quantitative research has not yet been fully applied to classical literature such as "The Tale of Genji" because, compared to modern literature, specialized knowledge is required to understand the sentence structure of classical works, and special care is required when dealing with revised texts. Therefore, in this study, the quantitative method has been applied to the authorship problem of "The Tale of Genji." There are many manuscript systems for "The Tale of Genji," and in this study we used the text data based on manuscripts called "Oshima Bon," which is considered to be the most reliable, in a system called "Ao Byoushi Bon."

2 Background

The quantitative analysis of "The Tale of Genji" using statistical methods has already been undertaken by others. Yasumoto (1957) analyzed "The Tale of Genji," which is thought to have been written by Murasaki Shikibu (973–1014) during the Heian period. Yasumoto divided "The Tale of Genji" into Uji Jujo as well as the other 44 volumes to perform a statistical hypothesis test. Twelve items including frequencies of nouns, declinable words, postpositional particles, and auxiliary verbs were used for the test. One hundred words were

extracted randomly from each volume, and the frequency of each item was obtained from these words, so it should be noted that not all sentences in "The Tale of Genji" were analyzed.

Yasumoto (1957) also performed a study based on the psychology of writing, and from the results he considered that the styles found in Uji Jujo were characterized by detailed descriptions that used fictional, declinable, close, and continuous concepts and that the styles of the other 44 volumes were characterized by intuitive descriptions that used melodramatic, indeclinable, dramatic, and intermittent configurations. Therefore, he concluded that the author of Uji Jujo was not the author of the other volumes.

In addition, Yasumoto (1977) used 12 variables that were obtained from a sampling similar to that used in Yasumoto (1957) to perform factor analysis and reexamine Uji Jujo from the perspective of the tale having multiple authors. As a result, he confirmed his previous conclusion that the styles found in Uji Jujo were different from those of the other volumes; however, based on this analysis, he did not conclude that the author was different.

Arai (1997) extracted samples from the central part in each volume of Uji Jujo, whose locations depended on the length of the volume, to perform a statistical test on the frequencies of syllable onset strings and vowel strings in the kana syllabary table. As a result, he concluded that the author of Uji Jujo was not different from the author of the other volumes.

All of these quantitative studies on "The Tale of Genji" have provided meaningful results; however, these analyses were not performed on the entire work. Murakami and Imanishi (1999) used multivariate analysis techniques to perform a quantitative study on all the sentences in "The Tale of Genji." In this study, the appearance ratios of specific function words such as auxiliary verbs were used to predict the order of creation of the volumes that comprise "The Tale of Genji" by quantification of the third part. Tsuchiyama and Murakami (2014) performed multivariate analysis such as principal component analysis (PCA) of "The Tale of Genji" and used the appearance ratios of auxiliary verbs and postpositional particles and word length distribution to clarify the quantitative structure of "The Tale of Genji." As a result, they concluded that Uji Jujo had a stylistic difference between the first five and the second five volumes.

3 Method of analysis and results

3.1 Target

For this analysis, we primarily used PCA based on correlation coefficient matrices. PCA is a method of dimension contraction that can be applied to multidimensional data. Information is contracted by seeking new composite variables from original data variables. Analysis results are expressed as a scatter plot of the principal component scores for the first and second principal components. In a scatter plot, first principal components are represented on the horizontal axis, while second principal components are represented on the vertical axis. In addition, the volume of information contained in each principal component is evaluated according to its contribution rate. The first and second principal component contribution rates are also shown on a scatter plot.

In this study, PCA is performed using the relative frequency of word n-grams and the relative frequency of word-length n-grams. Additionally, in the word frequency analysis, we investigated two types of functional words: postpositional particles and auxiliary verbs, which have been determined to be effective for quantitatively identifying authors. Frequencies are the frequency of which each item appears in the target text.

3.2 Analysis

In this study, the volumes with fewer words than the others were excluded from the analysis. Specifically, three volumes with less than 1,000 words were excluded. Table 1 shows the total number of words in each volume of "The Tale of Genji." As shown in Tab. 1, Volumes 11, 16, and 27 are excluded.

We approach the authorship problem that was discussed about Uji Jujo as well as that discussed about Niomiya Sanjo. In the analysis, 41 volumes that do not include Uji Jujo and Niomiya Sanjo are included in one group. First, the analysis is performed using this group and Uji Jujo, and then the analysis is performed using this group and Niomiya Sanjo. By analyzing this way, both authorship problems were examined.

3.2.1 Analysis of postpositional particle n-grams

First, PCA was performed on the 41 volumes excluding Uji Jujo. For the analysis, we used the 14 most frequently occurring words. These postpositional

Tab. 1: Total number of words in each volume of "The Tale of Genji".

Title	Total number of words	Title	Total number of words	Title	Total number of words
01 Kiritsubo	4804	19 Usukumo	6023	37 Yokobue	3694
02 Hahakigi	9383	20 Asagao	3993	38 Suzumushi	2748
03 Utsusemi	2187	21 Otome	10040	39 Yugiri	14021
04 Yugao	9566	22 Tamakazura	8186	40 Minori	3720
05 Wakamurasaki	9406	23 Hatsune	2689	41 Maboroshi	4286
06 Suetsumuhana	6139	24 Kocho	4041	42 Niomiya	2696
07 Momiji no Ga	5559	25 Hotaru	3767	43 Koubai	2517
08 Hana no En	2009	26 Tokonatsu	4351	44 Takekawa	8065
09 Aoi	9167	27 Kagaribi	653	45 Hashihime	7299
10 Sakaki	9664	28 Nowaki	3510	46 Shiigamoto	7290
11 Hanachirusato	724	29 Miyuki	5235	47 Agemaki	17437
12 Suma	8391	30 Fujibakama	2795	48 Sawarabi	3557
13 Akashi	7865	31 Makibashira	7234	49 Yadorigi	18839
14 Miotsukushi	6296	32 Umegae	3638	50 Azumaya	12999
15 Yomogiu	4605	33 Fuji no Uraha	4430	51 Ukifune	14415
16 Sekiya	934	34 Wakana Jo	20199	52 Kagero	11798
17 Eawase	3656	35 Wakana Ge	20223	53 Tenarai	14221
18 Matsukaze	4030	36 Kashiwagi	7926	54 Yume no Ukihashi	3556

particles in terms of frequency constitute 53.4% of the total frequency, and they are the minimum number of variables above 90%. The results of the top 14 postpositional particles in terms of frequency are shown in Fig. 1. In Fig. 1, the 10 volumes of Uji Jujo are mixed with the other volumes.

Next, PCA was performed on the 41 volumes excluding Niomiya Sanjo. For the analysis, we used the 13 most frequently occurring words of postpositional particles. Thirteen words are the minimum number of variables above 90%. Figure 2 shows the analysis results, and no remarkable difference is observed between Niomiya Sanjo and the other 41 volumes.

Finally, PCA was performed using postpositional particle bigrams as variables. The analysis on Uji Jujo used 232 variables with highest frequency of appearance, and the analysis for Niomiya Sanjo used 228 variables with highest frequency of appearance. These are the minimum number of variables above 90%. The results of the analysis are shown in Figs. 3 and 4 and do not suggest that Uji Jujo and Niomiya Sanjo are different from the other 41 volumes.

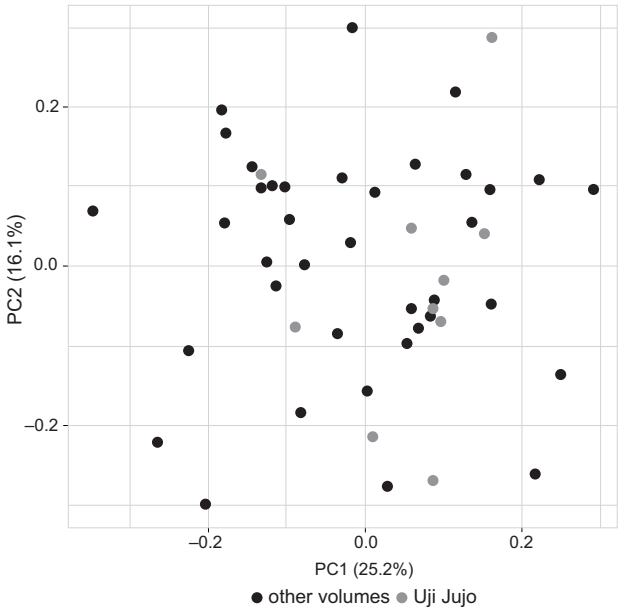


Fig. 1: PCA results of Uji Jujo using top 14 postpositional particles.

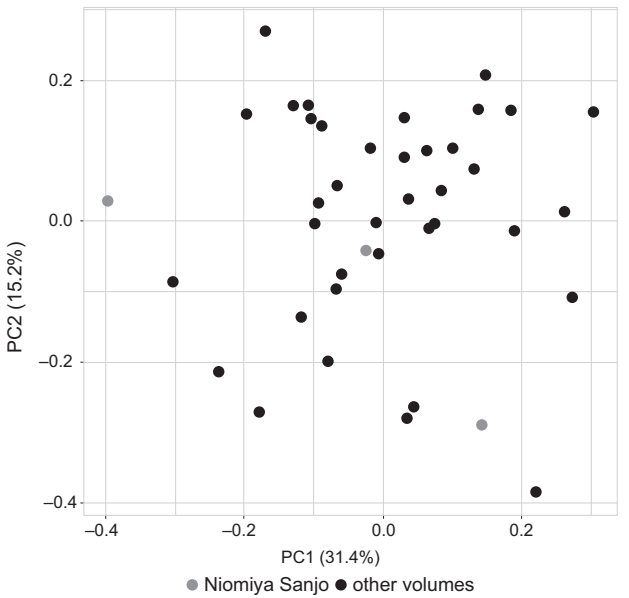


Fig. 2: PCA results of Niomiya Sanjo using top 13 postpositional particles.

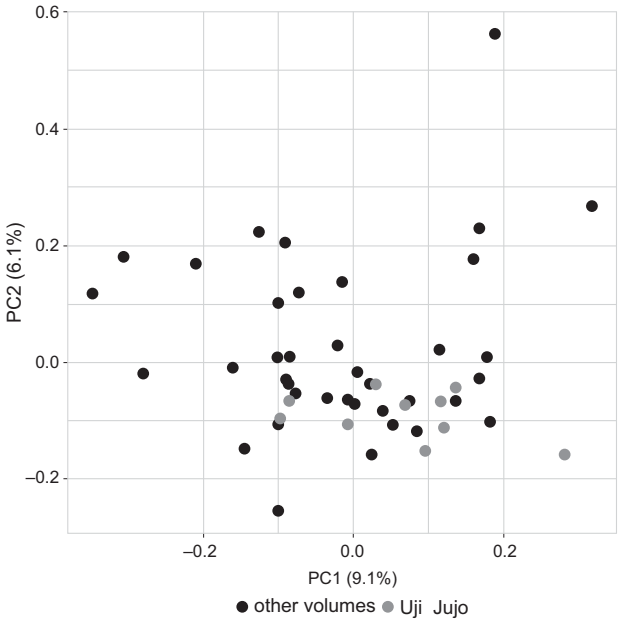


Fig. 3: PCA results of Uji Jujo using top 232 postpositional particle bigrams.

