Issues in Quantitative Linguistics

4

edited by

Emmerich Kelih
Róisín Knight
Ján Mačutek
Andrew Wilson

Dedicated to Reinhard Köhler on the occasion of his 65th birthday

2016

RAM-Verlag


Lecturer
Prof. Dr.
Köhler

Studies in Quantitative Linguistics

Editors

Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@univie.ac.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language.* 2014, VIII + 123 pp.

17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.
19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional occurrences in texts: Weighted Consensus Strings.* 2015. II+178 pp.

ISBN: 978-3-942303-44-6

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

Contents

The first steps

To learn quantitative linguistics means to collect the works written by Reinhard Köhler and read them thoroughly. Needless to say, one must already know what other (= normal) linguists know, i.e. definitions, classifications, rules, history, etc., but, if one ventures to take a step further, one necessarily bumps against the door of a world one wanted to avoid. It is surely not the infinite paradise but rather the hell in which Köhler, dressed as Lucifer, leads the innocent linguist through the labyrinth of new concepts, formulas, systems, synergetics, statistics, hypotheses, theories. The world of language begins to change its form and the linguist begins (very) slowly to see that there is more light in this world than outside. Köhler loses slowly his Luciferian shape and at once he seems to be the angel responsible for this world. He leads the linguist along ways that join the individual properties of language. One imagines that one is in phonology – very far from syntax – but the angel shows one that there is a very short way between them, and, what is more, one can express it by a formula. That means one need not walk the way; it is sufficient to think it.

Of course, this is rather an esoteric world that exists also in physics and biology, and there is nobody who could prohibit you from entering it (the only exception is the dean of your faculty or head of your department!), but fortunately there is an angel looking sharply at the dean (or head) and leading you into a world in which human language looks like a self-organized system. You may settle on a concept and the system shows you immediately all the links to other concepts. What is more, it brings you to concepts developed by the "Lucifer" himself. As a (pure) linguist, you never heard of them before.

If you succeed in abandoning this world, you will realize that, in front of the door, the same Lucifer stays and wishes you – smiling politely – good moral conscience and much success in repeating what you must teach the students in order to become an "expert" yourself. Linguistics is not about teaching a language but an immersion into worlds which are abstract and similar to the fifth (or higher) dimension of physics. At each step, you can see the smiling Lucifer who created it himself – perhaps only in order to irritate classical linguists who thought that they knew everything already.

But, if you want to stay in this world, you will soon see that you must learn a lot of mathematics, a lot of philosophy of science, forget grammatical rules and all degenerative drawing of trees, and ask the masked Lucifer for help. He will merely smile and, since he wants to retire this year, he will show you a monumental heap of paper and say: "Read all the papers and books in this heap. You can read them more quickly than I wrote them!"

Nevertheless, we hope that he will make the heap higher.

Gabriel Altmann

On Köhlerian Motifs

Gabriel Altmann

Abstract. The article brings some methodological issues concerning language units, especially their relativity and shows the place of Köhlerian motifs in linguistics. Problems concerning motifs and the perspectives are shown in ten points.

Keywords: *Linguistic units, motifs, synergetics*

We live in a linguistic world in which the units discovered centuries ago are still *believed* to be real entities and the only ones. There has been a move away from Plato, but F. de Saussure brought him back through the back door into linguistics. Further units were added, like the phoneme, the morpheme, the mora; diachrony was separated from synchrony; and two realities were conjectured: langue and parole (later on competence and performance, etc.). Spoken language was declared to be the realization of something sitting in the head. It was the time of the hegemony of noumenon and phenomenon, essentialism, etc.

However, in the course of time something changed. In the 20th century the systemic view of reality appeared and the philosophy of science shed light to any kind of scientific research. One was also forced to capture some frequently appearing regularities with the help of formulas. But again, Plato returned. One believed that mathematical models represent truth. The most extreme exaggeration has been produced by "algebraic linguistics" which did not allow any exceptions and presented everything as deterministic realities. Peculiarly enough, this came after G.K. Zipf (1935) already demonstrated that everything in language is linked with something else probabilistically, and changes in language are caused by efforts to make language production and decoding as easy as possible. Looking at other sciences one saw that even if there are no two absolutely identical roses or cats or even stones, there is something that is common to individual classes. Transferred into language, one sees that no word pronounced a million times by a million people is absolutely identical; there is some difference, at least in the phonetics. If pronounced in a sentence, the diversification is still greater. It may consist in phonetic, grammatical, semantic, intentional, dialectal, contextual, idiolectal, etc. differences. But there is something constant behind all this, common to all languages, dialects, historical epochs, speakers, hearers,… making communication possible. At the beginning, one thought that it was grammatical rules which can be represented by algorithms, trees, etc. Everybody speaking a language obeys the rules…! But rules change in the course of time; they are different for different languages, even for idiolects. An ancient Latin speaker would not understand even a single French word but French speakers do not have problems with their own language.

Thus we see that changes occur but there is something like equilibrium in language enabling us to understand and to be understood. The equilibrium is

maintained not by fixed forms or rules but by some intrinsic relations concealed behind them. These relations cannot be learned or constructed by the speaker, nevertheless, they exist somewhere in the background and can be captured by some means taken from mathematics. The mathematical models help us to understand the mechanisms. These mechanisms cannot be constructed – just as in natural sciences or biology -, they cannot be learned consciously to be followed, but they exist and are obeyed unconsciously just as the laws of physics.

If we accept this philosophy, we must ask which phenomena are linked by these laws. Starting from the classical view of language, the relevant phenomena are the properties of some units. But according to school grammar (!) there are sounds, phonemes, morphemes, syllables, moras, words, stems, phrases, clauses and sentences. And there are rules(!) allowing us to segment them. The number of definitions of these units is enormous. This is a quite normal beginning of a scientific discipline. But as time goes on, one sees that this is not all. From time to time one discovers a new unit, a new property, a relation between them and on this long journey one sees that it can never be finished. Nothing is simply given, everything must be defined. Unfortunately, definitions do not have any truth value, they are conventions. Their only task is to enable us to set up lists of units and find them in texts. Everything else, e.g. the overall relations between them, can be captured by mathematical models which do not express truth, but bring order in our view of languages and texts.

In recent years, four events corroborated this view: The defining of new units, namely motifs, hrebs and Belza chains; and presenting language as a control cycle, i.e. as an open dynamic system in which laws rule! Hrebs were defined by L. Hřebíček (2000) under the name "sentence aggregates" but they were re-baptized later on. A hreb contains all sentences in which a given entity (or its synonym, reference, metaphor etc.) can be found. Belza-chains (Belza 1971, Skorochod`ko 1981, Chen, Altmann 2015) are similar to hrebs but the sentences must be consecutive. Motifs were defined first rather intuitively in music; a formal definition was given to them by M. Boroda (1973, 1982a,b, 1988). They were introduced into linguistics by Köhler (2006) who is also the constructor, so to say, the father of language synergetics (1986, 2005) based on systems theory.

Our aim is to look here at the motifs. Since they are legal entities having some properties, they must follow some laws. Seen from the opposite direction: the definition of a unit may be accepted if it allows the formulation of a law. Even if the level of abstraction of motifs is very high, we conjecture that they have the same properties as e.g. words and these properties are measurable. There are two kinds of motifs: quantitative ones in which the degrees of the property are ordered in a non-decreasing sequence; and qualitative ones in which a new motif begins if some entity would be repeated from the previous motif (cf. Köhler, Naumann 2008). Two examples can illustrate this approach: In the sequence [2,3,3,5,2,4,5] there are two motifs, namely [2,3,3,5] and [2,4,5]; in the sequence [A,B,D,G.B,C,F] there are two qualitative motifs: [A,B,D,G] and [B,C,F].

Now, motifs have several properties and behaviours of which we can present here the first ten:

(1) Each motif has a length given by the number of entities occurring in it. Length is a well known property subject to laws, hence every text can be evaluated in this way (cf. Popescu, Best, Altmann 2014), computing the frequency of individual lengths.

(2) In quantitative motifs, the difference between the last and the first member, namely the range, is a characteristic feature whose distribution may be sought. In the same way, one can compute the mean of the numbers in a motif and study the distribution of means or the sequence of means.

(3) If one prepares a list of motifs and counts their frequencies, one may obtain a rank-order distribution of motifs.

(4) The richness of motifs in a text can be computed and the results can be compared statistically with other texts (Mačutek 2009).

(5) The distance between equal motifs can be computed and a model may be developed.

(6) One can define motifs at any linguistic level. In phonetics/phonemics, one can distinguish pairs of phonemes, syllables, poetic feet, phonetic verse motifs, etc. In morphology, one can define kinds of morphemes, parts of speech, semantic classes of nouns, verbs, adjectives, adverbs, etc.; in grammar, sequences of grammatical rules; in semantics one can measure the concreteness/abstractness of words, one can apply Osgood's (1957) semantic differential, ascribe to each entity its value and present the sequence in motifs. In a stage play one can state the classes of speech acts, observe the sequence and subdivide it into motifs. The individual actors can be characterized, the acts of the stage plays themselves can be characterized and compared with other ones.

(7) Having e.g. a survey of motifs in a text in one language, the text can be translated, its motifs can be evaluated and the languages can be compared.

(8) In the same way the development of a language can be observed and described. The usual ways are comparing the words and rules, the daily bread of historical linguistics performed on a very concrete level.

(9) Motifs of one kind have their properties and motifs of another kind have their own ones, perhaps parallel or different. One can conjecture that at least some of the properties are linked. That is, motifs are not isolated entities, simply defined and stated, but they also form a kind of Köhlerian control cycle, though the links may be somewhat more complex than those of the "well known" units.

(10) Motifs may help to distinguish text types and perhaps also express the degree of language type (synthetism, analytism, etc.).

The field has already been strongly extended by the solutions of Köhler and his collaborators (cf. Beliankou, Köhler, Naumann 2013; Köhler 2006; Köhler 2008a,b; Köhler, Naumann 2008, 2009, 2010) and other researchers (cf. Sanada 2010; Mačutek, Mikros 2015; Milička 2015; Čech, Vincze, Altmann 2016).

Future research will bring both new types of motifs and many new problems. Concerning some of the above problems, one may develop an enormous number of investigations, all parallel to, e.g., those solved for words. To mention only some of them: For some of the above problems one can characterize texts by entropy, repeat rate, moments and their functions, e.g. Ord's criterion, various

indicators used in text analysis (cf. Popescu, Mačutek, Altmann 2009; Popescu et al. 2010; Tuzzi, Popescu, Altmann 2010; Popescu, Čech, Altmann 2011; Altmann, Köhler 2015), compute motif richness, the h-point, etc.; for their sequences one can compute Minkowski sausages, Hurst coefficient, autocorrelation, Markov chains, the fractal dimension and many others. For now, the investigations have been rather modest because of the methodological problems associated with motifs.

Formerly, in linguistics, one considered some units as given, "natural", really existing, though we know that e.g. in spoken language there is no punctuation; but the sentence was a holy grail, even if it was taken from the written, i.e. secondary, language. There are no phonemes but there is a discipline called phonemics. The number of definitions of word and sentence is merely a sign of efforts to capture something intellectually, to set up criteria for segmentation, counting and measurement. To be sure, we have some vague *concepts* in our mind, and by language we try to convey them to the hearer, but the means by which we do it is a convention. We are not aware of the fact that we do it whilst obeying some laws which are not known to us. The case of the green colour of grass, mentioned several times by the present author is a very good example of our approach: the biologist considers it a property of plants, the physicist considers it a property of light, the physiologist considers it a property of our eyes (other animals may receive it differently) and a linguist considers it a roughly expressed concept that did not exist in some languages at all. It is not important what it "is", but how to analyze it from the given point of view. Now, Köhler's introduction of motifs is not only a linguistic property: one can define it also for other "real" objects, e.g. a row of houses, sequence of trees in a forest. Nevertheless, for many linguists it is not easy to accept a new entity that has little importance for learning a language or for describing its grammar. It is an abstraction, and practitioners do not know what to do with it. However, from a theoretical viewpoint it shows that we can construct a hierarchy of abstractions and search for their regularities. The same is done e.g. in physics where some entities are conjectured but their discovery comes much later. Motifs are theoretical entities but they can be measured just like some psychological entities whose existence will never be captured by means of machines.

Up to the discovery of motifs we had merely various classifications of entities. But motifs can be defined for all of them and there are various forms of motifs. At the moment we stay at the descriptive level, namely we want to show what kinds of motifs there are, what kinds of properties they have and how they behave. This is a sufficient task for an extensive discipline. But we know that, after having answered at least some of these questions, two new problems will arise: Firstly, *why* do the motifs behave in the given way, and secondly, are there some *higher* abstractions?

The first problem forces us to set up hypotheses, test them and search for some statements which can be expressed mathematically and are linked with other statements, i.e. we shall be forced to search for laws belonging to some theories. And even when we have found some of them, we shall incessantly ask whether there are still higher abstractions. This question is the same as that in

physics: what is the smallest element of matter? Does the "smallest" element consist of still smaller parts? In linguistics we go in the opposite direction and ask: do we already have the highest abstraction or should we continue?

Recently, it has been discovered by Zörnig et al. (2016) that texts can be analyzed also vertically, yielding consensus strings, and from this perspective we obtain quite new types of motifs. They have their own properties, may be used for characterizing texts, and later on, it will surely be possible to insert them in a Köhlerian control circuit. A problem that seems to be a dream of the future is the finding of links between motifs, hrebs, Belza chains and consensus strings. Are there some, or is each of them a very high abstraction in a different direction, in a quite different subsystem of language?

Already motifs, though they seem to be very simple, easily definable and explorable, represent something like quantum physics: a quite different view of linguistic reality.

References

Altmann, G., Köhler, R. (2015). *Forms and Degrees of Repetitions in Texts. Detection and Analysis*. Berlin-Munich-Boston: de Gruyter.

Beliankou, Andrei, Köhler, Reinhard, Naumann, Sven (2013). Quantitative Properties of Argumentation Motifs. In: Obradović, I.; Kelih, E.; Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012: 33-43.* Belgrade.

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych textov. In: *Semantičeskie problemy avtomatizacii i informacionnogo potoka: 58-73*. Kiev.

Boroda, M.G. (1973). K voprosu o metroritmičeski elementarnoj edinice v muzyke. *Bulletin of the Academy of Sciences of the Georgian SSR 71(3), 745-748.*

Boroda, M. (1982a). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J., Boroda, G.M., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen: 231-262.* Bochum: Brockmeyer.

Boroda, M.G. (1982a). Die melodische Elementareinheit. In: Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen: 205-222.* Bochum: Brockmeyer.

Boroda, M.G. (1988). Towards a problem of basic structural units of musical texts. *Musikometrika 1, 11-69.* Bochum: Brockmeyer.

Čech, R., Vincze, V., Altmann, G. (2016). On motifs and verb valency. In print.

Chen, R., Altmann, G. (2015). Conceptual inertia in texts. *Glottometrics 30, 73-88.*

Hrebiček, L. (2000). *Variation in sequences. Contributions to general text theory*. Oriental Institute, Prague.

Köhler, R. (2000). A study on the informational content of sequences of syntactic units. In: L.A. Kuz'min (ed.), *Jazyk, glagol, predloženie. K 70-letiju G. G.Sil'nitskogo.* Smolensk, 51-61.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: J. Genzor and M. Bucková (eds.), *Favete linguis. Studies in honour of Victor Krupa: 145-152*. Bratislava: Slovak Academic Press,

Köhler, R. (2008a). *Word length in text. A study in the syntagmatic dimension*. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v prohybe*: *416-421*. Bratislava: VEDA vydavatel'stvo SAV.

Köhler, R. (2008b). Sequences of linguistic quantities. Report on a new unit of investtigation. *Glottotheory 1(1), 115-119.*

Köhler, R., Altmann, G. (1996). "Language Forces" and synergetic modelling of language phenomena. In: Schmidt, P. (ed.), *Glottometrika 15: 63-76*. Trier: WVT.

Köhler, R., Naumann, S. (2008c). *Quantitative text analysis using L-, F- and T-segments*. In: Preisach, B.; Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications*: *637-646.* Berlin, Heidelberg: Springer,

Köhler, R., Naumann, S. (2009). *A contribution to quantitative studies on the sentence level*. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 34-57*. Lüdenscheid: RAM-Verlag.

Köhler, R., Naumann, S. (2010). A syntagmatic approach to automatic text clas sification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, P.; Kelih, E.; Mačutek, J. (eds.), *Text and Language: 81-89.* Wien: Praesens.

Mačutek, J. (2009). Motif richness. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM-Verlag.

Mačutek, J., Mikros, G. (2015). Menzerath-Altmann law for word length motifs. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 125-131.* Berlin/Boston: de Gryuter.

Milička, J. (2015). Is the distribution of L-motifs inherited from the word length distribution? In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 133-145.* Berlin/Boston: de Gryuter.

Osgood, C.E., Suci, G.J., Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana: Univ. Illinois Press.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The Lambda-structure of zexts.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and Codes of Text.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Mačutek, J.,Altmann, G. (2009). *Aspects of Word Frequencies.* Lüdenscheid: RAM-Verlag.

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives: 183-194.* Wien: Praesens.

Skorochod´ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten*. Bochum: Brockmeyer.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM-Verlag.

Zipf, G.K. (1935). *The Psycho-biology of language: an introduction to dynamic philology*. Boston: Houghtton Mifflin 1868[2].

Zörnig, P., Stachowski, K., Popescu, I.-I., Mosavi Miangah, T., Chen, R., Altmann, G. (2016). *Positional occurrences in texts: Weighted Consensus Strings.* Lüdenscheid: RAM-Verlag.

Quantitative Aspects of Hierarchical Motifs

Haitao Liu[*], *Yu Fang*

Abstract. Motif, put forward by Köhler, is useful in measuring sequential properties of linguistic units. Previous studies have shown its helpfulness in measuring words in linear order, but its usage in hierarchical order is ignored. In this study, motif is applied to hierarchical order of words, thus the hierarchical motif (HM) is formed. Reversed hierarchical motif (RHM) is also constructed by modifying motif's definition a little. Texts in English, Czech and Chinese are selected from PCEDT 2.0 and PKU Treebank as the focus of our analysis. HM and RHM of those texts are calculated and fitted to various distribution patterns. The results show that rank frequency distributions of HM and RHM do share some distribution patterns, while differences also exist and those differences have distinct forms among the three languages. This paper also shows that Zipf-Mandelbrot distribution can also be fitted into all HM and RHM, although significant differences in values of parameter a or parameter b between HM and RHM cannot be ignored.

Keywords: *motifs, hierarchical order, language types, Zipf-Mandelbrot distribution*

1. Introduction

Quantitative studies in linguistics are often concerned with units, properties and their relations (Tuzzi et al. 2009; Chen, Liu 2016; Buk, Rovenchak 2008; Liu, Huang 2012), but few of them have thus far paid attention to the sequential properties of those objects. Köhler and Naumann (2008: 637) claimed: "sequences in a text are organized in lawful patterns rather than chaotically or according to a uniform distribution". In this sense, sequences of those units also need special attention. Inspired by F-motif, put forward by Boroda (1982) for frequency studies in musical pieces, Köhler constructed a new unit called motif (originally named as segment) that was defined as "the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit" (2015: 108).

Measuring linguistic units from different aspects like their length, frequency and polysemy, Köhler further subdivided motif into L-motif, F-motif, P-motif, etc. The usefulness of this unit has been proved by several researchers, who have followed two major lines. Some researchers are concerned with the distribution

[*] Address correspondence to: Haitao Liu, Department of Linguistics, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com

patterns of motifs, especially on word level (Mačutek, Mikros 2015; Milička 2015). For example, Köhler (2006) tested the frequency distribution of the lengths of word L-motifs. He found that if length is measured in numbers of words, then Hyper-Pascal distribution can be fitted, and if it is measured in syllables, then truncated negative binomial distribution can be fitted. Some researchers expect such lawful patterns could be utilized for text classification or authorship attribution. Collecting a small text corpus consisting of 55 documents from five different text sorts, Köhler, Naumann (2010) first fitted the rank frequency distribution of L- and F-motifs of word length to Zipf-Mandelbrot law and reached good fitting results. Then nine attributes related to motifs and parameters of the Zipf-Mandelbrot law were formed, and a pairwise examination of those attributes revealed that some of them can be used for text classification. The capability of motifs in text genre classification has also been tested by Köhler, Naumann (2008), as they considered the L-motif TTR of some poems and pieces of prose by fitting the Menzerath-Altmann Law.

Those studies mainly focus on measuring words in linear order, while motifs in hierarchical order are ignored. Hierarchical order, according to Tesnière (1959), is one of the syntactic structures of human languages, with the other one being linear order. Hierarchical order not only reflects the significance of a word in a sentence, but also measures complexity of a sentence (Culicover 2013: 19). Jing and Liu (2015) provided two metrics - mean dependency distance and mean hierarchical distance - to calculate the syntactic complexity in linear and hierarchical order. Liu (2016) found that lawful distributions exist in frequencies of hierarchical order based on an investigation into Chinese, English and Czech dependency treebanks. These two studies could be regarded as pioneer attempts to explore hierarchical order and highlight the possibility and necessity of further investigation on this level.

For this reason, we will apply motif into hierarchical order and construct hierarchical motifs (HM), a notion introduced and employed here for the first time. Hierarchical order can be further classified into three categories: component, dependency and valence (Blidschun 2011: 29-31), and in this study, we only pay attention to dependency that emphases word relations. Syntactic dependency relations usually share two properties: a binary relation between two linguistic units with a unit as the governor and the other as dependent; a labeling link between the two units (Liu 2009: 256). Linguists often represent hierarchical order in tree diagrams and in a dependency tree. The root is often the finite verb which follows dependents of the root, and then further dependents of the dependents follow until all words in a sentences are included in a tree. Take *The nice teacher gives a book to me* as an example, whose hierarchical syntactic structure is shown in Figure 1.

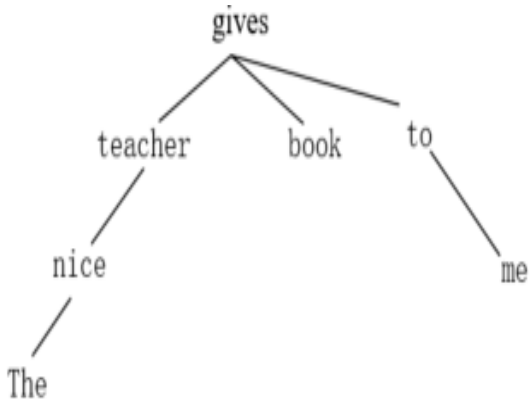


Figure 1. Hierarchical syntactic structure of *The nice teacher gives a book to me*

In Figure 1, the finite verb (*gives*) is on the top of the tree and every arc represents a dependency relation between two words in which a governor is followed by its dependents. In this sense, a hierarchical structure is formed. First, a root is given a certain number, and then every node gets its number that relates to its position in the tree. Thus HM can be defined as the longest continuous sequence of equal or increasing values of words' numbers in a sentence. "Variants of investigations based on motifs can be generated by changing the direction" (Köhler 2015: 108), but this type of study is, to date, rarely found. In this study, we tempt to replace "increasing" with "decreasing" in the definition and obtain a reversed hierarchical motif (RHM) that can be defined as the longest continuous sequence of equal or decreasing values of words' numbers in a sentence.

From observing previous studies, we can conclude that one advantageous property of a motif is that it displays lawful distributions in nearly all levels of linguistic units. Köhler (2015: 110) claimed "a rank-frequency distribution of the Zipf-Mandelbrot type". There are also other various distribution patterns like the Hyper-Pascal distribution, the Left-truncated negative binomial distribution and the Menzerath-Altmann distribution. All of these indicate the wide existence of distributions in motifs. As a result, we assume that the rank frequency distribution of HM and RHM can also be fitted by some models and may be not limited to Zipf-Mandelbrot distribution. Köhler (2015) suggested analyzing sentences from right to left when a language with syntactic left branching preference, which indicates a potential relationship between motifs and language types. Similarly, words can be organized in different orders to express the same meaning in left- or right-branching languages and thus result in different hierarchical syntactic structure. Consequently, distribution patterns of HM and RHM may differ in different language types. Here, since HM and RHM are formed differently, divergence may also occur in their distribution patterns. There is little research comparing the distribution patterns among languages, which seems to be another factor influencing results. In this study, PCEDT 2.0, a Czech-English parallel corpus, and PMT 1.0, a Chinese corpus, are selected as the data and we will keep the following three research questions in mind:

(a) Will rank frequency distributions of HM and RHM show the same distribution

patterns?

(b) Are the distribution patterns different among languages? In other words, do linguistic types have an impact on those distribution patterns?

(c) Can the Zipf-Mandelbrot distribution be fitted to all HM and RHM? If the answer is yes, are there any differences among parameters of HM and RHM in the three languages? And is it possible to use such difference to distinguish language types?

2. Materials and Method

PCEDT 2.0 and PMT 1.0 are chosen as the data for consideration in this study. PCEDT 2.0 is a manually parsed Czech-English parallel corpus that is 1:1 sentence-aligned, in which the English part contains the entire Penn Treebank-Wall Street Journal Section (Linguistic Data Consortium, 1999) and the Czech part consists of Czech translations of all of the Penn Treebank-WSJ texts. The annotation of the corpus contains an analytical layer and a tectogrammatical layer. In the English part, manual tectogrammatical annotation was built after transforming the original phrase-structure annotation of the Penn Treebank and was put into surface dependency representations automatically. In the Czech part, sentences were automatically morphologically annotated and parsed into surface-syntax dependency trees[1]. In this study, only the analytical layer is used. The whole corpus has over 1.2 million words for each language and, to reduce the workload while still reach our objectives, we selected 20 texts with 350 to 500 running words in the English part and find their counterparts in the Czech part. PMT 1.0, a multi-view Chinese Treebank, contains all the articles of People's Daily newspaper from January 1st to January 10th, 1998 (Qiu et al. 2013: 263) and about 14,463 sentences and 366,000 words are included in this Treebank. Sentences in this corpus are annotated into dependency trees automatically using a statistical dependency parser and then checked on a visualized annotation platform by annotators. With the same standard, we also choose 20 texts in PKU Treebank. These two corpora both contain news reports annotated into dependency trees, which ensures comparability of the study. Table 1 shows information about all of the texts that were selected.

[1] More information about this treebank can be accessed from http://ufal.mff.cuni.cz/pcedt2.0/.

Table 1
Number of Running Word in All Selected Texts

Text	English	Czech	Chinese
1	398	378	484
2	514	508	502
3	428	442	408
4	594	574	600
5	350	326	368
6	421	380	414
7	484	412	463
8	560	595	419
9	424	412	492
10	508	529	433
11	518	527	454
12	356	354	467
13	514	470	367
14	595	599	356
15	448	438	389
16	358	355	380
17	588	560	565
18	515	468	359
19	542	540	367
20	570	559	375

As we can see, the number of running words in Czech translations is nearly equal to that in English texts and falls in the range of 350 to 500, with the exception of Text 5. Assisted by surface dependency information in the corpora, here, we label the root of a syntactic tree as 1 and designate each node a number on the basis of its hierarchical order. The sentence *ShareData has about 4.1 million common shares outstanding* is taken from one text of PCEDT 2.0 as an example and its HM and RHM, as shown in Figure 2, are (2) (1-2-4) (4) (3-2) (3) and (2-1) (2) (4-4-3-2) (3) respectively.

Figure 2. Hierarchical syntactic structure of *ShareData has about 4.1 million common shares outstanding*

In this way, each word in a sentence gets a unique number. Then we calculate their motifs according to the definition we gave above. Saving the result of each text as a profile, we finally get the rank frequency distribution of HM and RHM in all 60 texts.

After all data are collected, the next step is to find proper distribution patterns for them, so Altmann-Fitter 3.1[1] is introduced. Altmann-Fitter is a software designed for the iterative fitting of univariate discrete probability distributions to frequency data. There are four modes of working, which are Selected Fitting, Automatic Fitting, Special Fitting and Batch Fitting. In this study, we will only use two of them. One is Automatic Fitting, which applies more than 200 probability distributions offered in this programme to the data. The result can be evaluated by values of X^2, $P(X^2)$, C and R^2. The other mode is Batch Fitting, which is quite useful provided a distribution hypothesis is tested on a large data set. Since the Zipf-Mandelbrot distribution is found in almost all rank frequency distributions, we also assume that this model can fit distributions of HM and RHM. As a result, this distribution is tested in Batch Fitting.

The Zipf–Mandelbrot law is a discrete probability distribution, named after the mathematician, Benoit Mandelbrot, which is a more generalized form of Zipf's law. We used the Zipf-Mandelbrot distribution as it is exposed in Altmann-Fitter (1994: 92), also see Wimmer, Altmann (1999: 666):

$$P_x = \frac{(b+x)^{-a}}{F(n)}, x = 1,2,\dots,n, a, b > 0, n \in \mathrm{N},\ \mathrm{F(n)} = \sum_{i=1}^{n}(b+i)^{-a}.$$

Here x refers to the rank of the data, P_x is the frequency of the data, a and b are two parameters of which parameter a depends on the number of units with high frequencies and parameter b is related to the total word number. Thus, significant tests of parameter a or b may reveal features of language types.

3. Results and Discussion

In this section, we will discuss the three research questions listed above. In section 3.1, we will present the fitting results of HM and RHM, and distributions of HM and RHM will be compared and contrasted to find whether they follow the same distribution patterns and whether the results differ among languages. In section 3.2, commons and differences of the fitting results of HM and RHM will be compared and contrasted in each language separately. In section 3.3, Zipf-Mandelbrot distribution will be fitted into all HM and RHM and significant tests will be carried out for the two parameters in this distribution.

[1] It can be accessed from http://www.ram-verlag.eu/software-neu/software/.

3.1. Distributions of HM and RHM

Using Automatic Fitting in the Altmann-Fitter for each text, we find that distributions of HM and RHM fit into several proper patterns, though some patterns are shared and some are different. We rank distribution patterns according to values of $P(X^2)$ in a descending order and those values above 0.05 are reserved because this indicates a large possibility of actual values equaling to empirical values. R^2 is also taken into consideration. Though R^2 is defined for linear functions only, it may also be used in connection with non-linear functions. In this study, those distributions with R^2 larger than 0.8 are left. Since each text has different fitting results, only those distributions that are shared by all texts in one language will be analyzed. Table 2, 3 and 4 show distribution patterns of HM and RHSM in English, Czech and Chinese respectively. We need to notice that no suitable distributions can be found in Text 13 of PMT 1.0 if $R^2 > 0.8$ is set as the criteria, so in this text, we include all distributions if their $P(X^2)$ is larger than 0.05.

Table 2
Distributions of HM and RHM in English

HM	Mean R^2	RHM	Mean R^2
Consul-Mittal-binomial with 3 parameters	0.9555	Consul-Mittal-binomial with 3 parameters	0.9770
Right truncated modified Zipf-Alekseev	0.9429	Right truncated modified Zipf-Alekseev	0.9508
Right truncated Waring	0.9403	Right truncated Waring	0.9481
Right truncated zeta	0.9469	Right truncated zeta	0.9128
Zipf-Mandelbrot	0.9381	Zipf-Mandelbrot	0.9626

Table 3
Distributions of HM and RHM in Czech

HM	Mean R^2	RHM	Mean R^2
Consul-Mittal-binomial with 3 parameters	0.9283	Consul-Mittal-binomial with 3 parameters	0.9753
Right truncated modified Zipf-Alekseev	0.9618	Right truncated modified Kemp2	0.9252
Right truncated negative binomial	0.9309	Right truncated modified Zipf-Alekseev	0.9547
Right truncated zeta	0.9489	Zipf-Mandelbrot	0.9564
Zipf-Mandelbrot	0.9586		

Table 4
Distributions of HM and RHM in Chinese

HM	Mean R^2	RHM	Mean R^2
Consul-Mittal-binomial with 3 parameters	0.9746	Right truncated modified Zipf-Alekseev	0.8970
Right truncated modified Zipf-Alekseev	0.9586	Zipf-Mandelbrot	0.8649
Right truncated Waring	0.9696		
Zipf-Mandelbrot	0.9635		

From these three tables, it is not difficult to see that distributions of HM and RHM do share some commonalities: firstly, the Zipf-Mandelbrot distribution are successfully fitted to all texts, which confirms Köhler's assumption about rank frequency distribution; secondly, Right truncated modified Zipf-Alekseev is another distribution that is successfully fitted to all texts; thirdly, though Consul-Mittal-binomial with 3 parameters cannot be applied to several texts in the distribution of RHM in Chinese, most texts do follow this distribution. So two conclusions can be made: distributions of HM follow certain lawful patterns, as do distributions of RHM. In other words, the notion of motif could not only apply to basic linguistic units like morphs, words or phrases, but also to more complicated ones like the hierarchical order of sentences. It is also shown that variants of motifs, here as HM and RHM, share some similarities.

Despite those commonalities, differences cannot be ignored. If distributions of HM are treated as a whole in all three languages, which is also the case with RHM, it is revealed that Right truncated negative binomial appears in distribution patterns of HM, especially in HM of all Czech texts, but is not useful in RHM. The Right truncated modified Kemp2, however, only has its appearance in RHM but not in HM. Parallel materials are chosen in English and Czech and all three languages share the same text types, but distributions of HM and RHM yield those differences; the reasons for this are worth exploring. Noticing that differences have various forms in different languages, we may contribute the differences to language types, thus we need to investigate this matter separately in different languages.

3.2. Comparing distribution differences of HM and RHM in three languages

In English, distributions of HM and RHM in all 20 selected texts follow the same patterns. To get a deeper understanding, rank frequency distribution of HM and RHM in one text is plotted and shown in Figure 3.

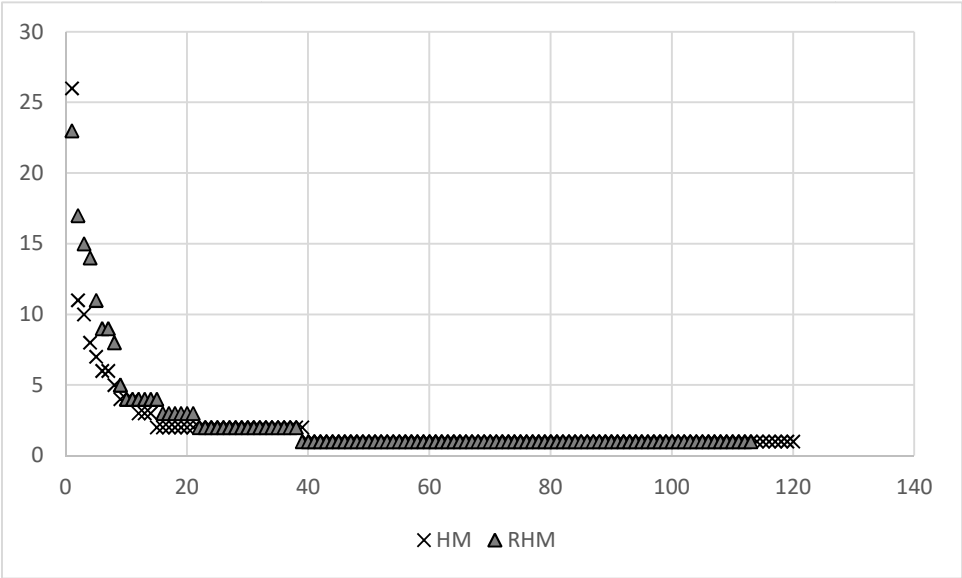


Figure 3. Rank frequency distribution of HM and RHM in Text 17 of English

Figure 3 tells us that rank frequency distributions of HM and RHM in this text almost coincide. Divergence does occur at the beginning: frequencies of HM are lower than those of RHM from Rank 2 to Rank 8. From Rank 9 to Rank 21, though frequencies of HM are still a little bit lower, the gap always stands at 1. To find out reasons for this divergence, one sentence was taken out from this text for further analysis. The dependency tree of this sentence is shown in Figure 4.

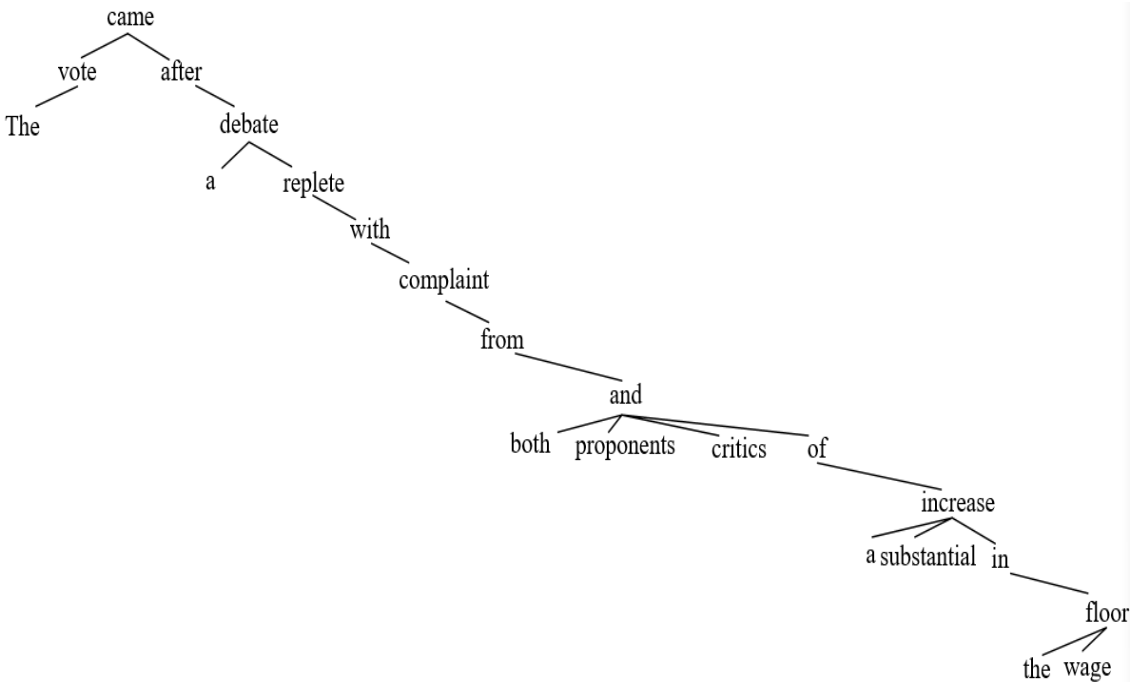


Figure 4 Hierarchical syntactic structure of *The vote came after a debate replete with complaint from both proponents and critics of a substantial increase in the wage floor*

In this sentence, HM should be like (3), (2), (1-2-4), (3-4-5-6-7-9-9), (8-9-9-11-11), (10-11-13-13), (12), and RHM is (3-2-1), (2), (4-3), (4), (5), (6), (7), (9-9-8), (9-9), (11-11-10), (11), (13-13-12). The longest length of HM is 7, followed by two HMs whose lengths are 5 and 4 respectively. The longest length of RHM is only 3, and the length of most RHMs are 1 or 2. The shorter the RHM is, the more possibility it will repeat. So at the beginning, frequencies of HM are lower than those of RHM. Probing into dependency tree of this sentence deeply, we can find that most words occur on the right of the finite verb, which is a typical tree pattern of a right-branching language, and the linear order conforms to the hierarchical order. These two reasons explain why length of HM is longer and why frequencies of RHM are higher.

In Czech, five distribution patterns can be fitted into HM of the 20 texts, while only four into RHM of those texts, among which, three distribution patterns are shared. The Right truncated negative binomial and Right truncated zeta cannot be applied to RHM of some texts, while the Right truncated modified Kemp2 cannot be applied to HM. Here one text whose RHM cannot be fitted by the Right truncated negative binomial was chosen for further analysis.

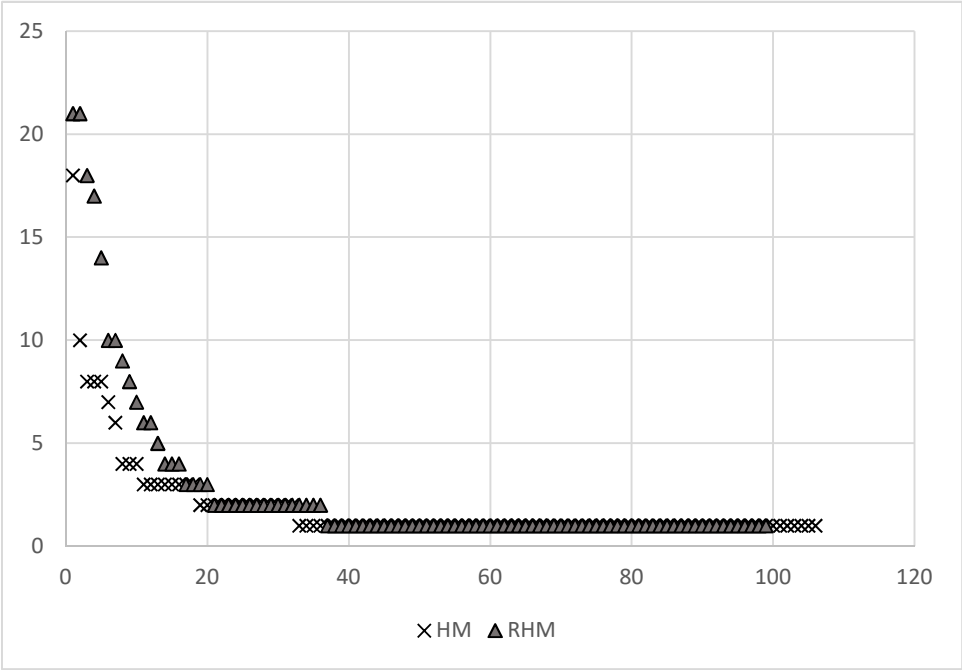


Figure 5 Rank frequency distribution of HM and RHM in Text 11 of Czech

From Figure 5, we can conclude that beginning from Rank 1 to Rank 17, frequencies of RHM in Text 11 are larger than those of HM and sometimes the gap is even larger than 10, like Rank 2 and Rank 3. In Rank 19 and Rank 20 as well as Rank 33 to Rank 36, frequencies of RHM are also larger though the gap is not as wide as before. HM and RHM from Rank 1 to Rank 17 and their frequencies are listed in Table 5

Table 5
Frequencies of HM and RHM in Text 11 of English

Rank	HM	Freq.	RHM	Freq.
1	1	18	4	21
2	1-3	10	5	21
3	2	8	3	18
4	2-3	8	6	17
5	5	8	3-2	14
6	6	7	2	10
7	3	6	7	10
8	3-4-5	4	9	9
9	4-5	4	5-4	8
10	7	4	8	7
11	2-3-5	3	10	6
12	3-4	3	4-3	6
13	3-5	3	3-1	5
14	4	3	11	4
15	4-4	3	2-1	4
16	4-5-6	3	7-6	4
17	4-6-7	3	12-1	3

It clearly shows that length of most RHMs is 1 and only seven RHMs is 2, but length of ten HMs is equal to or larger than 2. It is those RHMs with length of 1 that contribute to high frequencies at the beginning. From this aspect, Czech and English are very similar to each other. At the same time, we should notice that divergence between HM and RHM is wider in Czech than in English, which implies differences between the two languages. Here we choose one sentence from Text 11: *Tento krok je naplánován, aby odradil nepřátelské pokusy o převzetí od dvou evropských lodních přepravních koncernů, společnosti Stena Holding AG a Tiphook PLC* (*The move is designed to ward off a hostile takeover attempt by two European shipping concerns, Stena Holding AG and Tiphook PLC*).

The HM of this sentence is (3), (2), (1-2-3), (2-3-5), (4-5-7-10-10-10-10), (9), (8-9-12-12), (11), (10-11-12) and RHM is (3-2-1), (2), (3-2), (3), (5-4), (5), (6), (7), (10-10-10-10-9-8), (9), (12-12-11-10), (11), (12). The average length of HM and RHM in this sentence is longer than that in an English sentence, and many neighboring words share the same number. This, on the one hand, indicates Czech has freer word order, and on the other hand, manifests Czech is an inflectional language.

Finally, Chinese texts will be scrutinized. Four kinds of distribution patterns are modeled into frequencies of HM, while only two patterns could be found in frequencies of RHM. Consul-Mittal-binomial with 3 parameters and Right truncated Waring are unique to the rank frequency distribution of HM. With the same method, we select one text and its frequencies of HM and RHM are compared in Figure 7.

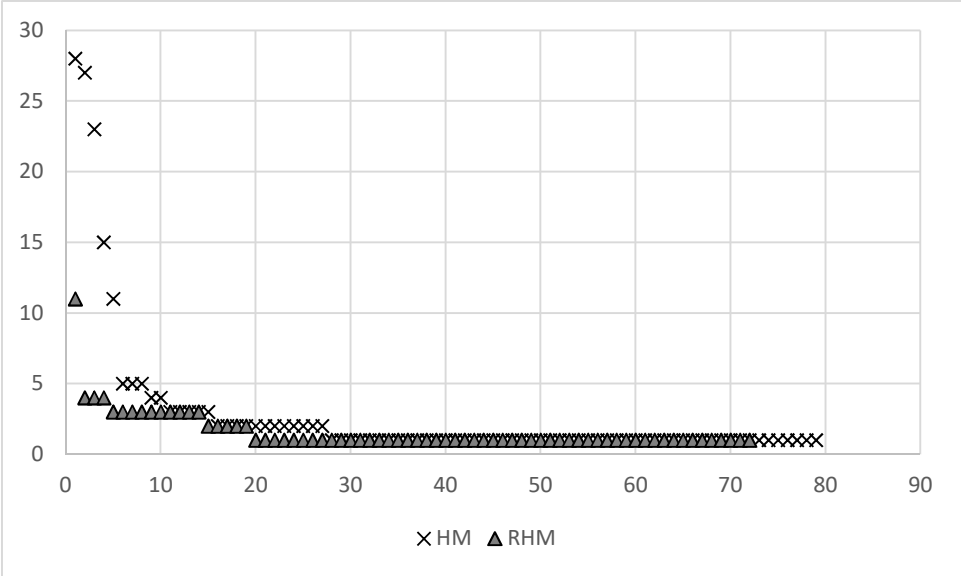


Figure 7 Rank frequency distribution of HM and RHM in Text 6 of Chinese

Compared with RHM of other texts in English and Czech, frequencies of RHM in this text are much lower. Observing directly from this figure, we can find that even the frequency of the highest rank is just a little above 10 and this feature is not limited to Text 6 but is shared by nearly all 20 Chinese texts. Obviously, it is an unfavorable condition in a power law distribution, thus only two distributions can be applied. We also select one sentence from this text: "两国建交为双方未来的合作打下了坚实的基础 (Pinyin: *Liang guo jian jiao wei shuang fang wei lai de he zuo da xia le jian shi de ji chu*; English: *The establishment of diplomatic relation between the two countries lays a solid foundation for future cooperation*)". Figure 8 shows the dependency tree of this sentence.

Figure 8 Hierarchical syntactic structure of a Chinese sentence

Many words are on the left of the finite verb, which is contrary to word ordering in English and indicates Chinese is a left-branching language. Moreover, modifiers are usually before center words, which increases the possibility of long

RHMs. The HM of this sentence is (4), (3), (2-2-5-5-4), (3), (1-2-4), (3), (2) and its RHM is (4-3-2-2), (5-5-4-3-1), (4), (4-3-2). This gives us a direct impression that length of the HM is shorter than that of the RHM. To get a more comprehensive understanding, frequencies of the HM and RHM ranked within the Top 10 were checked, and the result is shown in Table 6. It is obviously that only three of the RHMs' lengths in this table is 1, thus frequencies of the RHM are low.

Table 6
Frequencies of HM and RHM in Text 6 of Chinese

Rank	HM	Freq.	RHM	Freq.
1	4	28	3	11
2	3	27	2-1	4
3	5	23	4	4
4	2	15	4-3-2-2	4
5	6	11	2	3
6	3-3	5	2-2	3
7	3-3-4	5	2-2-2	3
8	7	5	3-2	3
9	2-2	4	4-2	3
10	5-5	4	4-3	3

3.3. Significance tests of parameters in Zipf-Mandelbrot distribution

The above result shows that rank frequency distributions of the HM and RHM in all three languages can be fitted using the Zipf-Mandelbrot distribution. However, at the same time, differences exist between their frequencies, so we expect significant difference in parameters a or/and parameter b.

In addition, the data above also shows different patterns have been found between distributions of the HM and RHM. In other words, there must be divergences between frequencies of HM and those of RHM. As a result, it is reasonable to assume a significant difference exists in a or b or both between HM and RHM. In this sense, the respective hypotheses are:

H_0: Values of parameters in HM and RHM are equal.

H_1: Values of parameters in HM and RHM are not equal.

One-sample Kolmogorov-Smirnov test shows the data are normally distributed: $p_{a(HM)} = 0.432 > 0.05$, $p_{a(RHM)} = 0.056 > 0.05$. Then, a paired sample t-test was conducted and a significant difference between parameter a of HM and RHM, $t_{(59)} = 2.466$, $p = 0.017 < 0.05$ can be seen. There is also a significant difference between parameter b of HM and RHM, $t_{(59)} = -2.884$, $p = 0.005 < 0.05$. Such results provide further support for the findings discussed above. Here, we still want to know whether this significant difference exists in the three languages separately, so a paired sample t-test needed to be carried out for each language. Table 5 shows results of the tests.

Table 5
Results of paired sample t-test of parameter a and b

	parameter a	parameter b
English	p = 0.248 > 0.05	p = 0.004 <0.05
Czech	p = 0.000 < 0.05	p = 0.439 > 0.05
Chinese	p = 0.000 < 0.05	p = 0.031 <0.05

Conclusions are thus reached: in English and Czech, similarities can be seen between parameter a and parameter b respectively, while in Chinese, significant difference is shown both in parameter a and parameter b. To this point, we should resort to implications of these two parameters: parameter a depends on the number of units with high frequencies and parameter b is related to the total word number. Therefore, the HM and RHM with high frequencies in English texts have little difference, while the opposite is true in Czech and Chinese. Word number shows its impact on parameter b of HM and RHM in English and Chinese, but not in Czech.

Such a result may be contributed to language types. Köhler and Naumann (2010) discovered the potential of parameter a and b of the Zipf-Mandelbrot distribution which could separate juridical texts from other texts like poems, narratives, scientific and journalistic texts. The powerfulness of the two parameters in text classification may extend to language classification. Here, values of parameter a are used as axis x and values of parameter b are used as axis y.

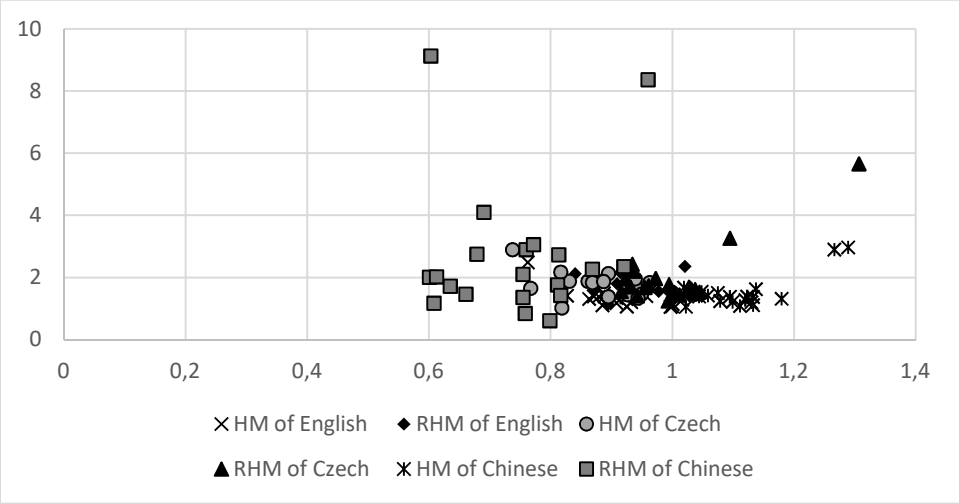


Figure 7 Parameters a and b of the Z-M distribution in HM and RHM

As we can see from Figure 7, the HM of Chinese and the RHM of Chinese are separated from others, which indicate the uniqueness of Chinese from other two languages. In other words, English and Czech stand more closely to each other. Moreover, there is no overlap between HM of Chinese and its RHM, which

reinforces the result that the differences between distribution of HM and RHM in Chinese are larger than that in English and Czech.

To explain this result, we may resort to linguistic typology, "a subject about language classification" (Liu, Li: 3458). English and Czech both belong to Indo-European languages whose features include rich inflection and free word order. This, to some extent, may explain similarities between the distributions of the HM and RHM for the two languages. Moreover, measured by dependency relations between two words, English and Czech show neither a head-final tendency nor a head-initial tendency, although the number of head-final constructions is slightly more than that of head-initial constructions in English while the opposite is true in Czech (Liu 2010: 1571). That is to say, in English or Czech grammatical units, heads could precede or follow dependents. This may be another reason for their similarities. However, since English and Czech belong to two sub-groups, differences thus occur. English, belonging to the Germanic subgroup, "is becoming more isolating and littered with inflectional changes" (Liu, Li 2010: 3462). Specifically, function words and a relatively fixed SVO syntax are used for meaning expression. Its syntactic features are characterized by a topic-comment structure, and to ensure this structure, passive construction and a dummy subject must be adopted such as *it* or *there*. In this way, there is fair amount of consensus that the syntactic structure of English is typically right-branching (Berg 2003; Levy, Manning 2003). Czech is a West Slavic language and a highly inflected fusional language. Because of this inflection, Czech word order is very flexible and words may be transposed to change emphasis or form questions.

Chinese does share some similarities with English for they are both isolating languages that rely on more rigid word order to encode functional relations. However, as a branch of the Sino-Tibetan language family, Mandarin Chinese is a kind of analytic language with very few grammatical inflections, that is to say, "Chinese makes less use of function words and morphology than English" (Levy, Manning 2003:439). As a result, it has a relatively strict word order. Meanwhile, it obeys SOV sentence structure which makes frequent use of the topic comment construction to form sentences, however, dummy subjects are rare in Chinese, so subjects of some sentences can be very long. Postpositive attributives are also not often used for head-final constructions and are preferred in Chinese sentences (Liu 2010). All of these lead to Chinese being a left-branching language.

Significant differences have been found in parameter a and b of the Zipf-Mandelbrot distribution, which explains the distribution divergence of HM and RHM. At the same time, it cannot be ignored that such divergence has different forms in the three languages and this phenomenon may be explained by types of language, which is clearly shown in the scatter diagram of the two parameters.

4. Conclusions

Based on the above analyses, we came to conclusions corresponding to the three research questions posed in the introduction section:

(1) Similarities and differences can be found among rank frequency distributions of the HM and RHM. On the one hand, rank frequency distributions of the HM and RHM share two distribution patterns: the Zipf-Mandelbrot distribution and the Right truncated modified Zipf-Alekseev. Moreover, the Consul-Mittal-binomial with 3 parameters yields excellent results in rank frequencies of most texts except the RHM in some Chinese texts. On the other hand, the right truncated negative binomial can only be applied to the HM but not to RHM; and the Right truncated modified Kemp2 only appears in the RHM but not in HM.

(2) Distribution patterns differ among the three languages which can be attributed to language types. In English, distributions of the HM and RHM in all 20 selected texts follow the same five patterns. In Czech, five patterns could be fitted into the HM and four could be fitted into the RHM, but only three of them are shared. The Right truncated negative binomial and the Right truncated zeta cannot be applied to the RHM, while the Right truncated modified Kemp2 cannot be applied to the HM. In Chinese, the HM can be fitted by four distribution patterns, while the RHM can be only fitted by two patterns and both of them appear in distributions of the HM.

(3) The Zipf-Mandelbrot distribution can be fitted to all HM and RHM although significant differences cannot be ignored in values of parameter a or parameter b between the HM and RHM when taking the three languages as a whole. If these values are considered in each language separately, divergence will occur: in English, significant differences can be found in parameter b, but not in parameter a. In Czech, the opposite is true. In Chinese, both parameters show significant differences. We attribute such differences to language types and by exploring the relationship between parameter a and parameter b, which ultimately separates the HM and RHM of Chinese from others, we further prove our assumption.

This study, by applying motif to hierarchical structure of sentences, on the one hand, shows the powerfulness of motif on sentence level; on the other hand, explores rank frequency distribution of this new unit. Similarities and differences between distributions patterns of the HM and RHM are fully explored, which may be attributed to types of language. However, why and how different types of languages show different distribution patterns are not fully explained, which needs our attention in further studies.

Acknowledgments

This work is partly supported by the National Social Science Foundation of China (Grant No. 11&ZD188).

References

Altmann-FITTER (1994). *Begleitbuch zu* Altmann-FITTER. *Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen*. Lüdenscheid: RAM–Verlag.

Berg, T. (2003). Right-branching in English derivational morphology. English Language and Linguistics, 7(02), 279-307.

Blidschun, Claudia (2011) *Systemstrukturen des Deutschen*. Würzburg: Lehrstuhl für deutsche Sprachwissenschaft.

Boroda, Moisei (1982): Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Jurij K./Boroda, Moisei G./Nadarejšvili, Isabela Š. [eds.]: *Sprache, Text, Kunst. Quantitative Analysen:* 231-262. Bochum: Brockmeyer,

Buk, S.,, Rovenchak, A. (2008). Menzerath–Altmann Law for Syntactic Structures in Ukrainian. *Glottotheory* 1(1), 10-17.

Chen, H., Liu, H. (2016). How to Measure Word Length in Spoken and Written Chinese. *Journal of Quantitative Linguistics* 23(1), 5-29.

Jing, Y., Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. *Depling 2015*, 161-170.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. *Favete linguis. Studies in honour of Victor Krupa*, 145-152.

Köhler, R. (2015). Linguistic Motifs. *Sequences in Language and Text*, 107-129.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F-and T-segments. In: *Data Analysis, Machine Learning and Applications:* 637-645. Berlin- Heidelberg: Springer.

Köhler, R., Naumann, S. (2010). A syntagmatic approach to automatic text classification. statistical properties of F-and L-motifs as text characteristics. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives*, 82-89. Wien: Prasens.

Levy, R., Manning, C. (2003). Is it harder to parse Chinese, or the Chinese Treebank? In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume* 1: 439-446. Association for Computational Linguistics.

Liu, Haitao (2009). Probability Distribution of Dependencies based on Chinese Dependency Treebank. *Journal of Quantitative Linguistics* 16(3), 256–273.

Liu, Haitao (2010). Dependency direction as a means of word-order typology: a method based on dependency treebanks. *Lingua* 120(6), 1567-1578.

Liu, Haitao (2016). Probability distribution of syntactic hierarchical structure in human languages. Manuscript.

Liu, Haitao, Li Wenwen (2010). Language Clusters based on Linguistic Complex Networks. *Chinese Science Bulletin* 55(30), 3458-3465.

Liu, Haitao, Huang Wei (2012). Quantitative Linguistics: State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Sciences)* 42(2), 178-192.

Mačutek, J., Mikros, G. K. (2015). Menzerath-Altmann Law for Word Length Motifs. *Sequences in Language and Text*, 107-129.

Milička, J. (2015). Is the Distribution of L-Motifs Inherited from the Word Lengths Distribution?. *Sequences in Language and Text*, 155-167.

Qiu, Likun, Yue Zhang, Peng Jin & Houfeng Wang. (2014) Multi-view Chinese Treebanking. *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. Dublin, Ireland, August.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris, Klincksieck.

Tuzzi, A., Popescu, I. I., Altmann, G. (2009). Zipf's laws in Italian Texts. *Journal of Quantitative Linguistics* 16(4), 354-367.