3.2.2 Analysis of auxiliary verb n-grams

PCA was performed using 13 auxiliary verbs with highest frequency of appearance in order to examine the authorship problem of Uji Jujo. Figure 5 shows the analysis results, and no remarkable difference is recognized between Uji Jujo and the other 41 volumes.

Then, PCA was performed on Niomiya Sanjo and the other 41 volumes. We used the top 13 auxiliary verbs. Thirteen words are the minimum number of variables above 90%. Figure 6 shows the analysis results, and no remarkable difference is observed between Niomiya Sanjo and the other 41 volumes.

Next, PCA was performed on the auxiliary verb bigram. Figure 7 shows the analysis results of Uji Jujo, and Fig. 8 shows the analysis results of Niomiya Sanjo. In the former, the top 216 variables were used; and in the latter, the top 212 variables were used. The results of the analysis do not suggest that Uji Jujo and Niomiya Sanjo are different from the other 41 volumes.

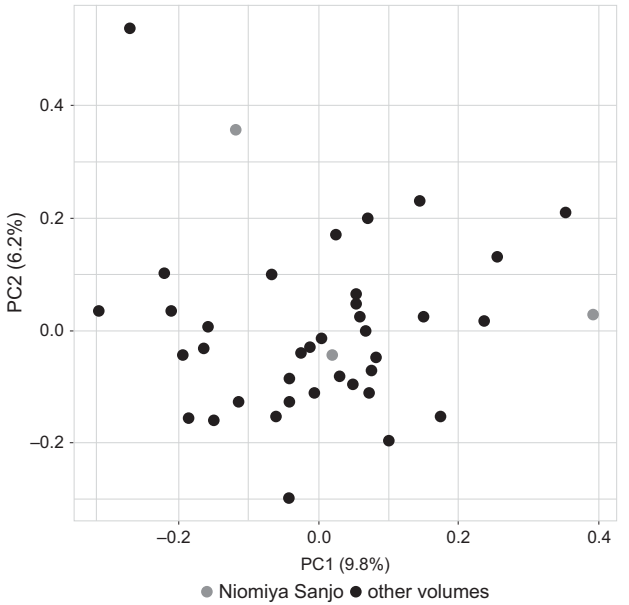


Fig. 4: PCA results of Niomiya Sanjo using top 228 postpositional particle bigrams.

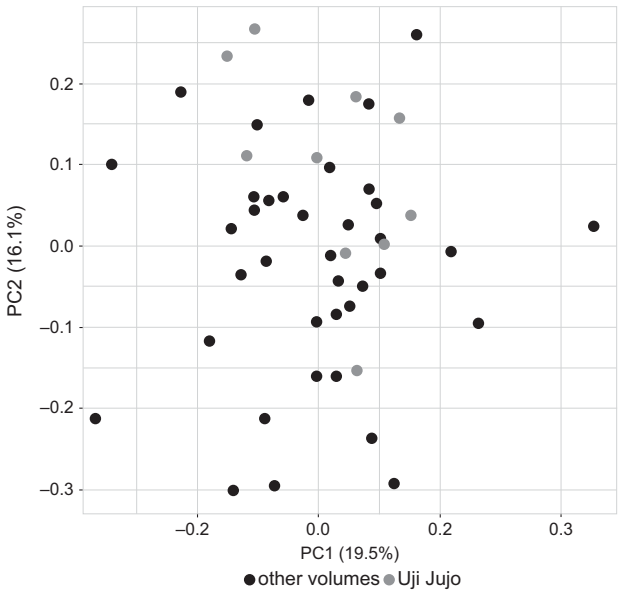


Fig. 5: PCA results of Uji Jujo using top 13 auxiliary verbs.

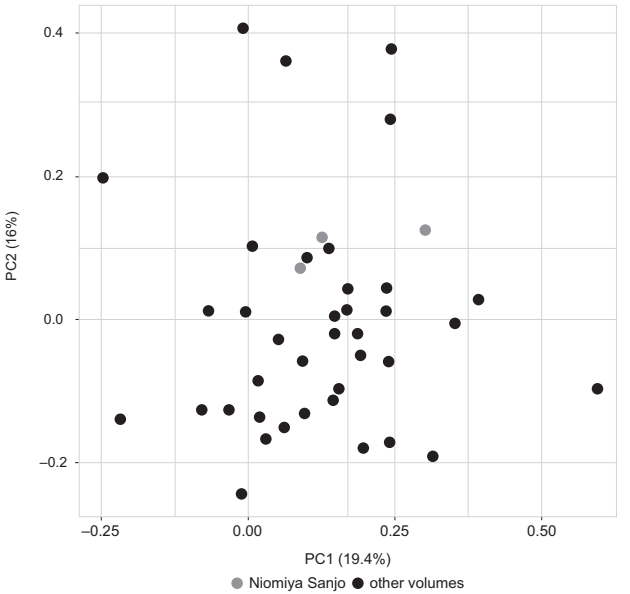


Fig. 6: PCA results of Niomiya Sanjo using top 13auxiliary verbs.

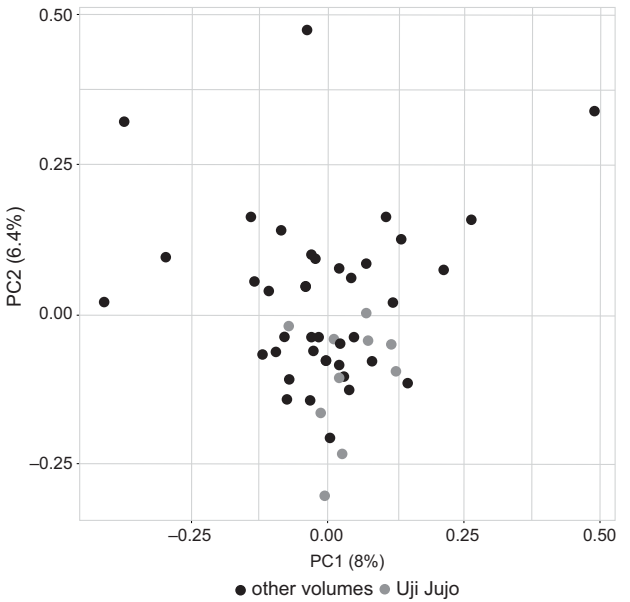


Fig. 7: PCA results of Uji Jujo using top 216 auxiliary verb bigrams.

3.2.3 Analysis of word length n-grams

Word length is the number of characters in a word. Therefore, analysis is performed using the number of characters of a word as variable. First, PCA was performed using the five variables with highest frequency to examine the authorship problem of Uji Jujo. Five variables are the minimum number of variables above 90%. Figure 9 shows the result of the analysis, in which Uji Jujo and the other 41 volumes are not separated into two groups.

Next, we investigate the authorship problem of Niomiya Sanjo using word length. The analysis was performed using the five variables with highest appearance frequency. The results of the analysis are shown in Fig. 10, and there is no difference between Niomiya Sanjo and the other 41 volumes.

Finally, PCA was performed on the word length bigrams. Figure 11 shows the analysis results of Uji Jujo, and Fig. 12 shows the analysis results of Niomiya Sanjo. In the former, the top 22 variables were used; and in the latter, the top 23 variables were used. The results show no separation between the two groups in either analysis.

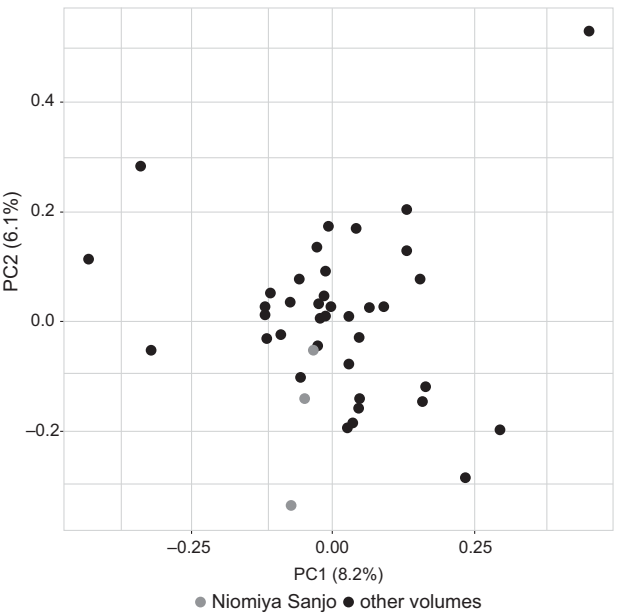


Fig. 8: PCA results of Niomiya Sanjo using top 212 auxiliary verb bigrams.

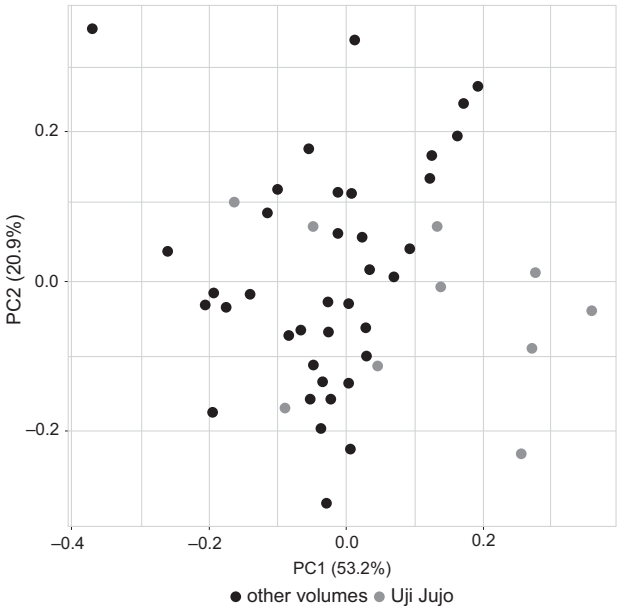


Fig. 9: PCA results of Uji Jujo using top five variables of word length.