Wimmer, G. Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen, Stamm.

Key Length Motifs in Czech and Arabic Texts

Jiří Milička

Institute of Comparative Linguistics
Faculty of Arts, Charles University, Prague
jiri@milicka.cz

Abstract. Length motifs (L-motifs) are defined as sequences of words whose lengths are monotonously increasing. In recent years, L-motifs have attracted well-deserved attention as they provide a new view of texts and their syntagmatic properties and nested structures. This study examines the key L-motifs, i.e. motifs that are over-represented in texts and negative key L-motifs that are underrepresented in texts. The data reveal motifs that are typical for Czech texts, motifs that are typical for Arabic texts, and motifs that are typical for both Czech and Arabic texts – their existence suggests that there are new general language-independent patterns waiting to be explored.

Keywords: *L-motifs, Czech, Arabic, language-independent patterns, text structure, keyness, sequences of units,*

1. Introduction

L-motifs have been defined as "[...] a continuous series of equal or increasing length values" (Köhler 2015: 90) and were introduced "in order to find a method which can give information about the sequential organisation of a text with respect to any linguistic unit and to any of its properties – without relying on a specific linguistic approach or grammar" (ibid: 89).

A typical example is the following Hungarian sentence (Köhler, 2008a: 416):

> *Azon a tájon, ahol most Budapest fekszik, már nagyon régen laknak emberek.*

The syllabic lengths of the words are:

> 2, 1, 2, 2, 1, 3, 2, 1, 2, 2, 2, 3

This sequence can be segmented into these five L-motifs:

> (2) (1 2 2) (1 3) (2) (1 2 2 2 3)

The segmentation can be applied iteratively, i.e. L-motifs of the L-motifs can be obtained (so-called LL-motifs). For example, the lengths of the L-motifs in the Hungarian sentence are:

1, 3, 2, 1, 5

Thus, the LL-motifs of the Hungarian sentence would be:

(1 3) (2) (1 5)

and the LLL- motifs would be

(2) (1 2)

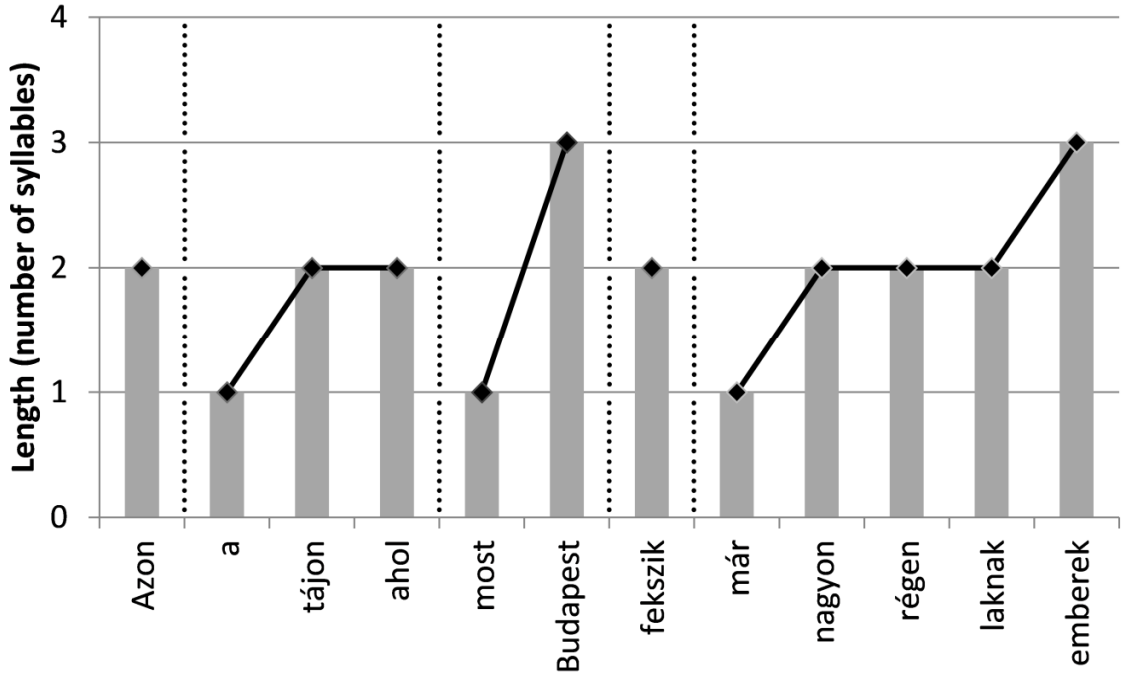


Figure 1. Example of segmentation of a Hungarian sentence into L-motifs.

The general properties of motifs in texts have already been examined: rank-frequency relation (Köhler 2008a; Köhler, Naumann 2009; Mačutek 2009; Sanada 2010), type-token relation (Köhler 2008b), Menzerath-Altmann Law (Köhler 2008b; Mačutek, Mikros 2015), length distribution (Köhler 2006; Köhler, Naumann 2008; Sanada 2010) etc., and there have been attempts to use them for some NLP tasks (Köhler – Naumann 2010) although little attention has been paid to individual motifs – which motifs are typical for a given text and why.

This study examines several Czech and Arabic texts from this point of view and is a follow-up to the research described in Milička (2015), specifically inspired by the chart depicted in Fig.2. The figure shows the frequency of the L-motif (1 1 2 2 2) in an actual Czech text and the distribution of the frequencies in random pseudo-texts that were created by one million randomizations of the original Czech text. This randomization enables us to determine which motifs in the texts are overrepresented and which are underrepresented; we will examine these key motifs and the negative key motifs more closely.

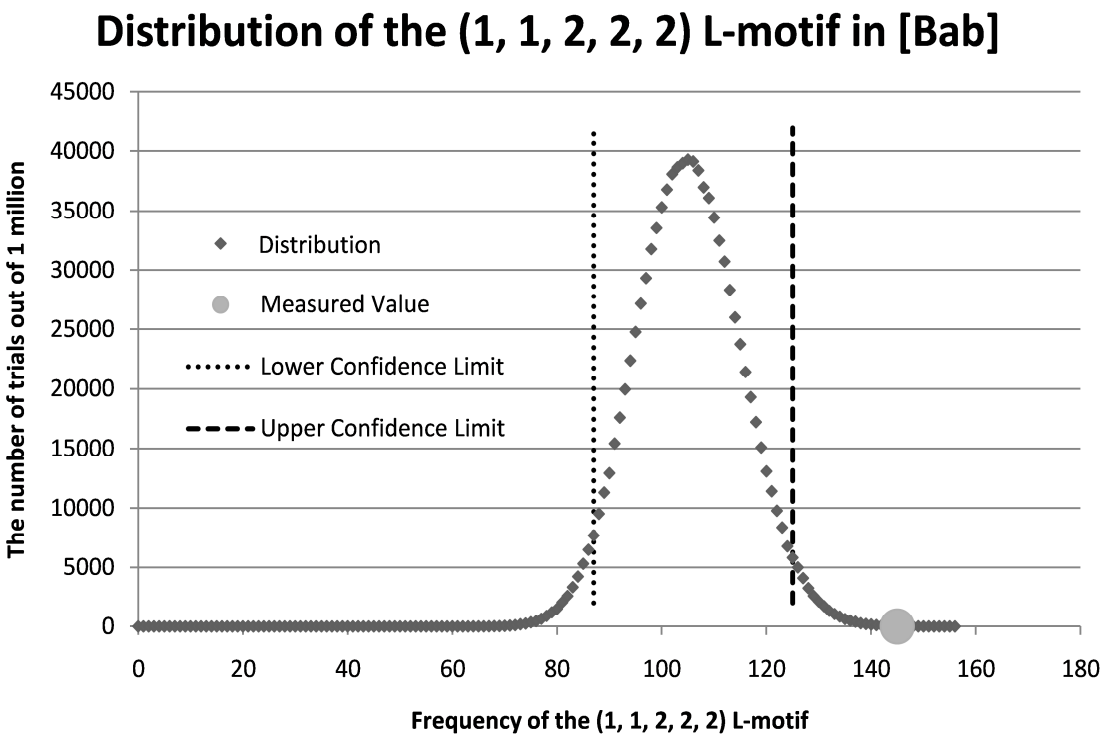


Figure 2. Distribution of the L-motif type (1 1 2 2 2) in one million pseudo-texts (randomized *Babička* by Božena Němcová) vs. the frequency of the L-motif in the actual text

2. Method

There is no general agreement on how keyness should be calculated. Our approach is inspired by the Minimal Ratio (Milička 2012) – the ratio between the measured value and the limit of the confidence interval that is closer to the measured value. In our case, confidence intervals are replaced by the α-th and $(1000-\alpha)$-th permille of the distribution of the motif in multiple randomized pseudo-texts. For our study, each text was randomized (i.e. words were resampled with replacement) a million times.

The definition of the Minimal Ratio (MR) of the motif *m* (which is represented in the examined text by f_m instances) at level α is as follows:

$$f_m < Pr(\alpha) \Rightarrow MR(\alpha) = \frac{f_\mathrm{m} + 1}{Pr(\alpha) + 1}$$

$$Pr(\alpha) \leq f_m \leq Pr(1000-\alpha) \Rightarrow MR(\alpha) = 1$$

$$f_\mathrm{m} > Pr(1000-\alpha) \Rightarrow MR(\alpha) = \frac{f_m + 1}{Pr(1000-\alpha) + 1}$$

As the value or the α-th permille may be zero, 1 is added to both f_{m} and Pr (this value is an arbitrarily chosen one). The formula implies that $MR(\alpha) < 1$ stands for negative key motifs or motifs that are represented in the text less than in the random model; $MR(\alpha) = 1$ means that the frequency of the given motif in the text is roughly as could be expected according to the random model, and $MR(\alpha) > 1$ means that the motif is overrepresented in the text. For every result reported in this study, $\alpha = 25$.

For example, the frequency of the motif (1 1 2 2 2) in *Babička* by Božena Němcová (as in Fig. 2) is 145 while the 975th permille is equal to 125, thus the Minimal Ratio is calculated as:

$$MR(25) = \frac{f_m + 1}{Pr(1000 - \alpha) + 1} = \frac{146}{126} = 1.16$$

3. Data

The key motifs were retrieved from four Czech and three Arabic texts.

Table 1
List of texts

Author	**Title**	**Century**	**Language**	**# of Tokens**
Milan Kundera	Žert	20	Czech	88435
Pavel Kohout	Katyně	20	Czech	99808
Božena Němcová	Babička	19	Czech	70140
Karel Čapek	Válka s mloky	20	Czech	62477
al-Hazimi al-Hamadani	Al-I‘tibar fi n-nasix wa-l-mansux	12	Arabic	71482
ibn as-Sallah	Ma‘rifatu anwā‘i ‘ulūmi l-hadith	13	Arabic	54915
ibn abi Zamanayn	Usulu s-sunna	11	Arabic	18607

The number of syllables in the Arabic words was determined according to traditional word segmentation. In the Czech texts, zero syllabic words (e.g. *s*, *z*, *v*, *k*) were merged with the following words and thus omitted, which is in accordance with the conclusion in Antić et al. (2006), and which makes the study compatible with other studies in this field (e.g. Köhler 2006b, Milička 2015).

All results and raw datasets are available on http://milicka.cz/kestazeni/keymotifs.zip. In addition to the data presented in the following sections, the lists

of LLLL-motifs and LLLLL-motifs can be found there along with the lists of motifs measured on longer collections of texts (compendia of one author), which were not included here as the paper does not deal with intertextual properties.

4. Results – Key L-Motifs

Czech texts

Table 2

List of key L-motifs in Czech texts. Only motifs that occur more than five times in the original text are included. The motifs that are overrepresented in more than one text (duplicates) are highlighted in grey.

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
1111223	40	1.46	23333	16	1.31	11111112	9	2.00	37	6	1.75
2444	6	1.40	11122233	8	1.29	345	8	1.80	**1111123**	31	1.60
11222233	7	1.33	1125	35	1.29	**222225**	6	1.40	11111122	8	1.29
2245	7	1.33	**11235**	9	1.25	112344	6	1.40	11122223	11	1.20
1111233	22	1.28	**11222**	145	1.17	**11235**	14	1.36	13444	6	1.17
11123	220	1.27	111111223	6	1.17	5	9	1.25	**222225**	7	1.14
11145	9	1.25	11225	13	1.17	34	140	1.22	**111123**	58	1.13
36	10	1.22	245	7	1.14	111113	40	1.21	1112223	23	1.09
113333	15	1.14	1123	402	1.12	**112222**	38	1.15	**112222**	47	1.07
11124	93	1.13	1122224	10	1.10	**2444**	7	1.14	1112233	16	1.06
111333	27	1.12	11111222	11	1.09	**344**	15	1.14	**11122**	116	1.05
111223	69	1.11	124	374	1.09	111112	32	1.14	1245	19	1.05
344	11	1.09	122	884	1.06	3333	16	1.13	111234	21	1.05
35	36	1.09	**1122**	348	1.05	**1111123**	16	1.13	22	1394	1.04
4	140	1.08	123	922	1.02	**11122**	88	1.11	**2**	3904	1.03
11133	113	1.08	1224	130	1.01	**111123**	40	1.11	111233	37	1.03
111222	63	1.07				17	10	1.10	**35**	56	1.02
1111133	22	1.05				1112	220	1.08			
11122	170	1.04				**11123**	101	1.05			
3	1144	1.03				**4**	116	1.04			
11222	139	1.02				**111223**	39	1.03			
1122	399	1.02				25	138	1.02			
2	3306	1.01				135	63	1.02			
						234	127	1.02			
						11112	75	1.01			
						11223	102	1.01			

There is a vast number of motifs in each text, so some are overrepresented solely by chance. To reduce the role of randomness, only motifs that occur more than five times in the original text are included in the list. There is an extremely low probability that a motif is overrepresented in more than one text out of four solely by chance and, at the same time, these "duplicates" are potentially important as they may show some general tendencies; these are highlighted in grey.

Arabic texts

Table 3

List of key L-motifs in Arabic texts. Only motifs that occur more than five times in the original text are included. The motifs that are overrepresented in more than one text (duplicates) are highlighted in grey

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usul (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
22234	536	7.78	133333444	3	4.00	**22234**	135	7.16
222345	22	4.60	**22234**	165	3.25	**222344**	11	2.40
222344	68	4.06	**222345**	13	2.80	**13**	962	2.09
1344	305	2.81	**222344**	25	2.17	113	139	1.57
2234	507	2.40	13445	10	1.57	**2234**	73	1.51
3455	6	2.33	**13**	1634	1.57	**45**	12	1.44
122245	6	2.33	**133**	582	1.52	2236	6	1.40
134	1090	2.15	**1345**	37	1.52	**2223**	77	1.28
122344	29	2.14	146	17	1.50	**1224**	55	1.27
2223444	9	2.00	**2455**	8	1.50	**255**	10	1.22
12244	63	2.00	**134**	515	1.43	11224	12	1.18
22345	19	1.82	256	6	1.40	2445	6	1.17
13	2573	1.81	5	29	1.36	**123**	181	1.17
2455	6	1.75	**45**	37	1.36	2244	29	1.15
1334	322	1.75	1445	17	1.29	222333	9	1.11
12344	60	1.65	**1344**	93	1.27	12	188	1.10
135	149	1.61	246	19	1.25	**236**	11	1.09
13444	40	1.58	**135**	137	1.24	24	434	1.07
1444	98	1.52	122	109	1.24	**135**	42	1.02
1355	6	1.40	**2234**	189	1.21	**14**	363	1.01
122	180	1.38	1333	171	1.19			
22344	64	1.35	**1224**	109	1.18			
144	394	1.35	333335	6	1.17			
23445	9	1.25	**1235**	42	1.16			
2222223	9	1.25	**123**	358	1.14			
13344	51	1.24	1223334	7	1.14			
255	15	1.23	**1355**	8	1.13			
1234	192	1.21	**145**	65	1.10			
123444	11	1.20	**2355**	10	1.10			
133444	12	1.18	**23**	1633	1.10			
444	20	1.17	**2223**	149	1.09			
2355	6	1.17	3445	11	1.09			
23	2154	1.13	33335	14	1.07			
1235	34	1.09	**1234**	116	1.04			
3444	46	1.09	**4**	430	1.04			
1345	24	1.09	1335	44	1.02			
2223	193	1.07	**223**	453	1.02			
236	15	1.07	**1334**	138	1.02			

4	356	1.06	**144**	193	1.02			
133	526	1.04	**14**	986	1.01			
12444	23	1.04	44	85	1.01			
235	126	1.04						
223	601	1.04						
145	57	1.04						
233	695	1.03						
1224	141	1.03						
123	441	1.02						

The main difference between the Czech and the Arabic data sets is that there are very few Arabic key L-motifs that contain more than one monosyllabic word (which is in accordance with Arabic grammar, as Arabic monosyllabic words are typically prepositions and some other synsemantic words). As we will see, such combinations are typical for negative key L-motifs.

Cross-language Duplicate Key L-Motifs

As mentioned above, clusters of monosyllabic words tend to be overrepresented in the Czech texts and underrepresented in the Arabic texts; in contrast, clusters of disyllabic words tend to be underrepresented in Czech texts and over-represented in Arabic texts. Are there any motifs that are shared across the languages? We can find some: (1 2 2), (1 2 3), (1 2 2 4), (5), (4), (1 3 5) and (1 3 4 4 4).

Results – Negative Key L-Motifs

Czech texts

Table 4

List of negative key L-motifs in Czech texts. Only motifs that have $Pr(25) > 5$ in resampled pseudo-texts are included. The motifs that are underrepresented in more than one text (duplicates) are highlighted in grey

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
2234	12	.65	122223	19	.80	**115**	29	.50	222233	6	.88
114	410	.87	**2223**	87	.91	222222	4	.71	**2223**	154	.90
113	1030	.94	1111112	9	.91	**114**	239	.81	**22222**	34	.90
112	980	.95	224	92	.91	**22222**	21	.85	**22223**	47	.91
2222	84	.96	**222**	251	.91	116	5	.86	**115**	101	.93
115	111	.96	111113	35	.92	1115	13	.88	**112**	709	.93
22	949	.98	**22**	786	.95	**113**	550	.89	144	111	.97
1133	188	.98	**223**	286	.95	15	187	.92	**114**	481	.97
223	315	.99	14	825	.99	1114	85	.96	1113	306	.98
133	504	.99	23	890	.99	**22223**	31	.97	**222**	398	.99
			233	163	.99						

Most of the negative key L-motifs can be found in more than one text, which means that there are possibly some general tendencies and that Czech speakers may generally avoid some motif patterns. The working hypothesis would be that Czech speakers tend to avoid clusters of disyllabic words.

Arabic texts

Table 5

List of negative key L-motifs in Arabic texts. Only motifs that have $Pr(25) > 5$ in resampled pseudo-texts are included. The motifs that are underrepresented in more than one text (duplicates) are highlighted in grey.

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usul (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
1113	0	.06	**1114**	0	.11	**114**	15	.37
1114	0	.06	**1123**	3	.15	**15**	46	.58
1124	2	.07	**115**	5	.25	**22**	93	.65
11333	0	.07	**1113**	2	.3	**1134**	6	.70
11234	0	.09	**1124**	8	.38	**34**	86	.73
112	3	.10	**114**	38	.38	**33**	124	.79
114	18	.11	**112**	8	.39	**2**	394	.80
1134	7	.16	**1133**	14	.44	**144**	29	.83
1123	6	.17	**1144**	6	.54	**125**	17	.86
113	28	.17	**11233**	3	.57	**1244**	6	.88
1144	4	.19	**113**	66	.61	**222**	34	.88
1133	9	.20	**11334**	4	.63	**4**	83	.94
115	4	.24	**11333**	5	.67	**114**	15	.37
11223	2	.33	**1134**	20	.68	**15**	46	.58
1122	2	.33	**15**	172	.75	**22**	93	.65
11233	3	.36	26	30	.76	**1134**	6	.70
11224	3	.44	**33**	516	.81	**34**	86	.73
223333	5	.46	**34**	482	.82	**33**	124	.79
334	126	.50	**334**	161	.82	**2**	394	.80
11334	6	.50	**333**	176	.83	**144**	29	.83
3334	45	.58	**125**	49	.85	**125**	17	.86
16	6	.58	**33333**	16	.85	**1244**	6	.88
333	154	.61	**16**	22	.88	**222**	34	.88
33334	13	.61	3	1645	.89	**4**	83	.94
15	119	.62	**2**	1089	.90			
34	495	.63	**24**	1145	.91			
24	1128	.64	**25**	311	.91			
33	521	.67	**225**	80	.92			
2	1171	.70	12	205	.92			
3333	58	.74	**23334**	40	.93			
22334	35	.75	**2334**	133	.95			
222	95	.76	35	149	.96			

244	245	.77	**124**	222	.96			
33333	17	.78	2333	145	.96			
233334	12	.81	234	411	.96			
224	404	.82	**3333**	65	.97			
12334	27	.82						
12333	28	.85						
22224	28	.85						
124	309	.86						
3344	36	.88						
25	236	.88						
225	63	.91						
2224	125	.94						
2244	79	.94						
23334	54	.95						
2334	179	.95						
344	133	.97						
1223	93	.98						
2444	51	.98						

Again, most of the negative key L-motifs can be found in more than one text and some of them in all three texts. Arabic speakers tend to avoid clusters of mono-syllabic words, which has a quite straightforward interpretation, as mentioned above.

Cross-language Duplicate Negative Key L-Motifs

The following motifs are underrepresented in both Czech and Arabic texts: (1 1 2), (1 1 3), (1 1 4), (1 1 5), (2 2 4), (2 2 2), (2 2), (1 4 4), (1 1 3 3), (1 5), (1 1 1 3) and (1 1 1 4). In particular, (1 1 4) deserves further research because of its strong effect size (low MR value) in six texts out of seven. Also (1 1 5) is strongly underrepresented in 5 texts and the (1 1 2) – (1 1 5) series suggests a more general pattern.

Results – Key LL-Motifs

Czech texts

Table 6

List of key LL-motifs in Czech texts. Only motifs that occur more than five times in the original text are included. The motifs that are overrepresented in more than one text (duplicates) are highlighted in grey

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
48	8	1.29	122225	7	1.33	2355	9	1.43	12255	6	1.40
345	27	1.17	1346	9	1.25	3334	17	1.29	1237	12	1.30

29	23	1.09	2334	43	1.16	**16**	135	1.05	222225	11	1.20
16	189	1.06	22444	7	1.14				2238	6	1.17
134	143	1.04	13334	10	1.10				222226	6	1.17
15	350	1.04	45	38	1.08				1344	30	1.15
5	126	1.03	14	432	1.02				22227	7	1.14
25	370	1.01	1223	57	1.02				12336	7	1.14
			144	67	1.01				122235	7	1.14
			15	268	1.01				12	343	1.09
									23333	18	1.06
									113	63	1.03
									2	1045	1.02

Unlike the L-motifs, there is no clear interpretation of these data, or at least the author of the study does not see any. There are only two duplicate key LL-motifs, which means that the LL-motifs are rather symptomatic of individual texts and that there is less room for generalization.

Arabic texts

Table 7

List of key LL-motifs in Arabic texts. Only motifs that occur more than five times in the original text are included. The motifs that are overrepresented in more than one text (duplicates) are highlighted in grey

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usulu (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
356	8	1.80	**22333**	20	1.31	1234	33	2.13
1222225	6	1.75	12344	8	1.29	12235	7	1.60
3344	12	1.63	23333	14	1.25	12234	11	1.50
222235	8	1.50	**22223**	36	1.19	22225	8	1.29
122235	8	1.50	13344	6	1.17	1223	36	1.23
444	10	1.38	**444**	6	1.17	**1225**	16	1.21
2255	10	1.22	138	7	1.14			
1222224	6	1.17	222333	7	1.14			
345	16	1.13	2246	8	1.13			
34	186	1.13	1235	24	1.09			
44	45	1.10	22222	17	1.06			
22244	13	1.08	335	21	1.05			
25	293	1.07	2223	79	1.04			
2235	31	1.07	1233	53	1.04			
22334	16	1.06	**1225**	33	1.03			
22223	40	1.05	35	71	1.01			
22333	20	1.05	**225**	83	1.01			
225	110	1.05						
224	195	1.03						
4	289	1.01						

There are more duplicate key LL-motifs than in the Czech texts, even though the Arabic texts are shorter and there are only three of them. This can mean either that there are some unknown general syntactical features in the Arabic language that need to be explored by means of LL-motifs or that the texts share some common passages – quotations, proverbs or other formulaic expressions. The latter interpretation is quite feasible since there is a strong tendency to formulaicity in Arabic literature.

Cross-language Duplicate Key LL-Motifs

It is remarkable that there are more cross-language duplicate key LL-motifs (namely (3 4 5), (2 5), (1 2 2 3), (1 2 2 2 3 5) and (2 3 3 3 3)) than duplicates within the Czech texts. The measurement must be repeated on further texts to check that the effect is not caused by random variation. If the duplicated LL-motifs were overrepresented in other texts, we would be on the verge of the discovery of new language independent syntactic patterns, because the cross-language duplicated key motifs cannot be explained by the formulaicity (we do not expect direct Arabic quotations in the Czech texts and vice versa).

Results – Negative Key LL-Motifs

Czech texts

Table 8

List of negative key LL-motifs in Czech texts. Only motifs that have $Pr(25) > 5$ in resampled pseudo-texts are included. The motifs that are underrepresented in more than one text (duplicates) are highlighted in grey.

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
2223	49	.83	1444	1	.67	**23**	424	.94	45	10	.61
24	509	.99	1333	13	.88	3	629	.97	233	158	.95
			26	100	.89				33	230	.98
			1233	18	.90						
			16	94	.94						
			34	134	.96						
			23	474	.99						

Arabic texts

Table 9
List of negative key LL-motifs in Arabic texts. Only motifs that have $Pr(25) > 5$ in resampled pseudo-texts are included. The motifs that are underrepresented in more than one text (duplicates) are highlighted in grey.

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usul (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
126	17	.69	**12**	146	.82	**15**	36	.66
2	509	.69	**26**	54	.83	**16**	15	.73
12	155	.70	**2**	513	.88	**12**	43	.79
22	177	.72	**15**	171	.92	**14**	91	.81
18	8	.75	**16**	72	.92	3	164	.92
16	80	.78	**14**	340	.95	**26**	15	.94
222	67	.84	**126**	17	.95	34	22	.96
17	33	.89	**17**	27	.97			
112	9	.91						
115	12	.93						
133	120	.94						

There is a substantial difference between the negative LL-motifs in the Czech and in the Arabic texts. Only one LL-motif was underrepresented in more than one Czech text while Arabic texts share the vast majority of negative key LL-motifs.

It seems that positive key LL-motifs are typical for individual texts while negative key LL-motifs are typical for a language code or at least for the discourse (as an intertextual property) . The question is why this rule is valid for Arabic texts and not for Czech texts.

Cross-language Duplicate Negative Key LL-Motifs

LL-motifs (2 6), (1 6), (3 4) and (3) are underrepresented in both Czech and Arabic texts. The same question is raised as for the positive LL-motifs: Why are there fewer duplicate negative LL-motifs in Czech texts than cross-language negative LL-motifs?

Results – Key LLL-Motifs

Czech texts

Table 10
List of key LLL-motifs in Czech texts. Only motifs that occur more than five times in the original text are included

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
122222	9	1.25	1444	6	1.40	112223	7	2	222225	6	1.17
22225	10	1.22	1345	6	1.40	246	7	1.33	222222	8	1.13
11222	6	1.17	24	160	1.01	12333	10	1.22	146	9	1.11
						33	66	1.06	122224	10	1.10
						14	178	1.01	2335	10	1.10
									2223	68	1.10
									1123	22	1.05

There are no duplicates (key LLL-motifs overrepresented in more than one text). This means that there are no long quotations or proverbs or other formulaic expressions shared between more than one Czech text.

Arabic texts

Table 11
List of key LLL-motifs in Arabic texts. Only motifs that occur more than five times in the original text are included. The motifs that are overrepresented in more than one text (duplicates) are highlighted in grey

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usulu (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
128	6	1.40	11224	6	1.40	**18**	8	2.25
235	32	1.32	334	17	1.38	144	12	1.30
155	9	1.25	5	19	1.25	**4**	34	1.30
227	10	1.22	**235**	25	1.24	34	20	1.24
237	6	1.17	255	6	1.17	234	16	1.06
27	21	1.16	**4**	77	1.15			
11223	8	1.13	116	7	1.14			
2226	9	1.11	2344	7	1.14			
1344	9	1.11	**18**	8	1.13			
4	88	1.05	1133	10	1.10			

There are still many LLL-motifs overrepresented in more than one Arabic text. We can expect that these duplicates are caused by some long quotations or other formulaic expressions.

Cross-language Duplicate LLL-Motifs

There are no LLL-motifs overrepresented in both the Czech and Arabic texts.

Results – Negative Key LLL-Motifs

Czech texts

Table 12
List of negative key LLL-motifs in Czech texts. Only motifs that have $Pr(25) > 5$ in the resampled pseudo-texts are included

Žert (Kundera)			**Babička (Němcová)**			**Válka s Mloky (Čapek)**			**Katyně (Kohout)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
									223	94	.91

Arabic texts

Table 13
List of negative key LLL-motifs in Arabic texts. Only motifs that have $Pr(25) > 5$ in resampled pseudo-texts are included. The motifs that are underrepresented in more than one text (duplicates) are highlighted in grey

Al-I'tibar (al-Hamadani)			**Ma'rifa (ibn as-Sallah)**			**Usul (ibn abi Zamanayn)**		
m	f_m	*MR*	*m*	f_m	*MR*	*m*	f_m	*MR*
134	20	.72	122	26	.93	**13**	51	.93
123	68	.83	2	212	.93	12	21	.96
13	224	.92	**13**	182	.97	14	33	.97
			123	61	.98			

There are still some LLL-motifs underrepresented in more than one Arabic text. The motif (1 3) especially attracts attention as it is underrepresented in all three texts.

Cross-language Duplicate Negative Key LLL-Motifs

There are no cross-language duplicate negative key LLL-motifs, which is not surprising since only one negative key LLL-motif was found in the Czech texts.

5. Conclusion

The paper shows that some length motifs are typical for certain single texts while others are overrepresented or underrepresented in more than one text. Some motifs are underrepresented or overrepresented in both the Czech and Arabic texts, which suggests some general cross-language patterns. A larger collection of texts in various languages is needed to explore these patterns more closely.

It is not clear whether the positive key length motifs are consequences of some (self-)quotations, or idiomatic or formulaic expressions, or whether they are a consequence of various grammatical patterns. Exploring typical key motifs one-by-one would be very helpful; therefore, a dedicated search tool or motif-processing extension for a corpus manager is needed in order to proceed with the research in this field. Such an extension would also enable us to research motif collocations and other typical lexicographical properties.

Acknowledgements

The research reflected in this article has been supported by the Grant Agency of the Czech Republic, project no. 13-28220S.

References

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

Köhler, Reinhard (2006). The frequency distribution of the lengths of length sequences. In: Jozef Genzor & Martina Bucková (eds.), *Favete linguis. Studies in honour of Victor Krupa*, 145–152. Bratislava: Slovak Academic Press.

Köhler, Reinhard (2008a). Word length in text. A study in the syntagmatic dimension. In: Sibyla Mislovičová (ed.), *Jazyk a jazykoveda v pohybe*, 416–421. Bratislava: VEDA.

Köhler, Reinhard (2008b). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory* 1(1). 115–119.

Köhler, Reinhard (2015). Motifs. In: Mikros, G. & Mačutek, J. (eds.). *Sequences in Language and Text,* 89–108. Berlin, Boston: De Gruyter Mouton.

Köhler, Reinhard, Sven Naumann (2008). Quantitative text analysis using L-, F- and T-segments. In: Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme & Reinhold Decker (eds.), *Data Analysis, Machine Learning and Applications*, 635–646. Berlin & Heidelberg: Springer.

Köhler, Reinhard, Sven Naumann (2009). A contribution to quantitative studies on the sentence level. In: Reinhard Köhler (ed.), *Issues in quantitative linguistics*, 34–57. Lüdenscheid: RAM-Verlag.

Köhler, Reinhard, Sven Naumann (2010). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.). *Text and language. Structures, functions,interrelations, quantitative perspectives*, 81–89. Wien: Praesens.

Liu, Haitao, Junying Liang (eds.). *Motifs in Language and Text*. Berlin, Boston: De Gruyter Mouton, 2017. In preparation.

Mačutek, Ján (2009). Motif richness. In: Reinhard Köhler (ed.), *Issues in Quantitative Linguistics*, 51–60. Lüdenscheid: RAM-Verlag.

Mačutek, Ján (2015). Type-token relation for word length motifs in Ukrainian texts. In: Tuzzi, A., Benešová, M. & Mačutek, J. (eds.), *Recent Contributions to Quantitative Linguistics*. 63–73. Berlin, Boston: De Gruyter Mouton.

Mačutek, Ján, George K. Mikros (2015). Menzerath-Altmann Law for Word Length Motifs. In: Mikros, G., Mačutek, J. (eds). *Sequences in Lan guage and Text*. 125–132. Berlin, Boston: De Gruyter Mouton.

Milička, Jiří (2012). Minimal ratio: an exact metric for keywords, collocations etc. *Czech and Slovak Linguistic Review*, 12(1), 62-7.

Milička, Jiří (2015). Is the Distribution of L-Motifs Inherited from the Word Lengths Distribution? In: Mikros, G., Mačutek, J. (eds.). *Sequences in Language and Text*. 133–146. Berlin, Boston: De Gruyter Mouton.

Sanada, Haruko (2010). Distribution of motifs in Japanese texts. In: Peter Grzybek, Emmerich Kelih, Ján Mačutek (eds.), *Text and language. Structures, functions, interrelations, quantitative perspectives*, 183–193. Wien: Praesens.

A Synergetic Regression Model of Language Complexity Trade-Offs

Germán Coloma
CEMA University, Buenos Aires, Argentina
gcoloma@cema.edu.ar

Abstract. In this paper we develop a statistical model of language complexity trade-offs using four typological measures (related to phonology, morphology, syntax and lexicon). The data come from the 100-language sample that appears in the World Atlas of Language Structures (WALS), and the trade-offs are calculated using different types of correlation coefficients. All those coefficients are statistically insignificant when they are computed using a standard (product-moment) methodology, but they become significant when we use simultaneous-equation regression methods, especially the ones based on seemingly unrelated regressions (SUR) and three-stage least squares (3SLS). These results are related to ideas suggested by the theoretical literature on synergetic linguistics.

Keywords*: complexity trade-off, WALS, correlation, simultaneous-equation regression, synergetic linguistics.*

1. Introduction

A language complexity trade-off is a situation in which a higher level of complexity for a certain language component appears in correspondence to a lower level of complexity for another component. The literature about this topic can be divided between papers that show that languages usually exhibit complexity trade-offs and papers that show that such trade-offs do not exist. Among the first group of papers we can cite contributions such as Nettle (1995) and Fenk-Oczlon, Fenk (2008), while in the second group there are articles like Shosted (2006) and Nichols (2009).

In general, the way in which the different authors assess the possible existence and significance of complexity trade-offs is some version of correlation analysis. Under that approach, two different measures of complexity (e.g., phonological and morphological complexity) are supposed to display a trade-off if they are negatively correlated between themselves, and the way to find that correlation (or its absence) is to calculate a coefficient based on the values of the different complexity measures in a sample of languages.[1]

[1] An alternative way to do that is to run a regression between the two variables under analysis. In that case, the relevant coefficient is the slope of the regression line obtained, which should also be negative and significant if we are looking for evidence of a complexity trade-off.

If we look at the main methodological difference between the literature that finds statistically significant complexity trade-offs and the literature that does not find them, we see that one important point is the type of measures that they use. While in the first group authors generally rely on "empirical measures" (i.e., measures of complexity calculated using data from actual words or texts written in different languages), in the second group they generally use theoretical or "typological" measures (i.e., measures obtained from the grammars of the sample languages).

Another feature that we have found in previous work (Coloma 2014, 2016) is that language complexity trade-offs seem to be more important and statistically more significant if we measure them using partial correlation coefficients instead of standard correlation coefficients. This is related to the fact that, when we use a partial correlation coefficient, we are also including information from factors besides the two correlating variables. It is also linked to the idea that complexity variables can be determined by a system in which there are interactions among them, so each partial measure of complexity can be correlated to several other measures at the same time.

This reference to a system of relationships between the different complexity measures can be related to a branch of the theoretical literature that sees language as a self-organizing and self-regulating system whose properties come from the interaction of several constitutive and control requirements. That branch is known as "synergetic linguistics", and its origins can be traced back to Köhler (1986, 1987). It is also related to another branch of the linguistic literature that sees language as a complex adaptive system (e.g., Beckner et al. 2009).

The aim of this paper is to look for the existence of trade-offs in a context in which language complexity is measured using theoretical variables (which is the one in which they have been harder to find), through a synergetic approach in which different factors interact. To do that we use a statistical methodology based on simultaneous-equation regressions, whose results allow us to calculate different types of partial correlation coefficients between our complexity measures. The analysis will be performed using the so-called "100-language sample" from the World Atlas of Language Structures (WALS), and complexity will be measured using binary variables that represent different concepts of phonological, morphological, syntactic and lexical complexity.

2. Description of the data

The WALS is a large database that compiles information about structural features from the grammars of the world's languages. In its current online version (Dryer, Haspelmath 2013), it contains data from 2679 languages and dialects, corresponding to 192 features that belong to different components of language structure.

The editors of the WALS have selected a sample of 100 languages which they ask the authors of the different chapters of the atlas to include in their

reports "if at all possible", and those languages are supposed to form a relatively balanced sample of genealogical and areal diversity.[2] Making use of the fact that we have more information about the languages that belong to this sample than the one available for the remaining languages, in this paper we use the 100-language WALS sample for a series of statistical analyses aimed at the detection of possible complexity trade-offs. To do that, we define four binary variables whose values can alternatively be "simple" or "complex", and those variables are built using information from certain features.[3]

The definitions of the abovementioned complexity variables are the following:

a) Phonology: A language is considered to be complex if it has more than 25 consonant phonemes, more than 6 vowel qualities, or uses tone as a distinctive phonological feature. This generates a division in which 60 languages are complex, and the remaining 40 languages are simple.

b) Morphology: A language is considered to be complex if it is polysynthetic, and simple if it is not. This implies that 32 languages in the sample are complex, and the remaining 68 ones are simple.

c) Syntax: A language is considered to be complex if it has no dominant word order for subject, object and verb, or if it uses relative pronouns to build relative clauses. Under this definition, 22 languages are complex and the remaining 78 ones are simple.

d) "Lexicon": A language is considered to be complex if it has definite articles and uses different verbs for nominal and locational predication. This implies that 33 languages are complex, and the remaining 67 languages are simple.[4]

The easiest way to detect possible trade-offs between these binary complexity variables is to calculate standard Pearson (product-moment) correlation coefficients, like the ones that appear in table 1. In that table there are five negative correlation coefficients and one positive correlation coefficient, but none of them is statistically significant at the 5% probability level.[5]

[2] The complete list of languages is reproduced in appendix 1.

[3] The WALS features used are: 1A (Consonant inventories), 2A (Vowel quality inventories), 13A (Tone), 20A (Fusion of selected inflectional formatives), 26A (Prefixing vs. suffixing in inflectional morphology), 37A (Definite articles), 81A (Order of subject, object and verb), 119A (Nominal and locational predication) and 122A (Relativization on subjects).

[4] The value of each complexity variable for each language is reported in appendix 2, where "simple" is denoted as "0" and "complex" is denoted as "1".

[5] For any two variables whose correlation is calculated using 100 observations, correlation coefficients are statistically significant at the 5% probability level if they are greater than 0.2 in absolute value.

Table 1
Correlation coefficients between complexity variables

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.1400	1.0000		
Syntax	-0.0591	0.0497	1.0000	
Lexicon	-0.1650	-0.0711	-0.0647	1.0000

3. Simultaneous equation regressions

The Pearson correlation coefficients reported on table 1 are in all cases calculated using information that covers two variables for each coefficient. In this case, however, it is possible to consider that our measures of phonological, morphological, syntactic and lexical complexity are somehow interrelated, in the sense that the relationship between any pair of those measures can be influenced by the other complexity variables.

One way to model a situation like the one described in the previous paragraph is to build a system of simultaneous equations like the following:

*Phonology = c(1) +c(2)*Morphology +c(3)*Syntax +c(4)*Lexicon* (1) ;
*Morphology = c(5) +c(6)*Phonology +c(7)* Syntax +c(8)* Lexicon* (2) ;
*Syntax = c(9) +c(10)*Phonology +c(11)*Morphology +c(12)*Lexicon* (3) ;
*Lexicon = c(13) +c(14)*Phonology +c(15)*Morphology +c(16)*Syntax* (4) ;

where *Phonology*, *Morphology*, *Syntax* and *Lexicon* are the complexity variables defined for the 100-language WALS sample, whose values can either be equal to 0 (if the language is simple in the corresponding domain) or equal to 1 (if the language is complex in that domain). Additionally, coefficients *c(1)* to *c(16)* are the values of the parameters that relate each complexity measure with the other measures.

One easy way to estimate coefficients *c(1)* to *c(16)* is to run a set of four separate ordinary least-square (OLS) regressions. If we do that, we get the following results:

*Phonology = 0.7286 –0.1572*Morphology –0.0749*Syntax –0.1872*Lexicon* (5) ;
*Morphology = 0.4300 –0.1462*Phonology +0.0389*Syntax –0.0935*Lexicon* (6) ;
*Syntax = 0.2649 –0.0560*Phonology +0.0313*Morphology –0.0644*Lexicon* (7) ;
*Lexicon = 0.4826 –0.1749*Phonology –0.0939*Morphology –0.0804*Syntax* (8) .

With these results, it is possible to calculate new (partial) correlation coefficients, defined as the square roots of the products of the corresponding pairwise regression coefficients.[6] For example, for the relationship between phonological

[6] For a more thorough explanation of the concept of partial correlation, and the available alternatives for its calculation, see Prokhorov (2002).

and morphological complexity, this is equal to the square root of "–0.1572" (which is the regression coefficient of *Morphology* as a determinant of *Phonology*) times "–0.1462" (which is the regression coefficient of *Phonology* as a determinant of *Morphology*). As both regression coefficients are negative, we must assign a negative sign to the corresponding correlation coefficient (i.e., to the corresponding square root), whose value is "r = –0.1516". If we make similar calculations for all the possible pairwise relationships that appear in our system, we will have a set of numbers like the ones reported in table 2.

Table 2
Partial correlation coefficients between complexity variables

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.1516	1.0000		
Syntax	-0.0648	0.0349	1.0000	
Lexicon	-0.1809	-0.0937	-0.0720	1.0000

The procedure used to calculate the regression coefficients that appear in equations 5 to 8 (which is the basis for the calculation of the partial correlation coefficients reported in table 2) estimates each equation independently. However, if we use a truly simultaneous procedure in which the four equations are estimated at the same time, we can also use the correlation coefficients between the residuals of the different equations, and derive a new set of regression coefficients like the following:

*Phonology = 0.8580 –0.3149*Morphology –0.1532*Syntax –0.3743*Lexicon* (9) ;
*Morphology = 0.5499 –0.2928*Phonology +0.0586*Syntax –0.2032*Lexicon* (10) ;
*Syntax = 0.3174 –0.1146*Phonology +0.0472*Morphology –0.1324*Lexicon* (11) ;
*Lexicon = 0.6414 –0.3496*Phonology –0.2040*Morphology –0.1652*Syntax* (12) .

Table 3
Partial correlation coefficients using SUR

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.3036	1.0000		
Syntax	-0.1325	0.0526	1.0000	
Lexicon	-0.3618	-0.2036	-0.1479	1.0000

This new set of regression coefficients comes from a statistical method known as the "seemingly unrelated regression" technique (SUR), originally proposed by Zellner (1962). This method is not very common in linguistics, but it

is relatively widespread in other social sciences such as economics (where it is standard for applications like demand estimation). In this case, however, its use generates an important increase in the magnitude of the estimated negative partial correlation coefficients, which can now be approximated by the numbers reported on table 3.

An additional variation that can be introduced is the elimination of the only positive correlation coefficient that we have obtained (which relates morphological and syntactic complexity), provided that its sign is counterintuitive and its absolute value (r = 0.0526) is small and statistically insignificant. If we do that, we can estimate a new restricted system of equations, whose results (using SUR) are the following:

*Phonology = 0.8591 –0.3161*Morphology –0.1563*Syntax –0.3745*Lexicon*(13)

*Morphology = 0.5686 –0.2989*Phonology –0.2099*Lexicon* (14);

*Syntax = 0.3421 –0.1263*Phonology –0.1404*Lexicon* (15);

*Lexicon = 0.6424–0.3498*Phonology –0.2053*Morphology –0.1671*Syntax* (16)

The new partial correlation coefficients implied by this system of regression equations appear on table 4, in which three out of the five estimated coefficients (phonology vs. morphology, phonology vs. lexicon, and morphology vs. lexicon) are now statistically significant at the 5% probability level.

Table 4
Partial correlation coefficients using a restricted version of SUR

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.3074	1.0000		
Syntax	-0.1405	0.0000	1.0000	
Lexicon	-0.3620	-0.2076	-0.1532	1.0000

4. Instrumental variables

The logic behind the equations used to estimate the partial correlation coefficients between phonological, morphological, syntactic and lexical complexity has to do with the idea that those complexity levels come from a system that generates them as the outcome of some unified procedure. That procedure may consist of the interaction between several constraints such as the ones proposed by the synergetic linguistics literature (e.g., Köhler 2005), or some kind of iterative learning mechanism like the one proposed by Smith, Kirby and Brighton (2003).

Those theoretical approaches share the common assumption that languages emerge in environments that can be influenced by a series of non-linguistic factors. Among those factors, the ones that are easier to analyze in

empirical work are the geographic, phylogenetic and demographic characteristics of the different languages. For example, as any language originated in a certain point in space, it can be classified as belonging to a certain region or area (e.g., one of the six large macro-areas that the WALS defines). The other major classification used by the linguistic literature is the phylogenetic one, which groups languages into families that share a common ancestor (e.g., Indo-European, Afro-Asiatic, Niger-Congo, etc.). A third element that we can use to classify languages is their relative size in terms of population, which is related to the geographic expansion that each language has had in history, and its alternative use as a first or second language by different people. Those characteristics have been used to analyze the relationship between language complexity and population, in papers such as Dahl (2011) or Bentz et al. (2015).

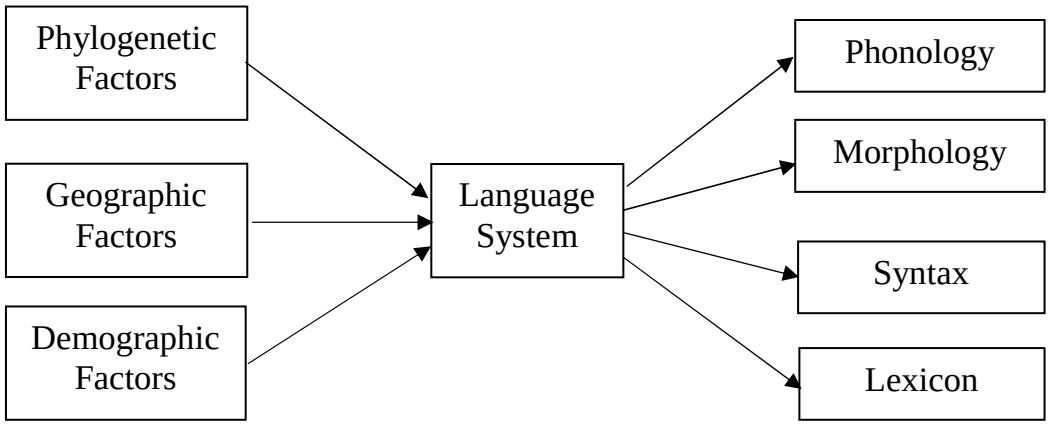


Figure 1: Relationships in a language system

One general way to think about the relationship between linguistic and non-linguistic variables is to assume that the latter are part of the environment in which the former arise. This implies that linguistic variables may be influenced by non-linguistic factors, buy not the other way round. If we use this type of reasoning, we may represent our relationships by a graph like the one that appears in figure 1. In it we see that each linguistic variable (phonological, morphological, syntactic and lexical complexity) is the outcome of a language system which has in turn been influenced by phylogenetic, geographic and demographic factors.

The application of this view to the statistical explanation of the levels of language complexity implies the possibility of running a system of equations where those levels of complexity are the dependent variables and the non-linguistic factors are the independent variables (on which the language complexity levels depend). In order to do that, we first need to encode the phylogenetic, geographic and demographic factors into numerical values, and the simplest way to do it is to create binary variables that take a value equal to one when a language belongs to a certain (geographic, phylogenetic or demographic) group and zero otherwise.

Using the six WALS macro-areas and three additional divisions for those areas, we have created nine binary geographic variables that correspond to Eurasia, South East Asia, Africa, Papunesia, Australia, North America, Meso-america, the Amazon basin, and (the rest of) South America. Due to the fact that in the 100-language WALS sample there are relatively many observations that belong to three particular families (Austronesian, Indo-European, and Niger-Congo), we have also created three variables related to those phylogenetic factors. Finally, we have classified languages according to their relative size, considering the ones with more than 5 million native speakers as "major" and the other ones as "minor".[7]

The next step in the estimation of the effect of non-linguistic factors on language variables was to run a system of OLS equations in which each of the four complexity measures used in the previous sections was regressed against the thirteen non-linguistic binary variables. As the sum of the geographic variables completely covers the whole sample of languages, we have used one region (Eurasia) as the default one (constant) and included the remaining geographic variables as explanatory variables. The results for each of the four regressions appear in table 5.

Table 5
Regression coefficients for the language complexity variables

Explanatory variables	Phonology	Morphology	Syntax	Lexicon
Constant	0.5919	0.2839	0.2471	0.3461
Africa	0.2445	-0.1906	-0.2413	0.3149
South East Asia	0.2117	-0.1674	-0.1817	0.2827
Papunesia	-0.1318	0.0834	-0.2281	0.0316
Australia	-0.5919	0.2876	0.0386	-0.2032
North America	0.0747	0.5495	0.2529	0.2373
Mesoamerica	0.0747	0.2161	-0.2471	-0.1794
South America	-0.4862	0.1720	-0.0157	0.3958
Amazon	0.0331	0.4661	0.0029	-0.3461
Austronesian	-0.4706	-0.2806	0.2621	0.4699
Indo-European	-0.2026	-0.1441	0.5814	0.1335
Niger-Congo	0.0626	-0.0334	0.0279	-0.2233
Major Language	0.2357	-0.1398	-0.0785	-0.3546

The outcome of this regression analysis shows results that are in line with received linguistic knowledge. We can see, for example, that Australian and Austronesian languages tend to have simpler phonologies, that North American

[7] The 33 major languages in the 100-language WALS sample are the following: Mandarin, English, Spanish, Hindi, Arabic, Russian, Japanese, German, French, Indonesian, Korean, Turkish, Vietnamese, Persian, Kannada, Hausa, Burmese, Tagalog, Yoruba, Swahili, Oromo, Thai, Malagasy, Greek, Zulu, Quechua, Berber, Hebrew, Khalkha, Finnish, Guarani, Georgian and Hmong Njua. The remaining 67 ones are considered to be "minor languages". To see which languages belong to the different geographic and phylogenetic groups, see appendix 1.

languages tend to have more complex morphologies, that Indo-European languages tend to have more complex levels of syntax, and that major languages tend to have a simpler lexicon (but a more complex phonology). These results may also be combined with the analysis that we performed in section 3, since the newly obtained regression coefficients can be the basis to build variables that are "instrumental" for that analysis.

The way to build these instrumental variables is to recover the predictions of the different regressions for each of the dependent variables of those regressions. With that we obtain four new variables (*Phonfit*, *Morphfit*, *Syntfit* and *Lexfit*), which are linear combinations of the values of our thirteen binary non-linguistic variables (multiplied by their respective regression coefficients). These instrumental variables have the property that they can replace the original variables of the regression systems run in section 3 and are at the same time completely exogenous to those systems.

Instrumental variables are a useful resource to solve a statistical problem known as the "endogeneity problem". This arises when we run a regression in which we know that both the dependent variable and (at least one of) the independent variables are somehow determined by the same mechanism. When this is the case, the obtained regression coefficients can be biased or inconsistent. If, however, we replace the endogenous independent variables by other variables that serve as exogenous instruments to approximate the value of those variables, then the estimation may become less precise but more consistent and unbiased.[8]

In the system of equations introduced in section 3, all variables seem to be endogenous in the sense described in the previous paragraph. This is because they are at the same time dependent variables in one equation and independent variables in other equations, and all the relationships are supposed to be generated by the same mechanism. If we add the idea that this mechanism is somehow influenced by non-linguistic factors like the ones represented by the set of phylogenetic, geographic and demographic variables included in the regressions performed in this section, we can think of those variables as good candidates to act as exogenous instruments to replace the original (endogenous) linguistic variables.

The statistical method that uses a set of instruments to estimate instrumental variables, and then uses those instrumental variables to replace the original endogenous variables in the context of a simultaneous-equation regression estimation, is known as "three-stage least squares" (3SLS). It was originally proposed by Zellner, Theil (1962), and is widely used in other social sciences such as economics (where it is standard for problems such as supply and demand estimation). If we use this method to run the system formed by equations 13 to 16, what we find is the following:

$$Phonology = 1.1276 - 0.5125 * Morphfit - 0.6073 * Syntfit - 0.6969 * Lexfit \quad (17)$$;
$$Morphology = 0.7413 - 0.5560 * Phonfit - 0.2657 * Lexfit \quad (18)$$;
$$Syntax = 0.5538 - 0.4416 * Phonfit - 0.2088 * Lexfit \quad (19)$$;

[8] For a more complete explanation of the endogeneity problem, see Kennedy (2008), chapter 9.

$$Lexicon = 0.7729 - 0.5569 * Phonfit - 0.1947 * Morphfit - 0.2112 * Syntfit \quad (20)\ .$$

With these regression coefficients, we can now derive new partial correlation coefficients, which are the ones that appear in table 6. There we can see that the five estimated coefficients are now negative and statistically significant at the 5% probability level, since all of them are higher than 0.2 in absolute value.

Table 6
Partial correlation coefficients using 3SLS

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.5338	1.0000		
Syntax	-0.5179	0.0000	1.0000	
Lexicon	-0.6230	-0.2275	-0.2100	1.0000

5. Concluding remarks

The analysis performed in this paper about possible complexity trade-offs in the 100-language WALS sample can be seen as a particular statistical exercise whose outcome is likely to change if we use other language samples or other definitions for the different types of language complexity. The message that we get from this analysis, however, is probably of a more general nature, since it seems to reconcile some contradictory results from previous literature.

In the very beginning, our analysis generates the standard result that, if we measure trade-offs using product-moment correlation coefficients between typological complexity measures, what we get is a set of statistically insignificant values which imply that language complexity trade-offs are either non-existent or unimportant (and this is equivalent to the conclusions of papers such as Shosted 2006). When we use non-linguistic variables related to phylogenetic, geographic and demographic factors, conversely, we obtain results that indicate that some complexity variables may indeed be influenced by those factors, and this seems to be in line with some contributions from sociolinguistic typology (e.g., Trudgill 2009).

What we do not get, if we restrict ourselves to standard correlation and regression techniques, is anything related to the logic behind the idea of language as a complex adaptive system, since that idea implies that language should evolve to be at the same time "compressed" (i.e., relatively simple and easy to learn) and "expressive" (i.e., relatively complex and capable to convey meanings for multiple concepts).[9] To reconcile these two requirements (or the alternative ones developed by the synergetic linguistics literature) we need to find some kind of trade-off between different levels of language complexity, such as the ones that typically appear in the literature that uses empirical measures of complexity (e.g., Fenk-Oczlon, Fenk 2008).

[9] For an interesting analysis of this dichotomy, see Kirby et al. (2015).

In one contribution that belongs to that literature (Coloma 2016) we got a result that shows that language complexity trade-offs seem to be more significant if we measure them using partial correlation coefficients instead of standard correlation coefficients, and they get even more significant if we use simultaneous-equation regression methods such as SUR. We therefore decided to apply the same logic to study the possible trade-offs between typological complexity measures, since simultaneous-equation regression methods have been designed to deal with statistical problems in which the different equations that we want to regress are generated by the same mechanism. And this is precisely the case here, because the synergetic approach to language has to do with the idea that complexity variables must come from some kind of unified generating process.

But, as we also have variables related to non-linguistic factors that may influence the language system from outside, we can use those variables to solve a statistical problem that simultaneous-equation models usually have, which is the endogeneity problem. To solve this we use non-linguistic factors to create instrumental variables, and then we use those instrumental variables as part of a 3SLS procedure. In this case, this can be seen as the statistical representation of a model in which non-linguistic factors are able to influence the system in which language is produced, and this system is in turn the one that generates the (interrelated) levels of complexity that correspond to its different sub-systems (i.e., phonology, morphology, syntax and lexicon).

When we did this, our results changed dramatically. Except for the coefficient that relates morphology and syntax, which is always insignificant, all the other correlation coefficients are negative and statistically significant when we estimate them using 3SLS. Moreover, their statistical significance increases when we move from standard to partial (OLS) coefficients, and the same occurs when we move from OLS to SUR, and from SUR to 3SLS coefficients.

This behaviour may be due to different causes, but one plausible one is the idea that the statistical sophistications included in our calculations are related to an increasing consideration of the interactions between language complexity variables. When we only use standard correlation coefficients, those interactions are computed pairwise, while the calculation of partial correlation coefficients through an OLS procedure implies considering multiple interactions as well. Using the SUR method is in turn equivalent to introducing relationships between the errors that arise when we estimate the different complexity equations, whereas 3SLS implies considering the effect of non-linguistic factors (and their influence on the system that is producing the different levels of language complexity).

As a final conclusion, therefore, we can say that language complexity trade-offs may be more pervasive than it seems when we measure them using simple statistical tools such as standard correlation coefficients or univariate regression equations. This is because there may be some interferences from other (linguistic and non-linguistic) factors, whose effects have to be taken into account using more sophisticated statistical procedures. But that is indeed the message implied by the synergetic approach to language, and the use of simultaneous-equation regression models can be a way to interpret the available data which is compatible with that approach.

References

Beckner, Clay, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, Tom Schoenemann (2009). Language Is a Complex Adaptive System. *Language Learning* 59, suppl. 1, 1-26.

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*, vol. 10(6), e0128254.

Coloma, Germán (2014). Towards a Synergetic Statistical Model of Language Phonology. *Journal of Quantitative Linguistics* 21, 100-122.

Coloma, Germán (2016). The Existence of Negative Correlation between Linguistic Measures Across Languages. *Corpus Linguistics and Linguistic Theory*, forthcoming.

Dahl, Osten (2011). Are Small Languages More or Less Complex than Big Ones? *Linguistic Typology* 15, 171-175.

Dryer, Matthew, Martin Haspelmath (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Fenk-Oczlon, Gertraud, August Fenk (2008). Complexity Trade-Offs Between the Subsystems of Language. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change:* 43-65. Amsterdam: John Benjamins.

Kennedy, Peter (2008). *A Guide to Econometrics*, 6th edition. New York: Wiley.

Kirby, Simon, Monica Tamariz, Hannah Cornish y Kenny Smith (2015). Compression and Communication in the Cultural Evolution of Linguistic Structure. *Cognition* 141, 87-102.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik: Stuktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, Reinhard (1987). System Theoretical Linguistics. *Theoretical Linguistics* 14, 241-257.

Köhler, Reinhard (2005). Synergetic Linguistics. In G. Altmann, R. Köhler & R. Piotrowski (eds.), *Quantitative Linguistics: An International Handbook:* 760-774. Berlin: De Gruyter.

Nettle, Daniel (1995). Segmental Inventory Size, Word Length and Communicative Efficiency. *Linguistics* 33, 359-367.

Nichols, Johanna (2009). Linguistic Complexity: A Comprehensive Definition and Survey. In: G. Sampson, D. Gil & P. Trudgill (eds.), *Language Complexity as an Evolving Variable:* 110-125. Oxford: Oxford University Press.

Prokhorov, A.V. (2002). Partial Correlation Coefficient. In: M. Hazewinkel (ed.), *Encyclopedia of Mathematics*. New York: Springer.

Shosted, Ryan (2006). Correlating Complexity: A Typological Approach. *Linguistic Typology* 10, 1-40.

Smith, Kenny, Simon Kirby, Henry Brighton (2003). Iterated Learning: A Framework for the Emergence of Language. *Artificial Life* 9, 371-386.

Trudgill, Peter (2009). Sociolinguistic Typology and Complexification. In G. Sampson, D. Gil & P. Trudgill (eds.), *Language Complexity as an Evolving Variable:* 98-109. Oxford: Oxford University Press.

Zellner, Arnold (1962). An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias. *Journal of the American Statistical Association* 57, 348–368.

Zellner, Arnold, Henri Theil (1962). Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica* 30, 54–78.

Appendix 1: List of languages in the WALS sample

Code	Language	Region	Family
1	Abkhaz	Eurasia	Northwest Caucasian
2	Acoma	North America	Keresan
3	Alamblak	Papunesia	Sepik
4	Amele	Papunesia	Trans-New Guinea
5	Apurina	Amazonia	Arawakan
6	Arabic (Egyptian)	Eurasia	Afro-Asiatic
7	Arapesh (Mountain)	Papunesia	Kombio
8	Asmat	Papunesia	Trans-New Guinea
9	Bagirmi	Africa	Nilo-Saharan
10	Barasano	Amazonia	Tucanoan
11	Basque	Eurasia	Vasconic
12	Berber (Middle Atlas)	Africa	Afro-Asiatic
13	Burmese	South East	Sino-Tibetan
14	Burushaski	Eurasia	Burushaskian
15	Canela-Kraho	Amazonia	Macro-Ge
16	Chamorro	Papunesia	Austronesian
17	Chukchi	Eurasia	Chukotkan
18	Cree (Plains)	North America	Algic
19	Daga	Papunesia	Dagan
20	Dani (Lower Valley)	Papunesia	Trans-New Guinea
21	English	Eurasia	Indo-European
22	Fijian	Papunesia	Austronesian
23	Finnish	Eurasia	Uralic
24	French	Eurasia	Indo-European
25	Georgian	Eurasia	Kartvelian
26	German	Eurasia	Indo-European
27	Gooniyandi	Australia	Bunuban
28	Grebo	Africa	Niger-Congo
29	Greek (Modern)	Eurasia	Indo-European
30	Greenlandic (West)	Eurasia	Eskimo-Aleut
31	Guarani	South America	Tupian
32	Hausa	Africa	Afro-Asiatic
33	Hebrew (Modern)	Eurasia	Afro-Asiatic

34	Hindi	Eurasia	Indo-European
35	Hixkaryana	Amazonia	Cariban
36	Hmong Njua	South East	Hmong-Mien
37	Imonda	Papunesia	Border
38	Indonesian	Papunesia	Austronesian
39	Jakaltek	Mesoamerica	Mayan
40	Japanese	Eurasia	Japonic
41	Kannada	Eurasia	Dravidian
42	Karok	North America	Karokian
43	Kayardild	Australia	Tangkic
44	Kewa	Papunesia	Trans-New Guinea
45	Khalkha	Eurasia	Altaic
46	Khoekhoe	Africa	Khoisan
47	Kiowa	North America	Tanoan
48	Koasati	North America	Muskogean
49	Korean	Eurasia	Koreanic
50	Koyraboro Senni	Africa	Nilo-Saharan
51	Krongo	Africa	Kaduglian
52	Kutenai	North America	Salish
53	Lakhota	North America	Siouan
54	Lango	Africa	Nilo-Saharan
55	Lavukaleve	Papunesia	East Papuan
56	Lezgian	Eurasia	Nakh-Daghestanian
57	Luvale	Africa	Niger-Congo
58	Makah	North America	Wakashan
59	Malagasy	Africa	Austronesian
60	Mandarin	South East	Sino-Tibetan
61	Mangarrayi	Australia	Mangarrayian
62	Mapudungun	South America	Araucanian
63	Maricopa	North America	Hokan
64	Martuthunira	Australia	Pama-Nyungan
65	Maung	Australia	Iwaidjan
66	Maybrat	Papunesia	West Papuan
67	Meithei	South East	Sino-Tibetan
68	Mixtec (Chalcatongo)	Mesoamerica	Oto-Manguean
69	Ngiyambaa	Australia	Pama-Nyungan
70	Oneida	North America	Iroquoian
71	Oromo (Harar)	Africa	Afro-Asiatic
72	Otomi (Mezquital)	Mesoamerica	Oto-Manguean
73	Paiwan	Papunesia	Austronesian
74	Persian	Eurasia	Indo-European
75	Piraha	Amazonia	Mura
76	Quechua (Imbabura)	South America	Quechuan
77	Rama	Mesoamerica	Chibchan
78	Rapanui	Papunesia	Austronesian

79	Russian	Eurasia	Indo-European
80	Sango	Africa	Niger-Congo
81	Sanuma	Amazonia	Yanomam
82	Slave	North America	Na-Dene
83	Spanish	Eurasia	Indo-European
84	Supyire	Africa	Niger-Congo
85	Swahili	Africa	Niger-Congo
86	Tagalog	Papunesia	Austronesian
87	Thai	South East	Tai-Kadai
88	Tiwi	Australia	Tiwian
89	Tukang Besi	Papunesia	Austronesian
90	Turkish	Eurasia	Altaic
91	Vietnamese	South East	Austro-Asiatic
92	Warao	South America	Waraoan
93	Wari	Amazonia	Chapacuran
94	Wichita	North America	Caddoan
95	Wichi	South America	Matacoan
96	Yagua	Amazonia	Peba-Yaguan
97	Yaqui	Mesoamerica	Uto-Aztecan
98	Yoruba	Africa	Niger-Congo
99	Zoque (Copainala)	Mesoamerica	Mixe-Zoque
100	Zulu	Africa	Niger-Congo

Appendix 2: Complexity variables

Code	Language	Phonology	Morphology	Syntax	Lexicon
1	Abkhaz	1	1	0	1
2	Acoma	1	1	1	0
3	Alamblak	1	0	0	1
4	Amele	0	1	0	0
5	Apurina	0	1	0	0
6	Arabic (Egyptian)	1	0	0	0
7	Arapesh (Mountain)	1	0	0	1
8	Asmat	0	1	0	0
9	Bagirmi	1	0	0	1
10	Barasano	1	0	1	0
11	Basque	0	0	0	1
12	Berber	1	0	0	0
13	Burmese	1	0	0	0
14	Burushaski	1	0	0	0
15	Canela-Kraho	1	1	0	0

16	Chamorro	0	0	0	1
17	Chukchi	0	1	1	0
18	Cree (Plains)	0	0	1	1
19	Daga	0	1	0	0
20	Dani (Lower Valley)	1	1	0	0
21	English	1	0	1	0
22	Fijian	0	0	1	0
23	Finnish	1	0	1	0
24	French	1	0	1	0
25	Georgian	1	0	1	0
26	German	1	0	1	0
27	Gooniyandi	0	0	1	0
28	Grebo	1	0	0	0
29	Greek (Modern)	0	0	1	0
30	Greenlandic (West)	0	1	0	0
31	Guarani	0	1	0	1
32	Hausa	1	0	0	1
33	Hebrew (Modern)	0	0	0	0
34	Hindi	1	0	0	0
35	Hixkaryana	0	1	0	0
36	Hmong Njua	1	0	0	0
37	Imonda	1	0	0	0
38	Indonesian	0	0	0	1
39	Jakaltek	1	1	0	0
40	Japanese	1	0	0	0
41	Kannada	1	0	0	0
42	Karok	1	1	1	1
43	Kayardild	0	1	1	0
44	Kewa	1	0	0	0
45	Khalkha	1	0	0	0
46	Khoekhoe	1	0	0	1
47	Kiowa	1	1	0	0
48	Koasati	1	1	0	1
49	Korean	1	0	0	0
50	Koyraboro Senni	0	0	0	1
51	Krongo	1	0	0	0
52	Kutenai	1	1	1	1
53	Lakhota	0	0	0	1
54	Lango	1	0	0	0
55	Lavukaleve	0	0	0	1
56	Lezgian	1	0	0	0
57	Luvale	1	0	0	0

58	Makah	1	1	0	1
59	Malagasy	1	0	0	1
60	Mandarin	1	0	0	0
61	Mangarrayi	0	1	0	0
62	Mapudungun	0	1	0	1
63	Maricopa	0	1	0	0
64	Martuthunira	0	0	0	1
65	Maung	0	1	0	0
66	Maybrat	0	0	0	1
67	Meithei	1	0	0	1
68	Mixtec	1	0	0	0
69	Ngiyambaa	0	0	0	0
70	Oneida	1	1	1	1
71	Oromo (Harar)	1	0	0	0
72	Otomi (Mezquital)	1	0	0	0
73	Paiwan	0	0	1	1
74	Persian	0	0	0	0
75	Piraha	1	1	0	0
76	Quechua (Imbabura)	1	0	0	0
77	Rama	0	0	0	0
78	Rapanui	0	0	0	1
79	Russian	1	0	1	0
80	Sango	1	0	0	1
81	Sanuma	1	1	0	0
82	Slave	1	1	0	0
83	Spanish	0	0	1	1
84	Supyire	1	0	0	1
85	Swahili	1	0	0	0
86	Tagalog	0	0	0	0
87	Thai	1	0	0	0
88	Tiwi	0	1	0	0
89	Tukang Besi	0	0	0	1
90	Turkish	1	0	0	0
91	Vietnamese	1	0	0	1
92	Warao	0	0	1	0
93	Wari	0	0	0	0
94	Wichita	0	1	1	0
95	Wichi	0	0	0	1
96	Yagua	1	1	1	0
97	Yaqui	1	1	0	0
98	Yoruba	1	0	0	0
99	Zoque (Copainala)	0	1	0	1

100	Zulu	1	0	0	0
Tot	**Total**	60	32	22	33

Synergetic Studies on Chinese Lexical Structure

Lu Wang

School of Foreign Languages, Dalian Maritime University, Dalian, 116026, China
wanglu-chn@hotmail.com

Abstract. This paper reports on a test of some aspects of the synergetic-linguistic model using data from Chinese. The lexical model, which is mainly based on the four properties word length, polysemy, frequency and polytextuality, is tested with data from the *People's Daily* news corpus. The results of this study demonstrate once more the cross-linguistic validity of the model. Furthermore, the properties of the model are classified into static type (word length and polysemy) and dynamic type (frequency and polytextuality). From this point of view, another static property, polyfunctionality (grammatical ambiguity), is adopted and its relationships with the above four properties are tested. Results show that polyfunctionality is strongly related with the two static properties, the decisive effect from polyfunctionality on dynamic properties is robust and the regulating effect from dynamic properties is relatively fluctuant.

***Keywords**: synergetic linguistics, Chinese, polyfunctionality*

1. Synergetic linguistics

Synergetic linguistics adopts quantitative conceptions and systems theoretical approach. It considers languages as self-organizing and self-regulating systems, focusing on the spontaneous rise and the development of structures, integrating existing mechanisms and processes into a dynamic system (Köhler, 2005b).

1.1 Synergetic linguistics as a scientific theory

Synergetic linguistics marks the first attempt to set up a scientific theory in the domain of linguistics. Theory, as a scientific term, is defined as a system of interrelated, universally valid laws without which explanation is not possible (Altmann 1993; Bunge 1967). Rather than descriptive approaches, individual concepts, or isolated laws and hypotheses, there is a first attempt at combining linguistic laws, which have been found in the frame work of quantitative linguistics, into a system of interconnected universal statements, thus forming a theory of language: synergetic linguistics (cf. Köhler 1986, 1987, 1993, 1999).

The main concerns of synergetic linguistics are to "provide a framework for linguistic theory building, i.e. a modeling approach which can be used to set up universal hypotheses by deduction from theoretical considerations, to test them, combine them into a network of laws and law-like statements, and explain the phenomena observed" (Köhler, 2005b); and to "re-establish the

view on language that has been lost for decades: the view of language as a psycho-social and as a biological-cognitive phenomenon at the same time (the emphasis that the cognitive paradigm has put on the latter aspects has almost completely displaced the former one in linguistics)" (Köhler, 2005b).

1.2 Synergetic approach and linguistic application

The synergetic approach, which is a specific branch of systems theory, is characterized as an interdisciplinary approach to the modeling of certain dynamic aspects of systems, which occur in different disciplines at different objects of investigation in an analogous way. Its focus is on the spontaneous rise and the development of structures (Köhler 2005a).

From the quantitative point of view, language is determined neither absolutely by chance nor absolutely by necessity, but by a combination of both aspects. The synergetic approach offers appropriate concepts and models in accordance with this perspective. "A characteristic property of self-organizing systems is the existence of cooperative (and competing) processes, which, together with external factors, constitute the dynamics of the system. Other crucial elements of synergetics are the enslaving principle and the order parameters: If a process A dynamically follows another process B it is called enslaved by B; order parameters are macroscopic entities which determine the behaviour of the microscopic mechanisms without being represented at their level themselves" (Köhler 2005b).

"The explanatory power of synergetic models is based on the process-oriented approach of synergetics. The modeling procedure starts from known or assumed mechanisms and processes of the object under study and formulates them by means of appropriate mathematical expressions (e. g. differential equations). The system's behavior can then be derived from the relations between the processes and the controlling order parameters. The possibility to form new structures is essentially connected with the existence of fluctuations, which make up the motor of evolution. The possible system states ("modes") which can occur (driven by those fluctuations) on the basis of the relations described by the equations are limited by the boundary conditions and order parameters. Only those modes can prevail in their competition with other ones which fit with these limitations. In self-organizing systems, the prevailing modes are those which contribute in some way or other to the function of the system" (Köhler 2005b).

1.3 The logic of explanation

As linguistic explanation is not likely to be possible by means of causal relations, synergetic linguistics aims at functional explanation (Köhler 2005b). According to the results of the philosophy of science, there is one widely accepted type of explanation: the deductive-nomologic one (Hempel 1965),

which can be illustrated by the following scheme:

$$\left.\begin{matrix} G_1, G_2, G_3, ..., G_n \\ S_1, S_2, S_3, ..., S_m \end{matrix}\right\} \quad \textit{Explanans}$$

$$E \qquad\qquad \textit{Explanandum}$$

where the G_i are laws, the S_i is boundary conditions and *E* is the proposition to be explained. The scheme shows that *E* is explained if it can be logically deduced from laws and boundary conditions. A functional explanation of a linguistic phenomenon E_f can then be pursued according to the following scheme:

(1) The system *S* is self-organizing. For each need, it possesses mechanisms to alter its state and structure in such a way that the need is met.

(2) The needs $N_1 \ldots N_k$ have to be met by the system.

(3) The need *N* can be met by the functional equivalents[1] $E_1 \ldots E_f \ldots E_n$.

(4) The interrelation between those functional equivalents which are able to meet the need *N* is given by the relation R_N $(E_{N1} \ldots E_{Nn})$.

(5) The structure of the system *S* can be expressed by means of the relation *Q* $(s_1 \ldots s_m)$ among the elements s_i of the system.

During the last few decades, a number of linguistic laws have been found, some of which could successfully be integrated into a general model by the abovementioned method. Thus, synergetic linguistics may be considered as a embryonic linguistic theory. In synergetic linguistics, the pre-conditions and the procedure of scientific explanation in linguistics are therefore reflected with particular attention.

1.4 The modeling

Within the framework of synergetic linguistics, modelling proceeds iteratively in refining phases, with each phase consisting of six individual steps (Köhler, 1986).

(1) In the first step, axioms are set up for the subsystem under consideration.

(2) The second step is the determination of system levels, units, and variables which are of interest to the current investigation.

(3) In step three, relevant consequences, effects, and interrelations are determined.

---

[1] E_f is an element of the system *S* with load R_{Nf}.

(4) The fourth step consists of the search for functional equivalents and multifunctionalities. In language, there are not only 1:1 correspondences — many relationships are of the 1:n or m:n type. This fact plays an important role in the logics of functional explanation.
(5) Step five is the mathematical formulation of the hypotheses set up so far — a precondition for any rigorous test.
(6) Step six is the empirical test of these mathematically formulated hypotheses.

2. The synergetic lexical model

2.1 The lexical model

The first synergetic-linguistic model was set up for a lexical subsystem and tested on German texts by Köhler (1986), as illustrated by Figure 1 and Table 1. Hereafter, this theory was tested not only on lexical structures, such as from Polish (Hammerl, 1991) and English (Gieseking, 1998) but also on morphological phenomena (Köhler, 1990a, 1990b, 1991; Krott 1996, 1998) and syntactic subsystems (Köhler, 1999, 2012). Moreover, the attempt to apply it in musicology again showed the wide applicability (Köhler, Martináková, 1998).

The original lexical control circuit (Köhler, 1986) is illustrated by Figure 1 and Table 1. In this modeling approach, the representation of structures and functions are expressed by a graphical notation, where rectangles correspond to system variables (state and control variables), circles symbolize requirements, squares represent operators, and arrows stand for effects or bonds. The squares contain symbols for the operator types, which are, in most cases, proportionality operators in form of either (symbols for) numerical values or only the signs (+ or -) of their values. Quantities which are arranged on a common edge are multiplied and junctions correspond to numerical addition (according to the rules of operator algebra and graph theory).

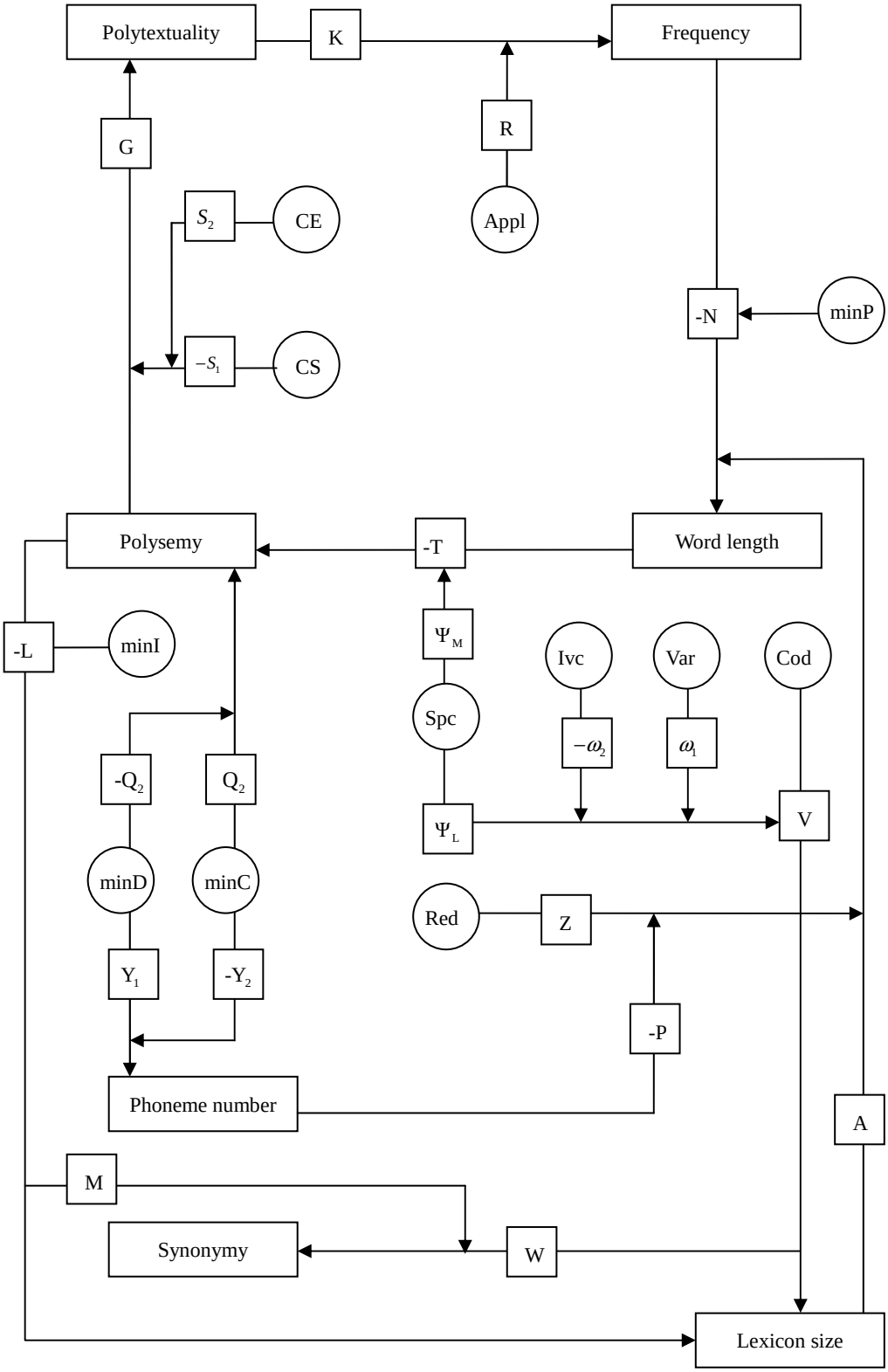


Figure 1. Diagram showing the structure of a lexical subsystem (taken from Köhler, 2005b).

Table 1
Requirements (taken from Köhler, 2005b)

Requirement	**Symbol**	**Influence on**
Coding	Cod	Size of inventories
Specification	Spc	Polysemy
De-specification	Dsp	Polysemy
Application	Usg	Frequency
Transmission security	Red	Length of units
Economy	Ec	Sub-requirements
Minimisation of production effort	minP	Length, complexity
Minimisation of encoding effort	minC	Size of inventories, polysemy
Minimisation of decoding effort	minD	Size of inventories, polysemy
Minimisation of inventories	minI	Size of inventories
Minimisation of memory effort	minM	Size of inventories
Context economy	CE	Polytextuality
Context specifity	CS	Polytextuality
Invariance of the expression-meaning-relation	Inv	Synonymy
Flexibility of the expression-meaning-relation	Var	Synonymy
Efficiency of coding	OC	Sub-requirements
Maximisation of complexity	maxC	Syntactic complexity
Preference of right branching	RB	Position
Limitation of embedding depth	LD	Depth of embedding
Minimisation of structural information	minS	Syntactic patterns
Adaptation	Adp	Degree of adaptation readiness
Stability	Stb	Degree of adaptation readiness

There are many elements linked to the lexical system. However, they influence the system by way of order parameters, instead of immediate impact. Therefore, those elements are not taken into consideration in the present investigation. The following four elements are the core properties of the lexical structure. The relationships between them are shown by a simplified lexical control circle in Figure 2. Arrows stand for effects and their direction. The plus sign indicates a positive effect, while minus indicates a negative effect. Solid lines represent direct relationships between two properties, while a dotted line indicates indirect ones.

Polysemy (PL): the number of different meanings a word carries;
Polytextuality (PT): the number of contexts in which a word occurs;
Frequency (F): the number of occurrences of a word;
Word length (L): the number of syllables of a word.

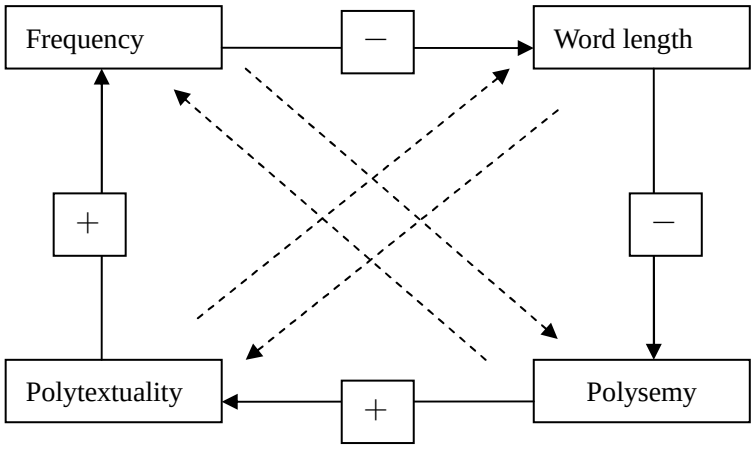


Figure 2. A simplified synergetic model on the lexical level
(taken from Wang, 2014).

2.2 Formulation

The mathematical relationship between each pair of directly linked properties (the quantities) is formulated by a differential equation (Köhler 1986, 2005b) of the form

$$\frac{dy}{y} = b \times \frac{dx}{x}.$$

It expresses that the relative rate of change of the variable y is proportional to the relative rate of change of the variable *x* (Altmann, Köhler, 1995). The solution yields

$$y = ax^b,$$

the well-known power function. To be specific, we will test the following hypotheses:

(1) Polysemy: $PL = aL^b$.
(2) Polytextuality: $PT = aPL^b$.
(3) Frequency: $F = aPT^b$.
(4) Length: $L = aF^b$.

2.3 Testing

The direct relationships between properties are relatively isolated. Therefore, the correlations of indirect properties will also be utilized to test the model further. We will evaluate the degree of disagreement between the theoretical predictions and empirically estimated results.

First, the theoretical function of indirect relationships will be derived from the direct ones. For example, to obtain the relationship between word length and polytextuality, we substitute *PL* of Function (1) $PL = aL^b$ into Function (2) $PT = aPL^b$:

$$\left.\begin{aligned} PL &= a_1 L^{b_1} \\ PT &= a_2 PL^{b_2} \end{aligned}\right\} \Rightarrow PT = a_2 (a_1 L^{b_1})^{b_2} \Rightarrow PT = a_2 a_1^{b_2} L^{b_1 b_2}$$

let $a_t = a_2 a_1^{b_2}$, $b_t = b_1 b_2$,

$$PT_t = a_t L^{b_t} .$$

Thus, the theoretical model function of the indirect relations should be:

$$Y_t = a_t X^{b_t} .$$

Second, we fit empirical data to power function to obtain the empirical function:

$$Y_e = a_e X^{b_e} \text{-}$$

Finally, the t-test will be employed to examine if the differences between the estimated parameters are significant.

3. Results of testing the model

We applied the lexical model to data from the *People's Daily* news corpus, a Chinese one-million words news corpus (Wang, 2014). The results are shown in Table 2.

Table 2
Results of the original synergetic lexical model on original data

Relationship	Functions	a	b	R^2	t-test
Direct	$PT = aPL^b$	0.0138	3.4286	0.7864	
	$PL = aL^b$	4.7892	-1.2924	0.9480	
	$F = aPT^b$	1.2653	1.0572	0.9790	
	$L = aF^b$	2.0795	-0.0383	0.3515	
Indirect	$PL = aF^b$	1.6440	0.1342	0.2292	0.000
	$F = aPL^b$	0.0094	4.0793	0.5292	0.193
	$L = aPT^b$	2.0960	-0.0479	0.5362	0.398

	$PT = aL^b$	20.5895	-1.488	0.9510	0.094
Double indirect	$L = aPL^b$	2.2726	-0.351	0.9354	0.000
	$F = aL^b$	52.9015	-2.0511	0.9711	0.173
	$PT = aF^b$	0.7357	0.9644	0.9798	0.000
	$PL = aPT^b$	1.6459	0.1638	0.2874	0.000

Among the direct relationships, $PL = aL^b$ and $F = aPT^b$ yield perfect fitting results with R^2 values up to 0.948 and 0.979. With a lower R^2 = 0.7864, $PT = aPL^b$ also abides by the power law, whereas $L = aF^b$ obtains R^2 = 0.3515 because of the fluctuation phenomenon which is found in the languages which were considered so far, i.e. German (Köhler, 1986), Polish (Hammerl, Sambor, 1993) and English (Gieseking, 1998).

The indirect relationships are adopted to test the model thoroughly, by way of evaluating the deviation between the theoretical functions and the empirically estimated functions. There will be some neglected independent variables if the parameters are opposite (increasing or decreasing effect). Despite a lower goodness-of-fit, the results do not confute the model.

Static vs. Dynamic

Considering the four properties, syllable length and polysemy are given in the dictionary and corpus-independent. They are static properties, which represent the two basic sides of words (sound and meaning). On the other hand, frequency and polytextuality are obtained only from running texts. They reflect the dynamic application of words. We will therefore evaluate our results from this point of view: hypothesis (1) $PL = aL^b$ (a static property acts on another one) and (3) $F = aPT^b$ (a dynamic property acts on another one) fit perfectly, 0.948 and 0.979 respectively; while hypothesis (2) $PT = aPL^b$ (a static property acts on a dynamic one) and (4) $L = aF^b$ (a dynamic property acts on a static one) are relatively poor with the goodness-of-fitting values 0.7864 and 0.3515. Therefore, the following interpretation seems to be possible: the interrelation between a pair of congeneric properties is stronger than a mixed pair.

For the mixed pairs, we still need to distinguish between two kinds: a dynamic property which depends on a static one (static → dynamic) or the reverse (dynamic → static). The former gets relatively better fitting results. However the latter always displays diverging data points around the theoretical line, regardless of whether it is a direct or indirect relationship. To be specific, we consider functions (4) $L = aF^b$, (6) $L = aPT^b$, (8) $PL = aF^b$ and (11) $PL = aPT^b$. Another observation is that the decisive effect of a static property on a dynamic property is stronger than the opposite effect. All the functions in this study are shown in Table 3 according to the types of their relationships.

Table 3
Relation categorization from static and dynamic viewpoints

Relation type	**Function**	**R^2**
Static → static	$PL = aL^b$ (1) $L = aPL^b$ (10)	0.948 0.9354
Dynamic → dynamic	$F = aPT^b$ (3) $PT = aF^b$ (12)	0.979 0.9798
Static → dynamic	$PT = aPL^b$ (2) $PT = aL^b$ (5) $F = aPL^b$ (7) $F = aL^b$ (9)	0.7864 0.95 0.5292 0.9711
Dynamic → static	$L = aF^b$ (4) $L = aPT^b$ (6) $PL = aF^b$ (8) $PL = aPT^b$ (11)	0.3515 0.5362 0.2292 0.2874

Reflection vs. Regulation

Actually, the static → dynamic type is the reflection relationship: a dynamic property reflects the feature of a static property in application. The dynamic → static relationship is not a simple dependent relation; it is connected with a regulating mechanism within the language system itself.

Because language is a psycho-social and biological-cognitive phenomenon (Köhler, 2005b), human speech acts change with the development of the language-external world. Language, however, is a self-regulating system. When the balance of this system is disturbed, a compensating behavior will be triggered by itself in order to reach a new steady state. In the lexical control circuit, at first, dynamic properties vary which then cause the words to change. Among others, word length and polysemy change under this influence. But words also resist the reduction to ensure transmission security and maintain lexicon size (as a negative feedback). This feedback always causes delays in the regulating process.

4. Polyfunctionality and properties in the synergetic lexical model

Polyfunctionality refers to the grammatical ambiguity of a word, which represents the number of parts of speech the word can act as. For example, the polyfunctionality values of *希望* (v. n.), *左右* (n. v. adv.) and *同* (v. adj. adv. prep. conj. n.) are two, three and six. Such polyfunctionality values are obtained by using the *Modern Chinese Dictionary* (5th edition).

Now we explore the relationships between polyfunctionality and the properties of the synergetic lexical model. From the "static vs. dynamic" point of view (see 3.2), polyfunctionality belongs to the static category. Therefore, its relationship with static properties (both forward and reversed directions) should abide by a power law. From the "reflection vs. regulation" point of view (see 3.2), its relationships with dynamic properties should also abide by a power function. Fluctuations will appear in the regulating process, namely the *F→PF* and *PT→PF* relationships.

4.1 Polyfunctionality and static properties

The results of the fittings are shown in Table 4 and in Figures 3 and 4, and the data in Appendix.

Table 4
Fitting the power law to the data of polyfunctionality-polysemy and polyfunctionality-length

Function	**A**	**b**	**R^2**
$PL = aPF^b$	1.3171	1.2657	0.9969
$PF = aPL^b$	0.9452	0.4971	0.9931
$L = aPF^b$	2.0425	-0.4595	0.9512
$PF = aL^b$	1.7284	-0.4559	0.8778

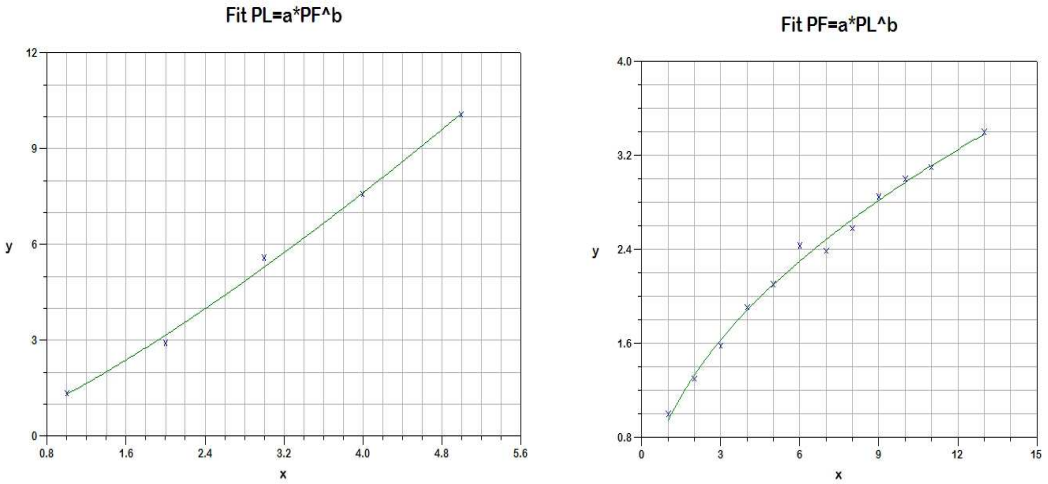


Figure 3. Fitting the power law to the data of polyfunctionality-polysemy.

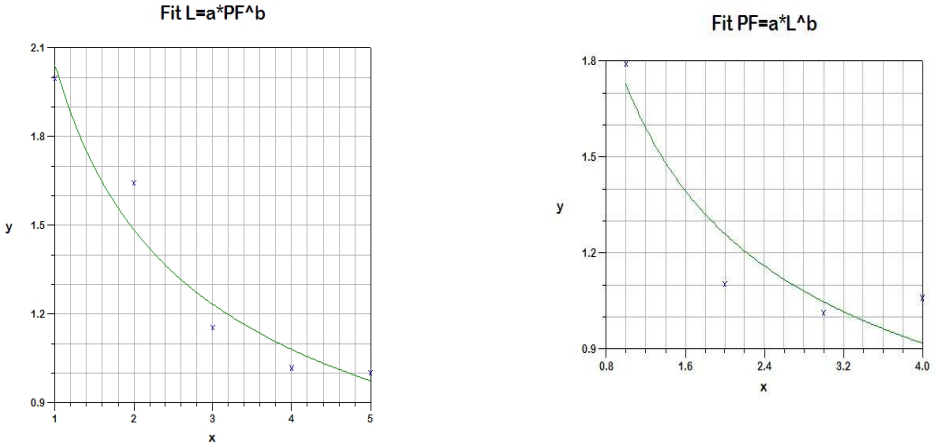


Figure 4. Fitting the power law to the data of polyfunctionality-length.

4.2 Polyfunctionality and dynamic properties

The results of the fittings are shown in Table 5 and in Figures 5 and 6, and the data in Appendix.

Table 5
Fitting the power law to the data of polyfunctionality-polytextuality and polyfunctionality-frequency

Function	A	b	R^2
$PT = aPF^b$	2.7220	2.4161	0.9613
$PF = aPT^b$	1.0470	0.0944	0.3898
$F = aPF^b$	9.7380	2.1764	0.9933
$PF = aF^b$	1.0311	0.0847	0.3815

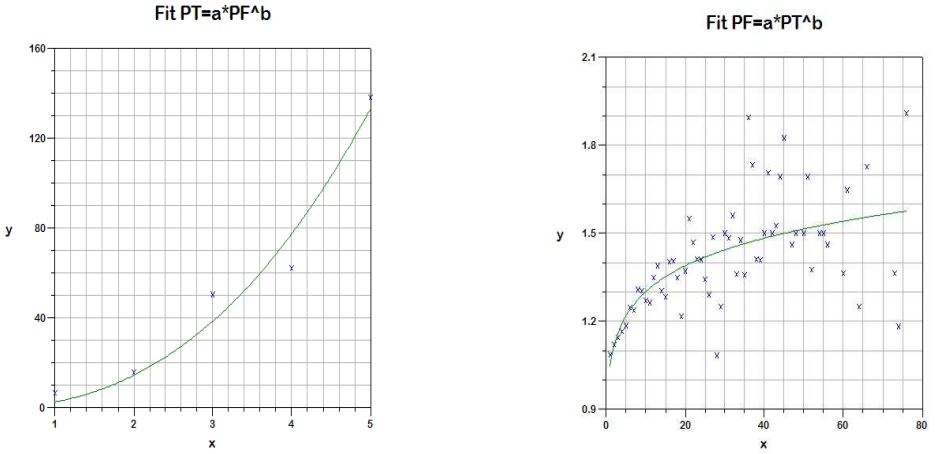


Figure 5. Fitting the power law to the data of polyfunctionality-polytextuality.

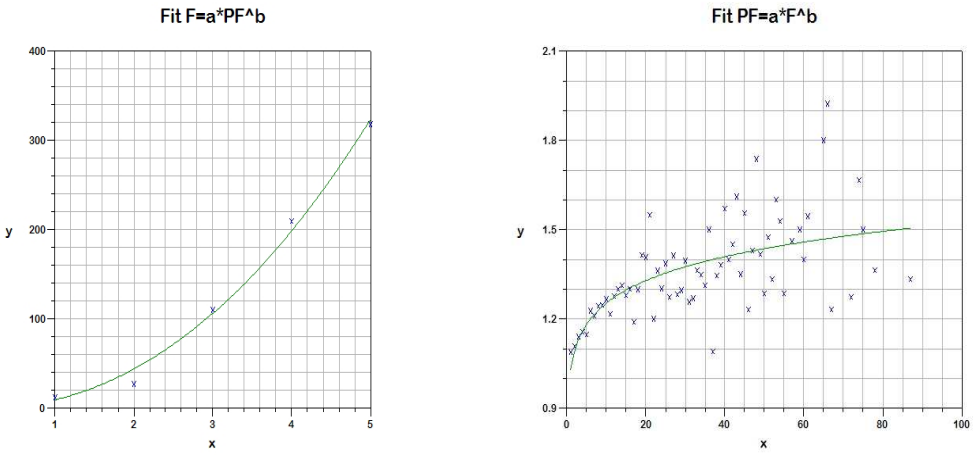


Figure 6. Fitting the power law to the data of polyfunctionality-frequency

The *PF* ↔ *PL* and *PF* ↔ *L* relationships all abide by the power law with very good results, which confirms to the static → static relationship type. Polyfunctionality increases polysemy and vice versa, whereas word length and polyfunctionality restrain the growth of each other. Unlike dynamic properties, there must be at least one property that restrains the growth among static → static relationships. Any static properties should not increase infinitely because of boundary conditions such as minD and minM (see Table 1, Figure 1). Here word length plays the role of restricting the growth of polysemy and polyfunctionality.

The relationships with dynamic properties also agree with our assumption. The reflection process shows excellent fitting results (*PF*→*PT* and *PF*→*F* obtain R^2 = 0.9613 and R^2 = 0.9933 respectively). The regulating processes show some fluctuations similar with the *F*→*L* and *PT*→*PL* relationships illustrated in 3.2. A possible reason for this result is that static properties have the utmost limit because of minM and minD, thus their values do not rise along with the growth of dynamic properties. When the utmost is reached, they rebound with a drop (sometimes a sharp drop) in value; then they will be raised by the dynamic ones until the limit is touched again.

The following Figure 7 shows the effect of polyfunctionality in the lexical control circuit.

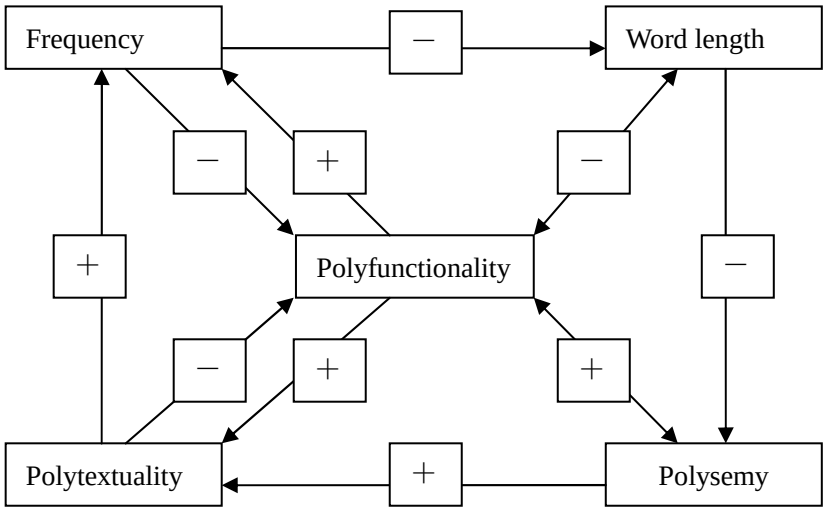


Figure 7. The effect of polyfunctionality in the lexical control circuit.

5. Discussion and Conclusions

In this study, the synergetic-linguistic model has been successfully tested on data from Chinese, a highly analytical language. Some fitting results seem to be unsatisfying if we only focus on the goodness-of-fit values. However, the deviations can be interpreted as being caused by the system's self-regulating behavior, which conforms exactly with the synergetic theory.

Polyfunctionality as a static property has been employed in the synergetic model. As assumed, polyfunctionality is highly dependent on the other static properties; this is reflected by dynamic properties with perfect results and regulated with dispersions.

From the static vs. dynamic point of view, the relationships are classified and listed in the following Table 6.

Table 6
Relation categorization from static and dynamic viewpoint

Relation type	**Function**	**Effect**	**R^2**
Static → static	$PL = aL^b$	-	0.948
	$L = aPL^b$	-	0.9354
	$L = aPF^b$	-	0.9512
	$PF = aL^b$	-	0.8778
	$PL = aPF^b$	+	0.9969
	$PF = aPL^b$	+	0.9931

Dynamic → dynamic	$F = aPT^b$	+	0.979
	$PT = aF^b$	+	0.9798
Static → dynamic (reflection)	$PT = aPL^b$	+	0.7864
	$PT = aL^b$	-	0.95
	$F = aPL^b$	+	0.5292
	$F = aL^b$	-	0.9711
	$PT = aPF^b$	+	0.9613
	$F = aPF^b$	+	0.9933
Dynamic → static (regulation)	$L = aF^b$	-	0.3515
	$L = aPT^b$	-	0.5362
	$PL = aF^b$	+	0.2292
	$PL = aPT^b$	+	0.2874
	$PF = aPT^b$	+	0.3898
	$PF = aF^b$	+	0.3815

(1) The static → static type

Among the 3 static properties, polysemy and polyfunctionality increase each other whereas word length decreases them. The 3 properties each stand for an aspect of static lexical knowledge of language. In the case of learning a new word, one should remember its written form, pronunciation, meaning(s) and part(s) of speech. Since a human brain's memory size must be finite, such static knowledge of a word should not exceed a certain boundary. The demands such as minC, minD and minM also require a limitation on the growth of static properties. Therefore, there must be one property (or at least one) that restricts it.

(2) The dynamic → dynamic type

Frequency and polytextuality are closely connected with each other.

(3) The static → dynamic type

With the adoption of polyfunctionality, it can be seen that the decisive effect of a static property on a dynamic property is robust.

(4) The dynamic → static type

The dynamic → static relationship is not a simple dependency relationship; it is connected with a regulating mechanism within the language system itself. When the balance of this system is disturbed, a compensating behavior will be triggered by itself in order to reach a new steady state. In the lexical control circuit, at first, dynamic properties vary which then causes the words to change. Among others, word length and polysemy change under this influence. But words also resist the reduction to ensure transmission security and maintain

lexicon size (as a negative feedback). This feedback always causes delays in the regulating process.

The synergetic linguistic theory studies language also from a diachronic point of view. However, in the individual empirical test steps we only obtained synchronic data. Therefore, what can be observed from such data is a section in the diachronic development of language. At that certain moment, language, as a dynamic system, is regulating itself (or some parts of it) to reach a steady state. This is another reason why some parts of the system are relatively fluctuant, while other parts seem to be relatively stable.

References

Altmann, G. (1993). Phoneme counts. *Glottometrika 14,* 55-70.

Altmann, G., Köhler, R. (1995). "Language Forces" and Synergetic Modelling of Language Phenomena. In: *Glottometrika 15*, 62-76.

Bunge, M. (1967). *Scientific Research I, II*. Berlin, Heidelberg, New York: Springer.

Gieseking, K. (1998). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (Eds.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik:* 387-433. Trier: Wissenschaftlicher Verlag Trier.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines Basismodells*. Trier: Wissenschaftlicher Verlag.

Hammerl, R., Sambor, J. (1993). Synergetic studies in Polish. In: Köhler, R., Rieger, B. (eds.), *Contributions to quantitative linguistics:* 11-31. Dordrecht: Klewer Academic Publishers.

Hempel, G. (Eds) (1965). *Aspects of Scientific Explanation.* New York: Free Press.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brokmeyer.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics 14*, 241-257.

Köhler, R. (1990a). Elemente der synergetischen Linguistik. In: Hammerl, R. (Eds.), *Glottometrika 12*, 179-187.

Köhler, R. (1990b). Synergetik und sprachliche Dynamik. In: Koch, W. A. (eds.), *Natürlichkeit der Sprache und Kultur:* 96-112, Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of Coding Methods in Grammar. In: Rothe, U. (ed.), *Diversification Processes in Language: Grammar* (pp. 47-55). Hagen: Rottmann.

Köhler, R. (1993). Synergetic linguistics. In: Köhler, R., Rieger, B. (eds.): *Contributions to Quantitative Linguistics:* 41-51. Dordrecht: Kluwer,.

Köhler, R. (1999). Syntactic Structures. Properties and Interrelations. *Journal of Quantitative Linguistics 6*, 46-57.

Köhler, R. (2005a). Introduction of quantitative linguistics In: Köhler, R., Altmann, G., Piotrowski, G. (eds.), *Quantitative Linguistics:* 1-16. Berlin: de Gruyter.

Köhler, R. (2005b). Synergetic linguistics In: Köhler, R., Altmann, G. & Piotrowski, G. (**e**ds.), *Quantitative Linguistics:* 760-774, Berlin: de Gruyter.

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

Köhler, R., Martináková, Z. (1998). A Systems Theoretical Approach to Language and Music. In: Altmann, G., Koch, W. A. (eds.), *Systems. A new paradigm for the human sciences:* 514-546, Berlin, New York: Walter de Gruyter.

Krott, A. (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics 3*, 29-37.

Krott, A. (1998). Ein funktionalanalytisches Modell der Wortbildung. In R. Köhler (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik,* Trier: Wissenschaftlicher Verlag Trier.

Wang, L. (2014). Synergetic Studies on Some Properties of Lexical Structures in Chinese. *Journal of Quantitative Linguistics 22*, 177-179.

Appendix

$PL = aPF^b$

PF	PL	Power function
1	1.3359	1.31717
2	2.9109	3.16709
3	5.5819	5.29105
4	7.5667	7.61516
5	10.0625	10.1004
a = 1.3171, b = 1.2657, R^2 = 0.9969		

$PF = aPL^b$

PL	PF	Power function
1	1.0000	0.9452
2	1.2981	1.3341
3	1.5804	1.6320
4	1.9086	1.8830
5	2.1044	2.1039
6	2.4324	2.3035
7	2.3875	2.4869
8	2.5789	2.6576

9	2.8485	2.8179
10	3.0000	2.9694
11	3.1000	3.1135
13	3.4000	3.3831
a = 0.9452, b = 0.4971, R^2 = 0.9931		

$L = aPF^b$

PF	L	Power function
1	1.9970	2.0425
2	1.6430	1.4853
3	1.1538	1.2327
4	1.0167	1.0800
5	1.0000	0.9748
a = 2.0425, b = -0.4595, R^2 = 0.9512		

$PF = aL^b$

L	PF	Power function
1	1.7905	1.7284
2	1.1021	1.2601
3	1.0123	1.0474
4	1.0588	0.9186
a = 1.7284, b = -0.4559, R^2 = 0.8778		

$PT = aPF^b$

PF	PT	Power function
1	6.4818	2.7220
2	15.9045	14.5288
3	50.4582	38.6979
4	62.1000	77.5453
5	138.3125	132.9540
a = 2.7220, b = 2.4161, R^2 = 0.9613		

$PF = aPT^b$

PT	PF	Power function	PT	PF	Power function
1	1.0857	1.0471	31	1.4839	1.4483
2	1.1198	1.1179	32	1.5600	1.4526
3	1.1441	1.1616	33	1.3600	1.4569
4	1.1658	1.1936	34	1.4762	1.4610
5	1.1851	1.2190	35	1.3571	1.4650
6	1.2459	1.2402	36	1.8947	1.4689
7	1.2380	1.2584	37	1.7333	1.4727
8	1.3079	1.2743	38	1.4118	1.4764
9	1.3042	1.2886	39	1.4091	1.4800
10	1.2706	1.3015	40	1.5000	1.4836
11	1.2635	1.3133	41	1.7059	1.4871
12	1.3478	1.3241	42	1.5000	1.4904
13	1.3893	1.3342	43	1.5263	1.4938
14	1.3039	1.3435	44	1.6923	1.4970
15	1.2816	1.3523	45	1.8235	1.5002
16	1.4023	1.3606	47	1.4615	1.5064
17	1.4054	1.3684	48	1.5000	1.5094
18	1.3492	1.3758	50	1.5000	1.5152
19	1.2174	1.3828	51	1.6923	1.5180
20	1.3710	1.3896	52	1.3750	1.5208
21	1.5500	1.3960	54	1.5000	1.5262
22	1.4688	1.4021	55	1.5000	1.5289
23	1.4107	1.4080	56	1.4615	1.5315
24	1.4103	1.4137	60	1.3636	1.5415
25	1.3421	1.4192	61	1.6471	1.5439
26	1.2903	1.4244	64	1.2500	1.5509
27	1.4865	1.4295	66	1.7273	1.5555
28	1.0833	1.4344	73	1.3636	1.5703
29	1.2500	1.4392	74	1.1818	1.5724
30	1.5000	1.4438	76	1.9091	1.5763
a = 1.0470, b = 0.0944, R^2 = 0.3898					

$F = aPF^b$

PF	F	Power function
1	12.1517	9.73805
2	26.8561	44.0189
3	110.0234	106.3870
4	209.2667	198.9790
5	318.1250	323.3880
a = 9.7380, b = 2.1764, R^2 = 0.9933		

$PF = aF^b$

F	PF	Power function	F	PF	Power function
1	1.0886	1.0312	35	1.3125	1.3939
2	1.1089	1.0936	36	1.5000	1.3972
3	1.1397	1.1318	37	1.0909	1.4005
4	1.1572	1.1598	38	1.3462	1.4036
5	1.1472	1.1819	39	1.3810	1.4067
6	1.2268	1.2003	40	1.5714	1.4098
7	1.2101	1.2161	41	1.4000	1.4127
8	1.2432	1.2299	42	1.4500	1.4156
9	1.2468	1.2423	43	1.6111	1.4184
10	1.2664	1.2534	44	1.3500	1.4212
11	1.2174	1.2636	45	1.5556	1.4239
12	1.2761	1.2730	46	1.2308	1.4266
13	1.3000	1.2816	47	1.4286	1.4292
14	1.3125	1.2897	48	1.7368	1.4317
15	1.2793	1.2973	49	1.4167	1.4342
16	1.3000	1.3044	50	1.2857	1.4367
17	1.1889	1.3111	51	1.4737	1.4391
18	1.2989	1.3175	52	1.3333	1.4415
19	1.4146	1.3235	53	1.6000	1.4438
20	1.4074	1.3293	54	1.5294	1.4461
21	1.5490	1.3348	55	1.2857	1.4483
22	1.2000	1.3401	57	1.4615	1.4527
23	1.3621	1.3451	59	1.5000	1.4570
24	1.3019	1.3500	60	1.4000	1.4591
25	1.3864	1.3547	61	1.5455	1.4611
26	1.2727	1.3592	65	1.8000	1.4690
27	1.4130	1.3636	66	1.9231	1.4709
28	1.2821	1.3678	67	1.2308	1.4728
29	1.2955	1.3718	72	1.2727	1.4818

30	1.3947	1.3758	74	1.6667	1.4852
31	1.2571	1.3796	75	1.5000	1.4869
32	1.2692	1.3833	78	1.3636	1.4919
33	1.3636	1.3869	87	1.3333	1.5057
34	1.3478	1.3905			
a = 1.0311, b = 0.0847, R^2 = 0.3815					

A Measurement of Parts of Speech in Texts Using the Noun-Based Proportion

Haruko Sanada

Abstract. The present study focuses on proportions of parts of speech in texts. Relationships between the increase of relative number of nouns (N) and those of verbs (V), adverbs or adjectives (A), conjunctions or interjections (I) are investigated, and linear functions are fitted to data from 10 Japanese texts. Noun-based proportions obtained from coefficients of these regression lines are not significantly different from empirical proportions of N, V, A, and I of the whole text. From our observation of regression lines, proportions of N, V, A and I for the whole text can be approximately obtained if we take 20% of high frequency words from the text.

Keywords: *Part of speech, noun, Kabashima's Hypothesis, Ohno-Mizutani's law, activity ratio.*

1. Outlining the aim of the paper and former studies on parts of speech

In this paper we focus on the proportion of parts of speech in texts. The proportion of parts of speech is known as an index to characterize a text, e.g. whether the text is descriptive to express facts, or rhetorical to express emotions. Busemann-Altmann's activity ratio (Busemann 1925, 1969, Altmann 1978, 1988), Ohno-Mizutani's law (Ohno, 1956, Mizutani 1965, 1989), and Kabashima's Hypothesis (Kabashima 1954, 1955, 1957, Mizutani 1983) are well known in quantitative studies on the proportions of parts of speech.

Busemann (1925, 1969) measured Q (activity ratio) with children's language data, which can be expressed as:

$$Q = \frac{a}{q} \qquad (1)$$

where a represents activity statements and q qualitative statements. However, Q shows only the ratio for a and q regardless of the size of the data, and it does not concern a percentage of the whole data. Altmann (1978, 1988) modified the index to

$$Q' = \frac{a}{a+v} \qquad (2)$$

where a is a number of adjectives and v is a number of active verbs in the text.

Ohno-Mizutani's law concerns the relationship between proportions of the parts of speech of the number of different words (type) and the style of texts. Ohno (1956) investigated the proportion of verbs and adjectives (including adjective verbs) comparing them with that of nouns using 9 classical texts of 4 types - poems, essays, diaries and novels - written from the 8th century to 14th century. This work shows two conclusions: (1) It is highly probable that the proportion of nouns has a certain relationship with the type of text. For an essay, the proportion is over 55 %, for a diary 50-55%, and for a novel under 50%. (2) The proportion of verbs and adjectives (including adjective-verbs) is greater if the proportion of nouns is smaller. Mizutani (1965, 1989) generalized Ohno's law, by stating that the proportion of verbs and adjectives (including adjective-verbs) (y) is approximately a linear function of the proportion of nouns (x) i.e.

$$y = b + ax\,. \qquad (3)$$

Kabashima's Hypothesis concerns a relationship between proportions of the parts of speech of the number of total words (token) and the style of texts. Kabashima (1954, 1955) investigated the proportion of (1) adjectives and adverbs (including adjective verbs) (p_a), (2) interjections and conjunctions (p_i), and (3) verbs (p_v) comparing them with that of nouns (p_n) using different types of texts: recorded conversations, conversations in the novel, novels whose conversations are eliminated, philosophical texts, texts on the natural science, 31-mora Japanese poems (*waka*), 17-mora Japanese poems (*haiku*), editorials of newspapers, articles of newspapers, headlines of newspapers, and articles of a dictionary. Kabashima explained that conversations contain more interjections and conjunctions (p_i) than other types of texts, and that novels contain more adjectives and adverbs (p_a) than articles of newspapers because novels describe emotions and scenes and newspapers mainly describe facts. Kabashima inductively concluded the following relationships: (1) the proportion of adjectives and adverbs (including adjective-verbs) (y_a) is approximately a linear function of the proportion of nouns (x_n), i.e.

$$y_a = b + ax_n \qquad (4)$$

where a and b are coefficients if $y_a = 100p_a$ and $x_n = 100p_n$. (2) The proportion of interjections and conjunctions (y_i) is approximately a function of the proportion of nouns (x_n), i.e.

$$\log_{10} y_i = c - d \log_{10} x_n , \qquad (5)$$

which can be expressed as

$$y_i = 10^c x_n^{-d} \qquad (6)$$

with coefficients c and d if $y_i = 100p_i$ and $x_n = 100p_n$. (3) The proportion of verbs (y_v) is approximately a function of the proportion of nouns (x_n), adjectives and adverbs (including adjective-verbs) (x_a), and interjections and conjunctions (x_i), i.e.

$$y_v = 100 - (x_n + x_a + x_i) \qquad (7)$$

if $y_v = 100p_v$, $x_n = 100p_n$, $x_a = 100p_a$ and $x_i = 100p_i$. Mizutani (1965) called these generalizations Kabashima's Indexes and tested these relationships with his own data.

In this study, these indexes are obtained for each of the 10 texts under consideration. Previous studies have shown that the indexes should not express the homogeneity of the distributions of parts of speech in a text when texts are categorized into certain genres using such indices, e.g. novels, essay, or newspaper articles. Neither should the problems of the author estimation express the homogeneity of the distributions of parts of speech in the text. However, no study has yet investigated whether distributions of parts of speech are homogeneous or not in the text.

Therefore, we investigate how the proportions of parts of speech are accumulated and reach to the indexes of the text. It is well known that the increase of tokens is correlated with that of types. Figure 1 shows an example of the curve from data of a Japanese prose. However, it is not known how the increase of verb, adjective and conjunction tokens corresponds to the increase of nominal tokens.