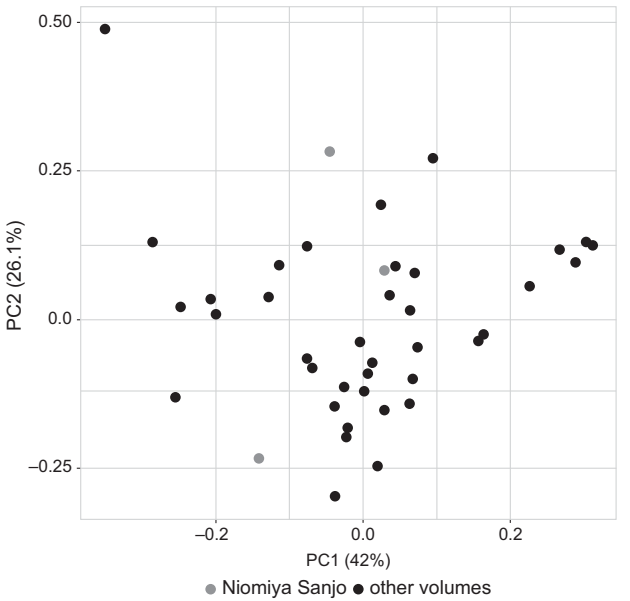


Fig. 10: PCA results of Niomiya Sanjo using top five variables of word length.

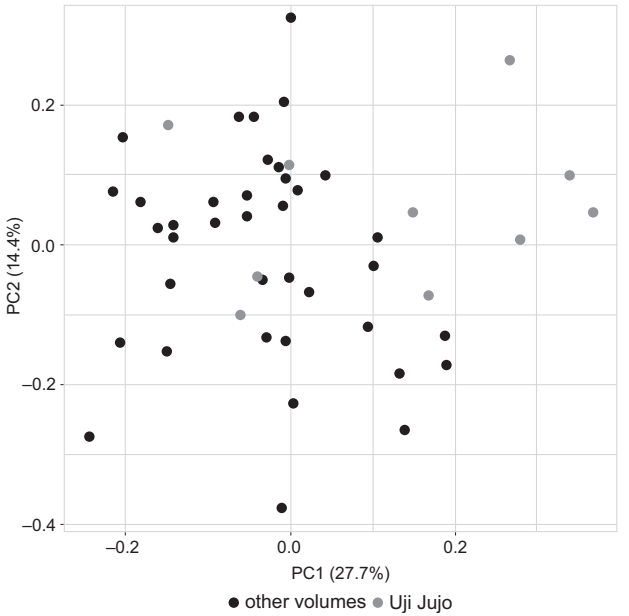


Fig. 11: PCA results of Uji Jujo using top five variables of word length bigrams.

4 Conclusion

In this study, a word n-gram and word length n-gram were used for analysis in order to resolve the authorship problem of "The Tale of Genji." The results show there was no difference in the measurement characteristics between the final 13 volumes and the other volumes of "The Tale of Genji." The authorship problem of this literary work has been discussed for a long time, but the quantitative analysis performed by this study shows that there is no evidence supporting the theory that there were multiple authors of "The Tale of Genji." It can, therefore, be concluded that "The Tale of Genji" was likely written by a single author.

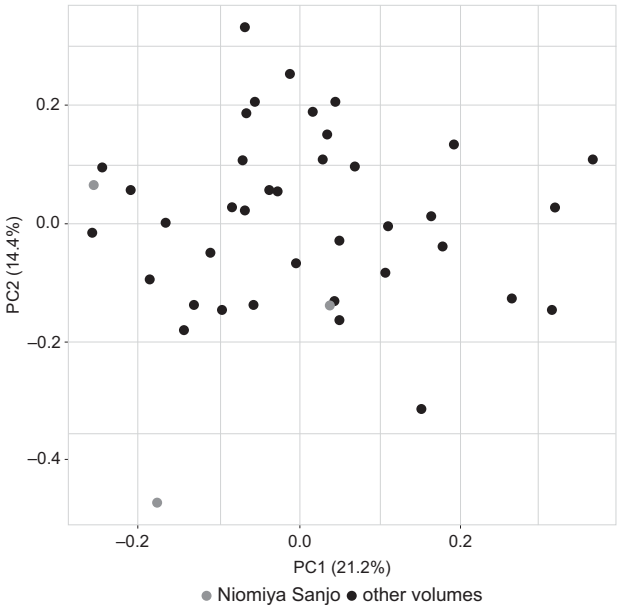


Fig. 12: PCA results of Niomiya Sanjo using top five variables of word length.

References

Ikeda, Kikan. 1951. *Genjimonogatari no kousei [Structure of the Tale of Genji]*. Shibundo (Published: “New Structure Tale of Genji (above)”).

Ohno, Ssumu. 1984. *Genjimonogatari* [The Tale of Genji]. Tokyo: Iwanami Shoten, Publishers.

Yasumoto, Biten. 1957. *Uji Jujo no sakusha: bunsho sinrigaku ni yoru sakusya suitei* [The author of Uji Jujo: Inferring the author through stylometry]. *Japanese Psychological Review* 2(1). 147–156.

Yasumoto, Biten. 1977. *Gendai no buntai kenkyu* [Stylometric research in recent years]. *Nihongo* [the Japanese language], 10. 395–423. Tokyo: Iwanami Shoten, Publishers.

Arai, Hiroshi. 1997. Genji monogatari/Uji Jujo no sakusha mondai: hitotsu no keiryou gengogakuteki approach [The Tale of Genji/Uji Jujo authorship problem: one quantitative linguistic approach]. *Hitotsubashi Ronshu [Hitotsubashi Treatise]*, 117(3). 397–413.

Murakami, Masayoshi & Yuichiro Imanishi. 1999. Quatitative analysis of auxiliary verb of The Tale of Genji. *Journal of Information Processing* 40(3). 774–782.

Tsuchiyama, Gen & Masayoshi Murakami. 2014. A quantitative analysis on the formation of the third part of “The Tale of Genji”, *Computer and Humanities Symposium Essays*, 2014 (3). 213–220.

Yawen Wang, Haitao Liu

Revisiting Zipf's law: A new indicator of lexical diversity

Abstract: In a given text, the occurrence of words follows a famously systematic frequency distribution, obeying a power law known as Zipf's law. Its most common form is the doubly logarithmic chart. This research demonstrates that the ignored parameter *C* in the original Zipf's law displays a unique pattern. Moreover, the parameter *C* exhibits some correlation with lexical diversity, due to the equilibrium between uniformity and diversity known as the principle of least effort. Parallel corpora are designed and built with 23 language translations of *Le Petit Prince*, 19 of *Alice's Adventures in Wonderland*, and their original versions. With Moving-Average Type–Token Ratio (MATTR) as a reliable indicator of lexical diversity, *C* indicates a great variety of lexical richness among different languages. The findings include: (a) *C* displays a close correlation with the lexical diversity of languages; (b) a higher *C* tightly correlates to a lower diversity to some degree; (c) *C* from nearly all words hardly exhibits such trend.

Keywords: Zipf's law, lexical diversity, parallel corpora, MATTR

1 Introduction

Zipf's law is widely adopted to illustrate the relationship between the frequency of a word and its rank when words are ranked with respect to the occurrence frequency (Zipf 1949). Specifically, the law was initially described based on data from James Joyce's novel *Ulysses* as follows, where $f(r_i)$ refers to a word's frequency of occurrence and r_i to the ranking of the word in *Ulysses*, and *C*, a constant.

$$f(r_i)^* r_i = C \tag{1}$$

Since Zipf (1935) analyzed the $r^2 \times f = C$ relationship, the parameter α has been introduced and analyzed for over 70 years according to equation (2) (Zörnig and Altmann 1995; Marco, Anke, and Merja 2009; Popescu 2009; Piantadosi

Yawen Wang, Department of Linguistics, Zhejiang University, China,
e-mail: mileywyw@zju.edu.cn
Haitao Liu, Department of Linguistics, Zhejiang University, China, e-mail: htliu@163.com

https://doi.org/10.1515/9783110763560-015

2014; Moreno-Sánchez, Font-Clos, and Corral 2016), while the parameter C has been almost neglected.

$$f(r_i) = \frac{C}{r_i^{\alpha}} \tag{2}$$

In the early analysis of Zipf's law, the parameter C still remains in the power law formula, though no one notices it (Simon 1955; Balasubrahmanyan and Naranan 2002; Naldi 2003). C is functionally important. It is either related to a y-intercept in linear regression or subtracted from the total number of the parameters by the normalization constraint (Li, Miramontes, and Cocho 2010).

In recent years, research on Zipf's law has shifted from the power law form to the probability density function as expressed in equation (3), coupled with the neglect of C (Ferrer-i-Cancho and Solé 2001; Ferrer-i-Cancho and Elvevåg 2010; Baixeries, Elvevåg, and Ferrer-i-Cancho 2013; Corral, Boleda, and Ferrer-i-Cancho 2015; Ferrer-i-Cancho 2016; Moreno-Sánchez, Font-Clos, and Corral 2016; Ferrer-i-Cancho, Bentz, and Seguin 2022).

$$f(r_i) \approx r_i^{-\alpha} \tag{3}$$

There is a need to understand the constant parameter C. In consideration of the insufficient study of the parameter C, this study tries to investigate C based on Zipf's law in equation (1). Specifically, the following issues will be addressed.

(1) Will the parameter C exhibit a unique pattern?
(2) Will the parameter C also indicate lexical richness, because Zipf's law is widely recognized as a measure of lexical diversity (Bentz et al. 2015)?

Parallel corpora are adopted to facilitate the exploration. The texts selected here are the original and translated texts of *Le Petit Prince* and *Alice's Adventures in Wonderland* in over 20 languages so as to ensure the universality of our results.

To this end, the paper is organized as follows: Section 2 provides materials and methods applied in the study. Section 3 describes the analysis of the results and discussion. We summarize our findings based on their implications and look forward to the future in Section 4.

2 Materials and methods

2.1 Materials

In order to explore the parameter C in Zipf's law, parallel corpora are first designed and constructed. The parallel corpora comprise the original and translated texts of *Le Petit Prince* and *Alice's Adventures in Wonderland*. Specifically, we collected 23 language translations of *Le Petit Prince* with its French original text and 19 language translations of *Alice's Adventures in Wonderland* with its original English version. These corpora cover a range of language families, including Turkic languages, Finno-Ugric languages, Japanese, Korean, and the constructed language Esperanto.

2.2 Methods

The methods for measuring lexical diversity have varied across the research area. The Moving-Average Type–Token Ratio (MATTR) is chosen for its reliability and validity (Covington and McFall 2010; Kubát 2014), which is calculated according to the following formula, where T refers to the text length, W represents the window size ($W < N$), and V_i is the number of the types in a window.

$$\mathrm{MATTR} = \frac{\sum_{i=1}^{T-W+1} V_i}{W(T-W+1)} \tag{4}$$

For texts over 10,000 words, the suggested window size is 500 words. SPSS 21 (Spss 2012) and R (R Core Team 2018) are employed for the sequentially statistical analysis. All the texts in the corpora are analyzed using QUITA (Quantitative Indicator Text Analyzer), which is a powerful program to analyze texts with many indicators by measuring their different characteristics (Kubát, Matlach, and Čech 2014). It directly provides the basic data and frequency table of the texts. After careful manual check and correction, the output of QUITA is converted to EXCEL formats for further analysis.

2.3 Calculation of the parameter *C*

The C values are obtained from the 24 language texts of *Le Petit Prince* according to Zipf's law in equation (1). In the past, Zipf's law has been examined with the double-logarithmic plot via linear regression. Nevertheless, some scholars have recently challenged this way by pointing out its limitations (Goldstein,

Morris, and Yen 2004; Bauke 2007; Clauset, Shalizi, and Newman 2009). In consideration of its limitation, this study did not use the double-logarithmic plot.

Figure 1 shows the *C* values of *Le Petit Prince and Alice's Adventures in Wonderland*, with the increase of rank. Obviously, the *C* values differ greatly. It is worth noting that *C* varies from below 300 to over 3000 in the beginning of the curve, and then fluctuates around one line from all the languages with the word rank increasing up to 500. Subsequently, the curves gradually come together with the frequency of word declining to 2 and 1. As the amount of hapax increases, all languages converge into one line.

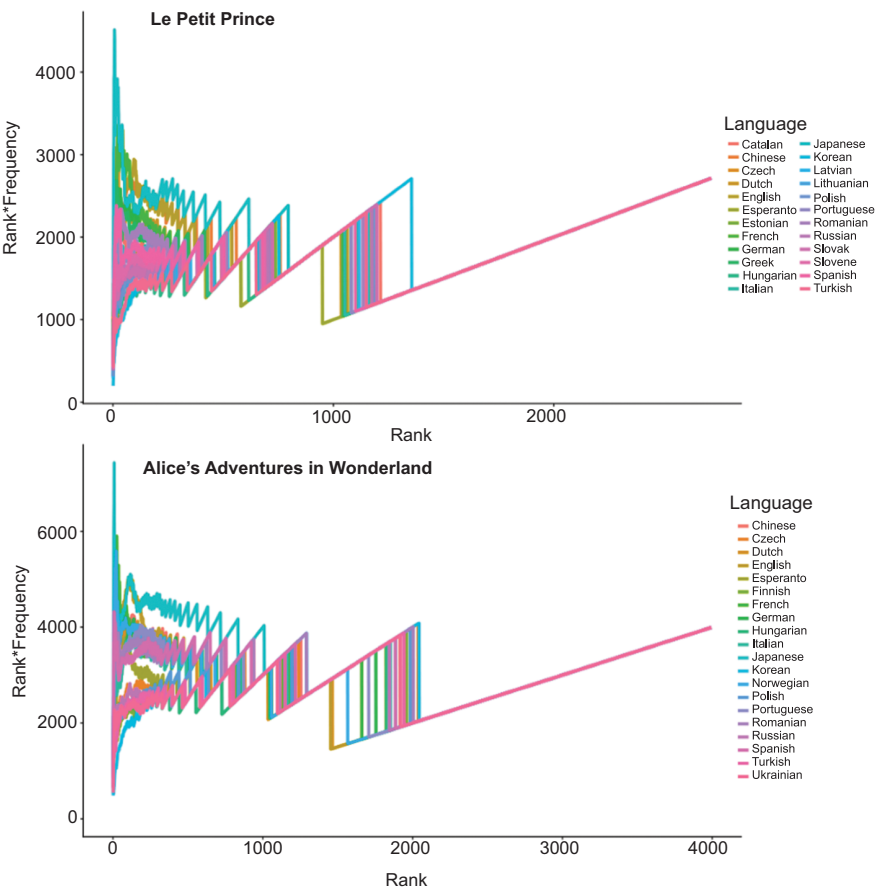


Fig. 1: The C values of *Le Petit Prince* and *Alice's Adventures in Wonderland*.

The C values exhibit obvious changes between the languages with gradual increase in the frequency rank. As the parameter C is far from a constant, C is obtained as the mean value by multiplying r_i with $f(r_i)$ in the most frequent words to compare the studied languages. In this research, the C values are calculated in the most frequent 50, 60, 70, . . ., and 2500 words.

3 Results

Zipf's law explicitly describes a universal relationship of one word's frequency with its rank across variation in languages, authors and genres, owing to the subtle balance between the Force of Unification and the Force of Diversification. This pilot study tries to depict the unique role of the parameter C with the abovementioned materials and methods. The research may shed some valuable guidelines on the similarities and differences in how languages organize and encode information.

3.1 Revisiting Zipf's law

As shown in Fig. 2a, the C values change from the most frequent 50 words to 2500 the words in a text of *Le Petit Prince*. As seen, C from the most frequent 200 words or so can largely differentiate these languages. Notably, the C values from the most frequent 1200 words ~ the most frequent 2500 words are concentrated around 2000 ~ 2500, due to the increase in the number of hapaxes involved.

Figure 2b captures the consistent pattern of C from the most frequent 50 words to 2500 words in a text of *Alice's Adventures in Wonderland*. Impressively, the corresponding turning point is postponed to the most frequent 500 words or so, which shows huge difference in the involved languages. Then, C from the most frequent 2500 languages appear in the range of 2500 ~ 3500 with more hapaxes. The disparity between *Le Petit Prince* and *Alice's Adventures in Wonderland* in the fluctuation range is mainly attributed to their diverse text lengths.

3.2 The relationship of the parameter C with the lexical diversity

In order to investigate the relationship of the parameter C with lexical diversity, their correlation is carefully examined. Figures 3 and 4 show the correlation between C and MATTR in *Le Petit Prince* and *Alice's Adventures in Wonderland*.

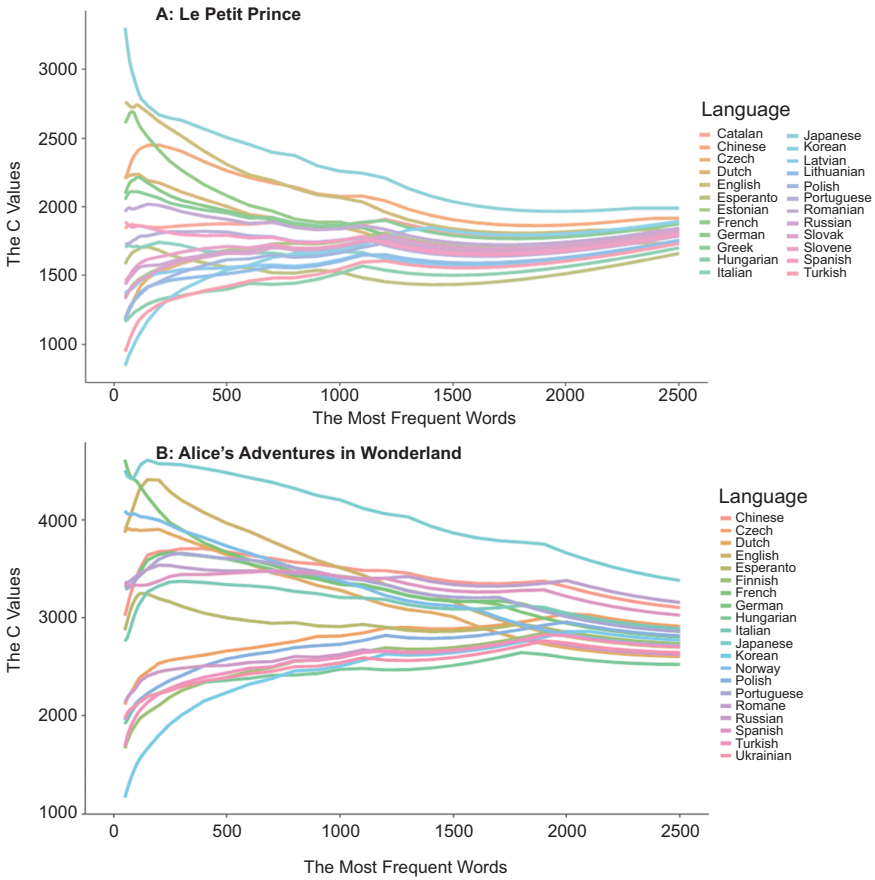


Fig. 2: The C values of *Le Petit Prince* and *Alice's Adventures in Wonderland*.

As clearly illustrated above, there is a generally negative correlation between C and MATTR (see details in Figs. 3 and 4). It shows that for the investigated languages, a larger C leads to a smaller MATTR. So, the C value also reveals the lexical diversity as expected. However, it has a positive correlation with MATTR when it comes to nearly all words in the texts.

Its correlation coefficient achieves the maximum at the most frequent 70 words ($r = -0.939$, $p_{2\text{-tailed}} < 0.01$). The following increase in the words results in a weaker correlation with MATTR. More words yield a weaker correlation between them. For the words ranked between 1500 to 2500, the C values scarcely follow such a trend, ultimately showing little trace of lexical diversity. These scenarios are a result of the hapax amount.

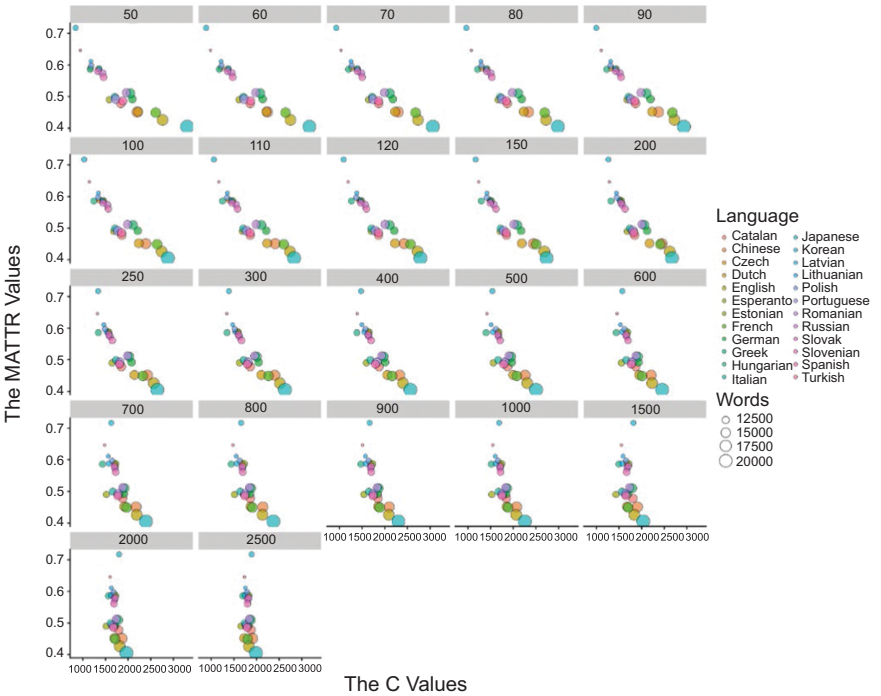


Fig. 3: The correlation between *C* and MATTR of *Le Petit Prince*.