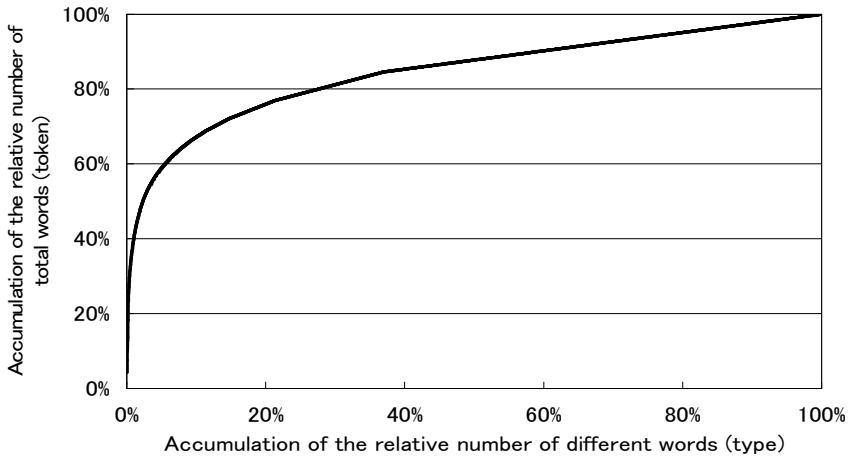


Figure 1. Increase of the proportion of tokens and types of words for the Japanese prose "Amerika Hijiki"

We follow Kabashima's Indexes which employ a proportion of nouns as a denominator. Kabashima's Indexes employ number of tokens while Ohno-Mizutani's Indexes employ the number of types, and Kabashima's Indexes are better to analyze whole texts. Kabashima's Indexes employ the number of verbs, adjectives or adverbs, and nouns while Busemann-Altmann's activity ratio employs only the number of verbs and adjectives (Altmann 1978: 93; 1988: 20), and we consider that the number of nouns must be a stable index regardless of the genre of the text. This can be called the noun-based proportion.

2. Defining parts of speech and introducing 10 Japanese texts

We analyze the proportions of parts of speech in 10 Japanese texts of prose which are given at the end of this paper. For the investigation softwares of a morphological analyzer are used, i.e. *MeCab* developed by Graduate Schools of Informatics in Kyoto University and NTT Communication Science Laboratories, and *UniDic* developed by the National Institute for Japanese Language and Linguistics. In this way we obtain vocabulary items as written in “conjugated” form, lexemes, parts of speech, the origin of the word (Japanese, Chinese, European), etc. Errors of the word boundary by the morphological analyzer are not corrected. Punctuations, blanks, parentheses, and other symbols are basically excluded except scripts which are incorrectly analyzed as "symbols" by the software. The exceptions are marked (*) in Table 1.

There is a discussion on the definition of parts of speech in Japanese. However, in the present study we employ definitions proposed by the National Institute for Japanese Language and Linguistics and the software *UniDic*. We summarize them into five groups: Noun (N), Verb (V), Adverb and Adjective (A), Interjection and conjunction (I), and others. These groups follow Kabashima's definitions. Table 1 shows a list of parts of speech and the groupings we have used for these.

Table 1
List of parts of speech and the groupings employed by the morphological analyzer *UniDic*
(N=Noun, V=Verb , A=Adverb and Adjective, I=Interjection and conjunction, O=others)

Group	Part of speech	Group	Part of speech
N	prenoun	I	interjection: filler
N	noun: proper noun: general	I	interjection: general
N	noun: proper noun:	I	conjunction

	personal name: general		
N	noun: proper noun: personal name: family name	O	postposition: case
N	noun: proper noun: personal name: first name	O	postposition: linking
N	noun: proper noun: organization	O	postposition: ending
N	noun: proper noun: place name: general	O	postposition: attributive
N	noun: proper noun: place name: country	O	postposition: conjunctive
N	noun: auxiliary verb type	O	postposition: adverbial
N	noun: numeral	O	auxiliary verb
N	noun: general: Suru ending type	O	prefix
N	noun: general: Suru ending & adjective verb type	O	suffix: adjective verb type
N	noun: general: general	O	suffix: adjective type
N	noun: general: adjective verb type	O	suffix: verb type
N	noun: general: adverb type	O	suffix: noun type: Suru ending
V	verb: general	O	suffix: noun type: general
V	verb: auxiliary verb type	O	suffix: noun type: adjective verb type
A	adjective verb: Tari ending	O	suffix: noun type: counter
A	adjective verb: general	O	suffix: noun type: adverb type
A	adjective verb: auxiliary verb type	O	sub symbol: general: script (*)
A	adjective: general	O	symbol: script (*)
A	adjective: auxiliary verb type	O	symbol: general (script) (*)
A	adverb		
A	prenoun adjectival		

(*) Scripts which are incorrectly analyzed as "symbols" by the software are included.

The selected 10 texts and their sizes are given in Table 2. The texts are collected from 100 digitalized paperbacks (Shinchosha 1995) as they are not very long or very short. In the present study the numbers of different words (types) are not lemmatized though Kabashima employed the number of lemmatized words. The numbers of total words (tokens) are categorized into Group N, V, A, I and O. We follow Kabashima's proportions employing our 10 texts. The number of words of N, V, A, I are respectively divided by the sum of N, V, A, and I. The proportions are also shown in

Table 2. The authors in the table are marked with numbers in order to easily distinguish them from each other.

Table 2
Japanese texts, their size, and their proportions of N, V, A, and I
(divided by the sum of N, V, A, and I) (Prop = proportion)

Author	**No of different words (Types)**	**No of all words (Tokens)**	**Group N (Tokens); Prop**	**Group V (Tokens); Prop**	**Group A (Tokens); Prop**	**Group I (Tokens) Prop**
(1) Nosaka	4216	17233	5333 0.56	2702 0.28	1353 0.14	184 0.02
(2) Kaiko	3716	21250	6098 0.57	3259 0.30	1278 0.12	78 0.01
(3) Miura	2509	13138	3373 0.51	2166 0.33	1013 0.15	104 0.02
(4) Kobayashi	1338	5613	1644 0.56	909 0.31	381 0.13	24 0.01
(5) Oe	3322	22467	6483 0.55	3901 0.33	1237 0.11	141 0.01
(6) Kawabata	5048	41922	10367 0.50	6780 0.32	3405 0.16	390 0.02
(7) Okamoto	3724	17700	5147 0.55	2798 0.30	1255 0.13	104 0.01
(8) Hori	3069	25013	6368 0.49	3863 0.30	2535 0.19	237 0.02
(9) Ariyoshi	2674	15316	4188 0.55	2373 0.31	999 0.13	91 0.01
(10) Mushakoji	3662	41278	11049 0.52	6728 0.31	2878 0.13	749 0.03

3. The noun-based proportion

We obtained a list of vocabulary from 10 texts analyzed by software. Words written in “conjugated” form were sorted by their frequency, their part of speech, and word forms, and accumulations of frequency by Group of the part of speech. Accumulations of frequency were divided by the sum of N, V, A, and I, and relative accumulations of frequency were also obtained. Excreted data from (7) Okamoto is shown in Table 3a and Table 3b.

Table 3a. Word frequency data and accumulations of frequencies for (7) Okamoto

No	Words written in "conjugated" form	Meaning	Part of speech	Group	Frequency	Accumul-ation of frequency	Accumul-ation of N	Accumul-ation of V	Accumul-ation of A	Accumul-ation of I
1	私	I	prenoun	N	660	660	660	0	0	0
2	その	its	prenoun adjectival	A	428	1088	660	0	428	0
3	し	do	verb	V	403	1491	660	403	428	0
4	いる	be	verb	V	281	1772	660	684	428	0
5	それ	it	prenoun	N	272	2044	932	684	428	0
6	い	be	verb	V	258	2302	932	942	428	0
7	よう	seem to be	adjective verb	A	231	2533	932	942	659	0
8	こと	matter	noun	N	151	2684	1083	942	659	0
9	そう	so	adverb	A	127	2811	1083	942	786	0
10	彼女	she	prenoun	N	118	2929	1201	942	786	0
11	もの	thing	noun	N	98	3027	1299	942	786	0
12	そんな	such	prenoun adjectival	A	92	3119	1299	942	878	0
13	方	manner	noun	N	87	3206	1386	942	878	0
14	何	what	prenoun	N	78	3284	1464	942	878	0
15	この	this	prenoun adjectival	A	77	3361	1464	942	955	0
	.....									
2774	鸚鵡	parrot	noun	N	1	13003	6368	3863	2535	237

Table 3b. Relative accumulations of frequencies for (7) Okamoto
(divided by the sum of N, V, A, and I)

No	Words written in "conjugated" form	Relative Accumulation of N	Relative Accumulation of V	Relative Accumulation of A	Relative Accumulation of I
1	私	0.0508 (=660/13003)	0.0000	0.0000	0.0000
2	その	0.0508	0.0000	0.0329 (=428/13003)	0.0000
3	し	0.0508	0.0310 (=428/13003)	0.0329	0.0000
4	いる	0.0508	0.0526	0.0329	0.0000
5	それ	0.0717	0.0526	0.0329	0.0000
6	い	0.0717	0.0724	0.0329	0.0000
7	よう	0.0717	0.0724	0.0507	0.0000
8	こと	0.0833	0.0724	0.0507	0.0000
9	そう	0.0833	0.0724	0.0604	0.0000
10	彼女	0.0924	0.0724	0.0604	0.0000
11	もの	0.0999	0.0724	0.0604	0.0000
12	そんな	0.0999	0.0724	0.0675	0.0000
13	方	0.1066	0.0724	0.0675	0.0000
14	何	0.1126	0.0724	0.0675	0.0000
15	この	0.1126	0.0724	0.0734	0.0000
	.....				
2774	鸚鵡	0.4897	0.2971	0.1950	0.0182

Relationships between a relative accumulation of N (x) and the relative accumulations of V ($y1$), A ($y2$) and I ($y3$) are shown in Figure 2 with regression lines. We tried to fit regression curves to the data, however, regression straight lines fit better. For (7) Okamoto we obtained the following regression lines:

$$y1 = 0.5305\,x + 0.0031, \qquad (8)$$

with $R^2 = 0.9904$ for the relative accumulation of V,

$$y2 = 0.2482\,x + 0.0018, \qquad (9)$$

with $R^2 = 0.9893$ for the relative accumulation of A, and

$$y3 = 0.0218\,x + 0.0002 \qquad (10)$$

with $R^2 = 0.9233$ for the relative accumulation of I.

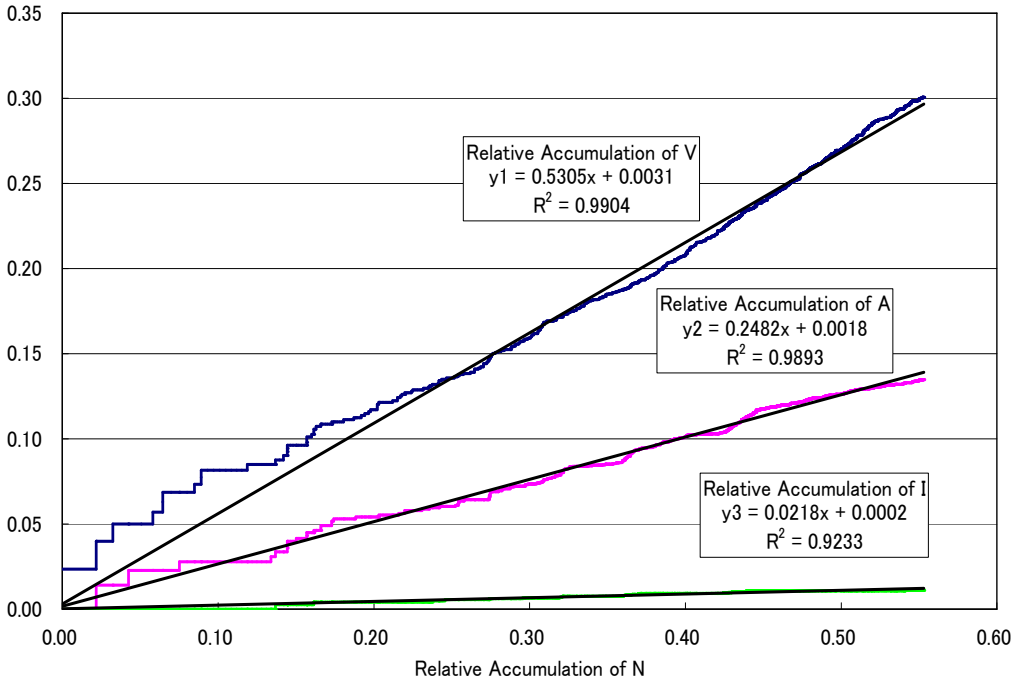


Figure 2. Relative accumulation of N (x) and relative accumulations of V ($y1$), A ($y2$) and I ($y3$) for (7) Okamoto

We investigated the relationships between the relative accumulation of N (x) and relative accumulations of V ($y1$), A ($y2$) and I ($y3$) for 10 texts shown above, and stated that straight regression lines fit to the data better than regression curves for these 10 cases. A summary of regression functions for 10 texts are shown in Table 4.

Table 4
Regression lines of relationships between a relative accumulation of N (x)
and relative accumulations of V ($y1$), A ($y2$) and I ($y3$) for 10 texts

Author	***y1*: Relative Accumulation of V**	***y2*: Relative Accumulation of A**	***y3*: Relative Accumulation of I**
(1) Nosaka	$y = 0.5116x + 0.0056$, $R^2 = 0.9899$	$y = 0.2309x + 0.0214$, $R^2 = 0.9776$	$y = 0.0335x + 0.0031$, $R^2 = 0.8719$
(2) Kaiko	$y = 0.5125x + 0.0115$, $R^2 = 0.9633$	$y = 0.2125x + 0.003$, $R^2 = 0.9808$	$y = 0.0136x + 0.0002$, $R^2 = 0.9243$
(3) Miura	$y = 0.5935x + 0.0259$, $R^2 = 0.9783$	$y = 0.3058x + 0.0047$, $R^2 = 0.9547$	$y = 0.0478x - 0.0065$, $R^2 = 0.9200$
(4) Kobayashi	$y = 0.3985x + 0.0817$, $R^2 = 0.9878$	$y = 0.2132x + 0.0104$, $R2 = 0.9881$	$y = 0.0181x - 0.001$, $R^2 = 0.8514$
(5) Oe	$y = 0.819x - 0.1262$, $R^2 = 0.9586$	$y = 0.2642x - 0.0372$, $R^2 = 0.9832$	$y = 0.0178x + 0.003$, $R^2 = 0.8469$
(6) Kawabata	$y = 0.6642x - 0.0112$, $R^2 = 0.9859$	$y = 0.3145x + 0.0101$, $R^2 = 0.9926$	$y = 0.0406x - 0.00005$, $R^2 = 0.9128$
(7) Okamoto	$y = 0.5305x + 0.0031$, $R^2 = 0.9904$	$y = 0.2482x + 0.0018$, $R^2 = 0.9893$	$y = 0.0218x + 0.0002$, $R^2 = 0.9233$
(8) Hori	$y = 0.6154x - 0.0115$, $R^2 = 0.9807$	$y = 0.3339x + 0.0334$, $R^2 = 0.9928$	$y = 0.043x - 0.0015$, $R^2 = 0.9380$
(9) Ariyoshi	$y = 0.5385x + 0.0136$, $R^2 = 0.9929$	$y = 0.2639x - 0.0098$, $R^2 = 0.9899$	$y = 0.036x - 0.006$, $R^2 = 0.8690$
(10) Mushakoji	$y = 0.5969x + 0.0069$, $R^2 = 0.9933$	$y = 0.2829x - 0.0091$, $R^2 = 0.9952$	$y = 0.0489x + 0.0111$, $R^2 = 0.8777$

4. Future tasks

It is observed in the 10 texts that some of empirical data points increase more than their regression line, and the regression line do not fit well if x is less than 0.2. This can occur if a limited number of words are repeatedly used. Figure 2 is shown in large-size as Figure 3. It can be regarded that proportions of N, V, A and I for a whole text is approximately obtained if we consider 20% of the high frequency words from the text. This point must be investigated in future studies.

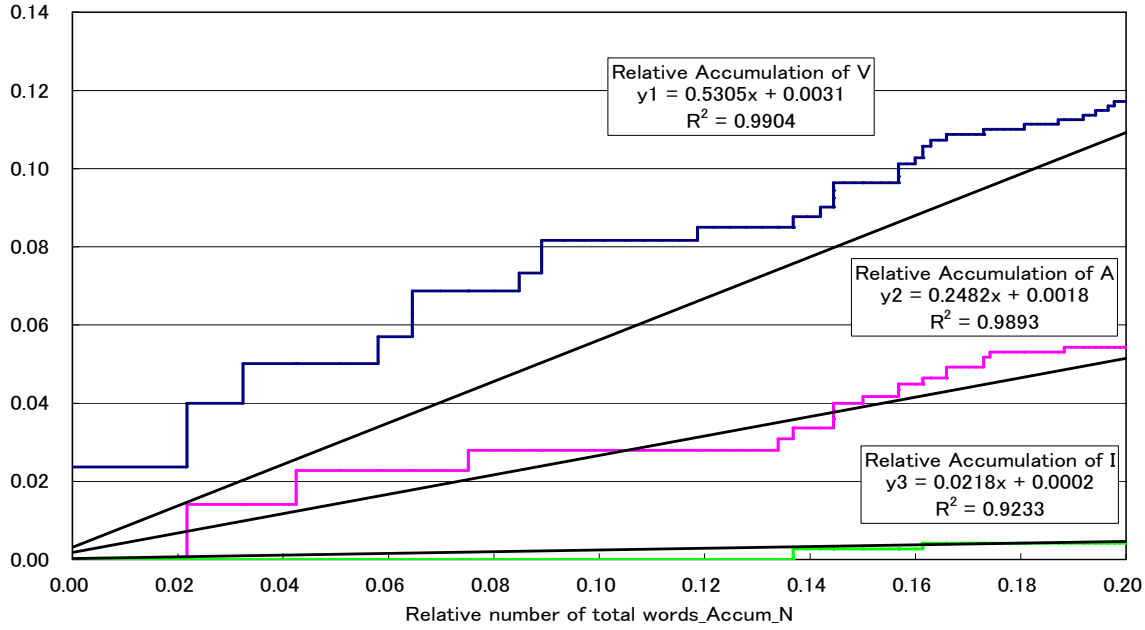


Figure 3. Relative accumulation of N (*x*) and relative accumulations of V (*y1*), A (*y2*) and I (*y3*) for (7) Okamoto
(A large-size of Figure 2 with *x* <=0.2)

References

Altmann, G. (1978). Zur Verwendung der Quotiente in der Textanalyse. In: *Glottometrika*, vol. 1, 91-106. Bochum: Brockmeyer.

Altmann, G. (1988). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Busemann, A. (1925, 1969). Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. In: *Zur Sprache des Kindes:* 1-59. Darmstadt: Wissenschaftliche Buchgesellschaft.

Kabashima, T. (1954). Gendaibun ni okeru hinshi no hiritsu to sono zogen no yoin ni tsuite (On proportions of the part of speeches in texts written in the present Japanese and factors which decide the proportion). *Kokugogaku* (Japanese linguistics) 18, 15-20.

Kabashima, T. (1955). Ruibetsu shita hinshi ni mirauru kisokusei (A regularity in the classified part of speeches). *Kokugo kokubun* (Japanese linguistics and literature) 24-6, 55-57.

Kabashima, T. (1957). Hyogenron no koso (2) (A concept of the art of writing (2)). *Saikyo daigaku gakujutsu hokoku jinbun* (The Scientific reports of Saikyo University, Humanistic science), 10, 23-48.

Mizutani, S. (1965). Ohno no goi hosoku ni tsuite (Notes on Ohno's vocabulary law). *Keiryo Kokugogaku* (Mathematical linguistics) 35, 1-13.

Mizutani, S. (1983). *Goi* (Vocabulary). Tokyo: Asakura Shoten.

Mizutani, S. (1989). Ohno's lexical law: its data adjustment by linear regression. In Mizutani, S. (ed.) *Japanese Quantitative Linguistics: 1-13* Bochum: Studienverlag Dr. N. Brockmeyer.

Ohno, S. (1956). Kihon goi ni kansuru ni-san no kenkyu (Studies on the basic vocabulary of Japanese). *Kokugogaku* (Japanese Linguistics) 24, 34-46.

Sanada, H. (to appear, a). Kabashima's Hypothesis. In: Reinhard Köhler, Peter Grzybek, Sven Naumann (eds.). Wörterbuecher zur Sprach- und Kommunikationswissenschaft (WSK), vol. 9. Quantitative und Formale Linguistik. Berlin, New York: de Gruyter.

Sanada, H. (to appear, b). Ohno-Mizutani's Law. In: Reinhard Köhler, Peter Grzybek, Sven Naumann (eds.). Wörterbuecher zur Sprach- und Kommunikationswissenschaft (WSK), vol. 9. Quantitative und Formale Linguistik. Berlin, New York: de Gruyter.

Texts used (All texts are included in Shinchosha (1995).)

Ariyoshi, S. (1961, 1995). Sumi (Ink stick).

Hori, T. (1933, 1995). Utsukushii Mura (Beautiful village).

Kaiko, K. (1957, 1995). Panikku (Panic).

Kawabata, Y. (1932, 1995). Yukiguni (Snow country).

Kobayashi H. (1950, 1995). Guzo suhai (Idolatry).

Miura, T. (1962, 1995). Kikyo (Going home).

Mushakoji, S. (1920, 1995). Yujo (Friendship).

Nosaka, K. (1967, 1995). Amerika Hijiki (American "Hijiki").

Oe, K. (1957, 1995). Shiiku (Prize Stock).

Okamoto, K. (1938, 1995). Pari Sai (Quatorze Juillet).

Shinchosha (1995). *Shincho Bunko no 100 satsu* (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Tokyo: Shinchosha.

Software

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

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

Remarks

The study was supported by the Alexander von Humboldt Foundation and the Japan Society for the Promotion of Science.

Testing Hypotheses on English Compounds

Hanna Gnatchuk

Abstract. The present article is devoted to testing four hypotheses on English compounds. The focus of our attention is the interrelationship between the number of the compounds and length, age, polysemy and word class. The material of the present research consists of two dictionaries – Longman Exams Dictionary (2007) and The Oxford Dictionary of English etymology (1966). The interrelationships can be modelled by the Zipf-Alekseev, Lorenz and exponential functions with an additive constant.

Keywords: *compound, hypothesis, English, function.*

1. Introduction

On the whole, a word is regarded as a basic and central language unit. It is a central entity in the hierarchy: phonemes, morphemes, word combinations and sentences. According to Ernst (1990), the word is the smallest bearer of meaning. It is codified in dictionaries and separated graphically by blanks in the texts of those languages that use an alphabetic script. Focusing on the phonetic aspect, one may state that a word is the combination of sounds. From the morphological point of view, a word is a unit that may be combined both with dependent morphemes (e.g. derivational or inflectional) and independent ones (e.g. to form compounds). As far as the syntactic aspect is concerned, a word is considered to be the least movable and unchangeable language unit. According to Kaempfert (1984), a word as a separate unit can only appear in the text. Text is defined in our approach as sequences of oral and written language elements. These language elements are based upon the combination of morphological and syntactic units (sentences, syntagms, words).

Dealing with the connection between *lexis and morphology*, Levickij, Lucak (2005) and Ivaniuk, Levickij (1989) explored the interrelation between morphological and semantic categories of words. In modern linguistics, the majority of studies are devoted to the relation between *words and syntax* (namely, syntactic semantics). Nevertheless, Levickij believes that this relation has not sufficiently been studied. The connection between lexis and text-type (*functional styles*) has been studied by Levickij, Pavlyčko, Semenyk (2001). In particular, they focused on the study of preferences of authors for certain semantic groups of words in their literary pieces.

Shedding light upon the connection between words and linguistic levels (phonemic, morphemic, syntactic), it is necessary to deal with the properties of the words which can lead to various hypotheses concerning the formation of compounds. Therefore it is important to consider the following features of a word: length, frequency, parts of speech, polysemy, morphological composition, production, etymology (age), number of associations, number of synonyms, abstract and concrete meaning aspect, etc. The above-mentioned features (variables) should be correlated with the number of compounds words may produce. It can help us to set up a theory of compounds. Compositionality is not

an isolated property; it is linked with other word properties, some of which will be scrutinized in this article. Today, it is well known that isolated entities do not exist in language; there is always at least another one with which they are linked.

For the testing of four hypotheses in English, we refer to the procedures proposed in the article "Synergetic linguistics" by R. Köhler (2005: 765): "1) to set up the axioms (in our case, the first language axiom is that the language is a self-organizing and self-regulating system and the requirements of speakers and hearers have an influence on the behavior of the language system); 2) to determine the system levels, units and variables (here e.g. the level of words with the following variables: frequency, length, polysemy,…); 3) to set up or systematize the hypotheses about the dependence of variables on others: e.g. the number of compounds decreases with their increasing length (Altmann, 1989); 4) to look for functional equivalents; 5) mathematical formulation of the hypothesis; 6) empirical testing".

2. Testing hypotheses on English compounds: length, age, polysemy, word class.

(a) Length

The word length has been investigated in many languages (cf. e.g. Piotrowski 1977; Grotjahn, Altmann 1993; Zörnig, Altmann 1993; Nemcová, Altmann 1994; Wimmer, Köhler, Grotjahn, Altmann 1994; Best 1996a; Köhler 2006; Grzybek 2006; Popescu, Best, Altmann 2014). On the whole, the length can be represented by the number of letters, sounds, phonemes, syllables or morphemes which constitute a word. However, the immediate constituents of the word are syllables or morphemes, hence measuring word length in terms of sounds, letters or phonemes means omitting an intermediate level and may lead to fractal structures. It is also worth mentioning that word length abides by Menzerath's law (Altmann, Schwibbe 1989): the bigger a construct is, the smaller its components. Nevertheless, the distribution of word length depends upon several factors. In particular, the length of words in the dictionary and text differs significantly, especially in strongly synthetic languages.

The task of the present research is to determine the dependence between the number of syllables in the word and the number of all English compounds in the dictionary of which it is a component. In this situation we put forward the following hypotheses:

H_1: The shorter a word is, the more compounds it produces
H_0: The shorter a word is, the fewer compounds it produces

The material for our research consists of a sample of 1752 English compounds taken from the Longman Exams Dictionary (2007). Aiming to find the connection between the word length in syllables and the number of compounds it produces, we have taken the following steps:

1) All the first components of compounds (specifying words of a compound) have been selected;
2) We have counted the number of syllables the first components of compounds have;
3) We have counted the number of compounds the first components have in the dictionary under consideration.

The model corresponding to the analyzed dependence is represented by Zipf-Alekseev function:

$$y = c*x^{a+b*ln(x)}$$

Here *a, b, c* are parameters
x – the number of syllables (length)
y – the number of compounds in the dictionary.

The results are presented Table 1.

Table 1
Dependence between word length (in syllables)
and the number of compounds in English

Length in syllables x	**Number of compounds y**	**Computed values ŷ**
1	721	720.78
2	687	688.09
3	250	246.64
4	76	76.16
5	16	23.71
6	2	7.74
N	1752	
	a = 1.4873, b = -2.2423, c = 720.7780, R^2 = 0.9998	

Table 1 shows that the fitting with R^2 = 0.9998 is almost perfect. The formula can be derived from the differential equation of the unified theory (cf. Wimmer, Altmann 2005). The determination coefficient R^2 = 0.9998 supports our above-formulated hypothesis. In this case, our attention should be drawn to similar research conducted by Rolf Hammerl (1990) on the basis of the Polish language. He has shown that disyllabic words are more productive in Polish than monosyllabic ones. However, monosyllabic English words are according to Table 1 more productive. Moreover, one should take into account different degrees of synthetism and analytism of Polish and English. In any case, it is necessary to conduct analogous research in other languages in order to reveal the laws according to which compounds are formed in different languages. Evidently, the degree of synthetism is a boundary condition that should be taken into account.

(b) Etymology (age)

It is possible to suppose that historically older words can diversify semantically. To make a word meaning more specific, a speaker has two possibilities: a) he can choose a certain environment; b) he can add certain derivative or inflectional affixes or construct compounds. Moreover, we may conjecture that older words are characterized by a higher propensity to construct compounds. Words diversify semantically and require specification. Therefore we may suppose that the tendency to build compounds decreases with decreasing age of words and vice versa.

The purpose of the present analysis is to determine the dependence between the first record (appearance) of a word and the number of compounds it may form in Modern English. Therefore, the following hypotheses have been set up:

H_1: The newer a word, the smaller the number of compounds it produces.
H_0: The age of the word does not influence compound building.

The material for the research consists of 8728 English words (which build compounds in Modern English). They have been taken from the Longman Exams Dictionary (2007). The dates of their first record have been found and noted in the etymological dictionary "The Oxford Dictionary of English Etymology" by C.T. Onions (1966). In order to find the connection between such variables as *age* and *the number of compounds,* the following steps have been taken:

1) We have selected 8728 initial components of English compounds from the whole dictionary;
2) We have noted the first record of each word (12^{th}, 13^{th}, 14^{th} centuries, etc);
3) The number of compounds of the word have been counted in the whole dictionary;

To give an example of the words which date back to the 20^{th} century (cf. Table 2).

Table 2
The example of the words dating back to the 20^{th} century

The first elements of the compounds	**The number of compounds it produces in English**
Garage	2
Internet	6
Laser	3
Mickey	1
Movie	4
Pep	5

Radio Sonic Goofy	10 1 1
9	30

In the etymological dictionary we look for the dates (centuries) when the initial component of compounds was first recorded. Simultaneously we count the number of compounds the first element can have. We can conjecture that the relative rate of change of the number of compounds is constantly decreasing, that is, we can suppose that the differential equation (considered in the form of relative rate of change) may be written in the form:

$$\frac{dy}{y-a} = -\frac{1}{c}dx$$

Here we obtain the exponential function with an additive constant:

$$y = a + b*exp(-x/c)$$

In this case we apply the resulting function to the data given in Table 3 and obtain the fitting displayed in the last column. In Table 3, N stands for the number of initial words dating back to a certain century, C – the number of compounds initial elements can have

Table 3
Dependence between the age of the first element of compound (century) and the number of compounds the initial word can have (C)

Century	N	C	N/C	Computed
Old English	4292	513	8.4	8.23
12	285	49	5.8	6.06
13	1246	265	4.7	4.80
14	1284	323	4.0	4.09
15	545	116	4.6	3.68
16	514	145	3.5	3.45
17	276	102	2.7	3.32
18	113	36	3.1	3.24
19	144	44	3.2	3.20
20	30	9	3.3	3.18
a = 3.1468, b = 2600.7720, c = 1.7662 R^2 = 0.9478				

The determination coefficient in the last column of Table 3 (R^2 = 0.9478) shows that the hypothesis in question is supported. If we manage to confirm this

hypothesis in other languages, we shall be able to deal with a historical language law. In this research we have made the first step.

(c) Polysemy

The connection between polysemy and compounding has been elaborated by U. Rothe (1988) and P. Steiner (1995) on the basis of German. In particular, U. Rothe has found the dependence between the average number of compounds and polysemy on the basis of 1858 German words taken from the "Wahrig Deutsches Wörterbuch". The hypothesis has been confirmed by Petra Steiner. She emphasized the importance of testing the hypothesis in other language in order to find the laws according to which the language functions. Therefore we intend here to deal with the testing of this hypothesis in the English language.

The aim of our investigation is to detect the connection between the number of meanings of a word (the first element of a compound) and the number of its compounds in the dictionary. Here we deal with the following hypotheses:

H_1: The more polysemic a word, the higher its ability is to build compounds

H_0: Polysemy has no influence on compounding.

The material for the research consists of 8785 English words (the first components of the compounds) taken from "Longman Exams Dictionary" (2007).

The following steps have been made in order to find the dependence between polysemy and the number of compounds:

a) We have chosen 8785 initial words which form English compounds;
b) We have counted both the number of meanings and the number of compounds for these initial words
c) The results are given in Table 4, which shows the connection between the mean number of compounds (y) and the number of meanings (x):

Table 4
The dependence between the means for the English compounds and their number of meanings

Number of meanings in intervals (x)	Average means of x	Means for the number of the compounds (y)	Computed $\hat{y}$
1-5	3	3.2	3.21
6-10	8	4.7	3.97
11-15	13	6.6	5.01
16-20	18	8.4	6.47
21-25	23	11.8	8.56
26-30	28	11.7	11.59
31-35	33	10.7	15.88

36-40	38	21.6	21.29
41-45	43	27.5	26.17
46-51	48	29.4	27.23
a = 27.4643; b = 46.5229; c = 15.8292; R^2 = 0.9353			

The data from the whole dictionary have been analyzed. Table 4 shows that the increase in the number of English compounds is linked with the degree of polysemy of the words. In this case, the Lorenz function can be proposed and defined as (cf. Popescu et al. 2009: 55):

$$y = \frac{a}{1+\left(\frac{x-b}{c}\right)^2} = \frac{ac^2}{c^2+(x-b)^2},$$

but even a simple power function would be sufficient.

We can see that the data in Table 3 testify to an excellent fitting (R^2 = 0.9353). The same conclusion was made by Rothe (1988) on the basis of German Therefore, it is necessary to do similar analyses in other languages (i.e. French, Russian, Spanish, Polish, Ukrainian, Slovenian, etc) in order to discover laws according to which the languages can function.

(d) Parts of speech

The objective of the investigation is to detect the interrelationship between the number of word classes a word belongs to (considering the first component of a compound) and the number of compounds the first component produces. In this case, it is possible to formulate the following hypotheses:

H_1: If a word belongs to more than one part of speech, then it possesses a higher propensity to form compounds;

H_0: Word classes have no influence on compounding

The material for the research consists of 10461 English words (that form the compounds) and 1489 English compounds taken from Longman Exams Dictionary.

The procedure of the research:

a) First of all, we have selected 10461 words for English compounds;
b) We have counted the number of compounds (1489) formed by initial words;
c) We have counted the number of word classes initial words belong to. For example, we have a word *down* which belongs to 5 parts of speech – *adverb, preposition, verb, adjective and noun.* It is the first component for 26 English compounds: *downfall, downhome, downgrade, down-market,* etc. The results are presented in Table 5;

d) For testing the trend, the simple power function is sufficient: $y = ax^b$, where x is the number of word classes (POS):

Table 5
The number of basic stems for the compounds (N)
and the number of compounds these stems produce (C)

Parts of speech	N	C	N/C	Computed
1	2581	661	3.9	3.07
2	5582	625	8.9	7.31
3	1540	163	9.4	12.15
4	594	33	18.0	17.41
5	164	7	23.4	23.01
a = 3.0719, b = 1.2513, R^2 = 0.9542				

It is worth mentioning that we deal with the data from the complete dictionary. It is clear that the purpose of the process of compounding is to specify the meaning. Therefore, we have to deal with two phenomena: 1) A word may possess different meanings if it belongs to several word classes; 2) Or a word may have a very general meaning if it belongs to several word classes. It brings us to the fact that the average (mean) compounding must increase with increasing word-class membership. We can observe this clearly in Table 5. The hypothesis has been supported. Nevertheless, it is necessary to analyze several other languages in order to discover some laws.

4. Conclusion and further perspectives

In the present article we have made a certain contribution to the development of a theory about English compounds. In particular, we have confirmed the existence of links between certain properties of a word and the number of compounds the first element can have. In such a way, the interrelationship between polysemy, age, length, and parts of speech, on the one hand, and the number of compounds, on the other hand, has been established statistically on the basis of English Monolingual and Etymological dictionaries. By testing hypotheses, we have created a control cycle which can lead to future investigations on the topic in question. Nevertheless, we emphasize the importance of testing hypotheses of this kind on different languages. In this case one should take into account the degree of synthetism and analytism of the analyzed languages.

References

Altmann, G. Hypotheses about compounds. In: Rolf Hammerl (ed.), *Glottometrika 10, 100-107. Bochum: Brockmeyer.*

Altmann, G., Beöthy, E., Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationswissenschaft 35(5), 537-543.*

Best, K-H. (1996). Word Length in Old Icelandic songs and Prose Texts. *Journal of Quantitative Linguistics 3*, 97-105.

Ernst P. (1990). *Einführung in die germanistische Sprachwissenschaft*: Skriptum. Wien.

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In: J. Boy, R. Köhler (eds.), *Glottometrika 6: 115-126*. Bochum: Brockmeyer

Grotjahn, R., Altmann, G. (1993). Modelling the distribution of word length: Some methodological problems. In: Köhler, Reinhard; Rieger, Burghard (eds.). *Contributions to Quantitative Linguistics: 141-153.* Dordrecht, NL: Kluwer.

Grzybek, Peter (2006). History and Methodology of Word Length Studies: The State of the Art. In: Grzybek, Peter (ed.) (2006): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 15-90.* Dordrecht, NL: Springer.

Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (an polnischem Sprachmaterial). In: Rolf Hammerl (ed.), *Glottometrika 12*: *73-83*. Bochum: Brockmeyer.

Ivanuk, V.J., Levitskij V.V. (1989). The influence of semantics on the usage of tense forms of verbs in German. Communicative aspects of units of language and speech. *Mezhvuzovskij sbornik*: 62-70. Izhevsk.

Kämpfert, M. (1984). *Wort und Wortverwendung. Probleme der semantischen Deskription anhand von Beobachtungen an der deutschen Gegenwartssprache*. Göppingen: Kümmerle

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative Linguistik. Ein internationales Handbuch: 760-774*. Berlin: Walter de Gruyter.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M (eds.), *Favete linguis. Studies in honour of Victor Krupa: 142-152*. Bratislava: Academic Press

Levickij, V., Lucak, M. (2005). Category of Tense and Verb Semantics in the English language. *Journal of Quantitative Linguistics 12, 2-3.*

Levickij, V., Pavlychko, O., Semenyuk. T. (2001). Sentence length and sentence structure as statistical characteristics of style in prose. In*:* Ludmila Uhlířová, Geiza Wimmer, Gabriel Altmann, Reinhard Köhler (eds.), *Text as a linguistic paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček*. Trier: Wissenschaftlicher Verlag, 177-186.

Longman Exams Dictionary (2007). Pearson Longman.

Nemcová, E., Altmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirische Textforschung 1, 40-44.*

Onions, C. T. (1966). *The Oxford Dictionary of English Etymology*. Oxford: At the Clarendon Press.

Piotrowskij R.G., Bektajev, A.A., Piotrowskaja, A.A. (1977). *Matematiches-kaja lingvistika*. Moskva: V'sshaja shkola.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid : RAM-Verlag

Rothe, U. (1988). Polylexy and compounding. *Glottometrika 9, 121-134.*

Rothe, U., Altmann, G., Wagner, K. (1992). Verteilung der Länge von Sprechakten in der Kindersprache. In: Wagner, K.R. (eds.), *Kinder-sprachstatistik*: 47-56. Essen: Die blaue Eule.

Sambor, J. (1984). Menzerath's Law and the Polysemy of Words. In: J. Boy, R. Köhler (eds.), *Glottometrika 6: 94-114*, Bochum: Brockmeyer.

Steiner, P. (1995). Effects of Polylexy on Compounding. *Journal of Quantitative Linguistics 2(2), 133-140.*

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quuantitative Linguistics. An International Handbook:791-807*. Berlin: de Gruyter.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. *Glottometrika 14, 190-196.*

Context-Specific Distribution of Word Meanings

Maria Rukk[1]

Abstract. Language is a self-organizing system where certain relationships between word frequencies and word lengths, word length and polysemy, polysemy and polytextuality, and polytextuality and word frequencies have been stated. This paper considers the hypothesis that the meanings of words with high polytextuality tend to be used context-specifically. The present study is a pilot study and focuses on the data preparation.

Keywords*: semantics, self-organizing system, meaning, polysemy, polytextuality.*

1. Introduction

Quantitative linguistics aims to investigate and clarify linguistic phenomena with the help of quantitative means. A special branch of quantitative linguistics is synergetic linguistics, where language is considered to be a self-organizing system characterized by cooperative and competitive processes; these processes, together with manifold external influences, constitute the dynamics of language systems (Köhler 1991).

Different aspects of language have been investigated so far, including lexis. However, this appears to be a complicated and somewhat controversial object, owing to a very large inventory of entities, limitless combinatorial ability, and the complexity and heterogeneity of interrelations in speech and language (Tuldava 1998). Nevertheless, every aspect of this complex system is being investigated, ranging from single units (e.g. word frequencies, word life cycle, analysis of compounds) up to the relationships between different subgroups of lexis, authorship problems, lexicographic problems, and the full analysis of a lexical system as a self-organizing one.

One of the most important universal features of a language system is polysemy, and more and more attempts are being made to systematize it. In general it is stated that:

> "Polysemy involves lexemes that are clearly united (share a common schema) as well as clearly seperable at the same time. Polysemous words are the result of lexemes gaining new usages over time which share the same phonological form and appear to have separate meanings to non-etymologists… Polysemous lexemes always share the same etymological background and/or are conceived of as being semantically related by speakers" (Tuggy 1993: 4-3).

[1] E-mail: roukkm@inbox.ru. Maria Rukk, 10-16-156, Lesnaya Str., Moscow, 125047 Russian Federation

In quantitative linguistics, polysemy is usually defined as the number of different meanings or functions of a linguistic unit[2]. Recently, many studies have been conducted within quantitative linguistics and many various aspects have been analysed; e.g., the polysemy in different languages (Lu Wang 2014), multi-dimensional polysemy networks (Glynn 2014), and polysemic distributions of language signs (Poddubnyy/Polikarpov 2014), etc. have been discussed.

Quantitative approaches to polysemy can furthermore be found in the works of Moskovich (1965, 1969), Shaykevich (1968), Krylov and Yakubovskaya (1977), Tuldava (1979), Köhler, Altmann (1986), Polikarpov (1987), Levickij, Kiiko and Spolnicka (1996), Levickij, Drebet and Kiiko (1999), Köhler (1990), Ziegler and Altmann (2001), Glynn (2014) and many others.

Some of these papers are explicitly devoted to hypotheses regarding certain inter-relationships between word frequencies and word lengths, word lengths and polysemy, polysemy and polytextuality[3], and polytextuality and word frequencies (e.g. the greater the polysemy, the greater the polytextuality, - i.e. the number of texts where this word is used at least once). Some of these regularities have later been accepted as general linguistic laws: e.g., the more polysemic a word is, the more synonyms it has (the hypothesis set up by R. Köhler (1990: 8) and tested by Ziegler and Altmann (2001)); or the more meanings a word has, the greater is the number of texts in a corpus where the word occurs at least once.

Meanwhile, one more hypothesis has been set up by R. Köhler[4], which needs to be tested. It seems to be true that the higher the polytextuality is, the higher the tendency is for the word to be used context-specifically, i.e., when some meanings are preferred in one kind of text, whereas other meanings tend to occur mainly in another group of texts. To put it in other words, our hypothesis is:

> The higher the polytextuality is, the stronger the tendency is for the word to be used context-specifically.

Our hypotheses will be tested based on Russian texts in the next section.

2. Analysis of Russian data

The data collected relate to the polysemy of prepositions in Russian fiction texts. The reasons for the particular focus on prepositions are: 1) Polytextuality in the case of prepositions is certain, as they occur in every text; 2) they do not have to be lemmatized. Our approach consists of these seven steps:

1. A corpus of texts is set up, the texts being taken from SynTagRus[5]. The filters applied were BASIC corpus, PR for prepositions, sub-corpora of fiction texts (from these, mainly drama texts were analyzed), released of

[2] http://www.glottopedia.org/index.php/Polysemy_in_quantitative_linguistics.
[3] Here: ability of the word to be used in different kinds of texts
[4] Personal communication.
[5] http://ruscorpora.ru/index.html

homonymy[6]. Texts of about 1000 words, ranging from 500 to 1500 word forms, with two exceptions, were taken for the analysis.

2. A list of all prepositions was prepared.
3. In dictionaries, the different meanings are usually denoted by numbers or letters. Counting them is the simplest way of measurement. The polysemy for each preposition for the purpose of this research was stated with the help of an on-line dictionary[7].
4. The textuality for each word was defined by counting the number of texts where the preposition was used.
5. The meaning of the preposition in each individual occurrence was determined.
6. For each text, the most frequent meaning, its frequency and the frequency of other meanings were noted.
7. For each preposition, the texts were noted where one and the same meaning is the most frequent (in a chart, cf. Table 4): e.g., in text 3 "Svidanie", meaning 1 of preposition "V" is used 8 times (v konsul'tatsii, v Chertanovke) and is the most frequent; meaning 6 is used only once (v den' po 30 abortov). The same meaning 1 of this preposition "V" was the most frequent in texts 2, 4, 7-10, 12-14. The goal of the following analysis is to see if these texts, where a meaning is the most frequent, have something in common, e.g. genre, functional style etc.

2.1. Retrieved data

The corpus of Russian fiction texts was set up and over 60 texts were selected according to the criterion of text length. Here, 15 texts were analysed only (Table 1).

Table 1
Texts analyzed at the preliminary stage
(Title, date of creation, number of word forms, text type and genre)

	Text	Date	WF	Text Type	GENRE
1	I. Antonova "Koza i zaychik"	2002	955	Play	Drama, children's
2	O. Tikhomirov "Bez repetitsii"	2001	960	Play	Drama, children's
3	L. Petrushevskaya "Svidanie"	1997	1356	Play	Drama
4	L. Petrushevskaya "Opyat' dvadtsat' pyat'"	1993	808	Play	Drama
5	L. Petrushevskaya	1985-	1384	Play	Drama

[6] Some prepositions and adverbs in Russian are homonyms. In the corpus used, this feature is taken into consideration, and applying the filter "released of homonymy" one gets only actual prepositions.
[7] http://www.gramota.ru/.

	"Stsena otravleniya Motsarta"	1990			
6	I. Bakhterev "Tsar' Makedon, ili Fenya i Chebolveki"	1950	1418	Play	Drama
7	V. Khlebnikov "Mirskontsa"	1912	1358	Play	Drama
8	A. Chekhov "O vrede tabaka"	1886-1902	1371	Play, miniature	Drama
9	A. Chekhov "Noch' pered sudom"	1890-1900	1729	Play	Drama, humor, satire
10	A. Chekhov "Lebedinaya pesnya (Kalkhas)"	1887	1938	Play	Drama
11	Ordinamenti	2004	1338	Drama	Humor, satire
12	S. Kozlov "Kak yezhik s Medvezhonkom spasli volka"	2003	1026	Fairy-tale	Children's
13	N. Teffi "Tanglefoot"	1911	536	Story, miniature	Humor, satire
14	. Radishchev "Pis'mo k drugu, zhitel'stvuyushchemu v Tobol'ske, po dolgu zvaniya svoego"	1790	1309	Letter	Documentary prose
15	D. Fonvisin "Krest'yanin, drakon I lisitsa"	1788	548	Fabel	Humor, satire

Though the corpus is very small, several groups can be defined here, e.g. children's books, drama vs other types of prose, authorship, date of writing, number of word forms (WF).

A total of 76 prepositions used in the selected texts were listed. From these, the prepositions with more than one meaning were chosen, and, from these, eight were not used in the texts analyzed. The resulting list includes the following prepositions (cf. Table 2).

Table 2
Prepositions analyzed, number of meanings each preposition has, and number of texts the preposition was used in

Preposition	No. of meanings	No. of texts	Preposition	No. of meanings	No. of texts
bez/bezo	3	11	na	11	15
v/ vo	12	15	nad/nado	3	8
vokrug	2	2	o/ob/obo	4	11
dlya	4	9	ot/oto	12	12
do	5	11	pered/peredo	7	8
za	26	15	po	26	10
ii/izo	10	12	pod/podo	9	9

Iz-za	4	4	prezhde	2	1
k/ko	13	14	pri	12	7
krome	2	2	pro	2	6
mezhdu	6	3	protiv	6	1
mimo	2	1	radi	4	3

For each preposition, the most frequent meaning, its frequency, and the frequency of other meanings in the same text are noted. The results for 22 words with the greatest amount number of meanings are summarized in a chart like the one presented in Table 3. Here, the letters have the following meaning:

(a) The most frequent meaning for this text[8]

(b) The frequency of this meaning in this text

(c) The frequency of all other meanings in the same text.

Table 3

Excerpt. Prepositions, the most frequent meanings of these prepositions for each text, the frequency of their usage, frequency of other meanings in the texts

	Text 1			**Text 2**			**Text 3**		
	(a)	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**
za	7, 12	1	0	17	1	0	12	3	1
po				16,II-1	1	0	12, 19	2	1
k				1	1	0	1	2	0
v	2	4	3,2	1	3	1,2	1	8	6,1
ot	1	1	0	1,5	1	0	1,6	1	0
pri									
s	3,4	2	1	2	7	4,1	4	8	1,3
na	2	4	1	1	5	1,2	2	5	1,2,4
iz	1	4	0	1	1	0	1	2	0
pod							I-2	1	0
pered				4	1	0			
mezhdu									
protiv									
cherez							4	1	0
do	2	1	0	3	4	1	1	3	1
dlya									
iz-za							4	1	0
o	II-1	1	0	II-1	1	0	II-1	3	0

[8] Each preposition analyzed has several meanings. Here, the meanings are numbered and designated in the same way as they are presented in the on-line dictionary www.gramota.ru. This refers to the usage of both Roman and Arabic numbers. Thus, e.g. preposition ”pod” has two major groups of meaning marked by Arabic numbers and differing in government, each group having about 10 sub-meanings marked by Roman numbers.

Then, all the texts are regrouped according to the criterion that one and the same meaning of this or that preposition is the most frequent (Table 4).

Table 4
Preposition meaning and texts, where it was used

	Meaning	**Texts**		**Meaning**	**Texts**
za	2	11, 12	s	1	11, 15
	7	1, 7, 11		2	2, 4, 9, 11, 15
	8	12, 13		3	1, 11, 15
	11	7, 9, 12, 14, 15		4	1, 3, 5, 6, 7, 11, 14, 15
	12	1, 3, 4, 6, 12		6	8, 11, 12, 13
	13	7, 8		7	9, 11, 14, 15
	15	10		8	10, 15
	17	2, 10		9	8, 11
	18	11	na	1	2, 4, 6, 7, 10, 13, 14
	19	5, 6, 13		2	1, 3, 5, 8, 11, 12, 13, 15
	23	11		4	5
po	1	10, 12, 15		7	12
	4	11, 12		8	9
	5	10		11	6
	12	3	iz	1	1-3, 9-13
	13	9		2	6
	15	5		4	8, 13, 14
	16	2, 5		6	7
	17	9	pod	I-1	6
	19	3		I-2	3
	II-1	2		II-1	4, 7, 8, 10, 11, 12
	II-5	8, 10		II-2	14
	III-1	14		II-4	8
k	1	2-7, 9, 11-15	pered	1	5, 6, 8, 12
	2	13		3	10
	7	6		4	2
	8	8, 10		5	5, 7
v	1	2-4, 7-14		7	9
	2	1, 11	mezhdu	1	12, 14
	9	6		4	15
	10	5, 15	protiv	1	11

ot	1	1-4, 8, 15	cherez	4	3
	4	6, 12, 14		6	9, 15
	5	2, 5, 9, 15	do	1	3, 7, 8, 10
	6	3		2	1, 4, 10, 14
	9	7		3	2, 11
pri	7	7, 10		5	5, 12
	8	7, 10	dlya	1	4, 5, 13
	9	14, 15		2	10
	11	5-7, 11		3	6, 8, 11, 14, 15
	12	11		4	10, 11
o	I-1	10, 11	iz-za	2	11
	I-2	11		3	9
	II-1	1-3, 5, 7-9, 11, 14, 15		4	3, 7

2.2. Results

From the chart, it can be seen that there are some groupings of prepositions. For instance, meaning 1 of preposition "V" is preferred in texts 2-4, 7-14, and meaning 10 in texts 5 and 15; and for meanings 3-9 there were no occurrences found.. The existence of such groupings is a good sign that the hypothesis (viz. the higher the polytextuality is, the stronger the tendency is for the word to be used context-specifically) could be confirmed. Nevertheless, statistical justification and provision of more data for different genres, functional styles, etc. is necessary as at present the corpus is still insufficient to identify a clear reason for differentiation. For example, children's texts (1, 2, 12), or subgroups of children's texts, sometimes come together but are accompanied by texts from other groups for one and the same meaning of a preposition. The same is true for the texts of a particular author (Chekhov or Petrushevskaya), or for texts written in a particular period (Radishchev and Fonvisin). A correlation analysis was carried out for several groups of texts and different meanings of prepositions:

- Children's texts vs "adult" texts, the correlation coefficient ranging from -0.25 (za, meaning 19) to 0.53 (Prepositions po (II-1), Na (7), pered (4)).
- Drama vs Prose –using the example of two prepositions, za and s. The correlation coefficient ranged from -0.6 (za, meaning 8) to 0.58 (za, meaning 7).
- Date of creation was tested using the example of preposition s. Four date ranges were taken, the 18th century (the maximum value was 0.66 for meaning 7), 19th century (the maximum negative value was -0.53 for meaning 4), 20th century (-0.49 for meaning 7) and 21st century (0.45 for meaning 3).
- Authorship – here we have 6 texts written by two authors: Chekhov and Petrushevskaya. The correlation for two prepositions was tested. The maximum values were reached at 0.35 (Petrushevskaya, Za, meaning 12

and meaning 11 (the same value, but negative)) and 0.53 (Chekhov, za, meaning 15, S, meaning 4, negative value).

The correlation coefficient was never high enough to provide a basis for a confident judgement about which criterion is decisive. The work should be continued in two directions: on the one hand, more data need to be analysed; on the other hand, other reasons for this or that distribution should be investigated.

Acknowledgement: Many thanks to Prof. Dr. R. Köhler for the idea of this article and to the members of the editorial board for the assistance and support.

References

Glynn D., Robinson J. A. (ed). Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy. *University of Paris VIII / University of Sussex. [Human Cognitive Processing, 43] 2014. viii, 545 pp.*

Glynn, Dylan (2014). Modelling multidimensional polysemy networks. The case of /over/. *QUALICO 2014. Book of abstracts. Philosophical Faculty of Palacký University, Olomouc, Czech Republic. May 29-June 1, 2014.p. 48-50.*

Köhler, Reinhard, Altmann, Gabriel (1986). Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft 5: 253-265.*

Köhler, R. (1990). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, Reinhard (1991). Synergetic linguistics. In: Reinhard Köhler, Burghard B. Rieger (Eds.), *Contributions to quantitative linguistics:. 41-51.* Kluwer Academic Publishers.

Krylov, Ju., Yakubovskaya, M. (1977). Statisticheskiy analiz polisemii kak yazykovoy universalii i problema semanticheskogo tozhdestva slova. *Nauchno-tekhnicheskaya informatsiya 2, Nr. 3,* 1-6.

Levickij V. V., Drebet V.V., Kiiko S.V. (1999). Some Quantitative Characteristics of Polysemy of Verbs, Nouns and Adjectives in the German Language. *Journal of Quantitative Linguistics 6(2): 172-187.*

Levickij V. V., Kiiko J.J., Spolnicka S.V. (1996). Quantitative analysis of verb polysemy in modern German. *Journal of Quantitative Linguistics 3(2): 132-135.*

L'vov, M.R. (1979). *Tendentsii razvitiya rechi uchashchikhsya. Vyp. 2: Posobie dlya studentov pedinstituta.* Ministerstvo prosveshcheniya RSFSR. MGPI Moskva 1979.

Moskovich, V. (1965). Opyt kvantitativnoy tipologii semanticheskogo polya. *Voprosy yazykoznaniya 4.* 80-91.

Moskovich, V. (1969*). Statistika i semantika*. Moskva. 304.

Oksaar, Els (1972). Stilstatistik und Textanalyse. In: Herbert Backes (ed.), *Festschrift fuer Hans Eggers zum 65. Geburtstag:* 630-648. Tübingen: Niemeyer.

Poddubny, V., Polikarpov, A. (2014). Evolutionary derivation of laws for polysemic and age-polysemic distribution of language sign ensembles. *QUALICO 2014. Book of abstracts: 94-96. Philosophical Faculty of Palacký University, Olomouc, Czech Republic. May 29-June 1, 2014.*

Polikarpov, A. (1987). Polisemiya: sistemno-kvantitativnye aspekty. *Uchenye zapiski TGU 774, 135-154*.

Polikarpov, A. (1991). A model of the verb life cycle. In: Reinhard Köhler, Burghard B. Rieger (Eds.), *Contributions to quantitative linguistics: Proceedings of the first International Conference on Quantitative Linguistics, QUALICO. Trier, 1991. 53-66.* Dordrecht: Springer .

Poljanskiy S.M. et al. (eds**.).** Kvantitativnaya lingvistika I semantika (Qualisem-99): II Mezhvuz.konf.: Tez. Dokl. - Novosibirsk: Novosibirsk State Paed. Univ, 1999. 79.

Shaikevich A. (1976). Distributivno-statisticheskiy analiz v semantike. In: *Printsipy i metody semanticheskikh issledovaniy:* 353-378. Moscow: Nauka.

Shaykevich, A. (1968). Opyt statisticheskogo vydeleniya funktsional'nykh stiley. *Voprosy yazykoznaniya 1*.64-76

Tuggy, David (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4-3.

Tuldava, Juhan (1979). O nekotorykh kvantitativno-sistemnykh kharakteristikakh polisemii. *Uchenye zapiski TGU 502*, 107-141.

Tuldava, Juhan (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. (Quantitative Linguistics 59). Trier: WVT Wissenschaftlicher Verlag .

Wang, Lu (2014). Polyfunctionality and polysemy in Chinese. *QUALICO 2014. Book of abstracts: 123-125. Philosophical Faculty of Palacký University, Olomouc, Czech Republic. May 29-June 1, 2014.*

Ziegler, A., Altmann, G. (2001). Die Beziehung zwischen Synonymie und Polysemie. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99: 226-229*. Bratislava: Veda.

The Quantitative Nature of *Working Maps* (WM) and *Taxatorial Areas* (TA). A Brief Look at two Basic Units of *Salzburg Dialectometry* (S-DM)

Hans Goebl, Pavel Smečka

1. Preliminary Remarks

Ultimately Salzburg Dialectometry (S-DM) is nothing other than a continuation of classical Romance linguistic geography by quantitative terms and means, obviously assisted by numerical and visual computing (visualistics). In this sense, we would underscore the great epistemic continuity between its theoretical and practical orientations and those of the founding fathers of Romance linguistic geography, such as Jules Gilliéron (1854-1926), the author of the French linguistic atlas ALF (published 1902-1910), and Karl Jaberg (1877-1958) and Jakob Jud (1882-1952), the authors of the Italian linguistic atlas AIS (published 1928-1940).

When speaking about linguistic geography in general, it's necessary to point out a few major methodological differences that characterize this discipline as it is practiced in different Modern Philologies (such as Romance, Germanic, English or Slavonic geolinguistics).

The central methodological key of all varieties of geolinguistics was *first* the study of the diffusion areas of single linguistic traits (features, attributes or characters) and their systematic collection in the form of "linguistic atlases", and *second* the subsequent discovery of their untamable spontaneity (independence or unpredictability) in space, which has given many linguists headaches over the last two centuries because of their chaotic or "irregular" nature. Their claim was therefore often that "dialects" cannot be classified (cf. H. Schuchardt 1870/1900) or simply do not exist (G. Paris 1881). On the contrary, S-DM has from the very beginning accepted fully the "spontaneous" properties of the above mentioned linguistic traits and consequently also the Protean nature of their geographic implementation. The Salzburg term for these allegedly chaotic surfaces is "taxatorial area(s)" (TA).

It should also be said that since the uprise of linguistic geography there have also been several elementary differences between the first linguistic atlases. Whereas Georg Wenker's (1852-1911) monumental "Deutscher Sprachatlas" (DSA) with its more than 50 000 inquiry points tended toward mainly data collection for its own sake, its Gallo-Romance counterpart the ALF originated in the old French tradition of geodetic measurement of the national territory by means of different variables. So, the geolinguistic "measurement" of France (and surrounding zones) - done by Jules Gilliéron and Edmond Edmont between 1897

and 1901 - can rather be qualified as "glotto-geodesy" than a mere collection of dialect data. Obviously this "geometric" underpinning of the ALF is an excellent prerequisite for later dialectometric processing.