It is therefore likely that connections exist between *C* and lexical diversity, primarily thanks to the principle of least effort and the human cognition mechanism (Zipf 1949). With the increase of ranking and hapax amount, the words reach equilibrium between the uniformity and diversity in the word usage.

Similarly, *Alice's Adventures in Wonderland* is explored to further examine whether the above principles are universal or not. Figure 4 illustrates the results of *Alice's Adventures in Wonderland*, which are consistent with those of *Le Petit Prince*. Specifically, a larger *C* for the languages brings less lexical diversity, and vice versa. Nevertheless, with the higher ranking, their correlation weakens. In addition, *C* from nearly all words adversely displays a positive correlation with MATTR.

The turning point in Fig. 4 appeared much later than that in Fig. 3. The correlation coefficient reaches the maximum at the most frequent 200 words ($r = -0.957$, $p_{2\text{-tailed}} < 0.01$). Its sluggish appearance is mainly ascribed to the different text lengths. As we know, the texts of *Le Petit Prince* are around 13,000 words, which are nearly half of *Alice's Adventures in Wonderland* (ca. 23,000 words). Moreover, the *C* values mostly appear in the range from 2600 to 3700. They

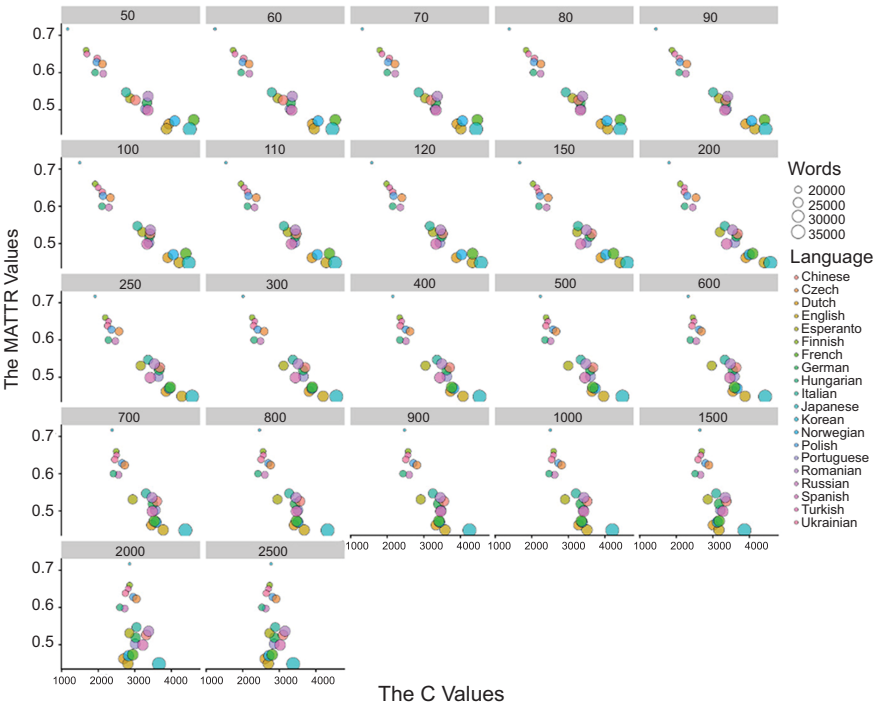


Fig. 4: The correlation between *C* and MATTR of *Alice's Adventures in Wonderland*.

exhibit an approximately double increase when compared to that of *Le Petit Prince*. Since the text length of *Alice's Adventures in Wonderland* is almost twice as long as that of *Le Petit Prince*, there exists the inextricable linkage of the text size with the parameter *C*.

Overall, this study has identified a close correlation between *C* and MATTR, whereas such a correlation is weakened when more hapaxes are involved. The fluctuation range of the *C* values results from the text size.

4 Conclusions

In summary, this research, as the first to pay great heed to the parameter *C*, brings a new perspective to the parameter *C* in the original Zipf's formula. With the corpora of *Le Petit Prince* and *Alice's Adventures in Wonderland* in more than 20 languages, *C* does exhibit a unique pattern. Moreover, *C* gives a clear indication of the lexical diversity of languages. Specifically, with the frequency

ranking of words lower than 1500 words, *C* has a close correlation with lexical diversity. Once the ranking is over 1500, their correlation weakens dramatically. Nevertheless, *C* based on the most frequent 2500 words even turns to a slightly positive relationship with MATTR. The text size also exerts influence on the *C* values. The improved performance of *C* in Zipf's law comes from the general cognitive mechanism, albeit with its close correlation with the text size. Multilingual analysis of languages will shed some valuable guidelines on the language universality and cognitive laws hidden behind human beings.

References

Baixeries, Jaume, Brita Elvevåg & Ramon Ferrer-i-Cancho. 2013. The evolution of the expo-nent of Zipf's law in language ontogeny. *PloS ONE* 8(3). e53227.

Balasubrahmanyan, Vriddhachalam K. & Sundaresan Naranan. 2002. Algorithmic information, complexity and Zipf's law. *Glottometrics* 4. 1–26.

Bauke, Heiko. 2007. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B* 58. 167–173. doi:10.1140/epjb/e2007-00219-y.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE* 10(6). e0128254.

Clauset, Aaron, Cosma Rohilla Shalizi & M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4). 661–703. doi:10.1137/070710111.

Corral, Álvaro, Gemma Boleda & Ramon Ferrer-i-Cancho. 2015. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE* 10 (7). doi:10.1371/journal.pone.0129031. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4497678/.

Covington, Michael A. & Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100.

Ferrer-i-Cancho, Ramon. 2016. Compression and the origins of Zipf's law for word frequencies. *Complexity* 21. 409–411. doi:10.1002/cplx.21820.

Ferrer-i-Cancho, Ramon, Christian Bentz & Caio Seguin. 2022. Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics* 29(2). 165–194.

Ferrer-i-Cancho, Ramon & Brita Elvevåg. 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* 5(3). e9411. doi:10.1371/journal.pone.0009411. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834740/.

Ferrer-i-Cancho, Ramon & Ricard V. Solé. 2001. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited*. *Journal of Quantitative Linguistics* 8(3). 165–173. doi:10.1076/jqul.8.3.165.4101.

Goldstein, Michel. L., Steven. A. Morris & Gary. G. Yen. 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B* 41. 255–258. doi:10.1140/epjb/e2004-00316-5.

Kubát, Miroslav. 2014. Moving window type-token ratio and text length. In Gabriel Altmann, Radek Čech, Ján Mačutek & Ludmila Uhlířová (eds.) *Empirical Approaches to Text and Language Analysis*. 105–113. Lüdenscheid: RAM-Verlag.

Kubát, Miroslav, Vladimír Matlach & Radek Čech. 2014. QUITA: *Quantitative Index Text Analyzer*. (Studies in Quantitative Linguistics 18). Lüdenscheid: RAM-Verlag.

Li, Wentian, Pedro Miramontes & Germinal Cocho. 2010. Fitting ranked linguistic data with two-parameter functions. *Entropy* 12(7). 1743–1764. doi:10.3390/e12071743.

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.) *Corpus Linguistics: An International Handbook Volume 2*. 803–821. Berlin: Mouton de Gruyter.

Moreno-Sánchez, Isabel, Francesc Font-Clos & Álvaro Corral. 2016. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* 11(1). e0147073. doi:10.1371/journal.pone.0147073.

Naldi, Maurizio. 2003. Concentration indices and Zipf's law. *Economics Letters* 78(3). 329–334. doi:10.1016/S0165-1765(02)00251-3.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21 (5). 1112–1130. doi:10.3758/s13423-014-0585-6.

Popescu, Ioan-Iovitz. 2009. *Word Frequency Studies. (Quantitative Linguistics Volume 64)*. Berlin: Mouton de Gruyter.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Simon, Herbert A. 1955. On a class of skew distribution functions. *Biometrika* 42(3/4). doi:10.2307/2333389.

SPSS, I. 2012. *IBM SPSS statistics version* 21. Boston: International Business Machines Corp 126.

Zipf, George Kingsley. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts Addison-Wesley.

Zörnig, Peter & Gabriel Altmann. 1995. Unified representation of Zipf distributions, *Computational Statistics & Data Analysis*, 19(4). 461–473. doi:10.1016/0167-9473(94)00009-8.

Makoto Yamazaki

A time-series analysis of vocabulary in Japanese texts: Non-characteristic words and topic words

Abstract: In this study, I analyzed the distribution of words in the text from a time-series perspective. The data were comprised of 635 texts, where a token ranges from 1950–2050 words from the Balanced Corpus of Contemporary Written Japanese. Each text was divided into 10 segments containing an equal number of words, and the distribution of words among them was investigated. The relationship between the frequency of appearances and the characteristics of the words was also analyzed. From the results, the following conclusions were drawn. (1) The distribution of words among the segments follows a decreasing curve, like a Zipf's curve, but starts to rise close to the end of the curve. (2) At the token level, as the word appearance ratio increases, the ratio of particles increases, and the ratio of nouns decreases. Additionally, the ratio of auxiliary verbs becomes slightly higher, and there is no considerable change in the ratio of verbs. (3) Conversely, at the type level, the proportion of parts of speech remains almost unchanged. (4) The average number of words that appear in all segments was about 12 words per text, and there was no significant difference between the registers. (5) Four hundred and seventy different words appeared in all segments. They were divided into topic words, scene words, function words, and non-characteristic words from the discourse structure point of view, and were classified according to the number of text appearances.

Keywords: time series analysis, vocabulary, distribution, topic word, non-characteristic word, Japanese

Acknowledgment: This paper is an 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.

Makoto Yamazaki, National Institute for Japanese Language and Linguistics,
e-mail: yamazaki@ninjal.ac.jp

https://doi.org/10.1515/9783110763560-016

1 Introduction

In Yamazaki (2021), I reported the characteristics of the quantitative distribution of words that occur commonly across different texts.

The results suggested that the degree of the common occurrence of words increases, that is, as the number of texts containing a given word increases, the number of co-occurring words increases. It was also demonstrated that the number of commonly occurring words follows a curve similar to Zipf's law and that neither the length of the text nor the number of texts affected the shape of the distribution. It was argued that the cause of this distribution is that a small number of function words appear repeatedly in different texts, while the content words that support the topic differ from text to text.

This study investigates the distributional characteristics of words that appear in partial texts created by dividing a single text into multiple parts.

2 Previous studies and research questions

Dividing the text and examining the distribution of the words in it is closely related to the time-series analysis of the text. Previous studies on the distribution of words in texts have mostly focused on "vocabulary growth" – the rate at which the number of different words increases as the context progresses – and on dividing the text into semantic paragraphs by using the increase in the number of words. Mizutani (1975), Yasue (1981), Yamazaki (1983, 2015), Youmans (1991), and Covington and McFall (2010) have all approached their studies from that perspective. Figure 1 shows the semantic coherence of the text from its shape, considering the similarity of the text by sliding through intervals of 40 words each (Yamazaki 2015: 187). In Fig. 1, R0, R1, R2, and R3 in Fig. 1 correspond to the plots for all words (R0), without punctuation (R1), then, without function words (R2), and furthermore, without non-characteristic words (R3). The shape of these plots remains the same, indicating that punctuation, function words, and non-characteristic words are not involved in the semantic organization of the discourse.

However, to my knowledge, no studies have quantitatively surveyed the distribution of words in a text.

In this study, the following research questions are analyzed:

RQ1 How are the words distributed in each segment?

RQ2 How does the distribution differ from the distribution among different texts?

RQ3 What is the relationship between the number of segments in which a word appears and the characteristics of the word?

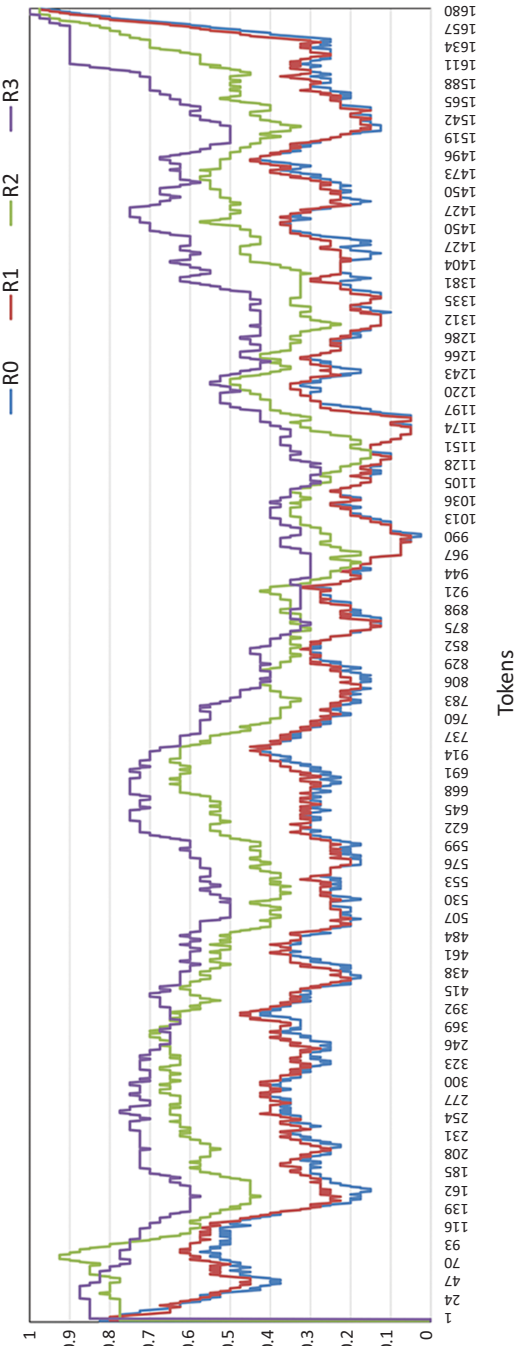


Fig. 1: Sliding similarity of the text.

3 Data and methods

The data used were comprised of 635 texts, where a token ranges from 1950–2050 words[1] from the Balanced Corpus of Contemporary Written Japanese[2] (BCCWJ). The breakdown of the texts is shown in Tab. 1. The tokens are about the same in different registers (it was designed that way), but the types vary widely from register to register. The greatest count of different words (in the magazines register PM) is more than twice as large as the smallest count (in the laws register OL).

Tab. 1: Data.

Register (Abbreviation)	Number of Texts	tokens (av.)	types (av.)
Library Books (LB)	259	1999.2	562.2
Best-selling Books (OB)	24	2011.0	563.1
Laws (OL)	10	1998.0	271.3
Minutes of the National Diet (OM)	1	2012.0	458.0
Textbooks (OT)	16	1994.4	459.5
White papers (OW)	34	1989.0	441.9
Yahoo! Blogs (OY)	38	1992.9	471.0
Publication Books (PB)	191	1997.3	523.4
Magazines (PM)	52	2002.3	612.6
Newspapers (PN)	10	1997.2	605.4
Total	635	1998.5	536.1

Each text was divided into 10 segments containing the same number of words[3] and for every word that appears in the text, the segments that the words appear in were examined. Table 2 shows the number of segments, word appearance ratio, number of tokens, and the number of types for each number of segments in which a word appears, as well as examples of words.

Word appearance ratio (WAR) is calculated by dividing the number of segments in which a word appears by the total number of segments. Therefore, in this study, WAR takes a value between 0.1 and 1 in increments of 0.1. In Tab. 2, we observe 546 words as types and 598 as tokens that appeared in only one segment, that is, words with a WAR of 0.1. Table 2 further indicates that about 75% of the number of types and 30% of the number of tokens are words that

1 The word count does not include spaces or punctuation.

2 For BCCWJ, see Maekawa et al. (2014).

3 Strictly speaking, the number of words in each segment is not the same. In case of fewer or more words, we adjusted the number of words in the last segment.

appeared in only one segment. Moreover, as this example is extracted from an economics book, some of the words in the example are related to the economy.

Tab. 2: Distribution of words in the text[4] (sample ID: LBa3_00006).

number of segments	word appearance ratio	types	tokens	word examples
1	0.1	546	598	kyuyo(salary), ten(dot), hachi(bee)
2	0.2	97	232	kinri(interest rate),kaiko(dismissal), sou (layer)
3	0.3	30	126	yokin(deposit), o(honorific prefix), kane (mone)
4	0.4	15	88	ni(two), nippon(Japan), rodo(labor)
5	0.5	6	41	kore(this), ga(case particle), tsuku(stick)
6	0.6	3	41	nai(auxiliary), koto(thing), ka(adverbial particle)
7	0.7	6	71	pasento(percent),iu(say), naru(become)
8	0.8	6	108	de(case particle), nen(year), nai(no)
9	0.9	1	41	ta(auxiliary)
10	1.0	13	671	no(case particle), wa(binding particle), ni (case particle)
Total	723	2017		

4 Results

4.1 Overview of distribution

First, I show an overview of the distribution of the number of segments in which a word appears. Figures 2 and 3 show the distribution of the number of words by WAR. Figures 2 and 3 correspond to the distribution of types and tokens, respectively. From Fig. 2, we observe that as the value of WAR increases, the number of words decreases.

A Tukey's HSD test for neighboring WAR in Figs. 2 and 3 was performed, and a significant difference at the 5% level was found where * was assigned. It is interesting to note that in both graphs, the curve is not monotonically decreasing; rather, the graph turns upward at the end.

4 The bibliographic information for this text is as follows. Naohiko Higashida. 1986. *Shakkin-koku no keizai-gaku: Buraziru mou hitotsu no keizai-genri* (The Economics of Debt: Brazil, Another Economic Principle), Tokyo: Nihon Keizai Shimbun Publishing.

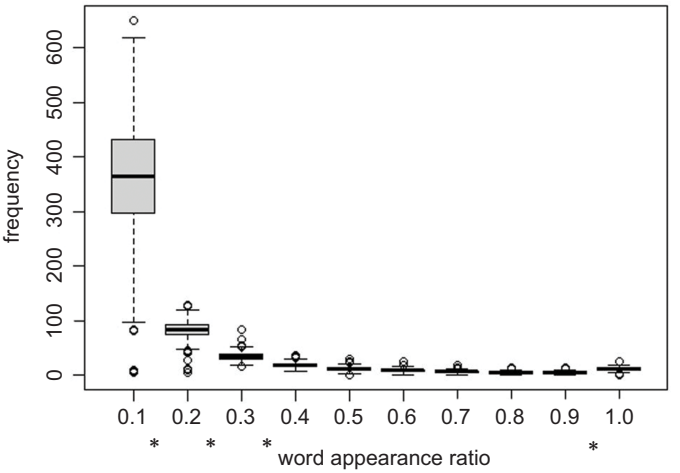


Fig. 2: Distribution of segments (type).

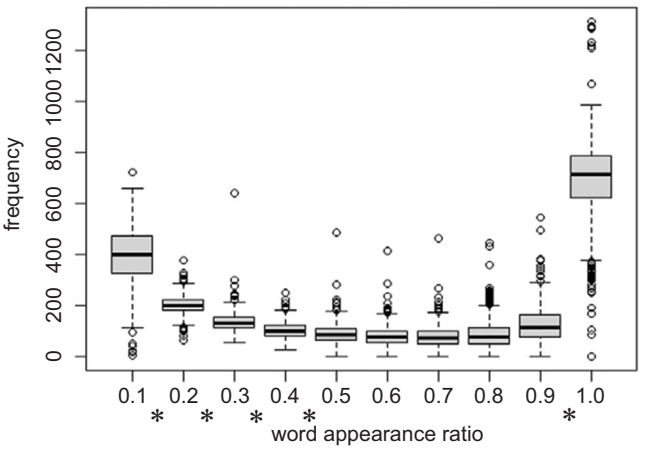


Fig. 3: Distribution of segments (token).

Figure 4 shows the relationship between the WAR and the type/token ratio of the words with that ratio. From Fig. 4, we observe that the value of TTR decreases as the WAR increases. This is related to the fact that the number of function words increases as the WAR increases, which will be discussed in section 4.3.

Herein, RQ1 is answered. The distribution of words among the segments follows a decreasing curve, like a Zipf's curve, but starts to rise near the end of the curve.

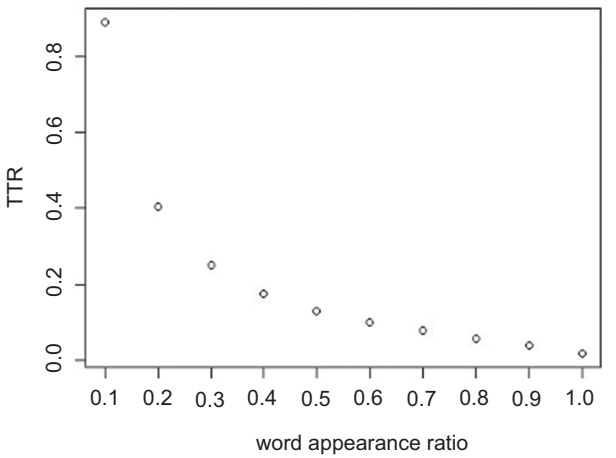


Fig. 4: Word appearance ratio and TTR.