One should be aware of the fact that in Gallo-Romance linguistics there was also a great hunger for interesting lexical and phonetic data, but that this hunger has not been stilled by compiling many linguistic atlases but mainly by elaborating on a great number of dialect dictionaries. Immediately after the publication of the ALF, the inner empirical differences between linguistic atlases and dialect vocabularies were discussed thoroughly[1]. It became clear that the proper value of the data of linguistic atlases is rather "relational" than purely documentary, and that they allow for numerous interdisciplinary comparisons, mainly with other space-related sciences[2].

Another peculiarity of Romance linguistic atlases should be underscored: since the publication of the ALF, their data is always exhibited in *full-text maps* and therefore in the original form, thus avoiding any visual or cartographic simplification. As a consequence Romance scholars have been forced, since the dawn of Romance geolinguistics, to use special cartographic techniques in order to get clear the spatial structure of the data laid out on the original atlas maps.

They did it by filling up *silent* (or: *mute*) *maps* of the respective atlas grid (see Jaberg 1906, passim). Obviously, this kind of work requires the training of some very useful linguistic skills such as *simplification* and *classification* of complex geolinguistic data, not to mention the practical challenge of a fairly readable *cartography*. As linguistic geography represented a central discipline in Romance linguistics during the first half of 20th century, many Romance scholars involuntarily became good data classifiers and map makers.

In contrast, Germanic (and other) linguistic atlases never contained their data in their original but instead in graphically encoded form. Their users were therefore not furnished with the same amount of practical and theoretical impulses and stimuli.

2. Taxatation: from the Original Atlas Maps to the Working Maps (WM)

Let us have a look at Map 1 (see Appendix) which is a good example of what has been done by many Romance scholars when classifying and discussing the content of single maps in the ALF (or other linguistic atlases). Even before WW I, Jules Gilliéron published colour maps of the same type (see e. g. Gilliéron/ Mongin 1905). Our map shows the spatial distribution of 15 different denominations of the *ewe* (Fr. *brebis*)[3]. Thus the linguistic nature of this map is *lexical*. The respective 15 diffusion areas vary greatly according to *size*, *shape* (compactness vs. coherence) and *geographic location*. Three major lexical types (in

[1] See Wartburg 1963, 159-163.
[2] See Goebl 2002b and 2006b.
[3] See Wartburg 1918.

Salzburg terminology: *taxates*) emerge: *brebis* (type/taxate 1), *ouaille* (type/ taxate 2) and *fedo* (type/taxate 3). Obviously, they can all be analyzed from an etymological and historical (diachronous) point of view. Etymologically they derive from three well-known Latin roots: VERVÍCE "mutton" > *brebis*, OVÍCULA "(little) sheep" > *ouaille*, and FÉTA "dam" > *fedo*. From a diachronic perspective, the spatial entanglement of the three great areas allows for the hypothetical reconstruction of the geographical spread of the three words over time. These kinds of reconstructive considerations have a very long tradition, not just in the field of Romance linguistics.

In Salzburg, similar colorful maps, provided with a well-defined linguistic interest, are traditionally called *working maps-WM* (Ger. *Arbeitskarten,* Fr. *cartes de travail,* It. *carte di lavoro* etc.). The "work" that has to be done while elaborating such a map relies fully on the theoretical competence of the dialectometrician.

In terms of Numerical Classification, the content of a WM correlates with a single row of the *data matrix*. Metrologically speaking the nature of these data is *qualitative*: they all lie on the *nominal* (or *cardinal*) measurement scale.

As "normal" linguistic atlases consist of several hundreds of *original* maps, each dialectometrization of such an atlas creates at least a similar number of *WM*. Note that the ratio 1:1 between *original map* and *WM* holds for *lexical* maps only. In the case of atlas maps that contain only one *lexical* type and are therefore *mononym*, one can derive several *phonetic* WM (each with different spatial structures) from *one original* atlas map. So the ALF map 233 *chanter* (< Lat. CANTÁRE "to sing") can be split up into several *phonetic* WM, showing respectively the geographic distribution of the Gallo-Romance results of initial C-, pretonic -A, intervocalic -NT-, stressed Á-, intervocalic -R-, and final -E.

Given that the ratio of one taxate/one WM does not create any variation, the taxatorial granulation (or *poly-nymy*) of a WM can theoretically vary between 2 (*bi-nymic* or *2-nymic* WM) and N. The grid of Map 1 (see Appendix) shows 641 (= N) inquiry points, of which 638 belong to the original atlas grid of the ALF, whereas 3 supplementary points (corresponding to the literary languages *French*, *Italian*, and *Catalan*) has been added for illustrative purposes. Incidentally Map 1 is *15-nymic*.

One should be aware of the fact that the whole taxatation process depends on the expert knowledge of the responsible linguist: see Table 1 (below). In the case of the ALF, which is a regular linguistic atlas containing geolinguistic *raw* data in *cartographic* arrangement, this responsibility was exclusively ours; in that of the below mentioned four English "atlases" AES, LAE, CLAE, and WGE the situation is quite more complex. On the one hand, they contain only classified (in Salzburg terminology: taxated) dialect material, and rely, on the other hand, on a source of dialect data ("Survey of English Dialects" – SED) that comprises the original raw data only in *tabular* – and therefore not in *cartographic* – form.

This circumstance complicates the data evaluation enormously, irrespective of the fact that the scientific responsibility is shared by 11 linguists who all pursued diverging scientific interests working independently from each other,

and even at different times (see Table 1, below). Nevertheless, these aggravating prerequisites could not eclipse the inner regularities of the whole data set.

3. The Protean Nature of Taxatorial Areas

Very early (geo)linguists discovered with great amazement the incredible multiformity of TA, even in cases where the categorical proximity of two etymologically related TA suggested a perfect coincidence of the two surfaces and their surrounding lines ("isoglosses"). Such a "categorical proximity" exists, e. g., among the ALF maps 250 *chat* (< *CATTU) "cat", 225 *champ* (< CAMPU) "field", 228 *Chandeleur* (< CANDELÓRU) "Candlemas", 229 *chandelle* (< *CANDÍLLA) "candle", 231 *chanson* (< CANTIÓNE) "song", and 221 *chaîne* (< CATÉNA) "chain", which all show results of initial Latin C+A. Against any (theoretical) expectation, the respective TA (and their surrounding lines) are far from being identical (or prone to coincide)[4].

This fact was first discovered and duly commented in 1889 by Georg Wenker, the author of the German linguistic atlas DSA: "Sind so sämtliche Formen, in denen das Wort erscheint, kartographisch verzeichnet, so werden die einzelnen zu Gruppen sich zusammenschließenden Abweichungen[5] durch Linien abgegrenzt, mit verschiedenen Farben kenntlich gemacht und so das Ganze zu einem übersichtlichen Bilde gestaltet. […] Dann geschieht die Uebertragung in die Grundkarten des Sprachatlas[6], zu denen ein erläuternder Text hinzutritt. Jedes einzelne Wort wird also ganz unabhängig von allen anderen, selbst von verwandten, zu Ende verarbeitet, dann erst werden seine Grenzlinien[7] und seine verschiedenen Formen verglichen mit verwandten Erscheinungen ähnlicher Wörter. Es ist dies eine Vorsicht, welche erst im Verlauf der Arbeit zum Grundsatz erhoben worden ist. *Anfänglich war ich wie wohl jeder allzusehr geneigt, von der bequemen und naheliegenden Vorstellung auszugehen, daß verwandte Wörter, etwa* Hund *und* Pfund, Wurst *und* Durst *auch in ihren mundartlichen Eigenheiten zusammenstimmen müßten. Indessen stellte sich heraus, daß dies nicht immer der Fall ist, daß zwar jedes einzelne Wort seine meist ganz festen Grenzlinien besitzt, daß die Grenzlinien verschiedener Wörter dagegen selbst da, wo man es ganz bestimmt erwartet, nicht immer zusammen fallen, sondern bald mehr bald weniger abweichen. Dies allgemeine Ergebniß muß zunächst, gerade wegen seines Gegensatzes zu den bisherigen Anschauungen, nachdrücklich betont werden, bis man sich an diese etwas unbequeme Thatsache gewöhnt hat.* [italics: HG/PS]" (Wenker 2013 [1889], 10).

A similar statement was made by Karl Jaberg in 1908 in his masterly presentation of the ALF: since that time the principle that "each word has its history of its own" (Fr. "chaque mot a son histoire", Ger. "Jedes Wort hat seine

[4] See Jaberg 1908, map III, Wartburg 1963, 22-24, and Berschin/Felixberger/Goebl 2008, 254-256.
[5] In Salzburg terminology: *taxates*.
[6] This base map ("Grundkarte") corresponds to a "silent map" as described above. The final version of the DSA base map had a grid with more than 50 000 inquiry points.
[7] I. e. *isoglosses*.

eigene Geschichte") has reigned, which in our case should be adapted in "each *taxate* has its *area* of its own"[8]. Starting from this principle, many linguists conjectured – unfortunately – that behind many TA there is absolutely no order or regularity whatsoever. We will see that this belief was pure superstition.

In summary it can be stated that

- TA vary enormously along *size*, *shape* and *location*.
- on mute maps (with N polygons) the *size* of TA oscillates between 1 and N-1.
- the quantitative measurement of their *shape* and *location* seems to be currently out of range.

4. The "Special Entanglement" of Taxatorial Areas

Obviously, the continuous change in *size*, *shape* and *location* of the different TA goes back to a wide range of intra- and extra-linguistic causes, some of which can be detected and even "explained". But what should now be clear is the fact that this variegated situation is not "unnatural", nor is it the consequence of a series of catastrophes that destroyed an assumed virgin regularity. Superposing a greater series of TA and controlling their spatial deviations from each other one discovers that all these TA are interlocked together like shingles on a roof[9]. As we found this phenomenon in all our dialectometric analyses, it seems appropriate to denominate it properly: we proposed for it the following terms: *special entanglement*, Ger. *spezielle Verzahnung*, Fr. *enchevêtrement particulier*, It. *intreccio particolare*, Sp. *entramado especial* etc.

It's highly probable that the special entanglement that also occurs in many other geo-based sciences belongs to the founding principles of all spatial networks. It's not less probable that it is a direct consequence of diversification processes that operate continuously in such reticulated structures. G. Altmann (1985) modeled these circumstances, referring to the two main Zipfian processes of *diversification* and *unification*, which permanently alter the inner structure of geolinguistic grids[10]. In particular, Altmann conjectured the interplay of *birth*-and-*death*-processes that created or annihilated the single components (or *taxates*) of the respective network.

From the standpoint of sociology and sociolinguistics the phenomenon of special entanglement can be interpreted as a consequence of a particular communicative behavior of man – generally conceived as HOMO *LOQUENS* – in space. For many years now[11] we've called it the "basilectal management of space by man". Note that similar concepts also exist in geography and anthropology

[8] For a thorough discussion of this question and the related problems see Malkiel 1967 and Christmann 1971.

[9] The same phenomenon was addressed in 1876 by the Italian linguist G. I. Ascoli (1829-1907) when he claimed that « dialects », conceived as geotypological concepts, are defined by a « particolar combinazione » of a set of linguistic traits.

[10] See also the bibliography of diversification compiled by K.-H. Best (2014).

[11] Cf. Goebl 1993, 277.

where the idea that the natural dimensions of space can be altered by human activity is very common.

A	B	C	D	E	F	G	H	I
Linguistic Atlas	**Number of Original Maps**	**Number of analyzed Original Maps**	**Number of classifiers (England : atlas authors)**	**Number of analyzed Working Maps**	**Number of Inquiry Points**	**Number of Taxates/ Taxatorial Areas**	**Average Size of the Taxatorial Areas (= E×F/G)**	**Range of Polynymy from 2 to *x***
ALF	1 421	626	1 (HG)	1 681	641	19 328	55,74	90
England TOT	1 711	1 516	11	1 524	313	16 810	28,37	108
AES	424	424	4	424	313	5 838	22,73	39
CLAE (I+II)	315×2 (2 taxatorial levels)	591	2	597	313+1	7 698	24,35	108
LAE	406	388	3	389	313+1	2 839	43,02	21
WGE	251	114	2	114	313+1	435	82,28	7

Table 1. Empirical and taxatorial characteristics of one French (ALF) and four English linguistic atlases (AES, CLAE [I and II], LAE, and WGE).

5. Evidence from French Dialects (ALF)

See Table 1 (above) and Appendix (Map 1, Figure 1 and Figure 2).

We will now show two law-like regularities, in the stock of WM and TA in the ALF that occurred in very similar form in all our dialectometric analyses of a great number of Romance, Germanic and English linguistic atlases.

The 1 421 maps of the series A of the ALF contain 638 inquiry points and show the results of 639 inquiries done by the fieldworker Edmond Edmont (1849-1926) between 1897 and 1901[12]. They cover all linguistic categories, from phonetics to syntax. Between 1996 and 1999, 626 out of these 1 421 maps were analyzed in Salzburg for dialectometric purposes[13]. The result is 1 681 WM containing 19 328 taxates and their respective areas. They still cover all linguistic categories.

Regarding the 1 681 WM, Figure 1 shows the very regular relationship between the increase in their inner polynymy, and the decrease in their absolute frequency. In other words, many WM have very simple structures; very few WM offer highly variegated structures.

[12] For the ALF in general see Brun-Trigaud/Le Berre/Le Dû 2005.

[13] The results of the dialectometrization of the ALF have been presented in a long series of articles published in different languages : see our contributions from 2000 to 2014b.

This relation was studied by G. Altmann in 1985 under the assumption of the permanent pull of self-regulating birth-death-processes. As a result, he defined the so-called "Goebl-Law", which applies to diversification-processes in geolinguistic data[14].

From a merely geolinguistic point of view, Figure 1 very clearly shows the percentages of "beautiful" and "chaotic" maps in an atlas corpus. Normally, geolinguistic handbooks and readers discuss only "beautiful" maps (with reduced polynymy between 2 to 10), neglecting completely highly polynymic maps for reasons of excessive "chaoticity". Nevertheless, it was shown with dialectometric means that corpuses with, one the one hand, *low*-polynymic and, on the other, *high*-polynymic WM contain exactly the same deep structures[15]. This proves that there is a great amount of *redundancy* in the global geolinguistic deep structures.

The same effect was demonstrated in the 1980s by manipulating systematically the *quantity* of WM to be combined dialectometrically. It was shown that the overall deep structures of the respective data stocks already appeared from the synthesis of approx. 200 randomly chosen WM[16].

Figure 2 shows the frequency distribution of 19 328 taxatorial areas whose geographic *size* oscillates between 1 (with high frequency) and 640 (= N – 1) (with low frequency). Our assumption is that the above mentioned Zipfian forces are also responsible in this case for the apparent regularity of the curve.

We remember that the aforementioned ALF data were elaborated in a special research project realized under the exclusive responsibility of H. Goebl. So, the collected data reflect what could be called his personal "geolinguistic bias". In the next chapter it will be shown that the same results emerge when combining different "personal geolinguistic biases".

6. Evidence from English Dialects (AES, CLAE, LAE, WGE)

See Table 1 (above) and Appendix (Map 1, Figure 3 and Figure 4).

The history of English geolinguistics is completely independent from French geolinguistics, both its practical experiences and its brilliant scientific achievements. The respective fieldwork began in France in 1897 and did not commence in England until 1950. It's very strange to see that the younger English initiative neglected completely the lessons of the older French one. The name of the English initiative is "Survey of English dialects" (SED). One of the strangest peculiarities of the English initiative was the publication of the collected data (embracing only 313 inquiry points) not in *cartographic* but exclusively in *tabular* form. So, the 12 data volumes of SED, published between 1962 and 1971, are far from being as suggestive as the large in folio maps of the ALF or

[14] In his study Altmann referred to analyses presented in Goebl 1984, based themselves on data drawn from the linguistic atlases AIS and ALF.
[15] See Goebl 2014a.
[16] See Goebl 1984 I, 206 ss.

AIS. So, English linguists could not benefit from the illuminating effects of *full-text maps* and their current (generalized) elaboration by means of *mute maps*.

As a consequence some English linguists became "privileged" interpreters of selected portions of the SED data, by publishing the results of their classificatory analyses of a certain amount of SED tables under the slightly misleading title "Linguistic atlas of…". This procedure holds for the WGE (1974), LAE (1978), AES (1979), and CLAE (1991, 1997) "atlases", none of which contains original dialect data, but instead coded maps.

CLAE had the benefit of being produced already by electronic means. Its author, Wolfgang Viereck (Bamberg), handed us over the electronic files of the two volumes of CLAE in the 1990s for further dialectometric analyses[17]. Given the particularly interesting results of the dialectometrization of these data, we subsequently decided to grasp the data of other similar English "linguistic atlases", all derived from SED. Although the respective data entry was rather laborious, it was well accomplished thanks to the precision and energy of our Salzburg collaborators.

It should be emphasized that the data collection generated in this way reflects the classificatory "philosophy" of 11 different Anglicists, embracing all linguistic categories.

Nevertheless, Figures 3 and 4 show exactly the same quantitative tendencies that we already saw in case of the ALF. In Figure 3 the polynymy of the 1 524 WM goes from 2 to 108[18], whereas in Figure 4 the size of the 16 810 TA varies between 1 and 310 (inquiry points or polygons). Obviously, the numbers shown in Table 1 (see above) for "England TOT" represent the sum of those of the AES, CLAE, LAE, and WGE.

Note that the 315 coded maps of the two volumes of the CLAE show two taxatation levels: a "lumped" (i.e. with a more coarse structure) and a "split" (i.e. with finer granulation) one. Thus, the number of the respective WM has been doubled.

In a nutshell, it seems to be evident that simply counting the frequency of the two basic units of dialectometric data matrices – *working maps* (WM) and *taxates* or *taxatorial areas* (TA) – produces clear-cut quantitative regularities that reflect some elementary properties of the dialectal behavior of man in space.

7. Concluding Remarks

Summing up, we like to emphasize some historical facts. One of the greatest discoveries of the last quarter of the 19th century was the theory of the general regularity of sound change. This remains true despite the many exaggerations and confusions that have been perpetrated since then. The central point of these discussions was the analysis of the change of linguistic utterances along the axis of *time*, done under the tacit assumption that *time* represents an absolute term.

[17] See Goebl 1997 and Goebl/Schiltz 1997.

[18] The polynymy 108 occurs on the map S9 ("I know a man [*who*] will do it for you.") of CLAE I (of 1991).

One of the errors committed during this time was the claim that a particular sound change in the language x must be valid also in all geographical varieties (dialects) of the same language. Many maps of the early linguistic atlases proved that this argument was wrong.

Even though the faultiness of this argument has been proved time and again, no serious discussions arose on the relationship between linguistic behavior in *time* and *space* and to what extent the famous linearity of *time* and the orderly structure of time-related linguistic utterances could have a counterpart in *space*. Obviously, coping with the challenges of *time* was much easier than coping with those of *space*. So almost hundred years passed between the beginning of the sound law discussions in the circle of the Leipzig neogrammarians (1876)[19] and the earliest publication of genuine dialectometry (Séguy 1971).

Nowadays, it should be taken for granted that a language evolves in *space* under the same constraints of "non-chaoticity" as it does in *time* (and perhaps also in other dimensions).

8. Acknowledgments

Drawing of Map 1: Yves Scherrer, Geneva,
Drawing of the Figures 1-4: Werner Goebl, Vienna,
Stylistic supervision: Benjamin Wright, Salzburg.

References

AES: Kolb, E., Glauser, B., Elmer, W., and Stamm, R. (eds.) (1979). *Atlas of English Sounds*, Bern: Francke.

AIS: Jaberg, K., Jud, J. (eds.) (1928-1940). *Sprach- und Sachatlas Italiens und der Südschweiz*, 8 vols., Zofingen: Ringier (reprint: Nendeln: Kraus, 1971) (web-based version: http://www3.pd.istc.cnr.it/navigais/).

ALF: Gilliéron, J., Edmont, E. (eds.) (1902-1910). *Atlas linguistique de la France*, 10 vols., Paris: Champion (reprint: Bologna: Forni, 1968) (two web-based versions: 1 : http://diglib.uibk.ac.at/urn:nbn:at:at-ubi:2-4568; 2 : http://cartodialect.imag.fr/cartoDialect/carteTheme).

Altmann, G. (1985). Die Entstehung diatopischer Varianten. *Zeitschrift für Sprachwissenschaft* 4, 139-155.

Ascoli, G. I. (1876). Paul Meyer e il franco-provenzale. *Archivio glottologico italiano* 2, 385-395.

Berschin, H., Felixberger, J., Goebl, H. (2008[2]). *Französische Sprachgeschichte. Lateinische Basis, interne und externe Geschichte, sprachliche Gliederung Frankreichs. Mit einer Einführung in die historische Sprachwissenschaft*, Hildesheim: Olms.

[19] This year saw the appearance of the famous treatise of August Leskien on the "Deklination im Slawisch-Litauischen und Germanischen" among others; for further details see Putschke 2001, 1508.

Best, K.-H. (2014). Diversification: Bibliography. *Glottometrics* 28, 87-91.

Brun-Trigaud, G., Le Berre, Y., Le Dû, J. (2005). *Lectures de l'Atlas linguistique de la France de Gilliéron et Edmont. Du temps dans l'espace. Essai d'interprétation des cartes de l'*Atlas linguistique de la France *de Jules Gilliéron et Edmond Edmont augmenté de quelques cartes de l'*Atlas linguistique de la Basse-Bretagne *de Pierre Le Roux*, Paris: CTHS.

CLAE: Viereck, W., Ramisch, H. (eds.) (1991, 1997). *The Computer Developed Linguistic Atlas of England*, Tübingen: Niemeyer, 2 vols.

Christmann, H. H. (1971). Lautgesetze und Wortgeschichte. Zu dem Satz „Jedes Wort hat seine eigene Geschichte". In: *Sprache und Geschichte. Festschrift für Harri Meier zum 65. Geburtstag*, Coseriu, E., Stempel, W.-D. (eds.), München: Fink, 111-124.

DSA: *Deutscher Sprachatlas, aufgrund des von Georg Wenker begründeten Sprachatlas des Deutschen Reiches in vereinfachter Form begonnen von Ferdinand Wrede, fortgesetzt von Walther Mitzka und Bernhard Martin*, Marburg/Lahn: Elwert 1927-1956, 4 vols. (23 fascicles with 128 maps).

Gilliéron, J., Mongin, J. (1905). *Scier dans la Gaule romane du sud et de l'est. Étude de géographie linguistique*, Paris: Champion (Italian translation: *« Segare » nella Gallia romanza meridionale e orientale*, by Lorenzo Massobrio, Novi Ligure: Grafica editoriale universitararia, 1990).

Goebl, H. (1984). *Dialektometrische Studien. Anhand italoromanischer, räto-romanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, 3 vols., Tübingen: Niemeyer.

Goebl, H. (1993). Dialectometry. A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data. In: Köhler, R., Rieger, B. B. (eds.). *Contributions to Quantitative Linguistics*, Dordrecht: Kluwer, 277-315.

Goebl, H. (1997). Some Dendrographic Classifications of the Data of CLAE 1 and CLAE 2. In: *CLAE* II, 23-32.

Goebl, H. (2000). La dialectométrisation de l'ALF: présentation des premiers résultats. *Linguistica* 40, 209-236.

Goebl, H. (2002a). Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de Linguistique Romane* 66, 5-63.

Goebl, H. (2002b). Sprachatlanten: woher? womit? wozu? Einige buntgemischte Reflexionen am Gartenzaun zwischen Romanistik und Germanistik. In: *Namen, Sprachen und Kulturen. Imena, Jeziki in Kulture. Festschrift für Heinz Dieter Pohl zum 60. Geburtstag*, Anreiter, P., Ernst, P., Hausner, I., and Kalb, H. (eds.), Wien: Edition Präsens, 257-274.

Goebl, H. (2003). Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur. *Estudis Romànics* XXV, 59-121.

Goebl, H. (2005). La dialectométrie corrélative. Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de Linguistique Romane* 69, 321-367.

Goebl, H. (2006a). Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21/4, 411-435.

Goebl, H. (2006b). Warum die Dialektometrie nur in einem roman(ist)ischen Forschungskontext entstehen konnte. In: Dahmen, W., Holtus, G., Kramer, J., Metzeltin, M., Schweickard, W., and Winkelmann, O. (eds.), *Was kann eine vergleichende romanische Sprachwissenschaft heute (noch) leisten?* Romanistisches Kolloquium XX, Tübingen: Narr, 291-317.

Goebl, H. (2007a). A Bunch of Dialectometric Flowers: a brief Introduction to Dialectometry. In*:* Smit, U., Dollinger, St., Hüttner, J., Kaltenböck, G., and Lutzky, U. (eds.), *Tracing English through Time. Explorations in Language Variation.* In Honour of Herbert Schendl on the Occasion of his 65th Birthday, Wien: Braumüller, 133-171.

Goebl, H. (2007b). Dialectometry: Theoretical Prerequisites, Practical Problems, and Concrete Applications (mainly with Examples drawn from the "Atlas linguistique de la France", 1902-1010). In*: Geolinguistics around the World.* Proceedings of the 14th NIJLA [National Institute of Japanese Language] International Symposium (Tokyo, August 22-23, 2007), Tokyo: NIJLA, 65-74.

Goebl, H. (2009). Quelques coups d'oeil dialectométriques sur l'Atlas linguistique de la France: structures de surface et structures de profondeur. In*:* Dalbera-Stefanaggi, M.-J., Simoni-Aurembou, M.-R. (eds.), *Images de la langue: représentations spatiales, sémantiques et graphiques*, Paris: Editions du CTHS, 39-60.

Goebl, H. (2010). Dialectometry and quantitative mapping. In: Lameli, A., Kehrein, R., and Rabanus, St. (eds.), *Language and Space. An International Handbook of Linguistic Variation*, vol. 2: *Language Mapping* (Handbücher der Sprach- und Kommunikationswissenschaft [HSK] 30.2.), Berlin: de Gruyter; 1st part: 433-457 (text), 2d part (maps): 2201-2212.

Goebl, H. (2012). Introduction aux problèmes et méthodes de l'»École dialectométrique de Salzbourg« (avec des exemples gallo-, italo- et ibéroromans). In: Álvarez Pérez, A., Carrilho, E., and Magro, C. (eds.), *Proceedings of the International Symposium on Limits and Areas in Dialectology (LimiAr)*, Lisbon 2011. www.http://limiar.clul.ul.pt", Lisboa: Centro de Linguística da Universidade de Lisboa, 117-166.

Goebl, H. (2014). L'impact de la polynymie des cartes d'atlas sur le résultat de calculs dialectométriques. In: *Linguistique romane et Linguistique indo-européenne. Mélanges offerts à Witold Mańczak à l'occasion de son 90e anniversaire,* Polska Akademia Umiejetnosci. Instytut Filologii Romanskiej Uniwersytetu Jagiellonskiego (ed.), Kraków: Polska Akademia Umiejetnosci. Instytut Filologii Romanskiej Uniwersytetu Jagiellonskiego, 243-260.

Goebl, H., Schiltz, G. (1997). Dialectometrical Compilation of CLAE 1 and CLAE 2. Isoglosses and Dialect Integration. In: *CLAE* II, 13-21.

Goebl, H., Smečka, P. (2014). L'analyse dialectométrique des cartes de la série B de l'ALF. *Revue de Linguistique Romane* 78, 439-497.

Jaberg, K. (1906). Zum Atlas linguistique de la France. *Zeitschrift für romanische Philologie* 30, 512.

Jaberg, K. (1908). *Sprachgeographie. Beitrag zum Verständnis des Atlas linguistique de la France*, Aarau: Sauerländer (Spanish translation: *Geografía lingüística. Ensayo de interpretación del „Atlas lingüístico de Francia“*, by Llorente, A., Alvar, M., Granada: Universidad de Granada. Secretariado de Publicaciones, 1959).

LAE: Orton, H., Sanderson, St., and Widdowson, J. (eds.) (1978). *The Linguistic Atlas of England*, London: Croom Helm.

Malkiel, Y. (1967). Each Word has a History of its Own. *Glossa* 1, 137-149.

Paris, G. (1888). Les parlers de France. In: Paris, G., *Mélanges linguistiques. Latin vulgaire, langues romanes, langue française, notes étymologiques*, publiés par Mario Roques, Paris: Champion, 1909, 432-448.

Putschke, W. (2001). Die Dialektologie, ihr Beitrag zur historischen Sprachwissenschaft im 19. Jahrhundert und ihre Kritik am junggrammatischen Programm. In: Auroux, S., Koerner, E.F.K., Niederehe, H.-J., and Versteegh, K. (eds.), *History of Language Sciences. Geschichte der Sprachwissenschaften. Histoire des sciences du langage. An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present. Ein internationales Handbuch zur Entwicklung der Sprachforschung von den Anfängen bis zur Gegenwart. Manuel international sur l'évolution de l'étude du langage des origines à nos jours*, Berlin, New York: Walter de Gruyter, vol. 2, 1498-1513.

Schuchardt, H. (1870). *Über die Klassifikation der romanischen Mundarten* [Probe-Vorlesung gehalten zu Leipzig am 30. April 1870], Graz: Styria 1900 (also published in: Spitzer, L. (1928[2]). *Hugo Schuchardt-Brevier. Ein Vademecum der allgemeinen Sprachwissenschaft*, Halle: Niemeyer, 166-188; reprint: Tübingen: Niemeyer, and Darmstadt: Wissenschaftliche Buchgesellschaft, 1976).

SED: Orton, H., Halliday, W. J., Dieth, E., and Wakelin, M. F. (eds.) (1962-1971). *Survey of English Dialects. The Basic Material*, Leeds: E. J. Arnold, 12 vols. (reprint: London: Routledge, 1998).

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35, 335-357.

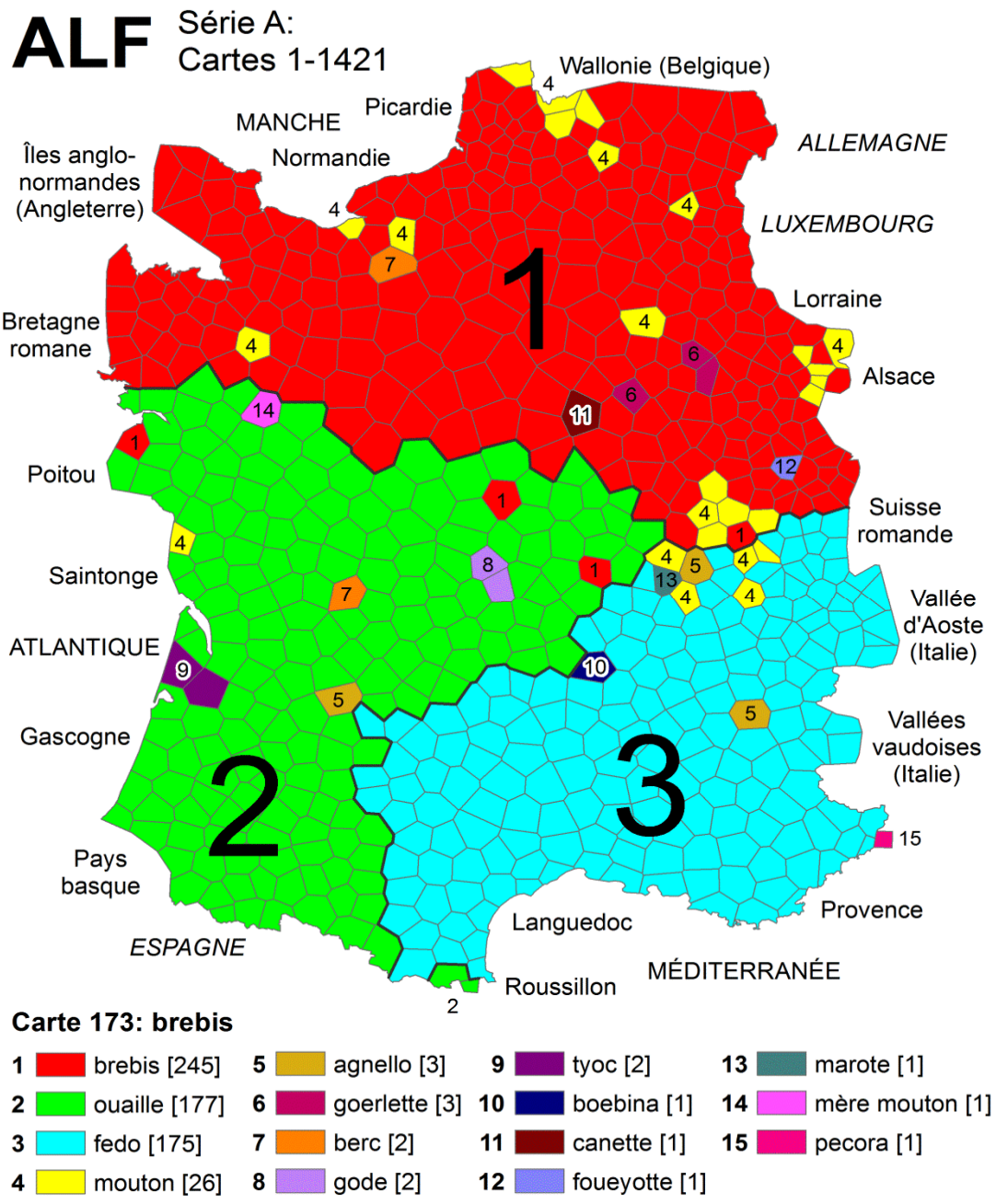
Wartburg, W. v. (1918). *Zur Benennung des Schafes in den romanischen* Sprachen : *ein Beitrag zur Frage der provinziellen Differenzierung des späten Lateins*, Berlin: Reimer.

Wartburg, W. v. (1963[2]). *Problèmes et méthodes de la linguistique*, Paris: Presses Universitaires de France, translated from German by Pierre Maillard with the participation of Stephen Ullmann, second edition.

Wenker, G. (2013). *Schriften zum Sprachatlas des Deutschen Reichs. Gesamtausgabe. Band I: Handschriften: Allgemeine Texte. Kartenkommentare 1889-1897*, published by Alfred Lameli, Hildesheim, Zürich, and New York: Olms.

WGE: Orton, H., Wright, N. (eds.) (1974). *A Word Geography of England*, London, New York: Seminar Press.

Appendix



Map 1: Taxatorial analysis (“working map”) of map 173 (*la brebis*) of the ALF showing the geographical distribution of fifteen Gallo-Romance denominations (“geo-synonyms”) of the “ewe”.

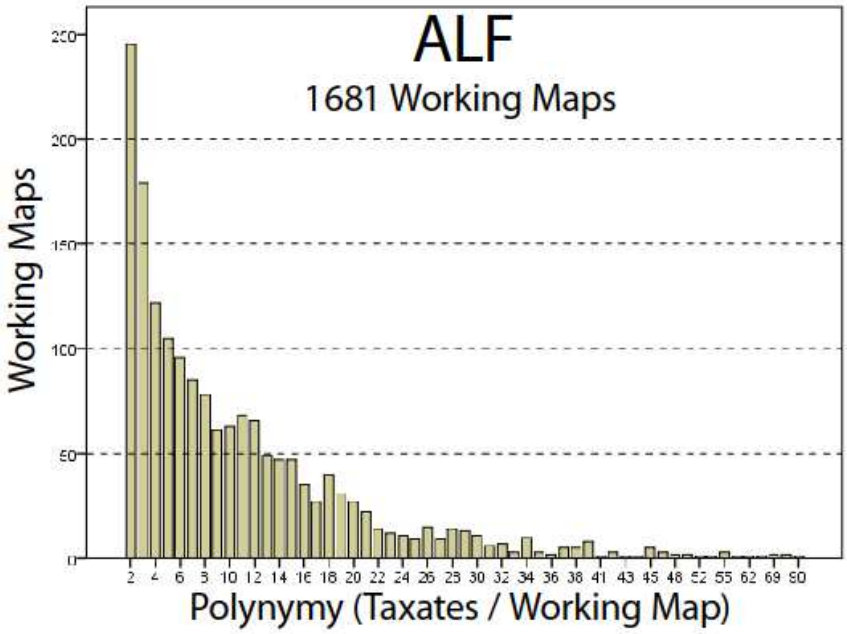


Figure 1. Histogram showing the relationship between geolinguistic polynymy and the number of working maps (WM). Data: 626 original maps of the ALF (1902-1910), taxatation (typification) encompassing all linguistic categories, 1 681 WM. The polynymy oscillates between 2 and 90 taxates per WM; the number of WM varies between 245 (2-nym WM) and 1 (90-nym WM).

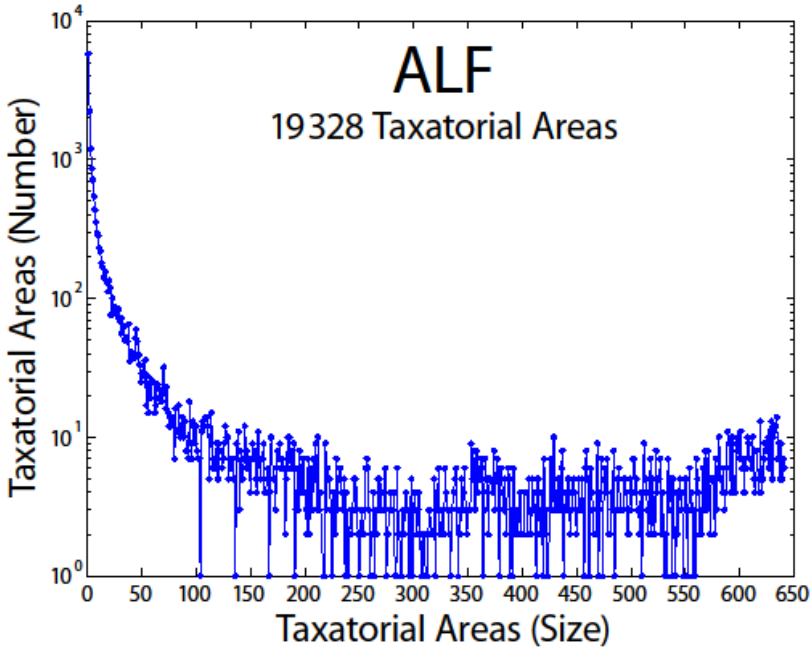


Figure 2. Diagram showing the relationship between size and number of 19 328 taxatorial areas (TA). Data: 626 original maps of the ALF (1902-1910), taxatation (typification) encompassing all linguistic categories, providing 1 681 WM, and 19 328 TA. The size of TA oscillates between 640 (inquiry points or polygons) and 1, their number between 5 743 and 1.

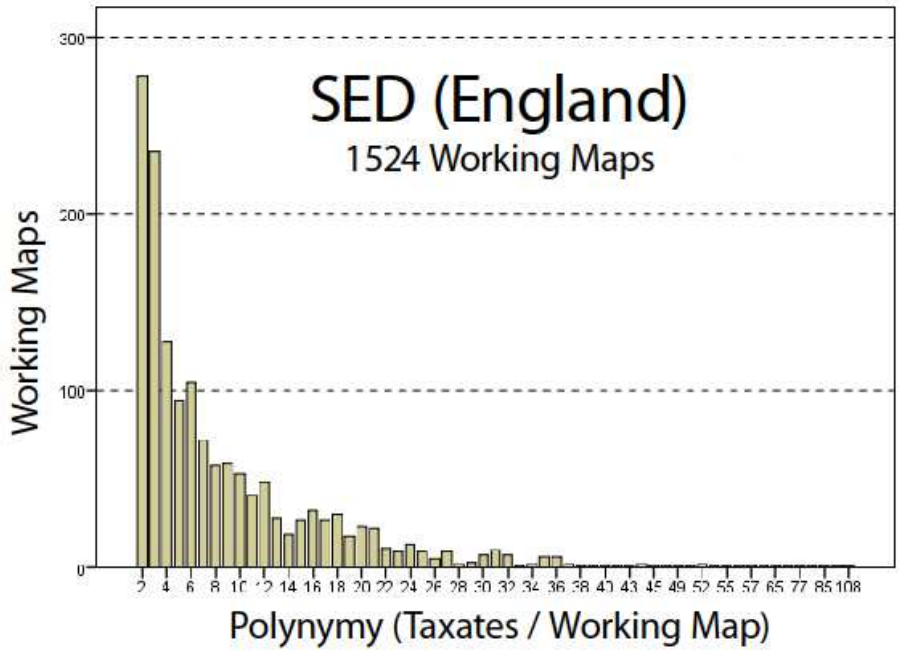


Figure 3. Histogram showing the relationship between geolinguistic polynymy and the number of working maps (WM). Data: 1 516 original maps of the AES, CLAE (I and II), LAE, and WGE, taxatation (typification) encompassing all linguistic categories, providing 1 524 WM. The polynymy oscillates between 2 and 108 taxates per WM; the number of WM varies between 278 and 1.

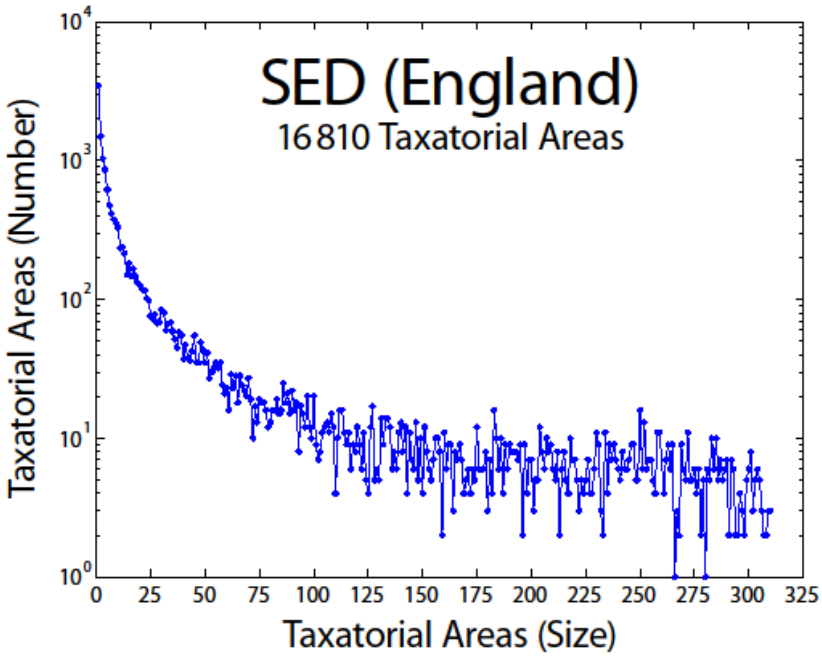


Figure 4: Diagram showing the relationship between the size and the number of 16 810 taxatorial areas (TA). Data: 1 516 original maps of AES, CLAE (I and II), LAE, and WGE, taxatation (typification) encompassing all linguistic categories, providing 1 524 WM, and 16 810 TA. The size of TA oscillates between 310 (inquiry points or polygons) and 1, their number between 3 477 and 1.

Play with the Data!

Sheila Embleton, Dorin Uritescu, Eric S. Wheeler
York University, Toronto, Canada

Abstract. A data point on a plot or graph sometimes is intended to represent not just one observation, but rather the result of processing a set of observations. In such cases, it pays to look at subsets of thoseobservations to see how much variation the data point implicitly represents. Playing with the data and exploring different possibilities can not only clarify this point, but also suggest interesting directions for further discovery. We illustrate with a fictitious example, as well as two real examples from our Finnish and Romanian data sets**.**

Keywords: *data variance, Finnish, Romanian, RODA, GODA, Romanian Online Dialect Atlas, General Online Dialect Atlas*

1. Introduction

Consider a comparative study of English, French and German, examining the proportion of times a grammatical subject precedes the verb in a given corpus, vs. the position of an adjective before or after the noun. The corpora having been chosen, the counts done, and the calculations made, we might chose to represent the results on a plot with English in one corner, German in another, and French somewhere in between (this being a fictitious study, the results are too!). It goes without saying that the point representing English is highly suspect because of the great variation in English dialects, style, genre, register and so on. But even if we control for such factors, and say that our “English” point represents the chosen corpus of English (or French or German) so that we can draw conclusions about the similarities and differences of these languages based on these selections, there is still an issue remaining. How much variation is in the corpus itself? That is to say, if the corpus were reduced, edited, corrected, or otherwise modified even a small amount, would the point on the plot move a little or a lot? Consider dividing the corpus into arbitrary subsets and repeating the analysis on each. Where would the English points lie? Close to the original point or widely distributed over the plot? (See Illustration 1.)

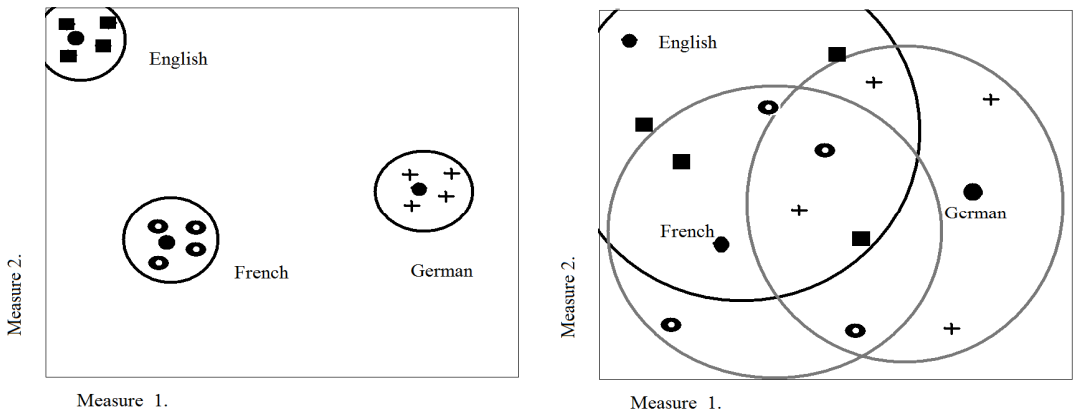


Illustration 1: Fictitious plot comparing 3 languages, with tight and wide variance

each. Where would the English points lie? Close to the original point or widely distributed over the plot? (See Illustration 1.)

In the first case, we might be justified in drawing conclusions about the differences among the 3 languages, but it would be much harder to justify such conclusions in the second case.

And yet, all too often what we have are simply the single points for the corpus as a whole. (We refrain from naming particular cases, because we are not critiquing such works in full, but we have come across many cases, published and unpublished, and perhaps have committed the sin ourselves.)

2. Finnish

To illustrate what we are saying, we took a second look at the Finnish data we have collected and analyzed using the General Online Dialect Atlas (GODA; see Embleton, Wheeler 1997, 2000). The Finnish data comes from a hard-copy atlas (Kettunen 1940), digitized by us. Using Kettunen's traditional labels for dialect regions, and the multidimensional scaling procedure (MDS) built into GODA, we are able to create a plot of the various field locations, arranged by traditional dialect group, and plotted according to their linguistic “closeness” to one another. See Illustration 2.

The point here is that each dot represents the MDS processing of 213 responses for a given location. What happens if we subset those responses? Will the dots move and if so, how much? In earlier parallel work done on Romanian, we found that subsetting the data according to function (phonology, morphology, syntax, etc.) made a difference (Embleton, Uritescu, Wheeler 2013). Here, we want to consider what happens when the data set is subdivided arbitrarily, so that we cannot attribute the variation to any particular factor.

We cheat a little. Instead of selecting subsets of the 213 files randomly, we simply divided the 213 into four “quarters” of 53 items (the last has 54) each. Since the files are not organized in any particular way, this is close enough to random to make our point.

Furthermore, instead of looking at all 529 locations, we concentrate on 5 traditional dialect groups (labelled “I 1a”, “II 1a”, “II 3”, “II 4” and “I-II”, all

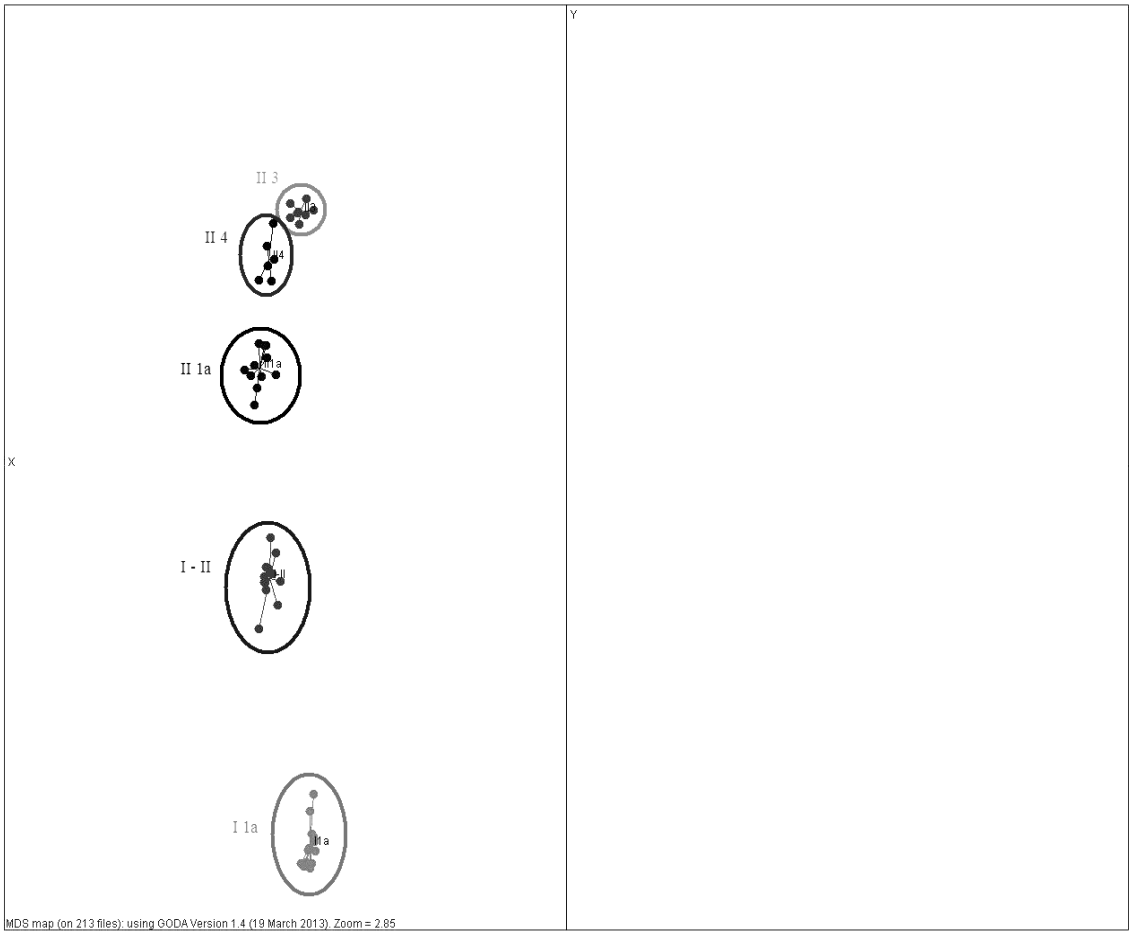


Illustration 2: Finnish dialects, selected regions. Circles added to show the unity of each region.

western dialects from the south western part of Finland, and each with about 6 to 10 field locations; GODA makes it easy to show only a selection of groups even though the MDS procedure works on all of them). In this case, the groups are compact (as shown by circles added to the map) and nicely spread apart, more or less in a line (Illustration 3) .

Illustration 3: Finnish dialects: MDS on all 213 selections for all field

Even here though, if we consider the circles to be the “points”, we see that each point has “variance”: each field location is a little different from the others in its group.

Multidimensional scaling is a process that uses a distance matrix (i.e. linguistic distances between each pair of field locations) to position the locations exactly at the right distance from one another -- but in a high-dimensional space. To visualize this high-dimensional space, the points are projected to a two-dimensional plot in an optimal way (or in the case of GODA, to a three-dimensional space, that can be rotated, and enlarged to give a desirable two-dimensional view). However, the projection means that when we do MDS on each of the “quarters” of the data, the views are not going to use precisely the same viewpoint. We show all four results side by side below (Illustration 4) so that you can see how the “nice” picture in Illustration 3 gets modified.

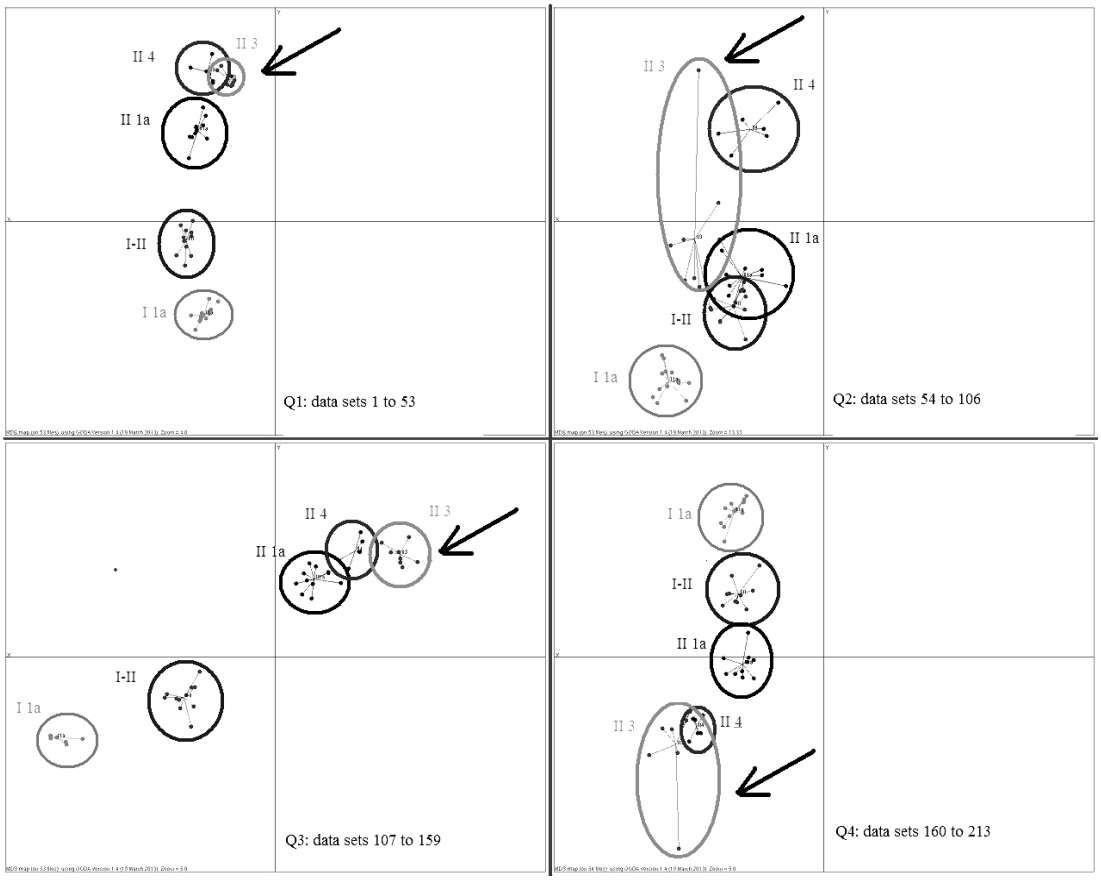


Illustration 4: Four MDS plots, one for each of the four "quarters" of the Finnish data. Circles added.

Q1 in Illustration 4 looks much the same as the full data set (Illustration 3), and so does Q3 (ignoring the rotation of the view). But in Q2 and Q4, the one circle for region “II 3” has got very much larger, and its position relative to the others has changed some. The neat full-data picture managed to hide some variation that was underlyingly there.

Furthermore, it appears that one of the field locations is an outlier (for that group – it may not be odd if it were assigned to another group). Outliers are worth looking at, to see what it is that makes them odd. This plot does not explain that oddness, but it can prompt the researcher to look for an explanation.

3. Romanian

For a second example, we consider some work-in-progress (Embleton, Uritescu and Wheeler 2016) where we intend to compare various measures of geographic distance to linguistic distance in our Romanian database. In one case, we have 377 interpretive maps that provide the linguistic distance. Following our advice

above, we created 5 randomly chosen subsets of the base, with approximately half the maps in each random sample (no cheating here; using a Java random-number generator, we gave each location a 50/50 chance of being included in a sample). We compared each random distance matrix to the base distance matrix using a Mantel test (a test that measures the correlation between two distance matrices of the same size). The results were that the base map correlated to itself with a value of 1.0 (as it should, by definition) and the random samples had correlations ranging from 0.965 to 0.975 with an average of 0.970. The random samples, then, were all very close to one another and to the base. More important, they are distinctly different from the correlations between geographic and linguistic distance (which were around 0.8). In other words, the 377 maps were a homogeneous set, and it is reasonable to use the set in further comparisons.

4. Conclusion

The moral to this story, then, is that it pays to play with the data – to try to use our analytical procedures in various ways, on various aspects of the data. In particular, we should try to understand what is behind a derived "fact" (i.e. a point on a plot) to be sure it is indeed what is seems to be. And, if along the way, we discover some pointers for further interesting exploration, all the better. Play with the data! One never knows what one might discover.

References

Embleton, Sheila, Eric S. Wheeler (1997). 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 (2013). Defining dialect regions with interpretations. Advancing the multidimensional scaling approach. Presentation at Methods 14, University of Western Ontario, London Canada, August 2011. Published in *Literary and Linguistics Computing* 2013, volume 28, number 1. Free access to PDF at http://llc.oxfordjournals.org/cgi/reprint/fqs048?ijkey=NzoVKzEF9cRtmO3&keytype=ref

Embleton, Sheila, Dorin Uritescu, Eric S. Wheeler (2016). An Expanded Quantitative Study of Linguistic vs. Geographic Distance Using Romanian Dialect Data. Work-in-progress presented at the International Quantitative Linguistics Conference, Trier Germany. August 2016

Kettunen, Lauri (1940). *Suomen murrekartasto* [The dialect atlas of Finland]. Helsinki: Suomalaisen kirjallisuuden seura.

The First End-Of-Year Address by the New President of the Italian Republic Sergio Mattarella

Michele A. Cortelazzo

cortmic@unipd.it, DISLL – University of Padova, Italy

Arjuna Tuzzi

arjuna.tuzzi@unipd.it, Dep.t FISPPA – University of Padova, Italy

Abstract. This study aimed at examining to what extent Mattarella's first End-of-Year address resembles those of his predecessors by comparing the lexical profiles of all available addresses using Labbé's intertextual distance, cluster analysis and correspondence analysis. A substantial continuity with Giorgio Napolitano emerged from all analyses and we could say that Mattarella's first End-of-Year discourse followed the example of his predecessor and, at least for the moment, he conformed to that model. These findings could be indicative that there has been a shift in the pattern of similarity noted in the presidents' addresses as a chronological contiguity, not noted at the beginning of the republic's history, was found in recent addresses.

Keywords: *correspondence analysis; intertextual distance; political discourse; Presidential address; similarity measure; text clustering*

1. Introduction

This work is part of a series[1] of quantitative analyses of the End-of-Year addresses of presidents of the Italian Republic (1949-2015). The study is part of an international research frame focusing on addresses by important heads of states of different countries; those most studied were presented by French (e.g. Finniss-Boursin 1992, Labbé 1990, Leblanc 2003, Leblanc, Martinez 2005, Marchand 2007, Mayaffre 2004, 2012, Teletin 2013; and for French-Canadian: Labbé, Monière 2003, 2008) and American statesmen (e.g. Giuliano 2014, Liu 2012, Nye 2013, Savoy 2010, 2015, Schlesinger 1997). While not expressly foreseen in the Italian Constitution, the president of the Italian Republic's End-of-Year address has a notable symbolic value because it refers to a presentation by the nation's most authoritative state official (and according to many surveys, to the most beloved one) who directly addresses his fellow citizens, and thereby instilling an enduring link between the country's institutions and its people. With

[1] For example, cfr. the book edited by Cortelazzo, Tuzzi (2007a) and the book written by Tuzzi, Popescu, Altmann (2010); cfr. the articles Bernardi, Tuzzi (2010), Köhler, Tuzzi (2015), Pauli, Tuzzi (2009), Trevisani, Tuzzi (2013), Tuzzi, Köhler (2015), Tuzzi et al. (2012).

the passing of time, the message has become the occasion for the president to extend traditional greetings to the Italian people and it has assumed the function of a civil ceremony and the characteristics of a media event.