4.2 Comparison with distributions between different texts

In the previous section, we demonstrated that the distribution of words among the segments follows a curve that initially decreases and then increases. This trend is also seen when observing the degree of common occurrence across different texts (Yamazaki, 2021). Figure 5 shows the percentage of words that were commonly used in 10 different texts of 100 words in length. In tokens, as well as in types, the same trend as in Figs. 2 and 3 can be seen.

4.3 Parts of speech

Next, let us analyze the distribution of WAR by parts of speech. In analyzing the parts of speech, only the main parts of speech, such as nouns and verbs, were used. Unknown words that failed morphological analysis, or words with attributes that cannot be regarded as parts of speech, were not included. The number of deleted words was 347, about 0.1% tokens and about 0.04% types.

Table 3 shows the percentage of parts of speech by WAR. The numbers in the first row of Tab. 3 represent the number of segments in which a word appears. For example, the figures in the first column show the percentage of the part of speech of the words that appear in any one of the ten segments. Similarly, the numbers in the second column represent the percentage of the part of speech of the words that appear in any two segments, and so on. The percentage is measured in tokens.

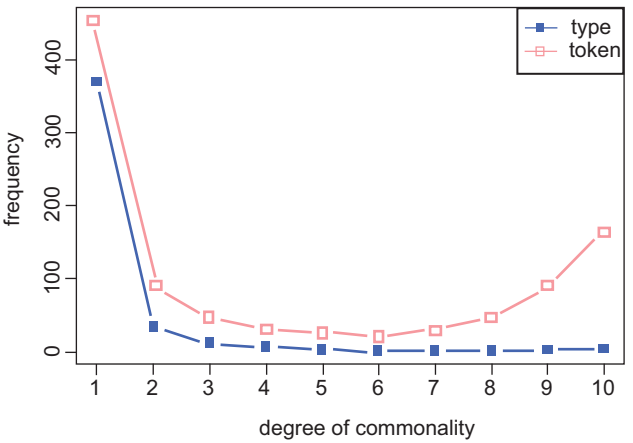


Fig. 5: The average number of commonly used words (type and token). Text length: 100, number of texts: 10.

It can be seen in Tab. 3 that the percentage of nouns decreases as the WAR increases, and in contrast, the percentage of particles increases. We can also see that the percentage of auxiliary verbs increases gradually, though not as much as that of particles and that the percentage of verbs remains almost constant.

Table 4 shows the percentages of parts of speech by the WAR in types. The percentage of parts of speech does not change as much in the case of tokens in Tab. 3. However, a closer look shows that the percentages of nouns, verbs, adverbs, and adjectives decrease, while the percentages of suffixes, particles, auxiliary verbs, pronouns, and symbols increase as the WAR increases.

4.4 Highest WAR words

In this section, we examine the characteristics of words that show the highest WAR, that is, words that appear in all segments. Figures 6 and 7 show the average number of words per text that showed the highest WAR in types and tokens respectively. Broadly speaking, the average frequency does not vary by the register, either in the types or in the tokens.

Table 5 shows the distribution of the words with the highest WAR by part of speech. In the breakdown of parts of speech, there are some differences between registers. For example, in law texts, the proportion of nouns is 38.5%, which is the highest among the registers. Moreover, in the Minutes of the National Diet, the percentage of auxiliary verbs is as high as 20%, which is also a

Tab. 3: The percentage of parts of speech by word appearance ratio (token).

POS	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Total
Verb	18.0	16.3	15.0	13.9	14.4	14.5	16.6	18.2	18.6	12.8	17.2
Adjective	2.8	2.6	2.2	2.1	2.2	2.8	3.2	2.4	1.5	0.3	2.6
Adjective Verb	3.0	2.1	1.7	1.8	1.8	2.0	2.1	1.8	0.8	0.1	2.6
Noun	60.3	57.1	53.9	50.1	45.8	40.8	36.8	33.9	24.6	8.5	56.5
Pronoun	1.0	1.9	2.9	3.4	3.8	3.9	3.7	2.5	1.5	0.5	1.5
Adverb	5.0	4.2	3.2	2.1	1.6	1.3	0.7	0.5	0.2	0.0	4.3
Adnominal	0.7	1.1	1.6	2.1	2.9	3.2	3.4	2.5	1.5	0.2	1.0
Conjunction	0.6	1.0	1.2	1.3	1.0	0.8	0.7	0.4	0.2	0.0	0.8
Interjection	0.5	0.4	0.4	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.4
Auxiliary	0.8	1.8	3.1	4.3	5.9	7.7	8.9	11.0	14.1	15.4	2.1
Particle	2.0	4.6	7.6	10.7	13.9	16.4	17.9	22.0	33.6	60.5	5.4
Prefix	1.0	1.2	1.2	1.4	1.2	0.9	1.0	0.8	0.7	0.3	1.0
Suffix	3.8	4.9	5.4	5.9	4.9	4.9	4.5	3.3	2.6	1.1	4.1
Symbol[5]	0.5	0.7	0.6	0.8	0.5	0.6	0.4	0.7	0.2	0.4	0.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Tab. 4: The percentage of parts of speech by word appearance ratio (type).

POS	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Total
Verb	10.3	10.0	8.9	8.0	7.7	7.3	6.7	6.5	7.3	5.4	9.7
Adjective	1.2	1.2	1.4	1.4	1.0	0.9	1.1	0.4	1.0	0.4	1.2
Adjective Verb	2.3	2.3	2.0	1.6	1.4	1.1	0.5	0.7	0.2	0.6	2.1
Noun	79.9	78.8	77.9	77.4	77.7	75.5	75.5	74.7	74.1	74.8	79.0
Pronoun	0.2	0.4	0.6	0.8	1.0	1.1	1.6	1.6	1.9	1.7	0.4
Adverb	2.9	2.1	2.1	1.7	1.2	1.2	0.7	0.6	0.6	0.0	2.4
Adnominal	0.1	0.2	0.3	0.4	0.4	0.4	0.5	0.4	0.6	0.6	0.2
Conjunction	0.1	0.2	0.3	0.5	0.4	0.5	0.5	0.4	0.4	0.2	0.2
Interjection	0.4	0.4	0.5	0.3	0.3	0.2	0.2	0.1	0.0	0.0	0.4
Auxiliary	0.1	0.3	0.5	0.7	0.8	1.4	1.6	2.1	2.1	2.2	0.4
Particle	0.2	0.5	0.9	1.3	1.8	2.2	2.9	3.7	4.6	4.3	0.6
Prefix	0.5	0.7	0.8	0.7	1.0	1.0	1.2	0.8	1.1	1.1	0.6
Suffix	1.5	2.5	3.4	4.4	4.5	5.8	5.3	5.9	5.2	6.5	2.4
Symbol	0.3	0.5	0.6	1.0	0.9	1.3	1.5	2.0	1.0	2.2	0.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

characteristic. The ratio of nouns seems to correspond to the stylistic scale of written and spoken language, but this is not a general statement, given that the ratio is almost identical for nouns in the white papers and the blogs.

5 Symbols in these papers do not refer to punctuation marks, but to non-Japanese letters such as A or B.

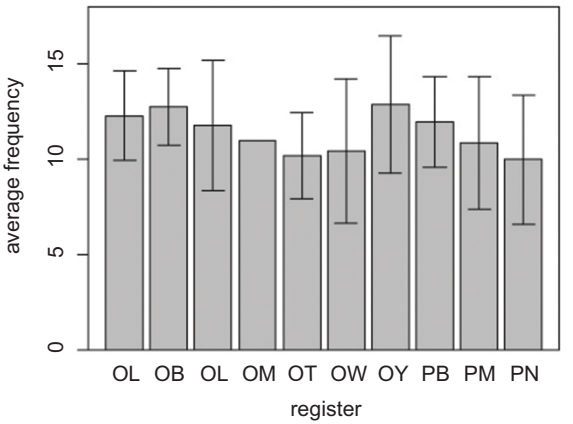


Fig. 6: Average number of words of the highest word appearance ratio (type).

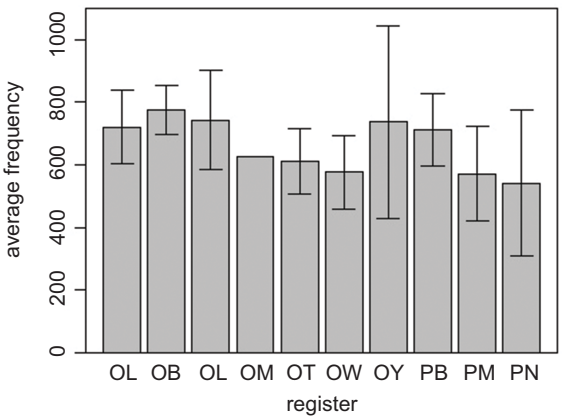


Fig. 7: Average number of words of the highest word appearance ratio (token).

Next, we consider the words in terms of their discourse constituting roles. The words that appeared in all segments were 470 words in types. Table 6 shows the classification of these words regarding whether or not they are topic words. Nouns and proper nouns are considered topic words, while other words are considered non-topic words. Table 6 shows that nouns account for about 70% of the total distribution, which is quite different from the distribution of parts of speech of words appearing in all the texts shown by Yamazaki (2021: 133). According to Yamazaki (2021: 133), 90% of the words that appear in all texts are function words, and only 9% are nouns.

Tab. 5: The percentage of parts of speech by register for the words that appeared in all segments (token).

POS	LB	OB	OL	OM	OT	OW	OY	PB	PM	PN
Verb	14.0	14.1	9.8	10.0	11.6	13.5	7.0	13.0	10.6	9.0
Adjective	0.3	0.3	0.0	0.0	0.0	0.0	0.4	0.3	0.2	0.0
Adjective Verb	0.1	0.3	0.0	0.0	1.2	0.0	0.0	0.1	0.2	0.0
Noun	4.7	3.2	38.5	10.0	7.9	21.0	21.1	7.3	12.6	13.0
Pronoun	0.8	0.3	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
Adnominal	0.2	0.3	1.6	0.0	0.0	0.3	0.0	0.1	0.0	0.0
Conjunction	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Auxiliary	16.3	16.9	1.6	20.0	12.8	6.9	15.4	16.3	14.5	14.0
Particle	62.8	63.9	39.3	60.0	65.2	54.2	48.3	61.2	60.8	58.0
Prefix	0.1	0.0	6.6	0.0	0.6	0.3	0.0	0.3	0.0	1.0
Suffix	0.6	0.0	2.5	0.0	0.6	3.7	3.5	0.9	0.5	5.0
Symbol	0.1	0.3	0.0	0.0	0.0	0.0	4.3	0.0	0.5	0.0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Tab. 6: Classification of POS.

POS	Frequency	ratio	topic/non-topic
verb	25	0.053	non-topic
adjective	2	0.004	non-topic
adjective verb	3	0.006	non-topic
noun	247	0.526	topic
proper name	80	0.170	topic
numerals	21	0.045	non-topic
pronoun	8	0.017	non-topic
adnominal	3	0.006	non-topic
conjunction	1	0.002	non-topic
auxiliary	10	0.021	non-topic
particle	25	0.053	non-topic
prefix	5	0.011	non-topic
suffix	30	0.064	non-topic
symbols	10	0.021	non-topic
Total	470	1.000	

Tanaka (1973) divided the high-frequency words that appear in the text into topic words, scene words, and non-characteristic words and analyzed their discourse role in the text. According to Tanaka (1973), the aforementioned groups of words have the following properties.

Topic words are the main subject of the text, such as the hero, heroine, and characters in novels, and appear very frequently. Scene words are words related

to a scene or background of the story and occur less frequently than topic words. Non-characteristic words are words that appear in any text and rarely reflect the character or characteristics of a particular text or document. They are based on the high-frequency words obtained from a vocabulary survey.[6]

Table 7 classifies the words that appear in all segments based on these classifications, primarily regarding their role in constructing the text. The average number of texts in Tab. 7 shows the number of texts in which a word appeared, on average. Topical words appear in almost only one text, whereas impersonal words appear in an average of 22 texts. Function words, such as particles and auxiliary verbs, have even higher values, appearing in about 164 texts. Therefore, we can classify the words that appear in all the segments into topical and non-topical words based on the number of occurrences.

Tab. 7: Classification of words by the discourse role.

classification	frequency	average number of texts	word example
topic word (proper noun)	79	1.0	Nakai(surname), Meiji(era), Syowa(era), Bz(band name), Ayani(person name)
topic word (general)	245	1.2	jou(article), ten(point), kodomo(child), kou(item), kitei(regulations)
scene word	51	1.98	i(class), watashi(I), ee(E), oo(O), gata(type), tou(etc.), bii(B)
non-characteristic word	60	22.6	suru(do), iru(stay), aru(exist), iu(say), ichi(one), ni(two), san(three)
function word	35	164.2	no(case particle), ni(case particle), o(case particle), ha(binding particle), te(conjunction particle)
Total	470	16.1	

Herein, RQ3 is answered. The number of segments in which a word appears is related to the part of speech and the role of the word in discourse structure.

6 Yamazaki (2012) presented a list of non-characteristic words by Tanaka (1973) adjusted for morphological analysis using the electronic dictionary UniDic.

4.5 Word distribution and lexical cohesion

Figure 8 shows the distances between the segments of words that appeared in only two segments. The distance between the segments is expressed as the difference in the number of segments. Therefore, if a word appears in adjacent segments, the distance is 1. From Fig. 8 we can see that as the distance between segments of a word increases, the frequency of the word decreases. This can be considered as a manifestation of lexical cohesion.

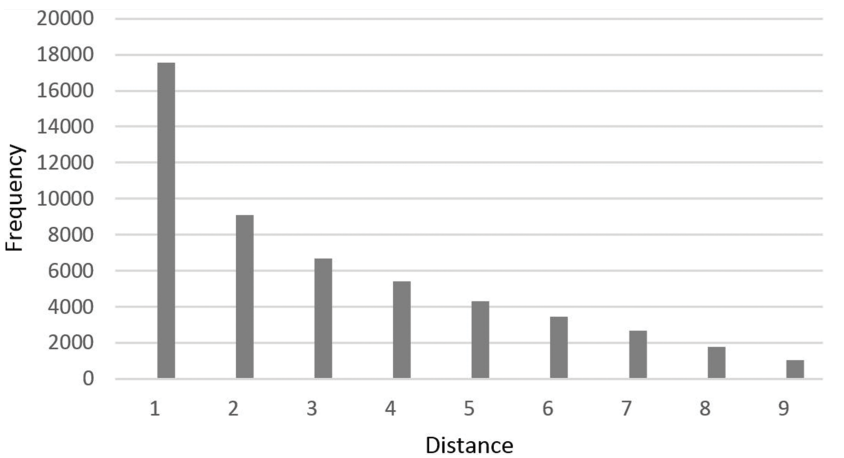


Fig. 8: The distance of the words that appeared in two segments.

5 Conclusions and further tasks

In this study, we explored how words are distributed in evenly spaced segments of the same text and analyzed the relationship between the frequency of appearances and the characteristics of the words. The results revealed that the shape of the distribution was similar to that between different texts. However, the breakdown of its parts of speech is very different. This may be because the words that support the topic of the text are widely distributed throughout the text. In the future, the relationship between the similarity of texts and the shape of the distribution could be explored. It would be interesting to find whether a high degree of similarity would lead to results like the present one, even if they were not extracted from the same text. It is also important to clarify the mathematical model of the shape of the distribution.

References

Covington, Michael A. & Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100.

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 & Evaluation* 48. 345–371. https://doi.org/10.1007/s10579-013-9261-0

Mizutani, Sizuo. 1975. Mijikai sakuhin no goi no ryoteki kozo: showa shoki ryukoka no chosa kara 1 [On the structure of vocabulary of the short work: From a survey of Japanese popular songs 1930s, (1)]. *Mathematical Linguistics* 72. 1–12.

Tanaka, Akio. 1973. Jido shoroku ni okeru ki wado no seikaku [Key-words for automatic abstracting of literary texts]. *Studies in Computational Linguistics* 5. 141–184.

Yamazaki, Makoto. 2012. *Danrakukan no ruijido o riyoshita tekusuto no kessokusei no sokutei* [Measurement of textual cohesion using similarity between paragraphs]. *Proceedings of the 2nd Corpus Linguistics Workshop*. 291–298.

Yamazaki, Makoto. 1983. *Bunsho no wadai no tenkai o hakaru shakudo* [A new index for topical change in the context], *Mathematical Linguistics* 13(7). 346–360.

Yamazaki, Makoto. 2015. *Tekisuto ni okeru goiteki kessokusei no keiryoteki kenkyu* [Lexical cohesion in Japanese texts: A quantitative approach]. Osaka: Izumi Shoin.

Yamazaki, Makoto. 2021. Distribution and characteristics of commonly used words across different texts in Japanese. In Adam Pawłowski, Ján Mačutek, Sheila Embleton & George Mikros (eds.) *Language and text: Data, models, information and applications*, 121–134. (Current Issues in Linguistic Theory 356) Amsterdam/Philadelphia: John Benjamins.

Yasue, Sawako. 1981. *Kobun o otta kotonari gosu no ug*oki [Different word count trends following the context]. *Tokyo Woman's Christian University Japanese Literature* 56. 32–45.

Youmans, Gilbert. 1991. A new tool for discourse analysis. Langauge 67(4). 763–789.

Authors' addresses

Benešová, Barbora
University of Ostrava
e-mail: benesovaba@seznam.cz

Čech, Radek
University of Ostrava
e-mail: cechradek@gmail.com

Chen, Xinying
Xi'an Jiaotong University
e-mail: chenxinying@mail.xjtu.edu.cn

Cortelazzo, Michele A.
Università degli Studi di Padova
e-mail: cortmic@unipd.it

Courtin, Marine
University Sorbonne Nouvelle
e-mail: rinema56@gmail.com

Cvrcek, Václav
Charles University
e-mail: vaclav.cvrcek@ff.cuni.cz

Embleton, Sheila
York University
e-mail: embleton@yorku.ca

Garrido, Juan María
Universidad Nacional de Educación a Distancia
e-mail: jmgarrido@flog.uned.es

Gatti, Franco M. T.
Università degli Studi di Padova
e-mail: franco.gatti.1@phd.unipd.it

Gerdes, Kim
University Paris Saclay
e-mail: kim@gerdes.fr

Hernández-Fernández, Antoni
Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia, Institut d'Estudis Catalans
e-mail: antonio.hernandez@upc.edu

Ján Mačutek
Mathematical Institute, Slovak Academy of Sciences and Constantine the Philosopher University in Nitra
e-mail: jmacutek@yahoo.com

Kahane, Sylvain
Université Paris Nanterre
e-mail: sylvain@kahane.fr

Kawasaki, Yoshifumi
The University of Tokyo
e-mail: ykawasaki@g.ecc.u-tokyo.ac.jp

Kelih, Emmerich
University of Vienna
e-mail: emmerich.kelih@univie.ac.at

Litvinova, Olga A.
Corpus Idiolectology Lab, Voronezh State Pedagogical University
e-mail: centr_rus_yaz@mail.ru

Litvinova, Tatiana A.
Corpus Idiolectology Lab, Voronezh State Pedagogical University
e-mail: centr_rus_yaz@mail.ru

Liu, Haitao
Department of Linguistics, Zhejiang University
e-mail: htliu@163.com

https://doi.org/10.1515/9783110763560-017

Lukeš, David
Charles University
e-mail: david.lukes@ff.cuni.cz

Luque, Bartolo
Universidad Politécnica de Madrid
e-mail: bartolome.luque@upm.es

Mikros, George K.
Hamad Bin Khalifa University
e-mail: gmikros@hbku.edu.qa

Milicka, Jirí
Charles University
e-mail: milicka@centrum.cz

Mačutek, Ján
Mathematical Institute, Slovak Academy of Sciences and Constantine the Philosopher University in Nitra
e-mail: jmacutek@yahoo.com

Motalova, Tereza
Palacký University Olomouc
e-mail: tereza.motalova@upol.cz

Pawłowskł, Adam
University of Wrocław
e-mail: adam.pawlowski@uwr.edu.pl

Pelegrinová, Kateřina
Department of the Czech language, Faculty of Arts, University of Ostrava
e-mail: pelegrinovak@gmail.com

Sanada, Haruko
Faculty of Economics, Rissho University
e-mail: hsanada@ris.ac.jp

Torre, Iván González
Universidad del País Vasco y Universidad Politécnica de Madrid
e-mail: ivan.gonzalez.torre@upm.es

Tsuchiyama, Gen
Center for Interdisciplinary AI and Data Science, Ochanomizu University
e-mail: tsuchiyama.gen@ocha.ac.jp

Tuzzi, Arjuna
Università degli Studi di Padova
e-mail: arjuna.tuzzi@unipd.it

Uritescu, Dorin
York University

Walkowłak, Tomasz
Wrocław University of Technology
e-mail: tomasz.walkowiak@pwr.edu.pl

Wang, Yawen
Department of Linguistics
Zhejiang University
e-mail: mileywyw@zju.edu.cn

Wheeler, Eric S.
York University
e-mail: eric.wheeler@sympatico.ca

Yamazaki, Makoto
National Institute for Japanese Language and Linguistics
e-mail: yamazaki@ninjal.ac.jp

Name index

https://doi.org/10.1515/9783110763560-018

Subject index

https://doi.org/10.1515/9783110763560-019