In 2015, President Giorgio Napolitano, who was serving the second year of a second seven-year term, filed his resignation and a new president was elected. During his first End-of-Year message later that year, President Sergio Mattarella displayed a style confirming in part the impression that he made when he first took office earlier in the year but also showing important variations. Mattarella's inauguration speech contained wise doses of wisdom and sophistication. The syntactic structure of the presentation was quite simple: brief and rarely complex phrases were counterbalanced by a refined lexicon. He utilized words such as *inverare* ('to make true'), *dispiegare* ('deploy'), *pervasivo* ('pervasive'), *inferto* ('dealt'). The speech was able to smooth its fragmentary nature linked to the brief phrases by an able use of rhetorical devices and in particular by frequent usage of lexical repetition, parallelisms, lists.

According to the Italian press, President Mattarella wanted to connect with his people in a more colloquial way during his first End-of-Year message. The construction of the text and the syntactic structure and lexicon achieved that aim. Moving from qualitative to quantitative considerations, we can now systematically compare Mattarella's first End-of-Year address with those of his predecessors and examine similarities and differences.

2. The corpus

The corpus of the End-of-Year addresses by presidents of the Italian Republic is made up of 67 speeches pronounced by 11 presidents; the first, by Luigi Einaudi, was given in 1949 at the second year of his term of office and the latest, by Sergio Mattarella, was delivered on December 31, 2015.

The lemmatised version of the corpus was analysed. Parts-of-speech (POS) tagging was carried out both automatically and in part corrected manually and each word was attributed to a pair (lemma, category). All of the calculations presented herein are based on a bag-of-words approach (lists of lemmas with their occurrences) and were carried out using Taltac software, a lexical and textual automatic processing for corpus and content analysis (Bolasco et al. 2009) and R, a free software environment for statistical computing and graphics (R Core Team, 2016).

As far as length is concerned (Table 1, Figures 1 and 2), Mattarella's address is slightly longer than the average ones, but given that speeches delivered over time show a pattern of progressive lengthening, its length, in terms of number of occurrences (N = 2,127), is consistent with that of the addresses of the most recent Italian presidents. Its length is similar to that of President Napolitano's first address (2006, N = 2,203); it is longer than the address by President Ciampi (1999, N = 1,939), but much shorter than that of President Scalfaro (1992, N = 2,772). If we examine, instead, the length of time it took to deliver the speech and calculate the velocity in terms of words per minute, it emerges that Mattarella presented his address at 106 words a minute compared to

a mean of 122 by his predecessor Giorgio Napolitano, 99 by Carlo Azeglio Ciampi, and 110 by Oscar Luigi Scalfaro.

It is important to remember that messages of the presidents of the Italian Republic are among the longest holiday greetings delivered by heads of states and government officials. Mattarella's address lasted 20 minutes (the standard length, with few deviations in recent End-of-Year Italian addresses)[2].

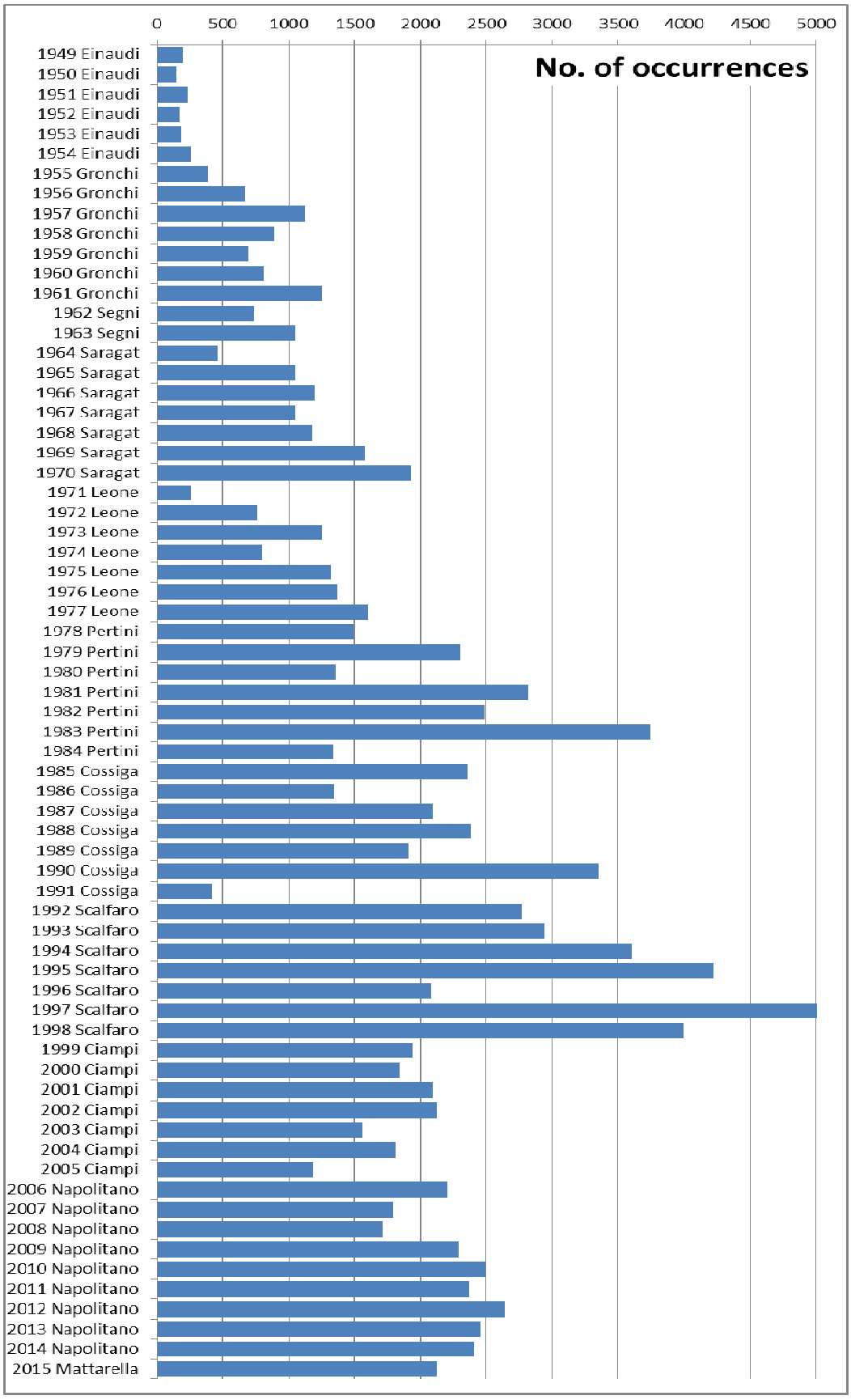
Mattarella's discourse can be defined as a simple, and for the most part, easily comprehensible text. The elements supporting this conclusion are, first and foremost, the construction of the phrases: the phrases tend for the most part to be short and with few subordinate conjunctions. Mattarella preferred to give a sense of completeness to his arguments by pronouncing independent clauses one after the other or, at the most, using coordinating conjunctions. The mean length of the sentences was 20.8 words, independent clauses represent 68.6% of the phrases, only 9.8% of the text is made up subordinating conjunctions. For comparison purposes, the address Napolitano gave the year before had sentences that contained an average of 34.8 words, independent clauses represented 55.9% of the phrases, and 17.1% of the text was made up of subordinating conjunctions meaning that the syntactic structure was completely different. There are sections, although not very many, such as the following phrase: *persone, quarantenni e cinquantenni, che il lavoro lo hanno perduto e che faticano a trovarne un altro* ('persons, forty and fifty-years old, who have lost their job/work and have difficulty in finding another') characterized by a dislocation to the left of the direct object (*lavoro*, 'job') and its repetition, obligatory in this context, with the pronoun *lo* ('it').

[2] In other countries, Christmas, End-of-Year or New Year's greetings are much shorter; in 2015, Russian president Vladimir Putin gave a holiday message that lasted 4 minutes, Queen Elizabeth II of Great Britain, together with Angela Merkel and Joachim Gauck, respectively the Chancellor and President of Germany, all gave addresses lasting less than 6 minutes. François Hollande, the president of France, spoke for less than 9 minutes and King Felipe VI of Spain spoke for approximately 12 minutes. Kim Jong-un, the leader of North Korea, was one of the heads of states whose message (lasting 30 minutes) was longer than that of President Mattarella.

Table 1
Discourse size in terms of No. of occurrences (tokens) and different words (lemma-types).

	tokens	lemma-types
1949 Einaudi	194	119
1950 Einaudi	150	91
1951 Einaudi	230	149
1952 Einaudi	179	123
1953 Einaudi	190	120
1954 Einaudi	260	154
1955 Gronchi	388	205
1956 Gronchi	665	320
1957 Gronchi	1130	459
1958 Gronchi	886	378
1959 Gronchi	697	331
1960 Gronchi	804	374
1961 Gronchi	1252	510
1962 Segni	738	328
1963 Segni	1057	448
1964 Saragat	465	224
1965 Saragat	1053	427
1966 Saragat	1199	499
1967 Saragat	1056	456
1968 Saragat	1174	465
1969 Saragat	1584	572
1970 Saragat	1929	670
1971 Leone	262	141
1972 Leone	767	327
1973 Leone	1250	503
1974 Leone	801	346
1975 Leone	1328	529
1976 Leone	1366	531
1977 Leone	1604	578
1978 Pertini	1493	480
1979 Pertini	2302	621
1980 Pertini	1360	425
1981 Pertini	2818	694
1982 Pertini	2486	664
1983 Pertini	3748	880
1984 Pertini	1340	396
1985 Cossiga	2359	698
1986 Cossiga	1349	471
1987 Cossiga	2091	755
1988 Cossiga	2385	708
1989 Cossiga	1912	646
1990 Cossiga	3346	974
1991 Cossiga	418	206
1992 Scalfaro	2772	781
1993 Scalfaro	2941	857
1994 Scalfaro	3605	922
1995 Scalfaro	4228	1036
1996 Scalfaro	2085	698
1997 Scalfaro	5013	1058
1998 Scalfaro	3995	910
1999 Ciampi	1939	654
2000 Ciampi	1844	668
2001 Ciampi	2094	724
2002 Ciampi	2126	743
2003 Ciampi	1562	574
2004 Ciampi	1806	650
2005 Ciampi	1191	436
2006 Napolitano	2203	759
2007 Napolitano	1793	660
2008 Napolitano	1712	616
2009 Napolitano	2293	764
2010 Napolitano	2498	844
2011 Napolitano	2365	832
2012 Napolitano	2642	901
2013 Napolitano	2453	848
2014 Napolitano	2409	871
2015 Mattarella	2127	749
Corpus	**113761**	**6700**

Figure 1. Discourses size in terms of No. of occurrences



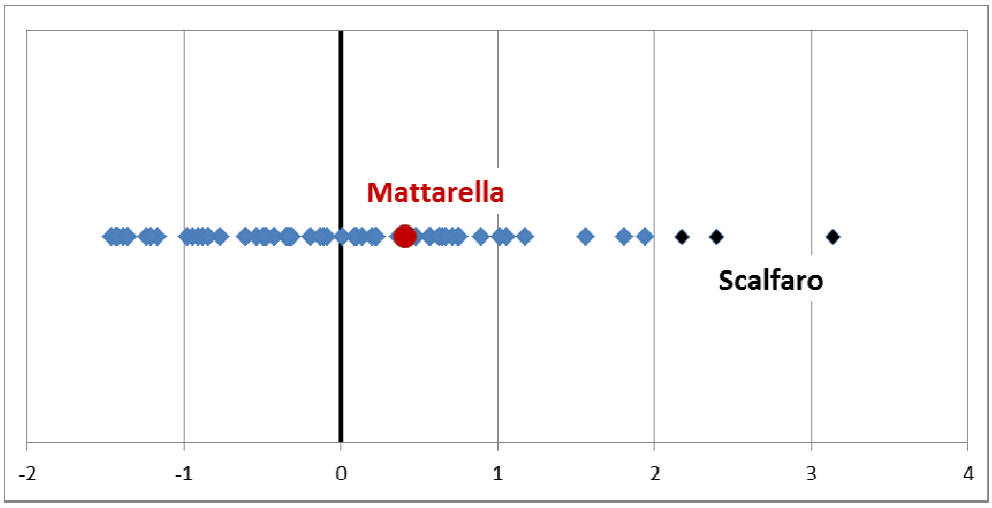


Figure 2. Discourses size in terms of z-scores.

3. Who does Sergio Mattarella resemble?

We have tried to examine to what extent Mattarella's first End-of-Year address resembles those of his predecessors by comparing the lexical profiles using Labbé intertextual distance formula (Labbé and Labbé 2001, 2007).

Given a pair of texts *A* and *B* of size N_A and N_B with $N_A \leq N_B$, the frequency $f_{i,B}$ of each word-type *i* in the larger text *B* is reduced ($f^*_{i,B}$) according to the size of the shorter text *A* by means of a simple proportion:

$$f^*_{i,B} = f_{i,B} \frac{N_A}{N_B}.$$

The distance d between *A* and *B* is:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^*_{i,B} \right|}{2N_A}$$

where $V_{A \cup B}$ represents the vocabulary of *A* and *B* as a whole. This dissimilarity measure satisfies the properties of a distance (non-negativity, symmetry, triangle inequality).

Relying on the results of precedent works (Cortelazzo et al. 2013, Tuzzi 2010), we decided to propose a general measure of similarity based on the entire words list (all the entries of the lemma-vocabulary) in order to compare Mattarella's address with the 66 addresses presented by the presidents who preceded him. We went on to examine if there were any similarities between the presid-

ents' addresses, and we compared the single address by Mattarella with the subcorpora of his predecessors, which contains all of the messages delivered by each of the presidents. We did this by using the iterative version of the intertextual distance based on the entire word list, but we did not include in this analysis the discourses of two presidents, Einaudi e Segni, because the subcorpus of 6 addresses of the former and two of the latter (as a whole) were shorter than Mattarella's first message.

For each replication $k = 1 \ldots m$, a sample is extracted including as many text chunks of size *n* as many the discourses are, and a 2D squared matrix of distances for each pair of text chunks is calculated. After *m* replications a 3D matrix is obtained and, then, a 2D matrix is calculated where the generic element is calculated as the mean out of *m* iterations:

$$\hat{d}_{ij} = \frac{\sum_{k=1}^{m} d_{ijk}}{m}$$

and is assumed as the estimated distance for the pair of discourses *i, j* with size *n* samples.

The intertextual distance (Figure 3) groups together the addresses delivered by each president. The general structure does not, however, follow a chronological or ideological similarity type of order. Both the intertextual distance and correspondence analysis (Figure 6, 7) show a pattern that is apparently explained by the strong personality of single presidents and not by any general features such as the historic time, the age of the president, his ideological position or background. The particularity and popularity of two figures emerged from the analysis: that of Sandro Pertini (president from 1978 to 1985) and Oscar Luigi Scalfaro (1992-1999). According to intertextual distance, the two presidents who preceded Mattarella, Carlo Azeglio Ciampi (1999-2006) and Giorgio Napolitano (2006-2015), appear quite similar although the clusters of discourses of the two statesmen are well separated. The only message delivered by Sergio Mattarella falls between those of his two predecessors and appears particularly close to the style of Giorgio Napolitano. A continuity and temporal contiguity were noted in the discourses of the latest presidents which might indicate that there has been a shift in a pattern according to which End-of-Year speeches tended to be related.

If we observe the intertextual distance of Mattarella's discourse and that of all the others (Figure 4) four of the closest ones, in ascending order, were found to be Napolitano's speeches (2006, 2013, 2009, 2011) and, in any case, all Napolitano's discourses are in the first 13 positions, a clear sign that he was Mattarella's reference model for his first discourse.

The result was confirmed (Figure 5) when the addresses of each president were analysed as a whole.

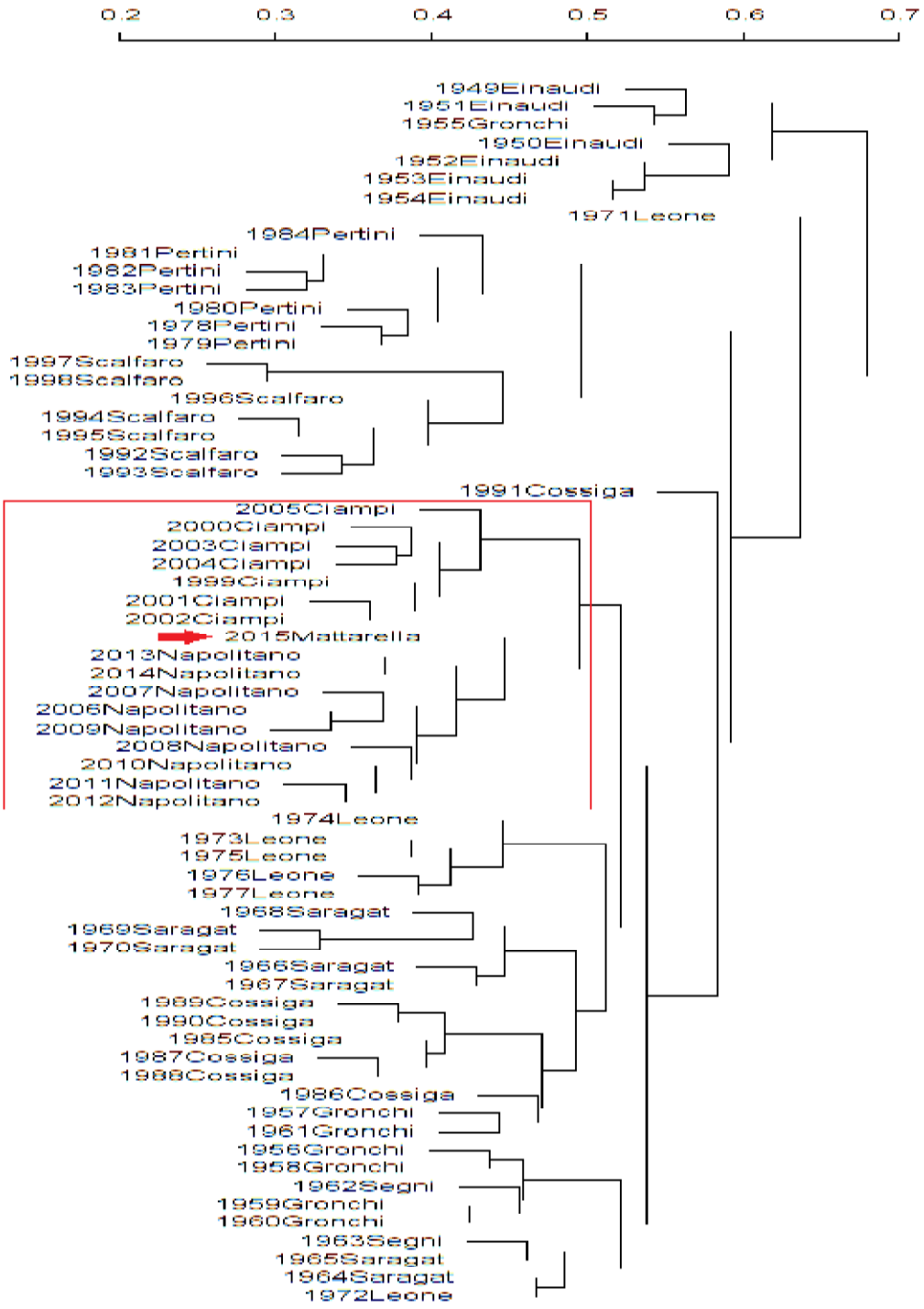


Figure 3. Cluster analysis of the 67 discourses based on intertextual distance. Dendrogram of the agglomerative hierarchical cluster algorithm with complete linkage.

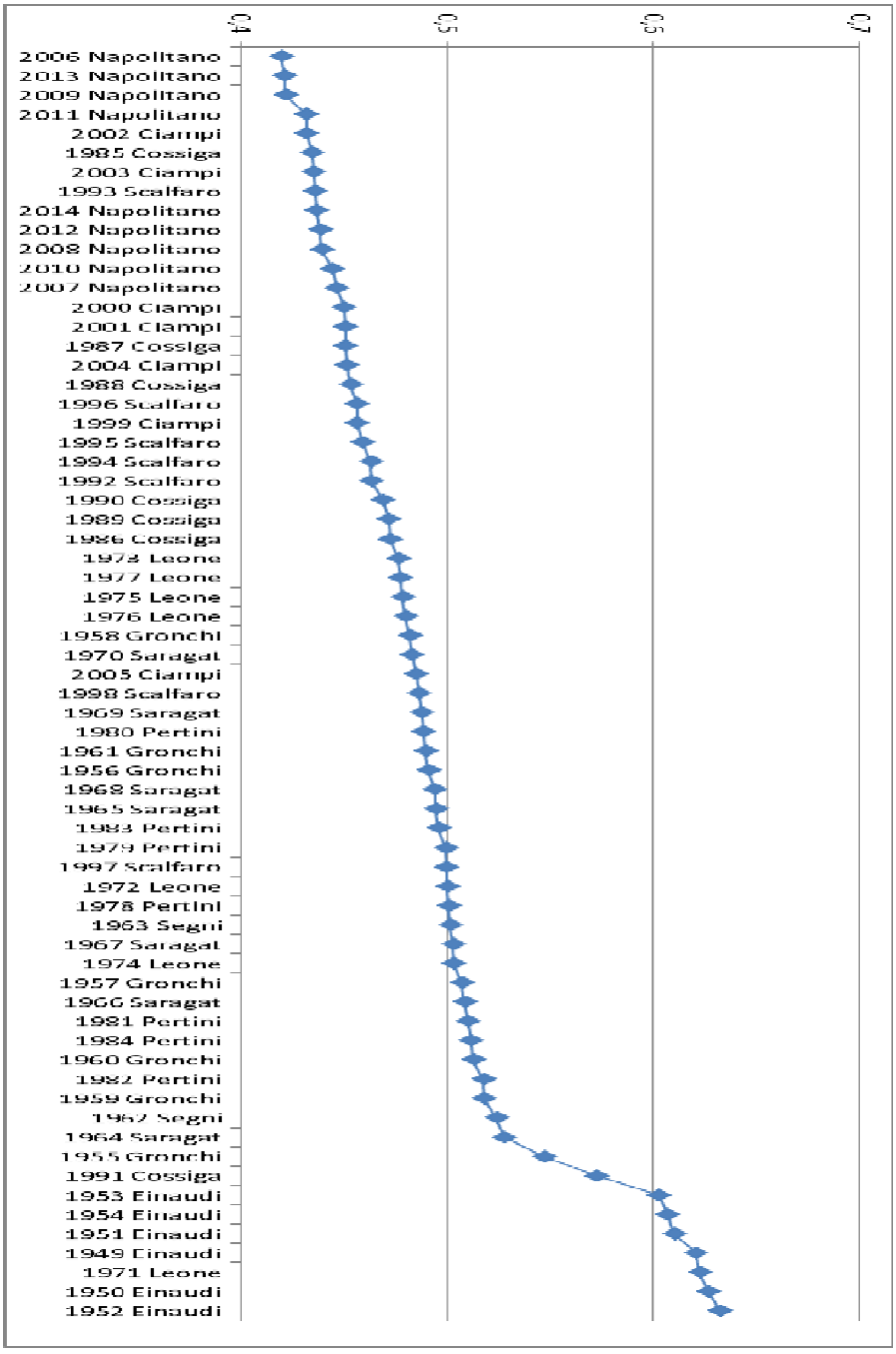


Figure 4. Intertextual distances of the address delivered by Mattarella in 2015 by all 66 precedent addresses in ascending order.

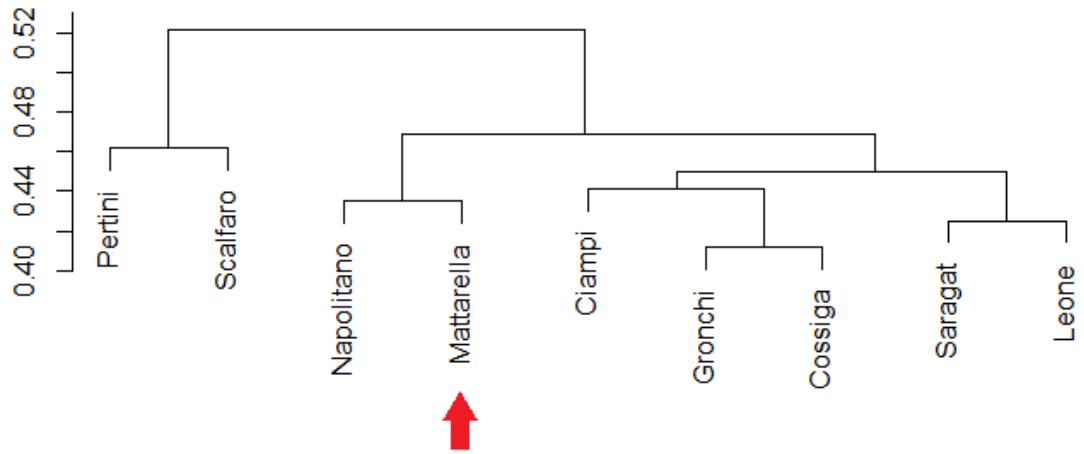


Figure 5. Cluster analysis of the 9 Presidents based on the iterative version of intertextual distance. Dendrogram of the agglomerative hierarchical cluster algorithm with complete linkage.

Table 2
Intertextual distances based on 200 replications and equal size samples (n = 2,000).

	Saragat	Leone	Pertini	Cossiga	Scalfaro	Ciampi	Napolitano	Mattarella
Gronchi	0.4141	0.4268	0.5105	0.4116	0.4984	0.4397	0.4460	0.4557
Saragat	-	0.4259	0.5080	0.4436	0.4915	0.4428	0.4533	0.4690
Leone		-	0.4891	0.4235	0.4612	0.4500	0.4317	0.4509
Pertini			-	0.5215	0.4618	0.4836	0.5203	0.5109
Cossiga				-	0.4874	0.4411	0.4420	0.4579
Scalfaro					-	0.4670	0.4894	0.4892
Ciampi						-	0.4407	0.4424
Napolitano							-	0.4351

Correspondence analysis (Greenacre 1984, 2007) offers a clear mapping of the 67 discourses that is consistent with the results emerging from intertextual distance. As far as the first factorial plane, obtained with the coordinates on the first and second axis, is concerned, no clear chronological pattern was found, but the relevance of the personalities of two presidents, Pertini and Scalfaro, clearly emerged (Figure 6). With regard to the second factorial plane, obtained with the coordinates of axes 3 and 4, two other figures emerged: Ciampi and Napolitano (Figure 7).

Mattarella's discourse fell into Napolitano's area which confirmed his affinity with his predecessor.

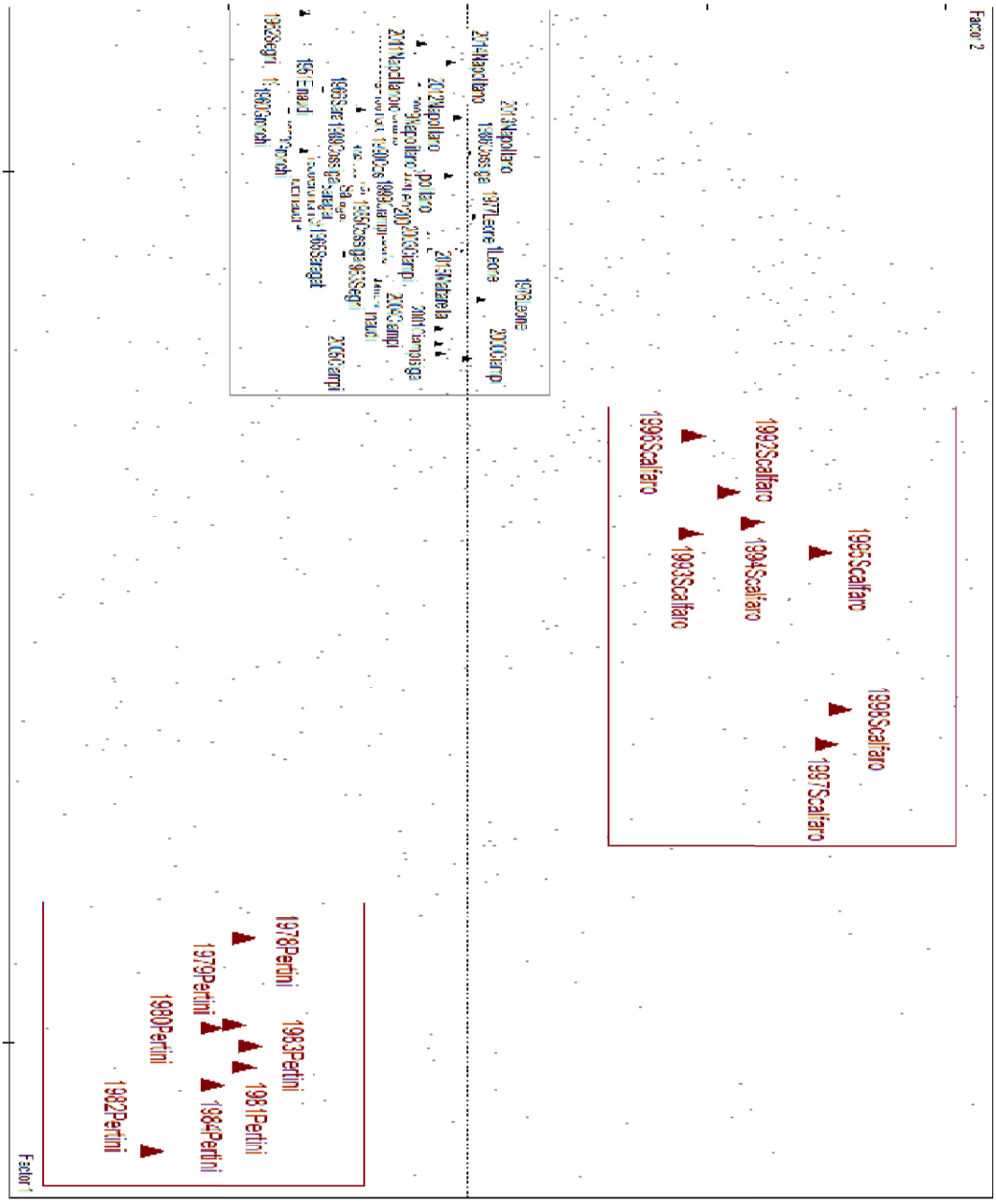


Figure 6. First factorial plane of correspondence analysis. Projection of discourses. Projection of discourses (triangles) and words (dots) in the background.

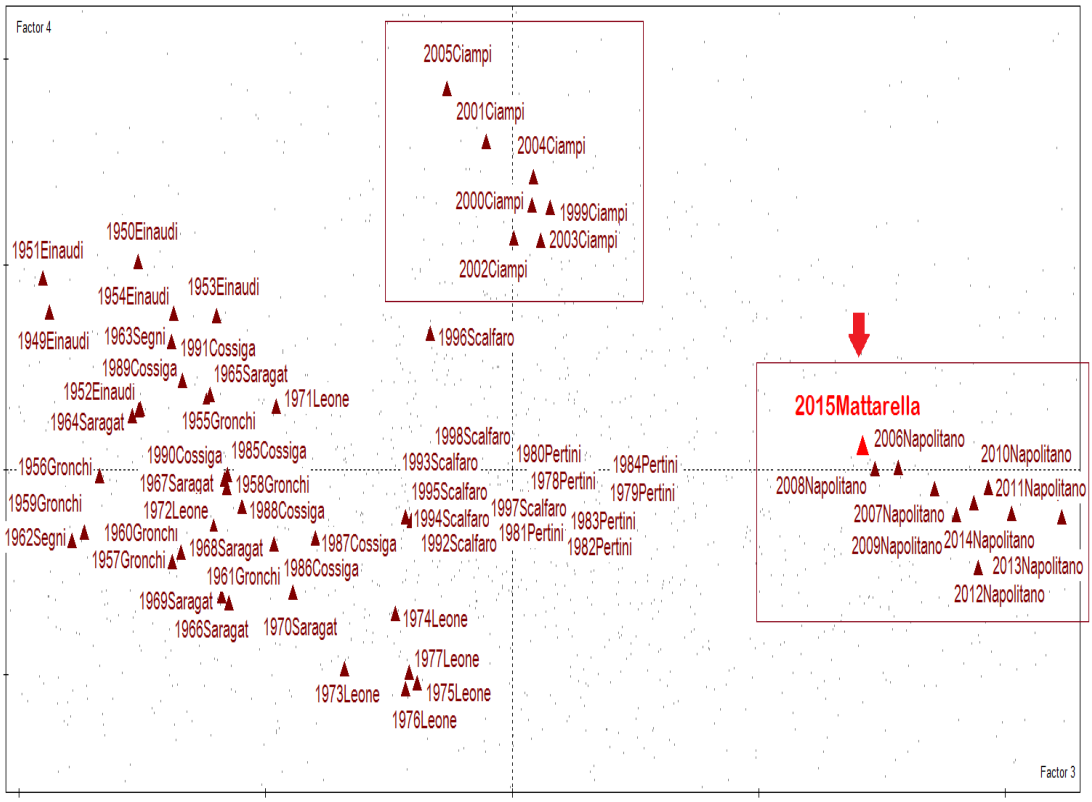


Figure 7. Second factorial plane of correspondence analysis. Projection of discourses (triangles) and words (dots) in the background.

4. Conclusions

A substantial continuity with his predecessors and in particular with Giorgio Napolitano emerged from an analysis of Sergio Mattarella's first End-of-Year address. This finding together with the similarity in the discourses by Napolitano and Ciampi, could be indicative that there has been a shift in the pattern of similarity noted in the presidents' addresses as a chronological contiguity, not noted at the beginning of the republic's history, was found in recent addresses.

At the moment we do not however have sufficient data to be able to sustain this hypothesis. An analysis of the End-of-Year messages of presidents of the Italian Republic has shown that the style of the first discourse of a new president does not always foreshadow the style of successive addresses. Almost all presidents make their debut in a measured, almost timid manner often following the sure path of their immediate predecessors. Only later do they familiarize with the experience and with mass communication and become increasingly spontaneous clearly showing their own personal style (Cortelazzo, Tuzzi 2007b: 233). When we have other samples by Sergio Mattarella at our disposition we will be able to identify his specific linguistic physiognomy.

For the time being, we can say that Mattarella's first End-of-Year discourse followed the example of his predecessor and, at least for the moment, he conformed to that model.

References

Bernardi L., Tuzzi A. (2010). L'autografo del Presidente della Repubblica: un archetipo del discorso di fine anno prodotto mediante ADT, in: S. Bolasco, I. Chiari, L. Giuliano (eds), *JADT2010 – Statistical Analysis of Textual Data. Proceedings of 10th International Conference*, vol. 2, 665-676. LED, Milano.

Bolasco S., Baiocchi F., Morrone A. (2009). TaLTaC2: Trattamento automatico Lessicale e Testuale per l'analisi del Contenuto di un Corpus, Roma (http://www.taltac.it).

Cortelazzo M.A., Tuzzi A. (2007a, eds). *Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica.* Marsilio Editori, Venezia.

Cortelazzo M.A., Tuzzi A. (2007b). Considerazioni finali, in: M.A. Cortelazzo, A. Tuzzi (eds), *Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica:* 231-237. Marsilio Editori, Venezia.

Cortelazzo M.A., Nadalutti P., Tuzzi A. (2013). Improving Labbé's Inter-textual Distance: Testing a Revised version on a Large Corpus of Italian Literature, *Journal of Quantitative Linguistics*, 20(2), 125-152.

Finniss-Boursin F. (1992). *Les discours de vœux des présidents de la République, Librairie générale de droit et de jurisprudence,*.Paris.

Giuliano L. (2014), *The value of words*, Dipartimento di Scienze statistiche, Roma.

Greenacre M.J. (1984). *Theory and application of correspondence analysis*, Academic Press, London.

Greenacre M. J. (2007). *Correspondence analysis in practice.* Chapman & Hall, London.

Köhler R., Tuzzi A. (2015). Linguistic modelling of sequential phenomena, in: Mikros G., Mačutek J. (eds), *Sequences in Language and Tex:* 109-124. Berlin,: De Gruyter Mouton (Quantitative Linguistics nr. 69).

Labbé C., Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Moliére. *Journal of Quantitative Linguistics* 8(4), 213-213.

Labbé D. (1990). *Le vocabulaire de François Mitterrand*, Presses de la Fondation nationale des sciences politiques. Paris.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English, *Journal of Quantitative Linguistics* 14(1), 33-80.

Labbé D., Monière D. (2003). *Le discours gouvernemental. Canada, Québec, France (1945-2000).* Champion, Paris.

Labbé D., Monière D. (2008). *Les mots qui nous gouvernent. Le discours des premiers ministers Québécois: 1960-2005.* Monière-Wollank Éditeurs, Montréal.

Leblanc J.-M. (2003). Les messages de vœux des présidents de la Cinquième République. L'ethos, la diachronie, deux facteurs de la variation lexicométrique. *Lexicometrica* 4.

Leblanc J.-M., Martinez W. (2005), Positionnements énonciatifs dans les vœux présidentiels sous la cinquième République. Analyse des marques personnelles par les méthodes de cooccurrence. *Corpus* 4, 105-128.

Liu F. (2012), Genre Analysis of American Presidential Inaugural Speech, *Theory and Practice in Language Studies* 2(11), 2407-2411.

Marchand P. (2007). *Le grand oral. Les discours de politique générale de la Ve République*, De Boeck Université, Bruxelles.

Mayaffre D. (2004). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République*, Champion, Paris.

Mayaffre D. (2012). *Le discours présidentiel sous la Ve République: Chirac, Mitterrand, Giscard, Pompidou, de Gaulle.* Presses de la Fondation Nationale des Sciences Politiques, Paris.

Nye J. S., Jr. (2013). *Presidential Leadership and the Creation of the American Era.* Princeton University Press, Princeton.

Pauli F., Tuzzi A. (2009). The end of year addresses of the Presidents of the Italian Republic (1948-2006): discoursal similarities and differences. *Glottometrics* 18, 40-51.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Wien, https://www.R-project.org.

Savoy J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics* 17(2), 123-141.

Savoy J. (2015). Vocabulary Growth Study: An Example with the State of the Union Addresses. *Journal of Quantitative Linguistics*, 22(4), pp. 289-310.

Schlesinger A.M. (1997). Rating the Presidents: Washington to Clinton. *Political Science Quarterly* 112(2), 179-190.

Teletin A. (2013), Les vœux présidentiels au Portugal, en France et en Roumanie, et la crise internationale. Les enjeux des formes d'adresse et des procédés d'atténuation/intensification. *Mots. Les langages du politique* 1(101), 31-46.

Trevisani M., Tuzzi A. (2013). Shaping the history of words. In: Obradović I., Kelih E., Köhler R. (eds.), *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics* (QUALICO): 84-95. Belgrade, Serbia, April 16-19, 2012, Akademska Misao, Beograd.

Tuzzi A. (2010), What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics / Statistica Applicata*, 22(1), 77-94.

Tuzzi A., Köhler R. (2015). Tracing the history of words. In: Tuzzi A., Benešová M., Mačutek J. (eds.), *Recent Contributions to Quantitative Linguistics* 203-214. Berlin: De Gruyter Mouton.

Tuzzi A., Popescu I.-I., Altmann G. (2010). *Quantitative analysis of Italian texts.* (= Studies in Quantitative Linguistics 6). Lüdenscheid: RAM-Verlag,

Tuzzi A., Popescu I.-I., Zörnig P., Altmann G. (2012). Aspects of the behaviour of parts-of-speech in Italian texts. *Glottometrics* 24, 41-69.

Thematic Concentration and Vocabulary Richness

Miroslav Kubát, Radek Čech

University of Ostrava, Czech Republic

Abstract. The contribution investigates a relation between two stylometric features with promising results in text classification: thematic concentration and vocabulary richness. Namely secondary thematic concentration (*STC*), moving average type-token ratio (*MATTR*), and repeat rate (RR_{MC}) are analysed. The main aim is to test the hypothesis that vocabulary richness negatively correlates with thematic concentration. The research is based on a corpus of more than 900 English texts from various genres. This study follows up a similar analysis (Čech 2016) which investigated Czech texts.

Keywords*: thematic concentration, vocabulary richness, correlation, type-token ratio, repeat rate*

1. Introduction

Several stylometric indices such as thematic concentration (Popescu et al. 2009), lambda-structure of text (Popescu et al. 2011), moving average type-token ratio (Covington, McFall 2010), nominality of text (Zörnig et al. 2016), or writer's view (Popescu, Altmann 2007) have been proposed in recent years. It seems reasonable to assume systematic relationships among these indices because they express text characteristics which are an output of a predictable (by means of a statistical hypothesis) verbal behaviour. Specifically, if majority of these indices are useful tools for a text classification, i.e. they are able to detect systematic properties of language production, they should be governed by the similar principles or mechanisms. It is a great challenge for the text linguistics to reveal these principles and, finally, to develop a text theory which could explain human language behaviour with regard to the text characteristics. Because there is no text theory of this kind, we can try to extend our knowledge of general text properties by an analysis of relationships among particular indices. This approach leads not only to better understanding of the indices but also it can be an important step in the theory building.

In this paper, we analyse the relationship between thematic concentration and vocabulary richness. These text properties have been analysed in several studies with promising results in terms of stylometry (e.g. Kubát, Čech 2016; Čech 2014; Popescu et al. 2012; Tuzzi et al. 2010). Both of them seem to be an effective tool of text classification with intelligible linguistic interpretation. As for the particular methods of analysis, secondary thematic concentration (*STC*), moving average type-token ratio (*MATTR*), and relative repeat rate (RR_{MC}) are

used in this study (for details, see below). This contribution follows up a similar research based on Czech data (Čech 2016).

The basic assumption of this study is that thematic concentration and vocabulary richness are interdependent. More specifically, thematic concentration is based on so called thematic words (*TW*). *TW* are highly frequent autosemantics above *h*-point in the rank-frequency distribution of a text (see chapter 3.1). One can therefore assume that text with poor vocabulary should generate more words with high frequency and, consequently, more thematic words. In other words, we expect a significant negative correlation between vocabulary richness and thematic concentration.

2. Language Material

There are two corpora in this study. The first corpus (hereinafter C1) consists of English fiction texts, specifically 400 individual chapters of several novels written by Mark Twain, Jack London, Arnold Bennet, Charles Dickens, Henry James, and Thomas Hardy were chosen. In addition to these texts, we collected also the second corpus (hereinafter C2) which comprises 516 English texts of 6 genres (letter, news, poem, political speech, scientific text, short story) in order to discover whether genre can affect the assumed correlation between thematic concentration and vocabulary richness. It is worth mentioning that the corpora are not lemmatized. Thus, a wordform is a basic unit in this research. The particular methods (see Section 3) are applied to individual texts in both corpora. For text processing software *QUITA – Quantitative Index text Analyzer* (Kubát et al. 2014) and *MaWaTaTaRaD* (Milička 2013) were used.

3. Methodology

3.1 Thematic Concentration

Every author of any text focuses on a topic or topics which are represented by several autosemantic words. Thematic concentration measures how intensively the author concentrates on the main theme(s) of the text. On the one hand, texts like scientific papers have usually high thematic concentration. On the other hand, e.g. informal letters or emails are not so thematically concentrated in general. There are several methods for measuring thematic concentration. In this study, we use secondary thematic concentration (*STC*) especially due to its effectiveness of text classification (Čech et al. 2015; Čech 2016).

STC is based on rank-frequency distribution and *h*-point (which represents a fuzzy boundary between synsemantic and autosemantic words; see formula 2). *STC* is calculated as follows:

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

$f(1)$…highest frequency
h…h-point
r'…rank of an autosemantic word above h-point
$f(r')$…frequency of autosemantic word

$$(2) \quad h = \begin{cases} r_i, & \text{if there is } r_i = f(r_i) \\ \dfrac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})} & \text{if there is } r \neq f(r) \end{cases},$$

r…rank
$f(r)$…frequency of the rank

STC is considered to be the stylomeric index which is independent on text length (Čech, Kubát 2016). In Figure 1, the relation between the text length and *STC* in 400 English texts (C1) can be seen. Both, the low value of the coefficient of determination $R^2 = 0.0016$ and almost horizontal line of the linear function expressing the relationship between these indices can be considered as a sufficient support of *STC* text length independence.

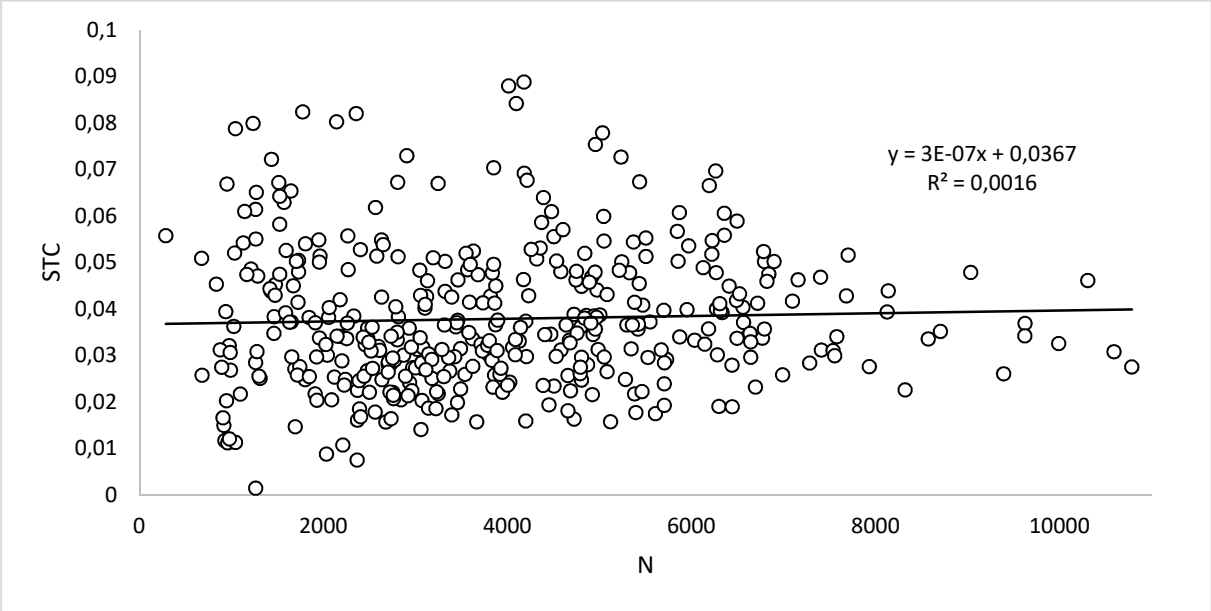


Figure 1. Relation between *STC* and text length (*N*) in 400 English texts (C1).

3.2 Vocabulary Richness

Vocabulary richness is one of the most traditional stylometric features. In this study, we decide to use two indices: (a) moving average type-token ratio (*MATTR*) and (b) relative repeat rate (RR_{MC}). These methods were chosen especially due to their strong resistance to the impact of the text length (Kubát 2014, McIntosh 1967).

3.2.1 Moving average type-token ratio (*MATTR*)

This vocabulary richness measure was proposed by Covington & McFall (2010) and further elaborated by Kubát & Milička (2013). *MATTR* is defined as follows; a text is divided into overlapped subtexts of the same length (so called "windows" with arbitrarily chosen size *L*; usually, the "window" moves forward one token at a time), next, type-token ratio is computed for every subtext and, finally, *MATTR* is defined as a mean of the particular values. For example, in the following sequence of characters: *a, b, c, a, a, d, f,* text length is 7 tokens ($N = 7$) and we choose the window size to 3 tokens ($L = 3$). We get 5 subsequent windows:

|*a, b, c* | *b, c, a* | *c, a, a* | *a, a, d* | *a, d, f*|,

and compute *MATTR* of the sequence as follows:

$$(3) \quad MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} = \frac{3+3+2+2+3}{3(7-3+1)} = 0.87$$

L…arbitrarily chosen length of a window, $L < N$
N…text length in tokens
V_i…number of types in an individual window

Although *MATTR* was proposed as an absolutely independent method on text length, in Figure 2 we can observe a slight dependence in our corpus (C1) (coefficient of determination R^2=0.031). Nevertheless, *MATTR* seems to be an appropriate index for the given purpose of this analysis.

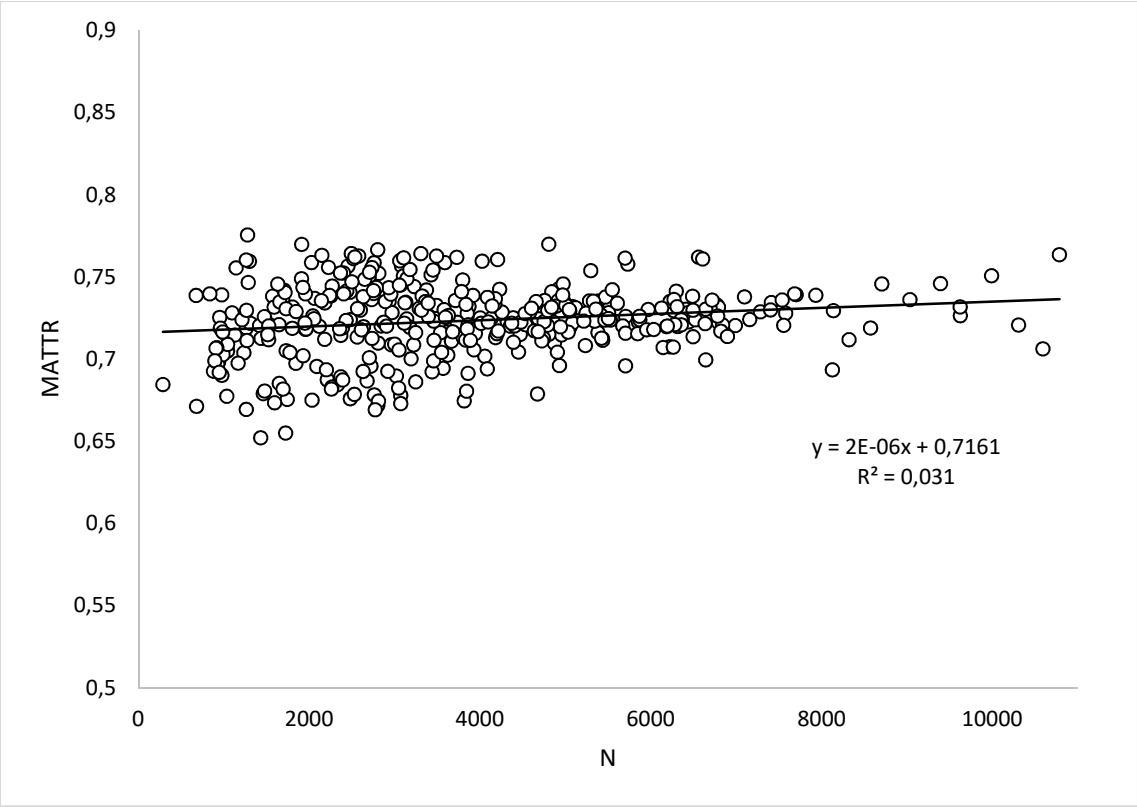


Figure 2. Relation between *MATTR* and text length (*N*) in 400 English texts (C1).

3.2.2 Relative Repeat Rate (RR_{MC})

Repeat rate (*RR*) is a simple indicator of a degree of vocabulary concentration. In fact, *RR* measures vocabulary richness inversely: the higher *RR* is, the less vocabulary diversity a text has. *RR* is defined as follows:

$$(4) \qquad RR = \frac{1}{N^2}\sum_{r=1}^{V} f_i^2$$

f_1…frequency of word *i* in a text
N…number of tokens
V…number of types

Given that the resulting values of *RR* lie within the interval <1/*V*;1>, McIntosh (1967) proposed the relative repeat rate (RR_{MC}). Since the results of RR_{MC} lie within the interval <0;1>, this relative repeat rate is comparable with other indicators such as *MATTR*. The formula is as follows:

$$(5) \quad RR_{MC} = \frac{1-\sqrt{RR}}{1-1/\sqrt{V}}$$

RR…repeat rate
V…number of types

As can be seen in Figure 3, RR_{MC} is not too much influenced by text length. The slight negative correlation with the coefficient of determination $R^2 = 0.031$ can be considered as an acceptable value for this research.

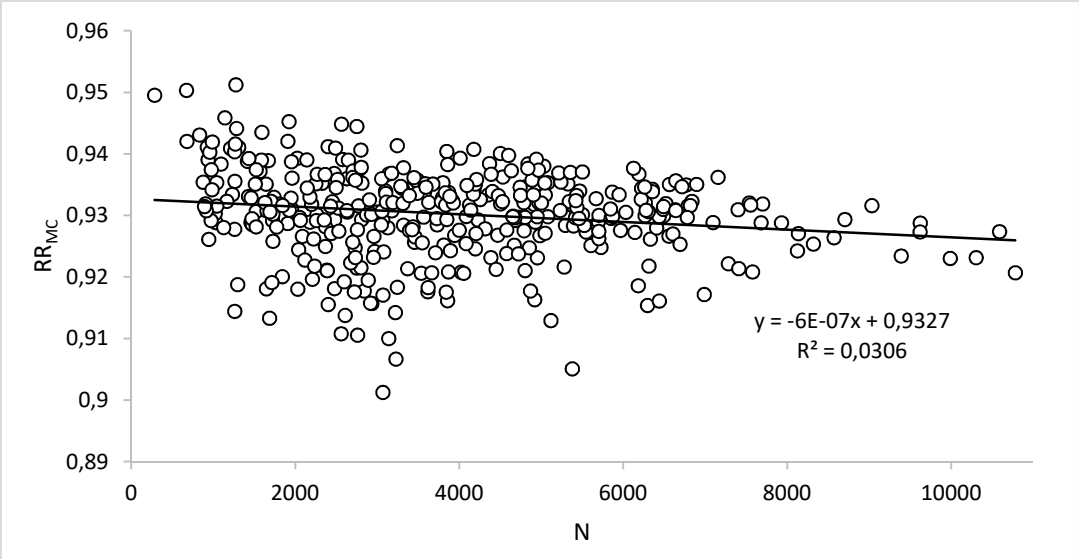


Figure 3. Relation between RR_{MC} and text length (N) in 400 English texts (C1).

4. Results

As can be seen in Figure 4 *MATTR* seems to be independent on *STC*, the coefficient of determination $R^2 = 0.0007$. To be more precise, we apply also Kendall's tau correlation coefficient with the results as follows: $\tau = -0.024$, $p = 0.466$. These results mean that there is non-significant ($\alpha = 0.05$) very slight negative correlation.

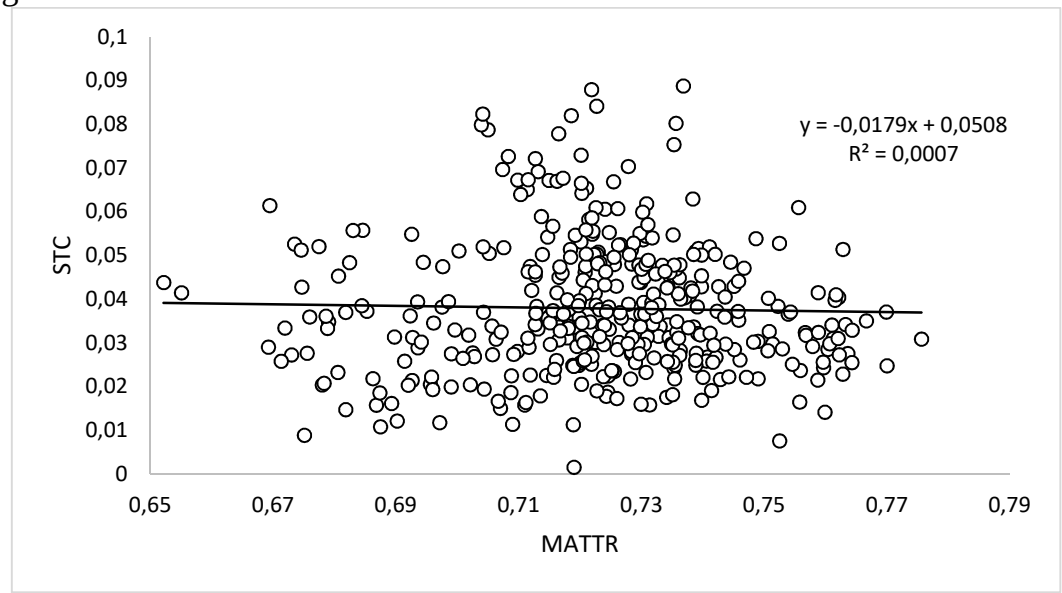


Figure 4. Correlation between *MATTR* and *STC* in 400 English texts (C1).

Contrary to *MATTR*, RR_{MC} significantly correlates with *STC*, see Figure 5 (R^2 = 0.2151, τ = 0.337, $p < 0.001$).

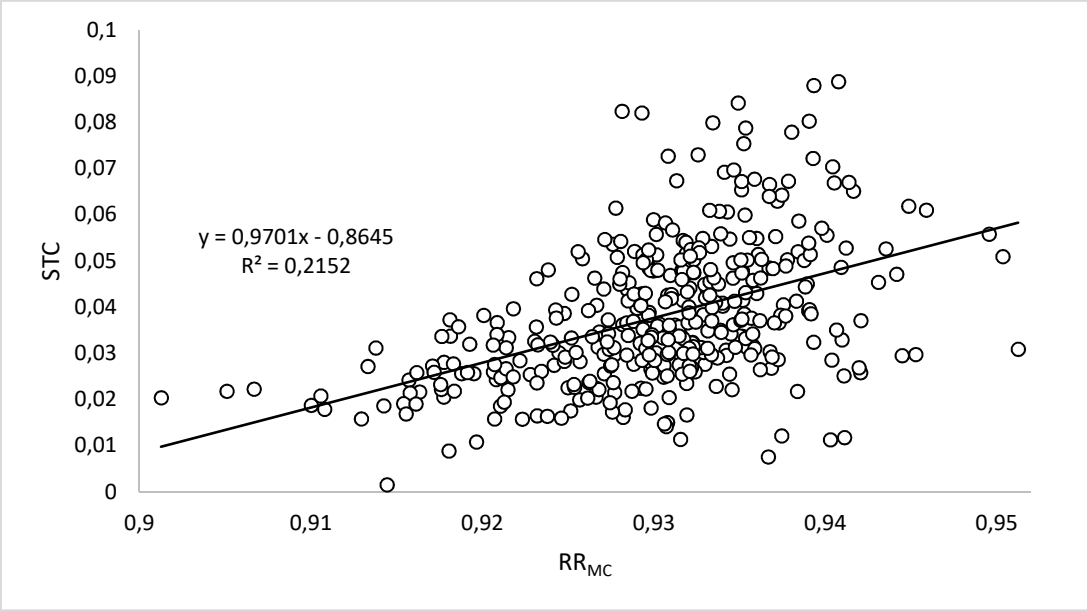


Figure 5. Correlation between RR_{MC} and *STC* in 400 English texts (C1).

In order to investigate the relation between vocabulary richness and thematic concentration in more detail, we compare several genres (letter, news, poem, political speech, scientific text, short story). The results of Kendall's tau correlation coefficient can be seen in Table 1 and Table 2. The obtained values mostly correspond to the previous ones. With the exception of poems, *MATTR* does not significantly correlates with *STC*, whereas RR_{MC} significantly correlates with *STC* in 5 of 6 genres. Consequently, it can be concluded that genre probably does not substantially influence the relation between vocabulary richness and thematic concentration.

Table 1
Correlations between *MATTR* and *STC*.

genre	**number of texts**	τ	***p*-value**
letter	100	0.067	0.327
news	100	-0.047	0.488
poem	100	-0.196	0.028
political speech	56	-0.154	0.375
scientific text	60	-0.081	0.66
short story	100	0.040	0.063

Table 2
Correlations between RR_{MC} and *STC*

genre	number of texts	τ	*p*-value
letter	100	-0.020	0.77
news	100	0.193	0.004
poem	100	0.476	< 0.001
political speech	56	0.358	0.004
scientific text	60	0.261	0.008
short story	100	0.236	< 0.001

Considering the obtained results, one can ask about a relation between *MATTR* and RR_{MC}. A significant positive correlation can be seen in Figure 6 (R^2 = 0.026, τ = 0.091, p < 0.007), however, the value of τ is very small.

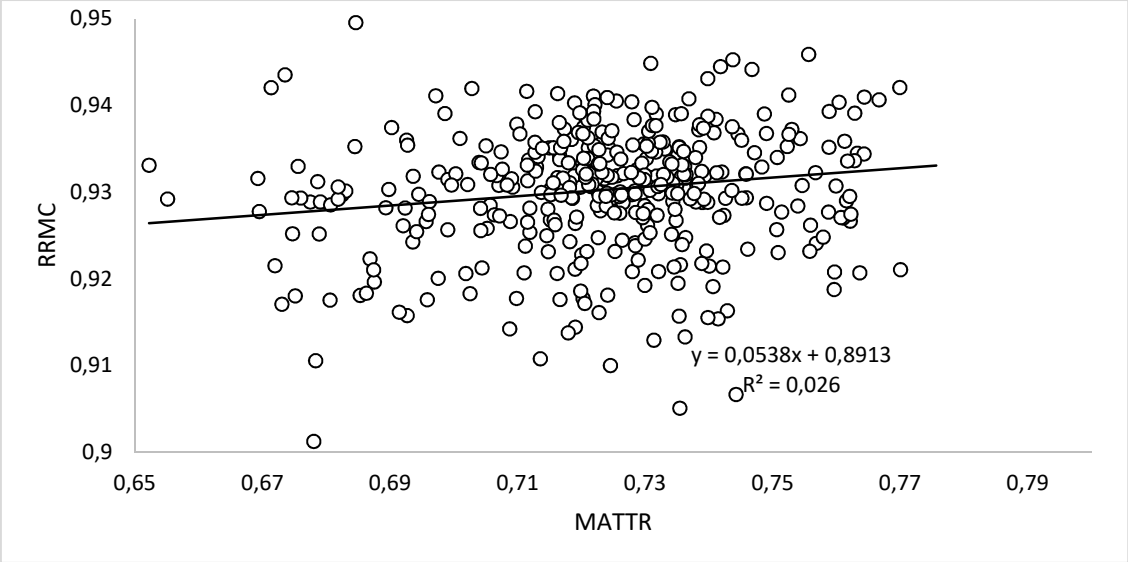


Figure 6. Correlation between *MATTR* and RR_{MC} in 400 English texts (C1).

5. Conclusion and Discussion

Considering the results, the final conclusion is quite ambiguous. Specifically, RR_{MC} significantly positively correlates with *STC*, while *MATTR* seems to be independent on *STC*. Moreover, both results do not support our assumption, i.e. the negative correlation between vocabulary richness and thematic concentration. To sum up, this study raised more questions than answers.

We suppose that one of possible explanations could be the fact that *STC* is based on a relatively small number of thematic words (with regard to number of all types used in the text). The number of these frequent autosemantics above *h*-

point is usually around 7 (but sometimes only 2 or 3; rarely even 0).[1] Thus, it seems reasonable to assume that frequencies of these few words cannot significantly affect a resulting value of vocabulary richness measure which is based on frequencies of all words in a text. Needless to say, this idea must be scrutinized empirically. Further, the concept of vocabulary richness itself is still not clear and well theoretically based, despite decades of research. For instance, some authors consider *TTR* to be a matter of information flow rather than vocabulary richness (e.g. Popescu et al. 2009; Wimmer 2005). Until vocabulary richness is thoroughly and deeply examined, it will be very difficult and problematic to deal with this concept in stylometry.

From a point of view of this study, our preliminary findings must be especially verified by (a) an application of more vocabulary richness indices, and (b) more texts, particularly in different languages[2].

References

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity*, 48, 899–910.

Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics*, 22, 215–232**.**

Čech, R., Kubát, M. (2016). Text length and the thematic concentration of text. *Mathematical Linguistics*, 2(1), 5–13.

Covington, M.A., McFall, J. D. (2010) Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.

Kubát, M. (2014). Moving window type-token ratio and text length. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM, 105–113.

Kubát, M., Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, 20(4), 339–349.

Kubát, M., Matlach, V., Čech, R. (2014). *QUITA - Quantitative Index text Analyzer*. Lüdensheid: RAM.

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

McIntosh, R. P. (1967). An indicator of diversity and the relation of certain concepts to diversity. *Ecology*, 48, 392–404.

[1] It is worth mentioning that a number of thematic words correlates with text length. However, *STC* is not influenced by text length due to the normalization by dividing each thematic unit by the sum of all the weights of all the units above the *h*-point and the highest frequency of the unit in the text (see Formula 1). For example, a correlation between thematic words and text length in 57 political speeches is displayed in a graph in the appendix of this paper.

[2] Czech was analysed by Čech (2016) with similar results which generally correspond to our findings.

Milička. J. (2013) *MaWaTaTaRaD* (software). Available at http://milicka.cz/en/mawatatarad/

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics*, 15, 71–81.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B . D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word frequency studies*. Berlin, New York: de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

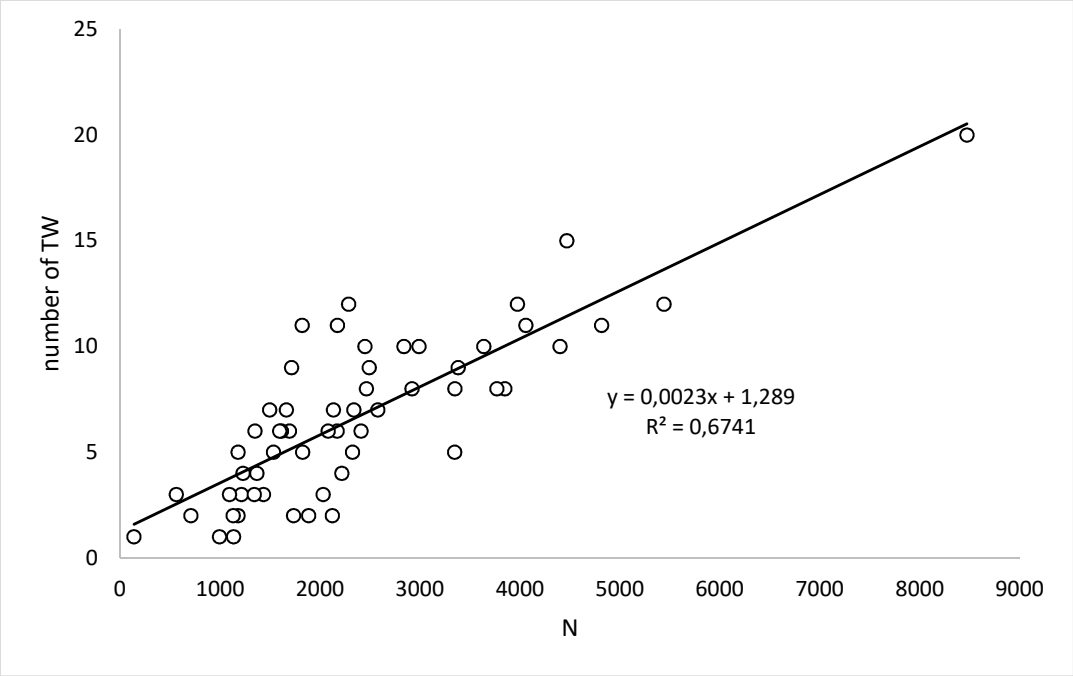
Popescu, I. I., Čech, R., Altmann, G. (2012). Some characterizations of Slovak poetry. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (Eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens, 187–196.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM.

Wimmer, G. (2005). The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, s. 361–368.

Zörnig, P., Stachowski, K., Popescu, I. I., Taybeh, M. N., Mohanty, P., Kelih, E., Chen, R., Altmann, G. (2016). *Descriptiveness, Activity and Nominality in Formalized Text Sequences*. Lüdensheid: RAM.

Appendix



Correlation between number of thematic words (*TW*) and text length (*N*) in 57 political speeches.

Probing the "Temperature" Approach on Ukrainian Texts: Long-prose Fiction by Ivan Franko

Solomija Buk, Andrij Rovenchak (Lviv)

Abstract. This paper will analyze Ivan Franko's long prose fiction, using the previously developed approach involving the "temperature" parameter defined from the frequency spectrum of texts. The scaling of this parameter with text size is studied and certain groupings of texts with respect to the scaling exponent and coefficient are observed. A new quantity is proposed to analyze relations between different sub-systems in texts.

Keywords: *"temperature" of text; frequency spectrum; Ivan Franko; long prose fiction; author's speech; direct speech.*

1. Introduction

At the current state of scientific development there is an increasing interest in intercultural and interdisciplinary fields. Researchers are able to observe common laws working in exact and natural sciences as well as in social sciences and humanities. Techniques from the domain of statistical physics are applied in mathematics (Tran et al. 2004), biology (Jin et al.2009), humanities and social sciences (Bohorquez et al. 2009; Palchykov et al. 2013; Colaiori et al. 2015; Mryglod et al. 2015). Linguistic studies are also well represented here (Eroglu 2014; Rodriguez et al. 2014), including approaches from the point of view of complex networks (Ferrer-i-Cancho et al. 2004; Čech et al. 2011; Mac Carrona, Kenna 2013).

The "temperature" concept with respect to languages was proposed in several ways (Mandelbrot 1953; Kosmidis et al. 2006), in particular, by analyzing the high-frequency words using the Boltzmann distribution (Miyazima, Yamamoto 2008; Rêgo et al. 2014).We introduced the "temperature" parameter in a different way, considering the low-frequency vocabulary (Rovenchak, Buk 2011a).

The idea of the present study is to extend the "temperature" approach to texts in Ukrainian written by Ivan Franko. The paper is organized as follows. Section 2 contains a brief description of Ivan Franko's works of fiction. In Section 3, the approach applied for the quantitative text analysis is explained. Results and discussion are given in Section 4 followed by conclusions in Section 5.

2. Ivan Franko's prose

Ivan Franko (1856–1916) was a famous Ukrainian writer, scholar, and public figure (Stech, Zhukovsky 2007). His long prose works are well described from the quantitative and statistical point of view (Buk, Rovenchak 2007; Buk 2010; 2011; 2012; 2013; Kelih et al. 2014). Note that there are also quantitative studies of Franko's letters and poetry (Best, Zinenko 1998; 1999) and fables (Holovatch, Palchykov 2017). We have an ongoing project of compiling a tagged corpus of Ivan Franko's prose (Buk 2007); this corpus could be a good base for further quantitative studies.

Ivan Franko authored over one hundred works of fiction in prose of different sizes, from short stories of some two pages to rather lengthy novels. There is no strict quantitative definition separating the types of narratives (namely, short stories, novellas, and novels) but both theorist and practitioners of literature generally agree on the upper limit of 20,000 words for a short story (cf. Thrall, Hibbard 1960: 457–458; Kotter 2008: 248–249; King 2010; Waas 2012: ix). However, this number is based on short stories written in English, and it should be somewhat lowered when dealing with a more synthetic language such as Ukrainian.

Within Franko's œuvre, ten works are usually referred to as long prose fiction (cf. Pastukh 1996; Denysiuk 2008):

1. *Boa constrictor* (1st edition: 1878–84; 2nd edition: 1905–07);
2. *Boryslav smijetsja* (Boryslav Laughs) (1880–81);
3. *Zakhar Berkut* (1883);
4. *Ne spytavšy brodu* (Without Asking a Wade) (1885–86);
5. *Dlja domašnjoho ohnyšča* (For the Hearth) (1892);
6. *Osnovy suspil'nosty* (Pillars of Society) (1894–95);
7. *Perekhresni stežky* (The Cross-paths) (1900);
8. *Velykyj šum* (The Great Noise) (1907);
9. *Petriji j Dovbuščuky* (2nd edition: 1909–12).
10. *Lelum i Polelum* [in Polish].

Figure 1. Title pages of some Ivan Franko's long prose works.

We will henceforth refer to the titles using the first letters of the Ukrainian transliteration, with this possibly followed by the edition number, i.e.: BC1, BC2, BSm, ZBe, NSB, DDO, OSu, PSt, VSh, and PD2. See Figure 1 for some title pages.

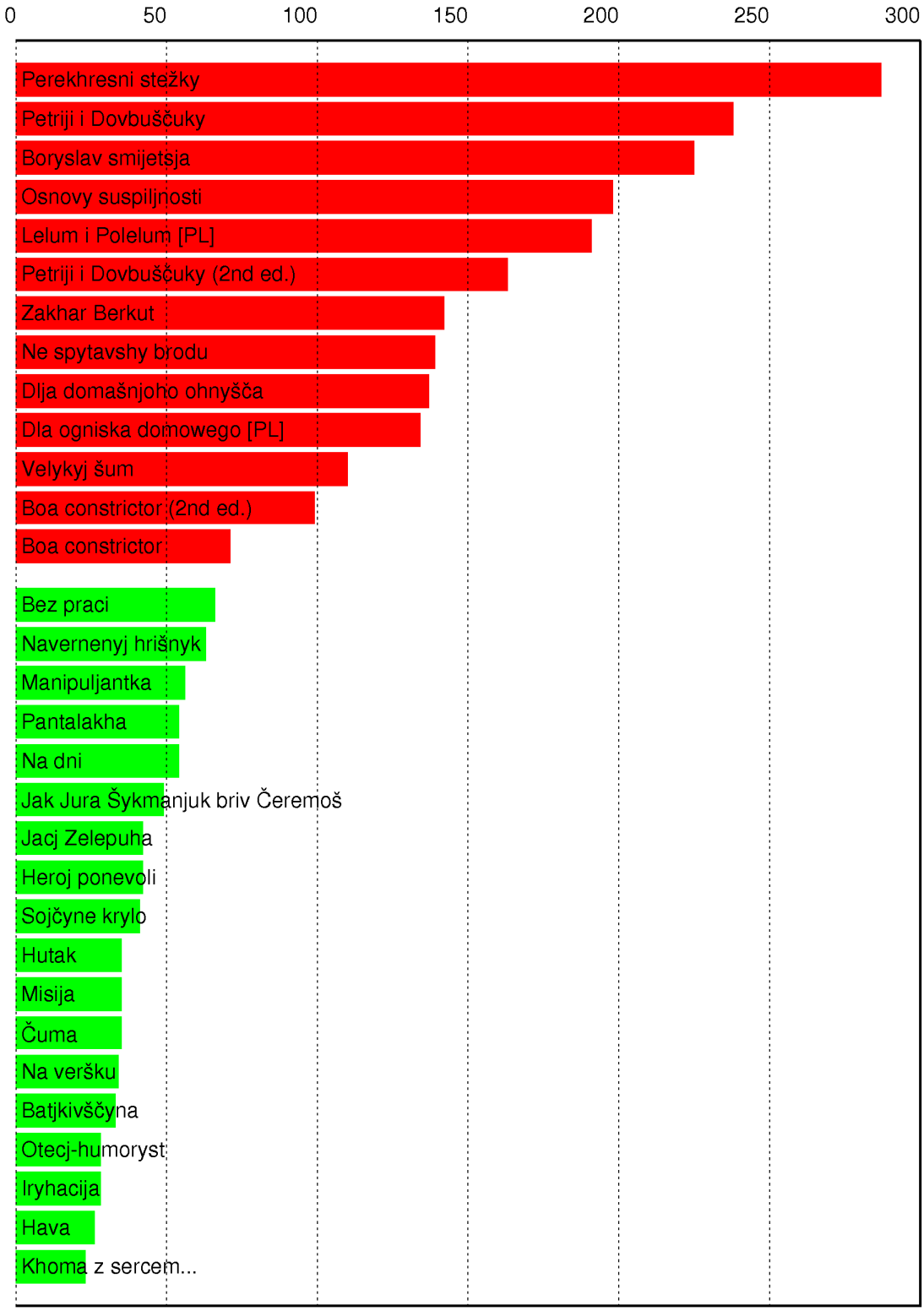


Figure 2. Ivan Franko's prose fiction, by size (in pages).
The mark [PL] denotes texts written in Polish

Interestingly enough, such a division of Franko's prose is confirmed by quantitative data. In Figure 2 the size of each piece of prose is shown, measured in number of pages (Franko 1976–86). On average, a printed page corresponds to

340 words. So, the first edition of *Boa constrictor* (71 pages, ca. 25 thousand words) can be attributed as a novel whereas its nearest lower neighbor *Bez praci* (Without work, 66 pages, ca. 22 thousand words) can be attributed as a short story with some reservations if considering solely quantitative data. A qualitatively significant jump is observed only for the second edition of *Boa constrictor* (99 pages, about 34 thousand words).

3. The "temperature" approach

The approach used for the analysis in the present work was previously developed by the authors (Rovenchak, Buk 2011a; 2011b) and then applied with several modifications to obtain some hints on language evolution (Rovenchak 2014) as well as to some contrastive studies (Rovenchak 2015a; 2015b).

The idea of this approach is based on the analogy between the frequency spectrum of texts and the so-called Bose-distribution in statistical physics (Huang 1987: 183). The frequency spectrum N_j is the number of tokens having absolute frequency exactly equal to j (Tuldava 1996; Popescu et al. 2009). The value of N_1 thus corresponds to the number of *hapax legomena*. The set of N_j is obtained for each text and subsequently fitted to the model of the Bose-distribution:

$$N_j = \frac{1}{z^{-1}e^{(j-1)^\alpha/T}-1}\,. \qquad (1)$$

Equation (1) contains three parameters: z, α, and T. The first parameter is fixed by the number of hapaxes,

$$z = \frac{N_1}{N_1+1}\,. \qquad (2)$$

The remaining parameters, α and T, are calculated by fitting the observed frequency spectrum to model (1) using the least-squares method. The relation $\tau = \ln T / \ln N_{\mathrm{tot}}$, where N_{tot} is the total number of words (tokens) in the given text, proved to be a parameter suitable for text classification alongside the exponent α (Rovenchak, Buk 2011a; 2011b; Rovenchak 2014).

Table 1
Fitting parameters of model (1) for Ivan Franko's long prose fiction

Title	**Abbr.**	N_{tot}	T	α	**ln T / ln N_{tot}**
Boa Constrictor (1st ed.)	BC1	25427	1554	1.66	0.724
Boa Constrictor (2nd ed.)	BC2	34215	1899	1.61	0.723
Boryslav smijetsja	BSm	77456	3192	1.54	0.717
Dlja domašnjoho ohnyšča	DDO	44841	2241	1.59	0.720
Ne spytavšy brodu	NSB	49170	2422	1.61	0.721
Osnovy suspiljnosty	OSu	67174	3038	1.58	0.721
Perekhresni stežky	PSt	93888	4099	1.56	0.727
Petriji i Dovbuščuky (2nded.)	PD2	52751	2730	1.58	0.728
Velykyj šum	VSh	37005	1923	1.66	0.719
Zakhar Berkut	ZBe	50206	2521	1.54	0.724
			Average:	**1.59**	**0.722**

We further study the behavior of the "temperature" parameter T in the course of text production, i.e. how T changes for the first 1,000 of words in each text, for the first 2,000 of words, and so on. For simplicity, we fix the parameter α at 1.6. It appears that the dependence of T on N is, to a good accuracy, described by a simple power law:

$$T = tN^{\beta}. \qquad (3)$$

Note that such dependences occur in linguistics in various contexts (Naranan, Balasubrahmanyan 1998; Köhler 2002; Kaniadakis 2009).

Figure 3 demonstrates the results of fitting for full texts (with no division on the author's and direct speech). Detailed results of calculations are given in the next section. The analysis was completed for texts as a whole as well as for the author's and direct speech separately. These types of speech differ in the values of some text parameters, e.g., number of hapax legomena, dictionary and text richness, etc. (Buk 2011).

We can observe minor oscillations around the fitting curve of model (3). The nature of such behavior could itself become the subject of more detailed studies in the future (cf. Zörnig et al. 1990).

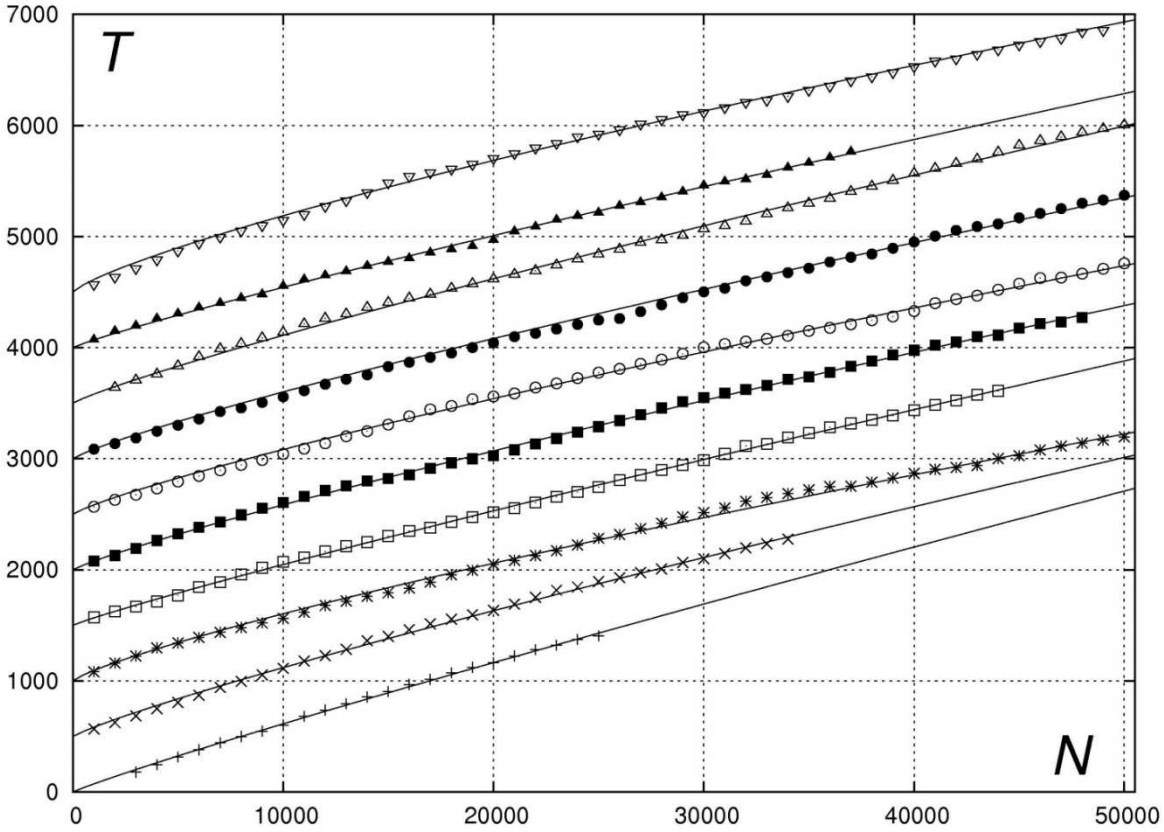


Figure 3. "Temperature" *T* as a function on text length *N* in Ivan Franko's prose fiction. For convenience, the values are shifted vertically by 500 for ten novels listed in Table 1, bottom to top: BC1 (+), BC2 (×), BSm (✳), DDO (□), NSB (■), OSu (○), PSt (●), PD2 (△), VSh (▲), and ZBe (▽).

4. Results and their interpretation

The following figures demonstrate the values of the *t* and *β* parameters of Ivan Franko's long prose fiction for full texts, author's speech, and direct speech of all the ten Ukrainian texts listed in Table 1. In Figure 4, the same scale is used for all the three panels while different scales (enlarged regions) are shown in Figure 5 for better visualization. Numerical values of the parameters and respective errors are given in Table 2.

In the direct speech data one can observe the highest concentration for PSt, PD2, and BC2; ZBe and VSh are located close together, as are BSm and OSu.

As expected, rather compact data are observedin author's speech. For full texts, the highest concentration is found for NSB, PD2, and BC2. A separate location of PSt and OSu is also clearly seen. A somewhat dispersed group containing DDO, BC1, and VSh can be distinguished as well.

It would be interesting to find some common features of texts reflecting the abovementioned grouping. Below, we will analyze the correlation of the calculated parameters with some other numerical text data.

To extend the previously applied apparatus, one can also define the quantity

$$\mu = T \ln z \qquad (4)$$

known in physics as the chemical potential. For sub-systems in equilibrium, the chemical potentials are equal. We can estimate to what extent such a claim is applicable to our model by considering the direct and author's speech as separate sub-systems within each text. To make the comparison, the following relative errors were calculated in each case:

$$\delta\mu = 2\,|\mu_a - \mu_d| \,/\, (\mu_a + \mu_d). \qquad (5)$$

The subscripts "a" and "d" correspond to the author's and direct speech, respectively.

The results are presented in Table 2. While there is quite strong inverse correlation between $\delta\mu$ and text size N (with the coefficient $R = -0.61$, see Table 3), we can still observe that the direct and author's speech is best balanced in *Perekhresni stežky* and worst in *Petriji i Dovbuščuky* (2nd ed.).

Curiously enough, such an estimation contradicts opinions of literary reviewers in the case of *Perekhresni stežky* (cf. Batsevych et al. 2007: 7-8).

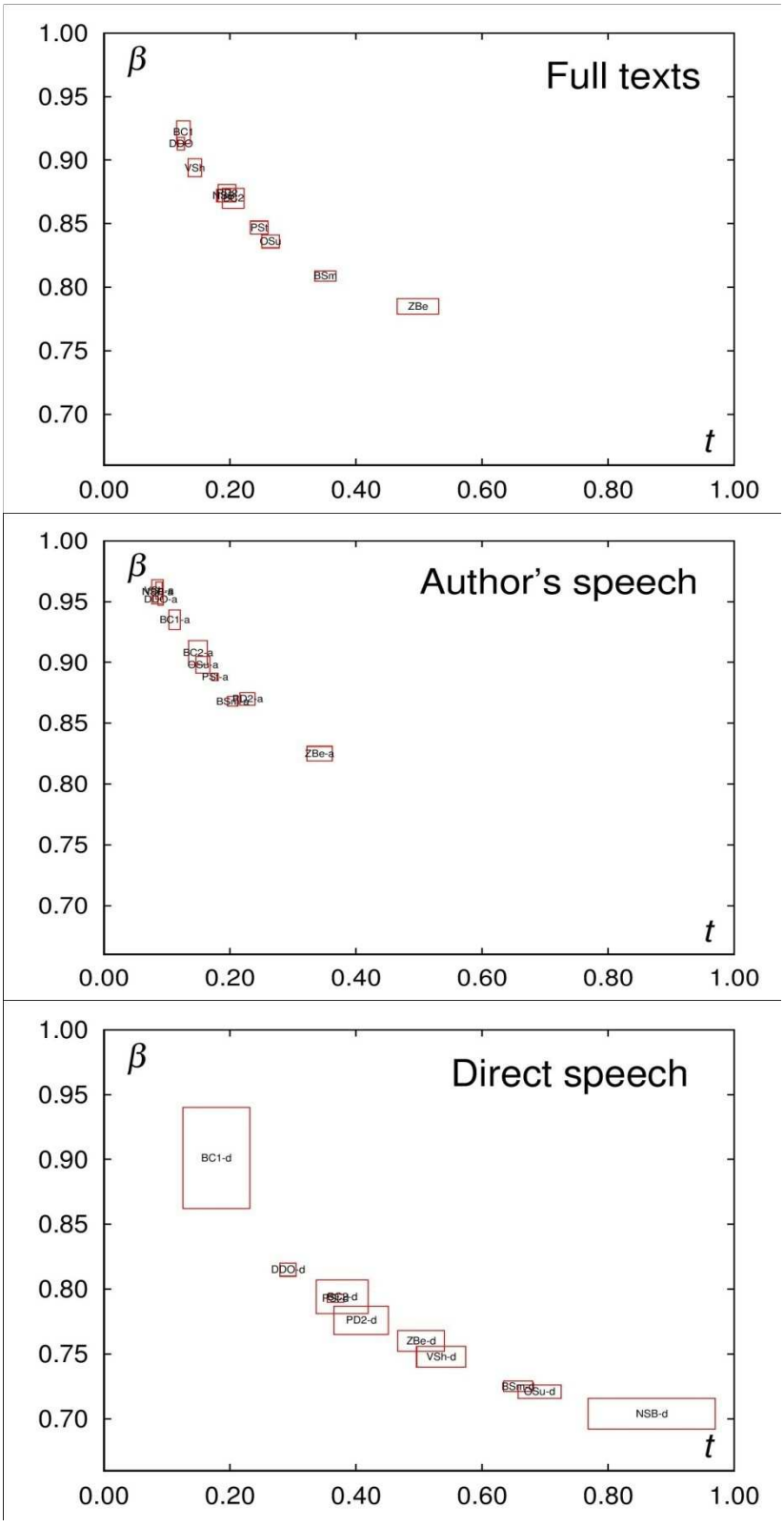


Figure 4. Parameters of Ivan Franko's long prose fiction for full texts, author's speech, and direct speech

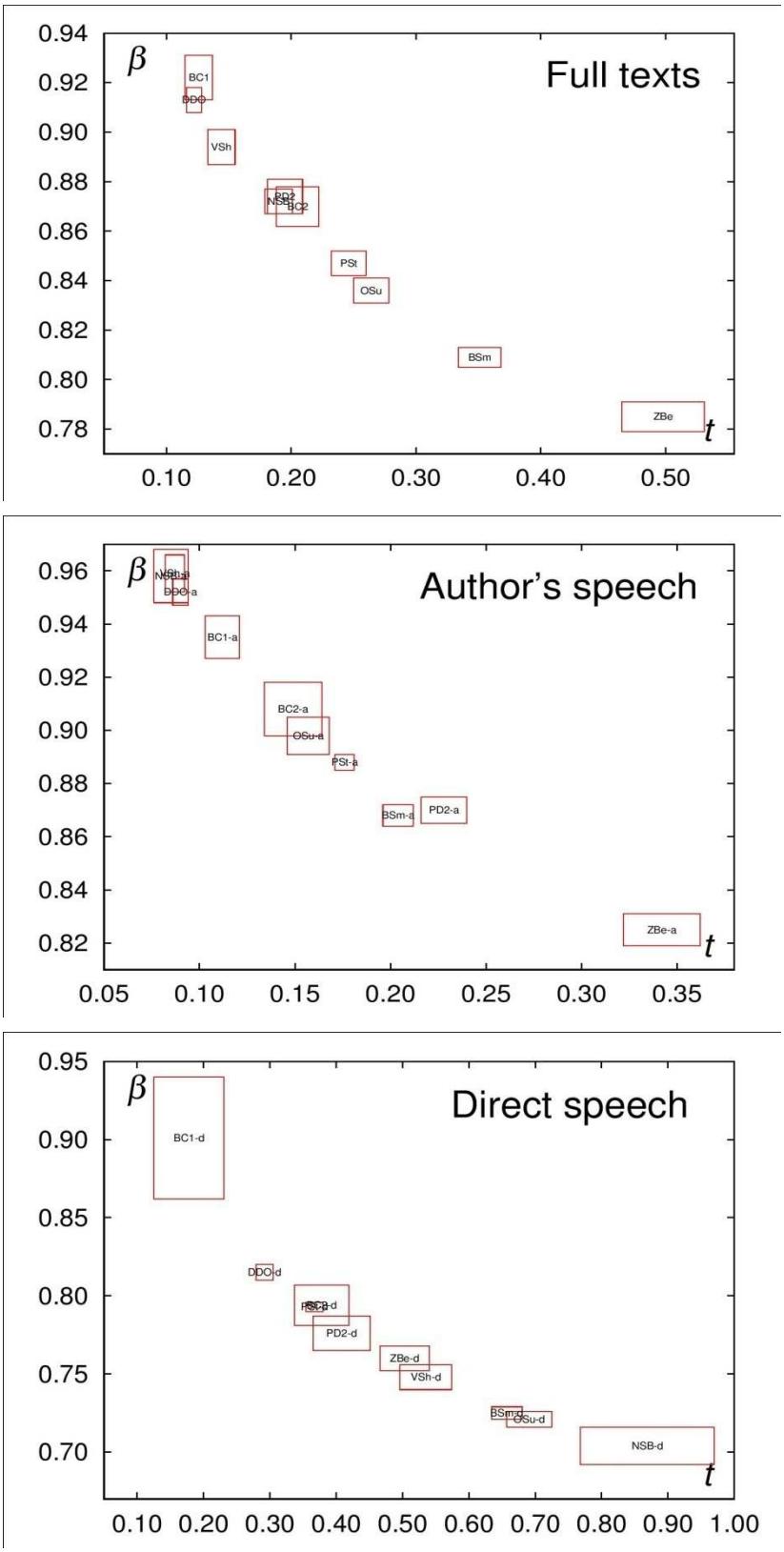


Figure 5. Parameters of Ivan Franko's long prose fiction for full texts, author's speech, and direct speech, enlarged regions compared to Figure 4.

Table 2
The values of parameters for Ivan Franko's long prose.
The extensions "-a" and "-d" correspond to the author's and direct speech, respectively. The "size" column lists the total number of tokens in thousands (mathematical floor is applied).

Text	t	Δt	β	$\Delta\beta$	T	N_1	μ	$\delta\mu$	size
BC1	0.126	0.011	0.922	0.009					
BC1-a	0.112	0.009	0.935	0.008	1270	5254	–0.242	13.4%	25
BC1-d	0.178	0.053	0.901	0.039	245	886	–0.276		
BC2	0.205	0.017	0.870	0.008					
BC2-a	0.149	0.015	0.908	0.010	1527	6522	–0.234	8.0%	34
BC2-d	0.378	0.041	0.794	0.013	443	1747	–0.254		
BSm	0.351	0.017	0.809	0.004					
BSm-a	0.204	0.008	0.868	0.004	2179	7930	–0.275	5.0%	76
BSm-d	0.657	0.023	0.725	0.004	1223	4231	–0.289		
DDO	0.122	0.006	0.913	0.005					
DDO-a	0.090	0.004	0.952	0.005	1393	5785	–0.241	8.7%	44
DDO-d	0.292	0.013	0.815	0.005	885	3368	–0.263		
NSB	0.190	0.011	0.872	0.005					
NSB-a	0.085	0.009	0.958	0.010	1479	6293	–0.235	4.9%	48
NSB-d	0.869	0.101	0.704	0.012	922	3738	–0.247		
OSu	0.264	0.014	0.836	0.005					
OSu-a	0.157	0.011	0.898	0.007	2051	7499	–0.274	8.8%	67
OSu-d	0.691	0.034	0.721	0.005	1121	4475	–0.251		
PSt	0.246	0.014	0.847	0.005					
PSt-a	0.176	0.005	0.888	0.003	2481	8640	–0.287	0.8%	93
PSt-d	0.367	0.013	0.793	0.003	1858	6524	–0.285		
PD2	0.195	0.014	0.874	0.007					
PD2-a	0.228	0.012	0.870	0.005	2035	6991	–0.291	18.3%	52
PD2-d	0.408	0.043	0.776	0.011	761	3142	–0.242		
VSh	0.144	0.011	0.894	0.007					
VSh-a	0.087	0.005	0.959	0.007	1176	5784	–0.203	13.2%	37
VSh-d	0.535	0.039	0.748	0.008	773	3332	–0.232		
ZBe	0.498	0.033	0.785	0.006					
ZBe-a	0.342	0.020	0.825	0.006	1769	6541	–0.270	6.6%	49
ZBe-d	0.503	0.037	0.760	0.008	829	2870	–0.289		

As there were no immediately apparent text properties reflecting the behavior of t, β,andμ, we analyzed their correlation with text size (N_{tot}), year of publication, epithetization index (EI), as well as the so-called vocabulary richness (TTR). The epithetization index is the number of adjectives divided by the number of nouns in a given text (Ruszkowski 2004; Buk 2012). The vocabulary richness is equal to the ratio of the number of different lemmas and the total number of words in a given text (type-token ratio or TTR).

Table 3
Correlation coefficients between various text parameters

	t	β	N_{tot}	μ	year	EI	TTR
t	—	–0.96	+0.42	–0.41	–0.42	–0.07	–0.57
β	–0.96	—	–0.59	+0.49	+0.31	–0.13	+0.70
N_{tot}	+0.42	–0.59	—	–0.61	–0.08	+0.63	–0.92

The highest correlation of β with TTR is in agreement with previously discovered relation between the "temperature"-related parameters and the level of language analyticity (Rovenchak, Buk 2011a; 2011b; Rovenchak 2014). Another interesting result is the absence of correlation between t, β and the epithetization index. The latter could signify, in particular, some complications in substantiating the argument that the groupings observed in Figures 4 and 5 were literary-based.

In view of a high correlation between t and β, it is interesting to find the fitting function as a simple power law:

$$\beta = bt^{-a} = 0.722t^{-0.114}. \qquad (6)$$

We thus obtain a very simple link between $\tau = \ln T/\ln N_{\text{tot}}$ and β:

$$\tau = \beta + \frac{1}{a}\frac{\ln b - \ln \beta}{\ln N_{\text{tot}}}, \qquad (7)$$

where $b = 0.722\ldots$ and $a = 0.114\ldots$. Obviously, the second item vanishes as the text size becomes large (cf. Rovenchak 2015b).

5. Prospects and discussion

In this paper we analyzed Ivan Franko's long prose fiction using a previously developed method inspired by a physical model. We were able to reconfirm previously revealed relations of the "temperature" parameter with the level of language analyticity. This time it is made indirectly, from the correlation between the "temperature" scaling exponent β and the vocabulary richness (type-token ratio with lemmas considered as types).

Moreover, some new results were also obtained. The observed small oscillations of the $T(N)$ dependence around the simple power-law fitting function might signify some supra-sentence peculiarities in the frequency structure of text (cf. Altmann 2014; Chen, Altmann 2015). In future studies it would be interesting to apply the analysis based on the introduced μ parameter for different sub-systems in texts, e.g. part-of-speech distribution. We also hope to obtain similar data for other languages and authors, to verify the approach developed in this study. It seems tempting to ascribe some unmeasurable properties to the analyzed texts on the basis of the calculated parameters. Whether it is possible – at least to a certain extent – remains so far an open issue. We expect that additional studies involving experts in the domain of literary criticism will clarify this problem.

Acknowledgements

A. R. acknowledges partial support by Project FF-30F (No. 0116U001539) from the Ministry of Educationand Science of Ukraine and by Grant F64/37-2016 (No. 0116U005055) from the State Fund For Fundamental Research of Ukraine.

References

Altmann, G. (2014). Supra-sentence levels. *Glottotheory 5: 25–39*.

Batsevych, F.S.; Buk, S. N.; Protsak, L.M.; Rovenchak; A.A.; Svarychevska, L. Yu.; Tsikhotskyj, I. L. (eds.) (2007). *Stežkamy Frankovoho tekstu (komunikatyvni, stylistyčni ta leksyčni vymiry romanu "Perekhresni stežky")*. Lviv: Lviv University Press.

Best, K.-H.; Zinenko S. (1998). Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. *Zeitschrift für Slavische Philologie 58: 107–123*.

Best, K.-H.; Zinenko S. (1999). Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In: Genzor, J.; Ondrejovič S. (eds.) *Pange lingua. Zborník na počest' Viktora Krupu: 201–213*. Bratislava: Veda.

Bohorquez, J.C., Gourley, S., Dixon, A.R., Spagat, M., Johnson, N.F. (2009). Common ecology quantifies human insurgency. *Nature 462: 911–914.*

Buk, S. (2007). Korpus tekstiv Ivana Franka: sproba vyznačennja osnovnykh parametriv. In: Shyrokov, V.A. (ed.), *Prykladna linhvistyka ta linhvistyčni tekhnolohiji: MegaLing 2006: 72–82*. Kyiv: Dovira.

Buk, S. (2010). Statystyčna structura romanu Ivana Franka "Boryslav smijetsja". *Učenye zapiski Tavričeskogo natsionalnogo universiteta im. V.I. Vernadskogo. Serija Filologija 23(62), № 3, 114–118*.

Buk, S. (2011). Prjama j avtorska mova velykoji prozy Ivana Franka: linhvostatystyčne doslidžennja u konteksti korpusnoji linhvistyky. *Visnyk Lvivskoho universytetu. Serija filolohična 52, 199–209.*

Buk, S. (2012). The epithetization index in a work of fiction (on the basis of the text corpus of Ivan Franko's long prose fiction). In: Obrębska, A. (ed.)

Practical Applications of Linguistic Research: 73–85. Łódź: Primum Verbum.

Buk, S. (2013). Kvantytatyvna parametryzacija tekstiv Ivana Franka: proekt ta joho realizacija. *Visnyk Lvivskoho universytetu. Serija filolohična 58, 290–307.*

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel *Perekhresni stežky (The Cross-Paths)***.** In: Grzybek, P., Köhler, R. (eds.) *Quantitative Linguistics 62: Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday*: *39-48*. Berlin–New York: Mouton de Gruyter.

Chen, R., Altmann, G. (2015). Conceptual inertia in texts. *Glottometrics 30, 73-85.*

Colaiori, F., Castellano, C., Cuskley, Ch. F., Loreto, V., Pugliese, M., Tria, F. (2015). General three-state model with biased population replacement: Analytical solution and application to language dynamics. *Physical Review E 91: 012808 [12 pages]*.

Čech, R., Mačutek, J., Žabokrtský, Z. (2011).The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A 390: 3614–3623*.

Denysiuk, I. (2008). Novatorstvo Franka-prozajika. *Ukrajinske literatureznavstvo 70: 138–152*.

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

Ferrer-i-Cancho, R., Solé, R.V., Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review A 69: 051915* .

Franko, I. (1976–86). *Zibrannja tvoriv u 50-ty tomakh*. Kyiv: Naukova Dumka.

Holovatch, Yu., Palchykov, V. (2017). Complex networks of words in fables. In: Kenna, R.; Mac Carron, M.; Mac Carron, P. (eds.) *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*: *159–175.* Springer.

Huang, K. (1987). *Statistical Mechanics.* 2nd edition. New York: Wiley.

Jin, Neng-zhi, Liu, Zi-xian, Qiu, Wen-yuan (2009). Frequency and correlation of nearest neighboring nucleotides in human genome. *Chinese Journal of Chemical Physics 22, 27–33.*

Kaniadakis, G. (2009). Maximum entropy principle and power-law tailed distributions. *European Physical Journal B 70, 3–13*.

Kelih, E., Rovenchak, A., Buk S. (2014). Analysing h-point in lemmatised and non-lemmatised texts. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.) *Studies in Quantitative Linguistics 17: Empirical Approaches to Text and Language Analysis; dedicated to Luděk Hřebiček on the occasion of his 80th birthday: 81–93*. Lüdenscheid: RAM-Verlag.

King, S. (2010). Afterword. In: *Different Seasons*. London: Hachette Livre UK Co.

Köhler, R. (2002). Power law models in linguistics: Hungarian. *Glottometrics 5, 51–61*.

Kosmidis, K., Kalampokis, A., Argyrakis, P. (2006). Statistical mechanical approach of human language. *Physica A 366, 495–502*.

Kotter, M.P. (2008). *English Literature: Modern World View*. New Delhi: Cyber Tech Publications.

Mac Carrona, P., Kenna, R. (2013). Network analysis of the Íslendinga sögur – the Sagas of Icelanders. *European Physical Journal B 86, 407*.

Mandelbrot, B. (1953). An informational theory of the statistical structure of languages. In: Jackson, W. (ed.) *Communication Theory 486–502*. New York: Academic.

Miyazima, S., Yamamoto, K. (2008). Measuring the temperature of texts. *Fractals 16: 25–32*.

Mryglod, O., Fuchs, B., Szell, M., Holovatch, Yu., Thurner, S. (2015). Interevent time distributions of human multi-level activity in a virtual world. *Physica A 419, 681–690.*

Naranan, S., Balasubrahmanyan, V. K. (1998). Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics 5, 35–61*.

Palchykov, V., Kaski, K., Kertész, J., Barabási, A.-L., Dunbar, R.I.M. (2012). Sex differences in intimate relationships. *Scientific Reports 2, 370.*

Pastukh, T. (1996). Roman u systemi prozovykh tvoriv Ivana Franka. *Ukrajins'ke literaturoznavstvo* 6*2, 100–108.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word Frequency Studies*. Berlin-New York: Mouton de Gruyter.

Rêgo, H.H.A., Braunstein, L.A., D'Agostino, G., Stanley, H.E., Miyazima, S. (2014). When a text is translated does the complexity of its vocabulary change? Translations and target readerships. *PLoS ONE 9(10): e110213.*

Rodriguez, E., Aguilar-Cornejo, M., Femat, R., Alvarez-Ramirez, J. (2014). Scale and time dependence of serial correlations in word-length time series of written texts. *Physica A 414: 378–386*.

Rovenchak, A. (2014). Trends in language evolution found from the frequency structure of texts mapped against the Bose-distribution. *Journal of Quantitative Linguistics 21(3), 281–294*.

Rovenchak, A. (2015a). Models of frequency spectrum in texts based on quantum distributions in fractional space dimensions. In: Dumitrache, I., Florea, A. M., Pop, F., Dumitraşcu, A. (eds.) *20th International Conference on Control Systems and Computer Science CSCS 2015: Proceedings, 27-29 May 2015, Bucharest, Romania, Vol. 2: 645–649*. Los Alamitos, CA: IEEE Computer Society.

Rovenchak, A. (2015b). Where Alice meets Little Prince: Another approach to study language relationships. In: Mikros, G. K., Mačutek, J. (eds.), *Sequences in Language and Text: 217–230*. Berlin–Boston: Mouton de Gruyter.

Rovenchak, A., Buk, S. (2011a). Application of a quantum ensemble model to linguistic analysis. *PhysicaA 390(7), 1326–1331*.

Rovenchak, A., Buk, S. (2011b). Defining thermodynamic parameters for texts from word rank-frequency distributions. *Journal of Physical Studies 15(1), 1005 [6 pages]*.

Ruszkowski, M. (2004). Wskaźnik epitetyzacji w badaniach stylistycznych. *Respectus Philologicus 5(10), 48–53.*

Stech, M. R., Zhukovsky, A. (2007). Franko, Ivan. In: Kubijovyč, V. (ed.) *Encyclopedia of Ukraine,* http://www.encyclopediaofukraine.com/pages/F/R/FrankoIvan.htm

Thrall, W.F., Hibbard, C.A. (1960). *A Handbook to Literature* (2nd ed, revised and enlarged by C. Hugh Holman). New York: Odyssey Press.

Tran, M.N., Murthy, M.V.N., Bhaduri, R.J. (2004). On the quantum density of states and partitioning an integer. *Annals of Physics 311, 204–219*.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3(1), 38–50*.

Waas, G. (2012). Introduction. In: *The Great American Short Story*. Bloomington, IN: AuthorHouse.

Zörnig, P., Köhler, R., Brinkmöller, R. (1990). Differential equation models for the oscillation of the word length as a function of frequency. *Glottometrika 12, 25–40.*

Can Pronouns Change the Dynamic Visualization of the Poetic World?

Sergey Andreev
Smolensk State University,
Russian Federation

Abstract. The article analyses linguistic features of style which reflect the author's type of visualization of the world. Traditionally the ratio between verbs, nouns and adjectives is used to estimate the type of style: static or dynamic. The question is whether the introduction of pronouns of nominal and adjectival character into analysis can change the estimation of the type of style. This problem addressed on the basis of long poems by A.S. Pushkin, one of the greatest authors in Russian literature.

Keywords: *style, static, dynamic, visualization, pronouns, motifs, long poems, Pushkin.*

Among many other important aspects of style there is an important question of the author's type of visualization of the world – whether it is static or dynamic. In the first case, description of the main themes takes the form of representing their permanent qualities with the help of adjectives; in the second case, by means of showing actions or states with verbs.

The type of relationship of these of these parts of speech (PoS) has been studied in different languages and for many authors. It is possible to compare the two above-mentioned world visualization types (WVT) by finding out in texts the proportion of adjectives against verbs or adjectives and the proportion of verbs against nouns (Gasparov 2012; Popescu et al. 2013; Andreev 2016).

Such studies, in a number of cases, exhibited unexpected results. Thus, in Russian literature, it was discovered that verses (even lyrics) in many cases contained a far greater proportion of verbs than might theoretically be supposed. It was found that, despite the belief that poetry in general, and lyrics in particular, should contain a far smaller proportion of verbs and a greater proportion of adjectives than prose, it actually does not. In Russian literature, many poets, including A.S. Pushkin, one of the best Russian poets, used a smaller proportion of adjectives than L. Tolstoy and A. Chekhov in their prose texts (Gasparov 2012).

Analyzing this phenomenon, one objection can be raised: is it possible that the proportion of adjectives and nouns, which shows a deviation in adjectives from the theoretically expected level, is due to the fact that other words with attributive function and the same morphological status are used instead. Strictly speaking, the same applies to nouns, expressing themes in the text.

Poetic text possesses a number of specific features (rhythm, specific syntactic structures, rhyme, etc.) which establish certain linear and vertical interrelations between the elements of the text. This imposes strict limitations on the choice of words, in many cases demanding the introduction of short words at some positions in lines. In many languages and in Russian in particular, it is the pronouns that – among other functions – perform the function of substituting for longer words.

Pronouns traditionally are subdivided into personal, demonstrative, possessive, reflexive, etc.; according to another categorization, they fall into substantival (nominal), adjectival and adverbial pronouns (Shvedova 1980). In Russian, nominal pronouns include the following traditional pronoun types: personal, reflexive, some interrogative and negative pronouns. To adjectival pronouns belong possessive, one type of reflexive, demonstrative, descriptive, indefinite, some interrogative and negative pronouns.

In studies devoted to WVT, these categorical subdivisions of pronouns are not, as a rule, taken into account. In this paper, we will estimate the changes that nominal and adjectival pronouns can introduce into WVT if they are included in counts of nouns and adjectives.

One of the main trends in quantitative linguistics is overcoming the borders of word-phrases and sentences (in poetry also lines) in order to analyze text structure and solve many other textometric problems (Köhler, Altmann 2014). In this line of work, there are two related methodologies: one using sequences, singled out on the basis of the number of certain PoS before each successive noun in the text (Naumann et al. 2012, Popescu et al. 2007, Tuzzi et al. 2009), the other establishing the borderlines by quantitative alternations of definite features in a sequence (Köhler 2006, 2008; Köhler, Naumann 2012, 2016).

The database for the study consisted of 5 long poems by A.S. Pushkin: *Vadim*; *Bratya razboyniki* (*The Robber Brothers*); *Tasit*; *Jezierski*; *Mednyj vsadnik* (*The Bronze Horseman*). In Andreev (2016), these poems were analyzed from the point of view of the above-mentioned dichotomy of static vs. dynamic character of poetic world visualization, without using pronouns in the analysis.

The observed (empirical) proportions of verbs and adjectives against nouns, now also including nominal and adjectival pronouns, were calculated and compared with expected values. To test the possible deviation of the observed proportion from its expectation the following criterion was used (Naumann et al. 2012: 29):

$$u = \frac{p(S) - E(p)}{\sqrt{p(1-p)/(S+N)}},$$

where p is the proportion of the given part of speech S, $E(p)$ is the expected proportion, $p(S)$ is the observed proportion in the text of the given part of speech S, N is the number of observed nouns (Naumann et al. 2012).

The expected proportion of (a) verbs and (b) adjectives against nouns is correspondingly 0.33846 and 0.25862 and is based on the counts of M. Gasparov for Russian literature (Andreev 2016; Gasparov 2012). In tables 1 and 2, the proportions of PoS and the *u*-criterion are given.

Table 1
Proportion of verbs

Long Poems	Without adjectival and nominal pronouns		Including nominal pronouns	
	V/(V+N)	u	V/(V+N)	u
Vadim	0.270	-0.53	0.291	-2.21
The Robber Brothers	0.384	2.17	0.318	-1.06
Tasit	0.376	1.86	0.310	-1.54
Jezierski	0.283	-2.39	0.251	-4.07
The Bronze Horseman	0.366	1.85	0.327	-0.83

Table 2
Proportion of adjectives

Long Poems	Without adjectival and nominal pronouns		Including adjectival and nominal pronouns	
	A/(A+N)	u	A/(A+N)	u
Vadim	0.288	1.38	0.291	1.67
The Robber Brothers	0.239	-0.91	0.233	-1.38
Tasit	0.258	-0.05	0.247	-0.65
Jezierski	0.216	-1.90	0.283	1.28
The Bronze Horseman	0.286	1.84	0.304	3.38

The results show certain differences in the obtained ratio of verbs to nouns and adjectives to nouns, but the scale of these differences is not the same. The dynamic of texts decreases considerably. Three texts formerly characterized by a high proportion of verbs now are neutral; the style of one formerly neutral text (*Vadim*) reveals nominality. This considerable change in the results due to the introduction of pronouns into the analysis demonstrates that the dynamic character of Pushkin's verse is not as strong as it was considered to be.

The situation with the adjectival type of style is remarkably different. In three poems, the new method of analysis has not led to any changes of style type, adjectival or neutral. In one poem (*Vadim*) the tendency towards adjectival style is strengthened. Perhaps the only real change in the estimation of style takes place in *Jezierski.* Now the style is neutral and not nominal ('deficiency' of adjectives disappears).

At the next stage of research, the method of motif analysis was used. Linguistic motifs are viewed as a sequence of equal or increasing values of some feature of a linguistic unit (Köhler, Naumann 2016). In this study, motifs are based on the morphological features – sequences of parts of speech – and reflect

the dynamics of (a) verb occurrence in relation to nouns and (b) adjective occurrence in relation to nouns. To form motifs, the number of verbs (or adjectives) preceding each noun in the text was counted. Then the borderlines for motifs were fixed in this sequence. The indicator for the borderline was a decrease in the number of verbs (or adjectives) in this sequence.

Thus, for example, the beginning *of The Bronze Horseman* gives the following sequence of verbs preceding nouns in the text:

0 0 1 1 1 1 1 0 0 0 0 0 1 3 0 1 0 0 0 1 0…

It means that before the first and the second nouns there were no verbs, before the third noun one verb was found, before the fourth also one, etc. There are five motifs in this sequence which are as follows (shown by slashes):

0 0 1 1 1 1 1 / 0 0 0 0 0 1 3 / 0 1 / 0 0 0 1 / 0…

Motifs are classified depending on how many members they include. In the example above, the sequence of motifs is as follows: 7-member motif; 7-member motif; 2-member motif; 4-member motif. These second-order motifs are used to characterize the texts of long poems, i.e., how many 1-member, 2-member, 3-member, etc. motifs each poem contains.

Due to the rare occurrence of sequences containing over 16 members, these are treated as one type of extra-long motif. Motifs were calculated for each long poem, both with and without taking into account nominal and adjectival pronouns. Thus, every text is described two times and the mark ‘Pr’ is used to show that pronouns were counted.

The obtained data was used as the basis for agglomerative hierarchical cluster analysis (complete linkage, Euclidean distances).

Figure 1 displays the tendency of grouping the texts, regardless of whether or not pronouns were included into the counts. Thus, *Jesierski* and *The Bronze Horseman* reveal similarity in the structure of motifs no matter whether pronouns were counted or not. Only *Vadim* shows a certain difference in this respect: *Pr_Vadim* and *Vadim* fall into different and rather distant clusters.

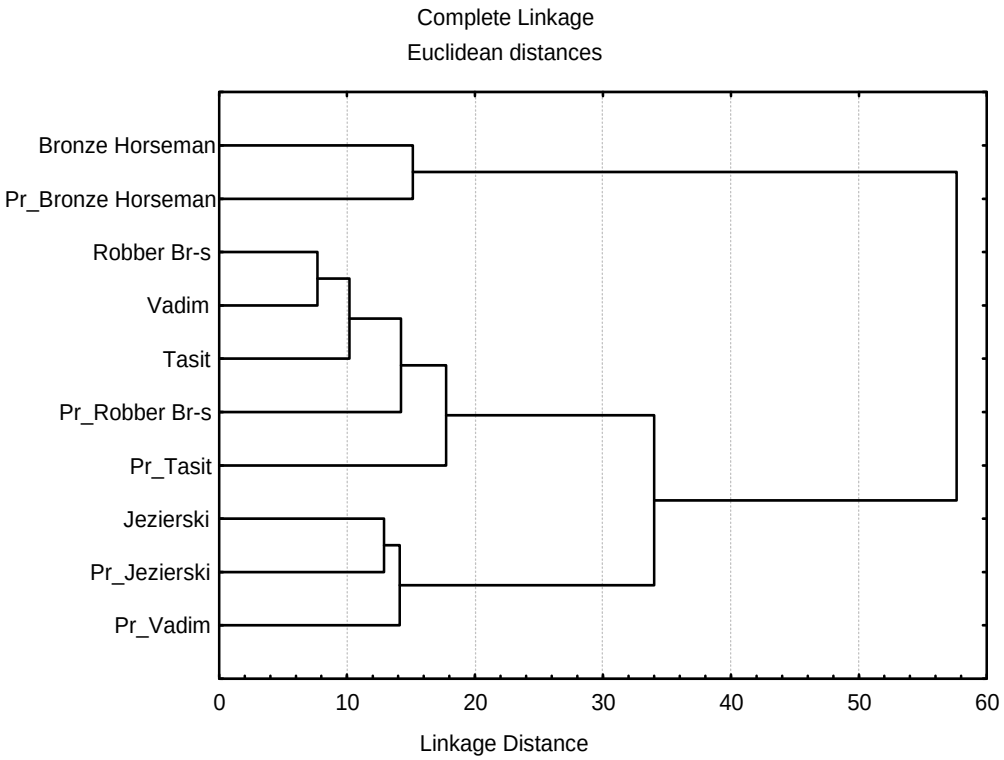


Fig. 1. Classification of texts on the basis of motifs in verb-noun sequences

The same tendency, to some extent, is revealed in Figure 2 for adjectival motifs, but it is somewhat weaker. The biggest difference is observed for *Jesierski*. This corresponds to the data in Table 2, where the image of the style of this poem after counting adjectival pronouns displayed a rather strong change, moving from nominal to adjectival.

On the whole, it is possible to conclude that including nominal and adjectival pronouns into the analysis changes, to some extent, the evaluation of Pushkin's style. But if the dynamic type of visualization becomes less strong, the adjectival tendency of his style is not noticeably changed after including adjectival pronouns in the counts. Of course, the use of pronouns in estimating WVT should be tested on far more extensive data for different poets, including short lyrical poems, on the one hand, and prose, on the other

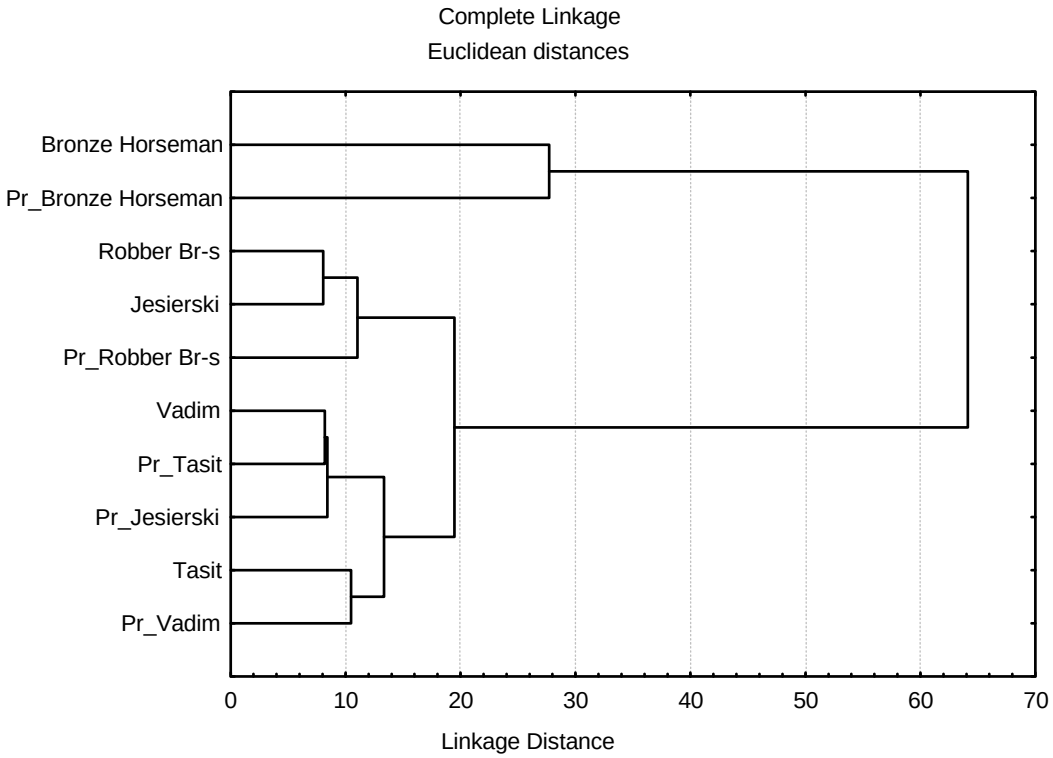


Fig. 2. Classification of texts on the basis of motives in adjective-noun sequences

The motif methodology reveals structural features of style that turn out to be very stable and may be taken as a good basis for the classification of verse texts.

References

Andreev, S. (2016). Verbal vs. adjectival styles in long poems by A.S. Pushkin. *Glottometrics 33,* 25–31.

Gasparov, M.L. (2012). Tochniye metody analyza grammatiki v stihe [Exact methods of verse analysis]. In: *V.L. Gasparov. Izbranniye trudy. Lingvistika stiha. Analyzy i interpretatsyy.* Vol. 4. Moscow: Yaziky Slavanskoy kul'tury, 23–35.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J.; Buckova, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa.* Bratislava: Slovak Academic Press, 145–152.

Köhler, R. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičova, Sibyla (ed.), *Jazyk a jazykoveda v prohybe.* Bratislava: VEDA vydavatel'stvo SAV, 416–421.

Köhler, R., Altmann, G. (2014). *Problems in quantitative linguistics*. Vol. 4. RAM-Verlag.

Köhler, R., Naumann, S. (2012) A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: *Proceedings of COLING 2012* (Mumbai, December 2012). Technical Papers, Mumbai, 263–278.

Köhler, R., Naumann, S. (2016) Syntactic text characterisation using linguistic S-motifs. *Glottometrics 34, 1-8.*

Naumann, S., Popescu, I.-I., Altmann, G. (2012). Aspects of nominal style. *Glottometrics 23*, 23–55.

Popescu, I.-I., Čech, R., **Best, K.-H., Altmann G. (**2013). Descriptivity in Slovak lyrics. *Glottotheory 4*, 92–104.

Shvedova N.Yu. (1980) (ed.). *Russkaja Grammatika* [Russian Grammar] / Vol. 1. Nauka, Moscow, 1980.

Popescu, I.-I., Best, K.-H., Altmann G. (2007). On the dynamics of word classes in text. *Glottometrics 14*, 58–71.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19*, 42–48.

A Study on Segmental *TTR*, Word Length and Sentence Length

Fan Fengxiang
School of Foreign languages, Dalian Maritime University
fanfengxiang@yahoo.com

Abstract. This research examines *TTR*, word length and sentence length at the sub-textual level. Texts were divided into three consecutive segments of equal length and their respective *TTR*, word length and sentence length were computed. The inter-textual rank distribution of the three metrics of the text segments follow an *S*-shaped curve and can be captured with the modified Nemcová and Serdelová model. More importantly, the *TTR*, word length and sentence length of the first segment of a text tend to be larger than those in the last segment.

Keywords: *text segments, TTR, word length, sentence length, ANOVA test.*

1. Introduction

TTR, word length (hereafter *WL*) and sentence length (hereafter *SL*) are popular old measures in quantitative linguistics, old in the sense that, unlike the new measures such as the motif (Köhler, 2006a, b, 2008a, b; Köhler, Naumann, 2008, 2009, 2010), arc length (Popescu et al. 2008), lambda structure (Popescu, Altmann, 2014) and so on, their applications can be traced back to the early stage of quantified language studies, now an important aspect of quantitative linguistics, which is relatively young and still developing at a steady stride. According to Köhler, *TTR* is an index traditionally used as a stylistic characteristic of texts, text sorts, or authors, and believed to represent vocabulary richness in a way which enabled researchers to compare texts to each other and even to identify individual authors (Köhler, 2012); while *WL* and *SL*, apart from the foregoing functions of *TTR*, have been considered from the very beginning key features whose systematic study can reveal important aspects of language structure, differentiations and text typology (Kalimeri et al, 2015). The applications of the three measures in modern quantitative linguistics are much wider and deeper, and more scientifically rigorous. Since 1980's there have been a large number of publications using *TTR*, *WL* or *SL* as pivotal metrics on different aspects of language (Köhler, 1982; Altmann, 1980, 1988; Wimmer et al, 1994; Wimmer, Altmann, 1996; Fan, 2007; Mikros, 2007; Grzybek et al., 2008; Fan, Grzybek, Altmann, 2010; Levitsky,& Melnyk, 2011; Fan, 2012; Kubát,& Milička, 2013; Kalimeri et al. 2015). In these studies the researchers were mainly interested in examining characteristics of texts or certain laws holding in language with these measures as differentiating indexes.

The present research intends to study the distributions of *TTR*, *WL* and *SL* at the sub-textual level of non-fictional texts. *TTR*, *WL* and *SL* of text segments are referred to as segmental *TTR*, *WL* and *SL*, in contrast to those of the entire text,

which are referred to as textual *TTR*, *WL* and *SL* here. Except for segmental *TTR*, segmental *WL* and *SL*, like textual ones, are in fact mean *WL* and mean *SL* of individual segments. According to Köhler, a text is an expression for a complex and multi-dimensionally structured, cognitive (conceptual, emotional, intentional) formation (Köhler, 2012, 2). Texts can be considered as a time series and have stratifications (Altmann et al, 2013). This means a text can be separated physically into its component structures or layers that have different discourse functions and strategies. According to Quirk et al. (1985), traditionally textual structures were studied from their internal relationships such as asyndetic connection, syndetic connection, thematic connection, rhematic connection and so on. Few researchers used measures such as *TTR*, *WL* or *SL* to characterize text structures. Naturally occurring texts can either be oral or written and their lengths range from one word to hundreds of thousands. In this research, only written texts from present-day magazines and newspapers were used. Generally a written text from the above mentioned source has three major sections: the initial section introducing to the reader the topic of the text; the body that develops or expounds on the topic and the concluding section. Specifically, this research focuses on whether the different functions of the major text segments can be reflected in the distribution of *TTR*, *WL* and *SL*. *WL* was measured in number of letters and *SL* in number of words.

2. Data and methods

Texts were from the following American and British magazines and newspapers:

1. *Newsweek*
2. *Scientific American Magazine*
3. *Smithsonian Magazine*
4. *The Economist*
5. *The Guardian*
6. *The Independent*
7. *The National Geographic*
8. *The New York Times*
9. *The New Yorker*
10. *The Washington Post*
11. *Time Magazine*

Altogether 285 texts were selected, published mostly between 2005 and 2015, with a few in the late 80's and early 90's. They were complete texts with lengths ranging from 545 to 15,150 words on politics, world affairs, finance, science and technology, education, entertainment, war, religion, history and so on. The total number of words of these articles was 639,172 words, with a mean text length of 2,173 words.

Each text was divided into three consecutive segments of roughly equal length, with individual segments beginning and ending in complete sentences. The *TTR*, *WL* and *SL* of the first, second and third segments are respectively

referred to as TTR_1, TTR_2, TTR_3, WL_1, WL_2, WL_3, SL_1, SL_2 and SL_3. The segmental *TTR*, *WL* and *SL* of the individual segments were first computed, followed by the overall mean TTR_1, TTR_2, TTR_3, WL_1, WL_2, WL_3, SL_1, SL_2 and SL_3 of the 285 texts.

There are several ways to compute *TTR*. The traditional computation $TTR = 100\frac{Types}{Tokens}$ (Laufer, Nation, 1995; Biber et al, 1999) cannot be used here since it is affected by text length. Köhler, Galle (1993) propose a method for calculating the *TTR* of a section of a text, *TTRx*:

$$TTR_x = \frac{t_x + T - \frac{xT}{N}}{N}$$

where x is the length of the section of the text, t_x the number of types of a section of a text, T the total number of types of the text, N the length of the entire text. *TTRx* proved to be a very good index for intra-textual vocabulary richness comparisons with text blocks of different lengths with the same N; however, the present research uses *TTR* across the entire text samples with N vastly different, so the use of *TTRx* is not appropriate here. Therefore, the moving average *TTR* (*MATTR*) were used, which uses a smoothly moving window with certain length. The window moves from the beginning of a text towards the end of the text one step, i.e. one word, a time, with the length of the window arbitrarily set, computing the *TTR* within the window length at the same time. The *TTR*'s thus obtained are added and then divided by the number of steps the window has moved after the text is exhausted to get the moving average *TTR* of the text. This way the *TTR* obtained is standardized and is not affected by the length of the text. The moving average *TTR* is computed with the following formula (Kubát, Milička, 2013):

$$MATTR = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L)},$$

where L is the length of the window, V_i is the size of vocabulary within the window and N the length of the text concerned. In this research L was set at 100 words. Hereafter *TTR* and *MATTR* are used interchangeably.

3. Result and analysis

3.1 Textual *TTR*, *WL* and *SL*

The general statistics on the textual *TTR*, *WL* and *SL* of the 285 texts are shown in Table 1.

Table 1
General statistics of textual *TTR*, *WL* and *SL* of the 285 texts

	TTR	**WL**	**SL**
Mean	0.7302	4.3909	21.2422
Median	0.7315	4.4057	20.8050
Mode	0.6975	4.1623	22.7250
Std. Deviation	0.0245	0.2101	3.8669
Minimum	0.6546	3.6756	11.4820
Maximum	0.8228	4.9982	31.0000

The distributions of the *TTR*, *WL* and *SL* of the 285 texts are shown in figure 1.

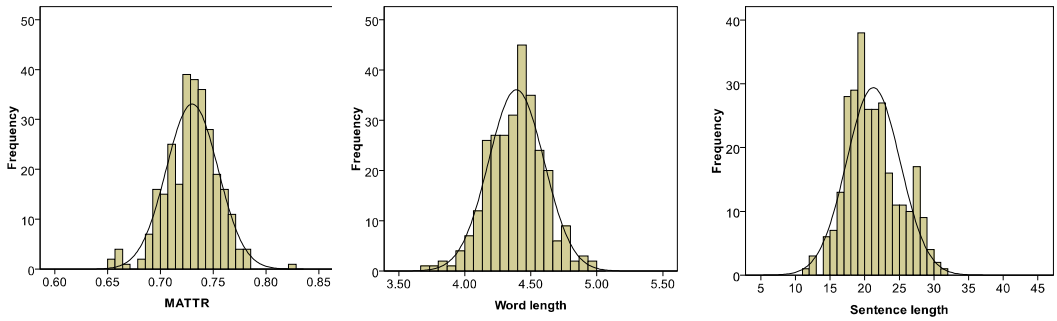


Fig. 1. Distribution of *TTR*, *WL* and *SL* of the 285 texts, with normal distribution curves.

As shown in the figure, the distributions of the *TTR*, *WL* and *SL* of the 285 texts are basically normal. They were then respectively ranked and the relationships between *TTR*, *WL* and *SL* and their respective ranks were examined and all displayed an *S*-shaped curve.

Nemcová Serdelová (2005) describe the relationship between the number of synonyms (*y*) of a word and the length of the word in syllables *x* with the following equation:

$$y = ax^b e^{cx} + 1,$$

which is a special case of Wimmer & Altmann (2005). It was modified to capture the relationship between *TTR*, *WL* and *SL* and their respective ranks, removing the addition of 1 and adding an exponent parameter *d*:

$$y = ax^b e^{cx^d},$$

where *y* is *TTR*, *WL* or *SL* while *x* is their respective ranks; *a*, *b*, *c* and *d* are parameters. The fit is good, as shown in Figure 2 and Table 2.

Table 2
Model parameters and R^2 of the rank distributions of the textual *TTR*, *WL* and *SL*

Parameters	TTR	WL	SL
a	0.8119	5.0933	31.8623
b	−0.0205	−0.0281	−0.0423
c	−1.00218867E-016	−1.83806339E-012	−0.0007
d	6.0633	4.3715	1.1465
R^2	0.9795	0.9812	0.9779

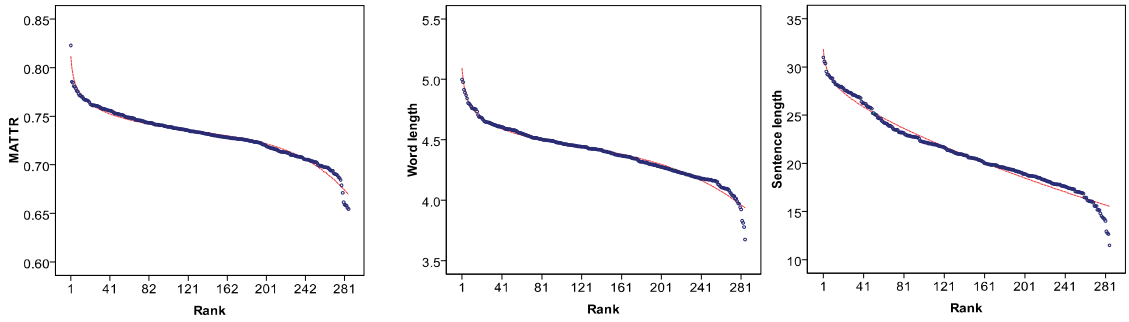


Figure 2. The relationship between *TTR*, *WL* and *SL* and their respective ranks and the model fit. The small circles are the observed values and the solid lines the model fit

Generally, the distributions of the quantity—rank relationship of any standardized or averaged textual measures such as the *TTR*, *WL* and *SL*, or quantities such as vocabulary size, number of sentences etc from texts of the same size, display *S*-shaped curves, with a sharp fall on either end caused by outliers, and can be captured by the modified Nemcová and Serdelová model.

3.2 The relationship between text length, *TTR*, *WL* and *SL*

A correlation test was performed to examine the relationships between text length (*TL*), *TTR*, *WL* and *SL*. The results are shown in Table 3.

Table 3
Correlations between *TL*, *SL*, *WL* and *TTR*

		TL	**SL**	**WL**	**TTR**
TL	Pearson Correlation	1	−0.158**	−0.063	−0.133*
	Sig. (2-tailed)		0.007	0.292	0.024
SL	Pearson Correlation	−0.158**	1	0.355**	0.025
	Sig. (2-tailed)	0.007		0.000	0.679
WL	Pearson Correlation	−0.063	0.355**	1	0.176**
	Sig. (2-tailed)	0.292	0.000		0.003
TTR	Pearson Correlation	−0.133*	0.025	0.176**	1
	Sig. (2-tailed)	0.024	0.679	0.003	

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

TL has a negative correlation with *SL* and *TTR*, significant at the 0.01 and 0.05 level respectively. This means the longer a text, the smaller the *SL* and *TTR* tend to be. *WL* has a positive correlation with *SL* and *TTR*, significant at the 0.01 level, suggesting that the longer the *WL*, the longer the *SL* and the larger the *TTR*.

3.3 Segmental *TTR*, *WL* and *SL*

The distributions of the segmental *TTR*, *WL* and *SL* are shown in figure 3. They are normally distributed. As with textual *TTR*, *WL* and *SL*, they were then respectively ranked and the relationships between the segmental *TTR*, *WL* and *SL* and their respective ranks are very similar to those of the textual *TTR*, *WL* and *SL*. The modified Nemcová and Serdelová model again provided very good fit to these relationships, as shown in Figure 4 and Table 4.

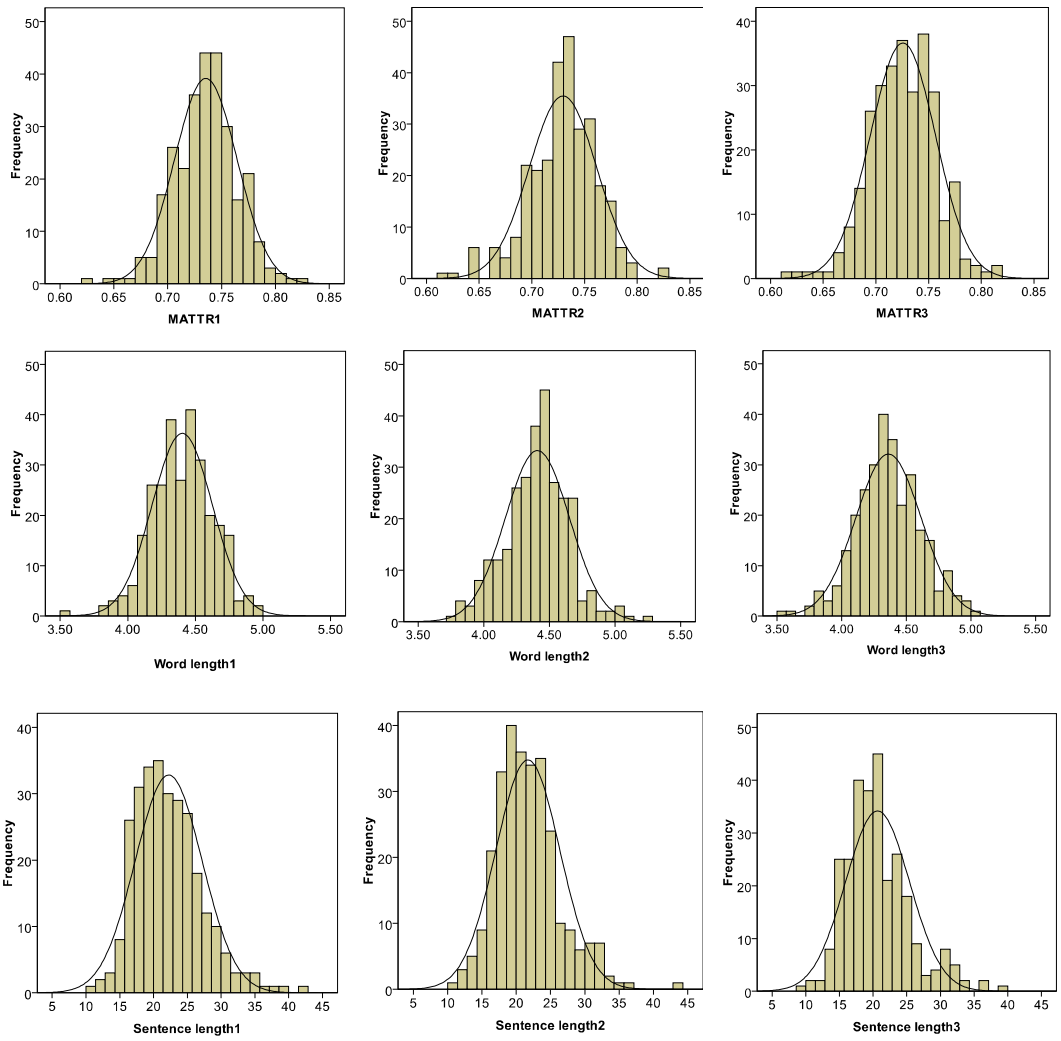


Fig. 3. The distributions of the segmental *TTR*, *WL* and *SL* with normal distribution curves

Table 4
Model parameters and R^2 of the rank distributions of the segmental *TTR*, *WL* and *SL*

	a	*b*	*c*	*d*	R^2
TTR_1	0.8321	−0.0236	−8.17415E-014	4.8926	0.9786
TTR_2	0.8357	−0.0263	−5.59484E-019	7.0461	0.9816
TTR_3	0.8276	−0.0248	−9.69297E-013	4.4681	0.9761
WL_1	5.1153	−0.0276	−3.57186E-011	3.8610	0.9737
WL_2	5.2718	−0.0347	−2.55565E-015	5.5693	0.9942
WL_3	5.2581	−0.0363	−3.59770E-014	5.0980	0.9770
SL_1	43.4237	−0.1256	−1.96670E-007	2.5597	0.9939
SL_2	43.0903	−0.1362	−2.08571E-011	4.1746	0.9854
SL_3	42.0611	−0.1377	−1.60964E-008	2.9976	0.9834

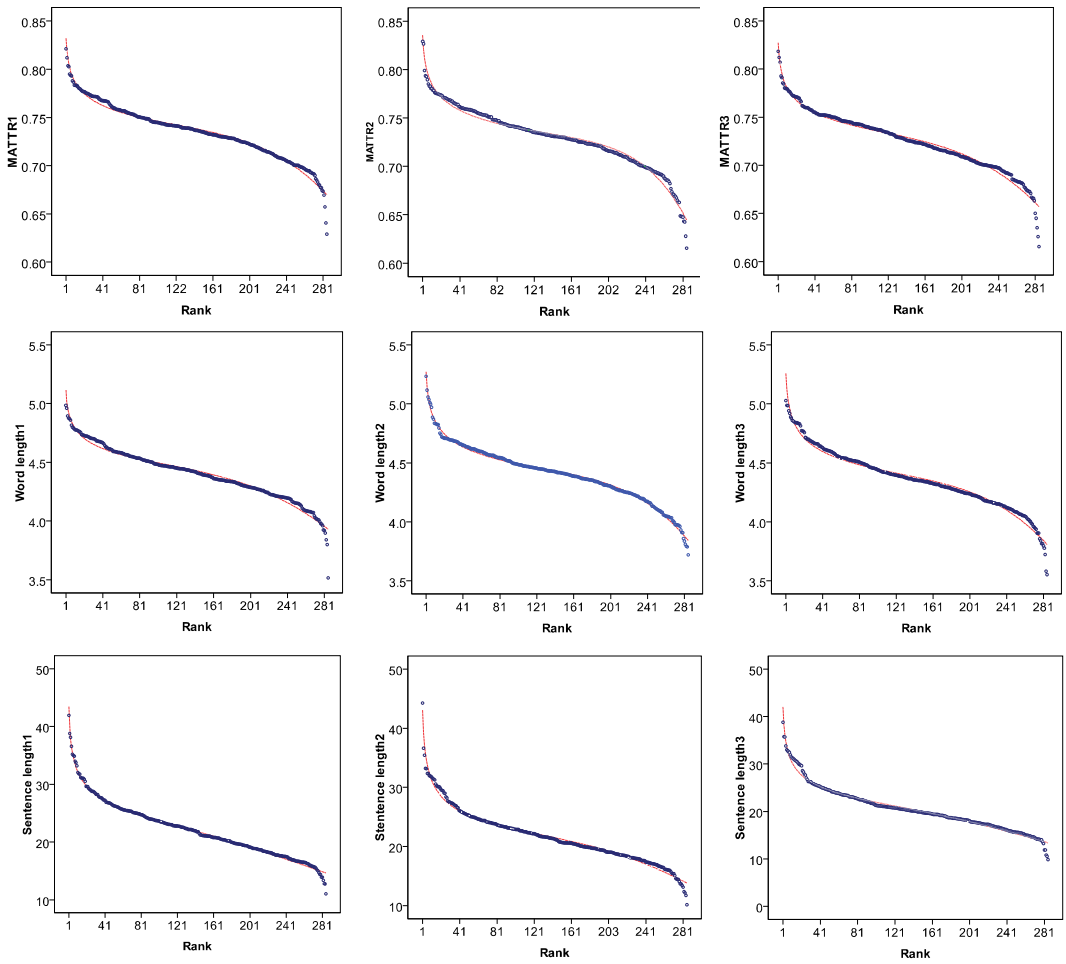


Figure 4. The relationship between segmental *TTR*, *WL* and *SL* and their respective ranks and the model fit. The small circles are the observed values and the solid lines the model fit

The general statistics on the segmental *TTR*, *WL* and *SL* of the 285 texts are shown in Tables 5 to 7. Note that the minimum and maximum values of the three measures may be smaller and larger respectively than those shown in Table 1 since those shown in Table 1 are textual mean values while the statistics shown below are segmental.

Table 5
Statistics of the segmental *TTR* of the 285 texts

	TTR_1	**TTR_2**	**TTR_3**
Mean	0.7354	0.7296	0.7256
Median	0.7366	0.7310	0.7257

Mode	0.7419	0.7232	0.7154
Std. Deviation	0.0290	0.0321	0.0311
Minimum	0.6289	0.6155	0.6156
Maximum	0.8211	0.8290	0.8184

Table 6
Statistics of segmental *WL* of the 285 texts

	WL_1	**WL_2**	**WL_3**
Mean	4.4030	4.4084	4.3612
Median	4.4086	4.4187	4.3529
Mode	4.2582	4.3852	3.9040
Std. Deviation	0.2237	0.2444	0.2528
Minimum	3.5154	3.7205	3.5534
Maximum	4.9840	5.2330	5.0269

Table 7
Statistics of the segmental *SL* of the 285 texts

	SL_1	**SL_2**	**SL_3**
Mean	22.2322	21.7012	20.6580
Median	21.8100	21.2550	20.0770
Mode	21.0000	19.0000	15.0000
Std. Deviation	4.9497	4.67090	4.7519
Minimum	11.0530	10.1150	9.8182
Maximum	41.9230	44.2500	38.8000

Tables 5 to 7 reveal very interesting phenomena. The mean, median and mode of TTR_1 and SL_1 are consistently larger than those of TTR_2, SL_2, TTR_3 and SL_3, and the mean, median and mode of TTR_2, and SL_2 are consistently larger than those of TTR_3 and SL_3. The mean, mode and median of WL_1 are all smaller than those of

WL_2 but larger than those of WL_3, while those of WL_2 are all larger than those of WL_3.

To see whether the differences between TTR_1, TTR_2 TTR_3; WL_1, WL_2 WL_3; and SL_1, SL_2 SL_1 are statistically significant, an ANOVA test was used to see whether the differences among the means of TTR_1—TTR_3, WL_1—WL_3 and SL_1—SL_3 are significantly different. The Levene test of homogeneity of variances was first performed on them. For TTR_1—TTR_3, WL_1—WL_3 and SL_1—SL_3, the Levene statistics were respectively 0.938, 1.114 and 0.959, with *p*-values respectively being 0.384, 0.392 and 0.329, showing their variances are homogeneous and the use of ANOVA test is appropriate. The ANOVA test result is shown in Table 8.

Table 8
ANOVA test result

		Sig.			Sig.			Sig.
TTR_1	TTR_2	0.062	WL_1	WL_2	0.788	SL_1	SL_2	0.186
	TTR_3	0.000		WL_3	0.039		SL_3	0.000
TTR_2	TTR_1	0.062	WL_2	WL_1	0.788	SL_2	SL_1	0.186
	TTR_3	0.272		WL_3	0.019		SL_3	0.010
TTR_3	TTR_1	0.000	WL_3	WL_1	0.039	SL_3	SL_1	0.000
	TTR_2	0.272		WL_2	0.019		SL_2	0.010

The differences between mean TTR_1 and TTR_2, TTR_2 and TTR_3, WL_1 and WL_2, SL_1 and SL_2 are not significant at the 0.05 level; the mean WL_2 and SL_2 are significantly larger respectively than WL_3 and SL_3. What is striking is that the mean TTR_1, WL_1 and SL_1 are all significantly larger at the 0.05 level than TTR_3 WL_3 and SL_3.

Table 9 shows the detailed information on the segmental *TTR*, *WL* and *SL* of the 285 texts.

Table 9
Segmental *TTR*, *WL* and *SL* comparison. *S1*: the first text segment, *S2*: the second text segment, *S3*: the third text segment

Segments	TTR	WL	SL
S1 > S2	169	143	154
S1 < S2	116	141	129
S1 = S2	0	1	2
S1 > S3	175	161	188
S1 < S3	110	122	97
S1 = S3	0	2	0

S2 > S3	154	161	171
S2 < S3	130	123	114
S2 = S3	1	1	0

Of all the segmental *TTR*, *WL* and *SL*, those between the first segments and the third are most noticeable; out of the 285 texts, the number of the first segments with *TTR*, *WL* and *SL* larger than those of the third segments are respectively 175, 161 and 188.

4. Conclusion and further work

The present research reveals three interesting phenomena. Firstly, the distribution of the segmental *TTR*, *WL* and *SL* and their respective ranks follow an *S*-shaped curve and can be described with the modified Nemcová and Serdelová model. Secondly, the longer the *WL*, the longer the *SL*. Thirdly, *TTR*, *WL* and *SL* of the first segment of a text all tend to be larger than those in the last segment of a text. The second phenomenon seems somewhat contradictive to Altmann's (1980) proposal that that the longer a language construct, the shorter its components; possible explanation of this contradiction might be that the word length and sentence length were measured respectively in number of letters and words in this research, instead of their immediate components—the syllable and clause. The third one is counter-intuitive since the initial section of a text is introductory while the middle part of a text is generally the key section in which the writer employs all possible devices to develop or expound on the first segment, entailing richer vocabulary, longer words and more complex sentences. However, this is not the case in the present research. Possibly such results might be due to the relatively small sample size (only 285 texts), narrow text source (11 magazines and newspapers) and fewer text sorts. The present research is only a pilot study, and research on a much larger scale on segmental *TTR*, *WL* and *SL* need to be carried out for the results to be generalized.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. In: *Glottometrika 2*, 1–10. Bochum: Brockmeyer.

Altmann, G. (1988). *Verteilungen der Satzlängen.* In: Schulz, K.-P. (ed.), *Glottometrika 9:* 147–169. Bochum: Brockmeyer.

Altmann, G., Popescu, I.-I., Zotta, D. (2013). Stratification in texts. *Glottometrics 25,* 85–93.

Biber, D., Johansson, S., Leech, D., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Fan, F. (2007). A Corpus based quantitative study on the change of TTR, word

length, and sentence length of the English language. In: Peter Grzybek and Reinhard Köhler (eds.). *Exact Methods in the Study of Language and Text:* 123–130. Berlin: Mouton de Gruyter.

Fan, F. (2012). A quantitative study on the lexical change of American English. *Journal of Quantitative Linguistics 20*(4), 288–300.

Fan, F., Grzybek, P., Altmann, G. (2010). Dynamics of word length in sentence. *Glottometrics* 20, 70–109.

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics* 16, 111–121.

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos. F., Papageorgiou, F. (2015) Word-length Entropies and Correlations of Natural Language Written Texts. *Journal of Quantitative Linguistics 22*(2), 101–118.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In: Lehfeldt, W.; Strauss, U. (eds.), *Glottometrika 4*, 103–113. Bochum: Brockmeyer.

Köhler, R. (2006a). The frequency distribution of the lengths of length sequences. In: Genzor, Josef; Bucková, Martina (eds.), *Favete linguis. Studies in honour of Viktor Krupa* : 145–152). Bratislava: Slovak Academic Press.

Köhler, R. (2006b). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, Reinhard, Sedlmeier, Peter, Zunker-Rapp, Gisela (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler:* 327–338. Lengerich: Pabst Science Publishers.

Köhler, R. (2008a). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, Sibyla (ed.), *Jazyk a jazykoveda v prohybe:* 416–421). Bratislava: VEDA vydavatel'stvo SAV.

Köhler, R. (2008b). Sequences of linguistic quantities. Report on a new unit of investigation. In: *Glottotheory* 1(1), 115–119.

Köhler, R. (2012). *Quantitative syntax analysis*. Walter de Gruyter GmbH & Co. KG: Berlin/Boston.

Köhler, R., Galle, M. (1993). Dynamic aspects of text characteristics. In: L. Hřebíček, G. Altmann (eds), *Quantitative Text Analysis:* 46–53. Trier: Wissenschaftlicher Verlag Trier.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using *L*-, *F*- and *T*-segments. In: Preisach, Burkhardt, Schmidt-Thieme, Decker (eds.), *Data Analysis, Machine Learning and Applications:* 637–646. Berlin, Heidelberg: Springer.

Köhler, R., Naumann, S. (2009). A contribution to quantitative studies on the sentence level." In: Köhler, Reinhard (ed.), *Issues in Quantitative Linguistics:* 34–57. Lüdenscheid: RAM-Verlag.

Köhler, R., Naumann, S. (2010). A syntagmatic approach to automatic text classification. Statistical properties of *F*- and *L*-motifs as text characteristics. In: Grzybek, Peter, Kelih, Emmerich, Mačutek, Ján (eds.), *Text and Language. Structures, Functions, Interrelations:* 81–90. Wien: Praesens Verlag.

Kubát, M., Milička, J. (2013) Vocabulary Richness Measure in Genres. *Journal*

of Quantitative Linguistics 20(4), 339–349.

Laufer, B., Nation, P. (1995) Vocabulary Size and Lexical Richness in L2 written Production. *Applied Linguistcs* 16(4), 307–322.

Levitsky, V., Melnyk, Y. (2011). Sentence length and sentence structure in English prose. *Glottometrics* 21, 14–24.

Mikros, G. (2007). Stylometric experiments in modern Greek: Investigating authorship in homogeneous newsier. In: Peter Grzybek, Reinhard Koehler, (eds). *Exact Methods in the Study of Language and Text:* 461-472. Berlin: Mouton de Gruyter.

Nemcová, E., Serdelová, K. (2005). On synonymy of Slovak. In: Altmann, G., Levickij, V. Perebyinis, V. (eds.), *Problems of Quantitative Linguistics:* 194–209. Chernivtsi: Ruta.

Popescu, I.-I., Altmann, G. (2014). The lambda structure of language levels. G*lottometrics 27,* 54–88

Popescu, I.-I., Mačutek, J., Altmann, G.. (2008). Word frequency and arc length. *Glottometrics 17*, 18–42.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman Group Limited: New York.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1*, 98–106.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (ed.), *Glottometrika 15, 112–133*. Trier: Wissenschaftlicher Verlag Trier.

Wimmer, G., Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.). *Contributions to the science of language: Word length and related issues:* 93–117. Boston: Kluver.

Statistical Analysis of the Diachronic Development of Terminal Rhyme in Chinese Poetry

Xiaxing Pan *

College of Chinese Language and Culture, Huaqiao University, Xiamen (361000), P. R. China, panxiaxing@hqu.edu.cn

Haitao Liu

School of International Studies, Zhejiang University, Hangzhou (310058), P. R. China, htliu@163.com

Abstract. Rhyming is always considered one of the most important phonaesthetics of Chinese poetry, especially the poetry of the Tang (618-907 AD) and Song (960-1279 AD) dynasties. However, poetry written in modern Chinese Mandarin tends to focus less on terminal rhyme. The present study investigates statistically the diachronic development of terminal rhyme in different kinds of Chinese poetry, including ancient poetry written in the Tang and Song dynasties, modern Chinese Mandarin poetry written since the1920s, modern Taiwanese poetry and Chinese translations of English poems. The results revealed that: 1) Terminal rhyme probability in three adjacent verse lines of ancient poetry is much higher than that in all the modern poetry 2) Though the probability of terminal rhyme in two adjacent verse lines of many Mandarin new poetry is lower than that of ancient poetry, there are still numbers of new Mandarin poetry preferring to the terminal rhyme 3) The rhyming patterns of Taiwanese poetry in two adjacent verse lines and three verse lines are very close to the new Mandarin poetry written and published in the1940s and the poetry created after 1978 4) The rhyming patterns of the translated poetry are similar to both the Taiwanese poetry and the new Mandarin poetry in 1940s as well as after 1978 5) It is reasonable for us to separate the short history of new Mandarin poetry into several periods.

Keywords: *Chinese poetry, terminal rhyme, rhyming of verse, statistical analysis*

1. Introduction

Poem or poetry is widely defined as composition highly relates to aesthetic and rhythmic qualities of language, e.g. meter, rhythm, phonaesthetics, etc. (Masters 1915: 308, Greene et al. 2012: 1046). Since Chinese is a tone language, both ancient Chinese poems[1] and modern Chinese poems[2] are not as much concerned

[1] These poems are mainly written from 618 A.D. to 1279 A.D and are defined as "Jinti Shi 近体诗" by Chinese poetic theorists

with syllables as English poetry is. Chinese poetry prominently relies on patterns of "level and oblique tones (平仄 píngzè)" as well as "terminal rhyme (押韵 yāyùn)" (Chen 1979: 372; Duanmu 2007: 275-277). Compared with the tones of ancient Chinese, the tones of Mandarin Chinese have changed a lot (Lin 2012: 109-110). Thus, the patterns of level and oblique tones, which are widely used in traditional Chinese poetry, are not inherited by new poetry at all. Fortunately, poetry of these two eras shares the other phonaesthetic feature – terminal rhyme – which makes it possible for us to quantitatively study the diachronic development of the metrical and rhythmical rules of Chinese poetry.

Generally, in Chinese poetry, if the vowels of the final syllables in two or several adjoining verse lines sound alike or similar to each other, they form a terminal rhyme, which is one of the most important phonetic features, and plays a dominant role in poetry rhythm. With the aid of terminal rhyme, poets could control the poetry rhythm easier, and create more enticing works (Qi 2009: 173; Kao, Jurafsky 2012: 10). However, nowadays, poetry writing is becoming less dependent on terminal rhyme. For instance, Kao, Jurafsky (2012) compare a collection of poetry texts written by both American amateur and professional poets. The study concludes that the frequency of terminal rhyme in professional poetry is much lower than that in amateur poetry. Voigt ,Jurafsky (2013) study terminal rhyme in Chinese poetry, including ancient poetry, Mandarin new poetry, and modern Taiwanese poetry. Through empirical evidence, their study indicates that, compared with terminal-rhyme ancient poetry, new poetry relies significantly less on terminal rhyme.

The frequency and the probability of the terminal rhyme can reveal how much the poets are concerned about the metrical and rhythmical beauties as well as the dominant genre of poetry in an era, i.e., free verse or rhythmic accentuated poetry. It is definitely true that a large number of new poems do not care much about the phonaesthetics like terminal rhyme, which is significantly different from ancient poetry (Voigt, Jurafsky 2013). However, in the short history of Mandarin new poetry in mainland China, debates over free verse and rhythmic accentuated poetry continue (Wang 2005, Sun 2010). For instance, in the very beginning of Mandarin new poetry (1910s-1920s), most of the poets preferred free verses. But in the late 1920s, rhythmic accentuated poems were more popular (Hong 2010a). Therefore, it would be worthwhile to divide the short history of new poetry into several periods to study the genre alternations.

Voigt, Jurafsky (2013) have quantitatively described the main phonaesthetic features of Mandarin new poetry. However, there are some shortcomings from the perspective of linguistics and literature:

1. The texts used in their study are not very convincing because some of the texts are not accepted as poetry for poets and poetic theorists.

[2] It is named as "XinShi 新诗" in China, and we'll use 'new poetry' for short below. New poetry contains Mandarin new poetry, Taiwanese poetry (Taiwan Shi 台湾诗) and Mandarin translated poetry (Yi Shi 译诗).

2. This study relies on automatic analysis too much and hardly pays any attention to linguistics and literary theory.
3. Since the debates over free verse and rhythmic accentuated poetry have never stopped, we should divide the texts into groups to discuss which the study did not.
4. The definition of "terminal rhyme" in this study is "we therefore qualify a given line as 'rhyming' if the last character of any line within a 3-line window shares its vowel final pronunciation, and for each poem calculate the proportion of rhyming lines" (Voigt, Jurafsky 2013: 3). This definition is not very accurate since the terminal rhyme of Chinese poetry is more complex (we will specify more below).

In view of these problems, our study aims to analyze the diachronic properties of terminal rhyme in Chinese poetry. The present study is arranged as follows:

1. We suppose that we have to select the poetry texts carefully, excluding those texts which are not widely accepted as "poetry". The first part of the study introduces the language materials and the approaches to choose the materials.
2. Terminal rhyme in Chinese poetry, especially in ancient poetry, has at least two types. The first one is terminal rhyme in two adjacent verse lines, and the second one is terminal rhyme in two verse lines which are not adjacent. The second part of the study introduces the types of Chinese vowels and the rhyming system of Chinese poetry.
3. As mentioned earlier, there is an important difference between terminal rhyme use in ancient poetry and in new poetry. The third section empirically investigates the differences with "SPSS 17.0".
4. The last part of the study discusses and interprets the language data linguistically and literarily, and describes the diachronic development of terminal rhyme in Chinese poetry.

2. Poetry texts and selection

Four poetry corpora (JT, TW, YS, XS) are built. The ancient poetry corpus coded as "JT" (an abbreviation of "Jinti Shi") consists of ancient poetry written in the Tang and Song dynasties. The Taiwanese poetry corpus coded as "TW" (an abbreviation of "Taiwan Shi") consists of Taiwanese modern poetry (mainly after 1949). The translated poetry corpus coded as "YS" (an abbreviation of "YiShi") includes Chinese translations of poetry that originally came from other languages. The Mandarin new poetry corpus coded as "XS" (an abbreviation of "XinShi") consists of new poetry written and published in mainland China. JT, TW, YS each has 100 texts while XS has 480 texts. All the texts in the four corpora are extracted from the internet. The ancient poetry texts are extracted from http://www.gushiwen.org/, while the others are from http://www.shigeku.org/.

Ancient poetry, Taiwanese poetry and translated poetry are selected randomly. However, since the automatic selection of Mandarin new poetry is not

very accurate, we select new poetry in two steps. In the first step, we extract a poetry text from http://www.shigeku.org/ randomly. In the second step, we manually checked whether the selected texts are "poetry" or not based on McDougall (1994), Hong (2010b), etc. If so, we keep it in the corpus. If not, we go back to step 1 and extract another piece of text to replace it.

The debates over free verse and rhythmic accentuated poetry continue to occur. Poets and poetic theorists often divide the history of Mandarin new poetry into some eras. We group the poetry texts into eight sub-corpora – 1917-1922 (vernacular poetry, with the corpus code as XS-1), 1923-1930 (1920s poetry, code as XS-2), 1931-1942 (1930s poetry, code as XS-3), 1943-1949 (1940s poetry, code as XS-4), 1950-1977 (Maoist time poetry, code as XS-5), 1978-1989 (1980s poetry, code as XS-6), 1990-1999 (1990s poetry, code as XS-7), 2000-2010 (new century poetry, code as XS-8). Every sub-corpus contains 60 texts. Basic information (including word tokens, number of texts, and average text length) on the corpora is listed in Table 1.

Table 1
Basic information on the corpora

ID	Tokens	Pieces	Average text length (words)
JT	2855.00	100	28.55
SW	58901.00	100	589.01
TW	10304.00	100	103.04
YS	13527.00	100	135.27
XS-1	7886.00	60	131.43
XS-2	10992.00	60	183.20
XS-3	9557.00	60	159.28
XS-4	8100.00	60	135.00
XS-5	11422.00	60	190.37
XS-6	9537.00	60	158.95
XS-7	14251.00	60	237.52
XS-8	7051.00	60	117.52

3. The terminal rhyme phonaesthetics in Chinese poetry and rhyming pattern extraction methods

In Chinese, every single syllable consists of one initial (声母 shēngmǔ), one final (韵母 yùnmǔ) and one tone (声调, shēngdiào) (Chao 2011: 47-79, Lee, Zee 2014: 369-399). Terminal rhyme is matter of finals, for only the finals have vowels. Though finals have different forms, they all share the same basic construction

shown in (1):

(1) final = medial + head vowel + ending.

Every final has a "head vowel (韵腹 yùnfù)", which is the main vowel of the whole syllable. But the "medial" and "ending" are not necessary for a final (Chao, 2011: 47). For instance, in the syllable "shū (/ʃu1/书, book)[3]", the vowel "u" is the main vowel and the final of the syllable at the same time. In the syllable "xià (/ɕia4/下, down)", besides the main vowel "a", there is another vowel "i" between "x" and "a", which is defined as "medial (韵头 yùntóu)". In the syllable "hēi (/hei1/黑, black)", the main vowel "e" is followed by another vowel "i", which is defined as "ending (韵尾 yùnwěi)". Apart from vowels, some of the consonants like "n" and "ng" could be the "ending" of a final. Accordingly, all the Chinese finals may be clustered into three main groups:

Finals consisting of head vowels only, e.g. a (他, tā /tʰa1/, he), o (伯, bó /po2/, uncle), e (歌, gē /kʌ1/, song), i (鸡, jī /dzi1/, chicken), (字, zì /dz✶4/, script), (汁, zhī /dzɹ1/, juice)[4], u (父, fù /fu4/, dad), ü (绿, lù /lʏ4/, green), and the retroflex suffixation (儿, ér /aɹ2/, son).

1. Finals consisting of more than one vowel, which can be further divided into three sub-groups.

Medial + Head vowel: ia (下), ie (切, qiē /tɕi.ɛ1/, cut), ua (花, huā /hwa1/, flower), uo (国, guó /gu卐2, country), üe (学, xué /②u☜:2, learn).

   a. Head vowel + Ending: ai (开, kāi /kʰai1/, open), ei (黑), ao (刀, dāo /dau1/, knife), ou (豆, dòu /do4/, bean).

   b. Medial + Head vowel + Ending: iao (小, xiǎo /ɕiau3/, tiny), iou (休, xiū

[3] The Chinese phoneme forms in the paper follow IPA-S (Warren 1994: 5-15). The number following the phonemes refers to the value of tone.

[4] The letter "i" in Chinese *pinyin* records three phonemes, /i/ (鸡), /✶/ (字), /ɹ/ (汁).

/ɕiu:1/, rest), uai (帅, shuài /♠wai4/, handsome), uei (追, zhuī /dC·u.i:1/, chase).

2. Finals consisting of not only vowels, but also consonants, such as the nasals "n" and "ng": an (蛋, dàn /dan4/, egg), ian (先, xiān /ɕi. ☜n1/, earlier), uan (玩, wán /wan2/, play), üan (犬, quǎn /tɕyan3/, dog), en (分, fēn /fΛn1/, score), in (近, jìn /ㄖd⇔n4/, close), uen (困, kùn /kw☩n4/, sleepy), ün (军, jūn /d⇔n1/, army), ang (房, fang /faŋ2/, house), iang (枪, qiāng /tɕaŋ1/, gun), uang (光, guāng /gwaŋ1/, light), eng (风, fēng /fΛŋ1/, wind), ing (请, qǐng /tɕiŋ3/, please), ueng (瓮, wēng /wΛŋ1/, urn), ong (冬, dōng /d☩ŋ1/, winter), iong (兄, xiōng /ɕjŋ/, brother).

Details of these three groups are listed in Table 2, and Table 3 presents all the types of finals that we found in the selected poems.

Table 2
The three main groups of Chinese finals

Head vowel only		a, o, e, i, u, ü, er
Multi-vowel finals	Medial + Head vowel	ia, ie, ua, uo, üe
	Head vowel + Ending	ai, ei, ao, ou
	Medial + Head vowel + Ending	iao, iou, uai, uei
Vowels and consonants		an, ian, uan, üan, en, in, uen, ün, ang, iang, uang, eng, ing, ueng, ong, iong

Table 3
Types of final extracted from all the poems

a，ai，an，ang，ao，e，ei，en，eng，er，i，ia，ian，iang，iao，ie，in，ing，iong，iu，o，ong，ou，u，ua，uai，uan，uang，ue，ui，un，uo，v

In Table 3, it should be noted that 1) the single letter "i" stands for the three finals --i/i/ (鸡), -i/✶/ (字) and the other –i/ɹ/ (汁); 2) the letters "eng" stand for the two finals --eng and ueng; 3) the single letter "u" stands for two finals actually--u/u/ (父) and ü/y/ (句, jù /dzy4/, sentence), (去, qù /t②y4/, go), (许, xǔ /②y:3/, permit); 4) the letters "ue" stand for the final üe /yɛ/ (学); 5) The letters "uan" stand for two finals –uan /wan/ (玩) and üan /yan/ (犬); 6) The letters "ui" stand for the final uei /wei/ (追); 7) The letters "un" stand for two finals -- un/wʌn/ (困) and ün /yɛn/ (军); 8) The letter "v" stands for the final "ü" following the initials "n" or "l", as "女 (nǚ /ny3/, daughter)" and "绿" for example. Consequently, the real number of types of finals in Table 3 should be 39.

Actually, in Chinese rhyming verses, finals in the last syllables are not necessarily to be identical. Traditionally, finals are grouped into "shíbāyùn (十八韵, eighteen rhymes)" or "shísānzhé (十三辙, thirteen rhymes)" (Hu 1995: 64-67, Huang & Liao 2002: 68-69). The present study is mainly based on "shísānzhé" (see Table 4), which defines that if finals in two or three adjacent verse lines sound similar or identical, there is a terminal rhyme[5]. For instances, in the poetry "静夜思 (jìngyèsī, *A Tranquil Night*)" written by Li Bai[6],

a. 床前明月光 (guang)

b. 疑是地上霜 (shuang)

c. 举头望明月 (yue)

d. 低头思故乡 (xiang)

[5] In the present study, we are not going to investigate the finals of the last syllables in four or more adjacent verse lines for two reasons: 1) If two similar finals are separated by at least two verse lines, it would be difficult to define a consonance between them. 2) The more verse lines we take into consideration at once, the more complex the rhyming should be, and then the more difficult and complex the statistical analysis would be.

[6] Li Bai (701-762), a very famous Chinese poet in Tang Dynasty. "静夜思" is describing the poet's homesickness in a moon-night. Here is one version of the English translation of the poem: "Abed, I see a silver light, // I wonder if it's frost aground. // Looking up, I find the moon bright; // Bowing, in homesickness I'm drowned."

Both of the finals in the last syllables of line “a.” and line “b.” are “uang”, which form a terminal rhyme in two adjacent verse lines. Besides, according to the “shísānzhé” in Table 4, the final “iang” in the last syllable of line “d.” can form a rhyme with the final “uang” of the last syllable of line “b.”, which could be defined as a terminal rhyme in three adjacent verse lines.

Table 4
“Shísānzhé” of Chinese finals#

Shísānzhé	Finals	eg.
1	a, ia, ua	他，下，花
2	o, uo, e	伯，国，歌
3	ie, üe	切，学
4	i, er, ü	鸡，字，汁，儿，绿
5	u	父
6	ai, uai	开，帅
7	ei, uei	黑，追
8	ao, iao	刀，小
9	ou, iou	豆，休
10	an, ian, uan, üan	蛋，先，玩，犬
11	en, in, uen, ün	分，近，困，军
12	ang, iang, uang	房，枪，光
13	eng, ing, ueng, ong, iong	风，请，瓮，冬，兄

3. Results

Fig. 1 and Fig. 2 are respectively the histograms of the terminal rhyme probabilities in two and three adjacent verse lines.

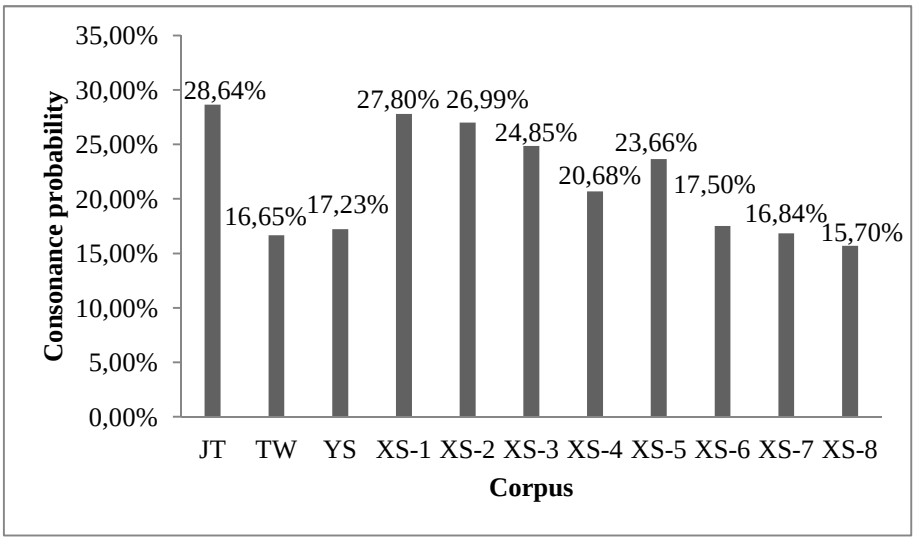


Figure 1. Histogram of the terminal rhyme probabilities in two adjacent verse lines

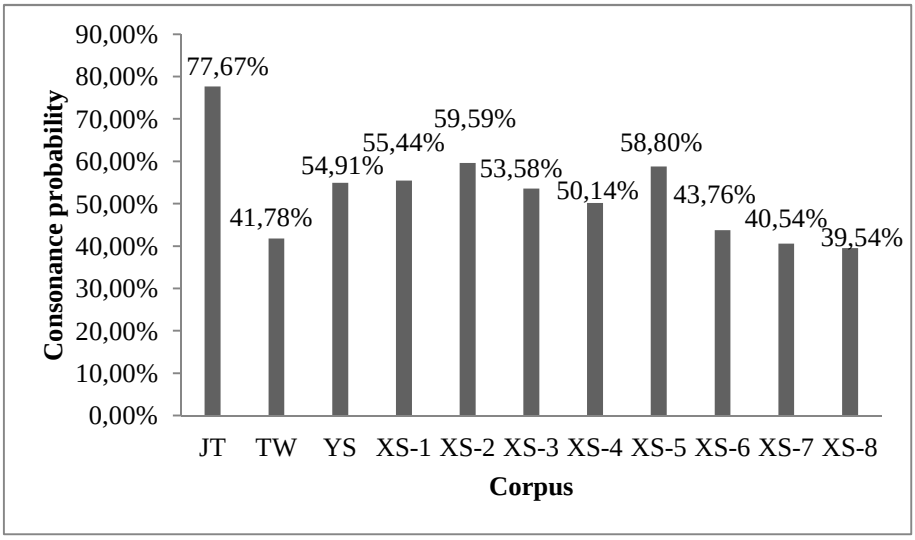


Figure 2. Histogram of the terminal rhyme probabilities in three adjacent verse lines

The results show that poems in different corpora exhibit two different kinds of rhyming patterns. For instance, Fig. 1 shows that the probabilities in corpora JT, XS-1, XS-2, XS-3, XS-4 and XS-5 are higher than 20%, while the probabilities in other corpora are lower than 20%. On the other hand, Fig.2 shows that the terminal rhyme probabilities in three verse lines of the poems in JT are higher than that in other poems, which confirms the great importance of terminal rhyme in ancients poetry. YS, XS-1, XS-2, XS-3, XS-4 and XS-5 are between 50% and

60%, while the probabilities in other corpora are lower than 50%.

The two figures have a number of common characteristics. 1) Both of the histograms reveal that the terminal rhyme probabilities in traditional poems are the highest. 2) The terminal rhyme probabilities of Taiwanese poems are as low as the poems in XS-6, XS-7 and XS-8. 3) The terminal rhyme probabilities of translated poems in two adjacent verse lines are also as low as the poems in TW, XS-6, XS-7 and XS-8, but the probabilities in three adjacent verse lines are much higher than these four corpora, which may mean that translated poems prefer rhyme in three adjacent lines. 4) In the new poetry corpora, poems in XS-1, XS-2 and XS-5 employ terminal rhyme more frequently than the others. 5) All the new poetry texts could be clustered into two groups according to the probabilities. The first group includes the corpora XS-1, XS-2, XS-3 and XS-4, in which the probabilities decrease gradually, but increase suddenly in XS-5. Thus, the other corpora XS-5, XS-6, XS-7 and XS-8 are clustered into the second group.

Fig. 1 and Fig. 2 confirm that different poems attach to the terminal rhyme with various degrees. However, the probabilities in some corpora also have something in common. We need to clarify that which corpora are significantly different in applying terminal rhyme, and which corpora are rather similar. For this we used an ANOVA test in "SPSS". Texts in every corpus are divided into five groups randomly, with 20 poetry texts per group in the three corpora JT, TW and YS, and 12 poetry texts per group in the eight new poetry sub-corpora. The H_0 is:

> *There is no difference in terminal rhyme probabilities in two and three adjacent verse lines among the selected poems*.

Table 5 and Table 6 list the mean terminal rhyme probabilities (%) in two and three adjacent verse lines of the five text groups.

Table 4
The mean terminal rhyme probabilities (%) in two adjacent verse lines of the five-text groups

	JT	**TW**	**YS**	**XS-1**	**XS-2**	**XS-3**	**XS-4**	**XS-5**	**XS-6**	**XS-7**	**XS-8**
1	25.00	20.90	16.46	23.78	31.55	25.54	29.00	21.18	15.71	16.81	18.23
2	33.33	19.19	13.74	23.19	32.27	30.38	17.95	25.23	18.85	17.15	14.92
3	26.32	15.16	17.68	34.36	22.27	33.73	21.03	28.25	12.88	18.23	14.29
4	34.18	14.74	20.48	28.86	21.88	18.15	15.60	16.00	20.57	11.54	18.23
5	24.39	13.28	17.78	28.80	26.96	16.43	19.80	27.63	19.50	20.45	12.82

Table 5
The mean terminal rhyme probabilities (%) in three adjacent verse lines of the five-text groups

	JT	**TW**	**YS**	**XS-1**	**XS-2**	**XS-3**	**XS-4**	**XS-5**	**XS-6**	**XS-7**	**XS-8**
1	71.05	46.09	54.19	51.27	67.45	49.25	63.47	73.60	42.17	40.37	45.03
2	81.82	47.41	49.59	50.95	64.44	64.92	54.10	53.47	44.58	41.69	36.09
3	73.21	37.95	57.39	56.29	52.88	64.20	48.87	63.51	39.84	41.78	38.54
4	86.44	32.53	54.58	59.40	54.84	46.43	38.35	50.41	48.86	31.19	42.01
5	75.81	44.92	58.82	59.30	58.33	43.08	45.95	52.99	43.33	47.65	36.04

4.1. ANOVA test in two adjacent verse lines

Table 7 lists the result of the ANOVA of the mean probabilities in two adjacent verse lines. The value of F is 6.031, df_1 is 10, df_2 is 44, and $p<0.001$, which means that there is a significant difference in the probabilities among the texts. So a following post hoc pairwise comparison is made to clarify which groups are significantly different and which groups are not.

Table 6
ANOVA test for the mean terminal rhyme probabilities in two adjacent verse lines

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	1252.316	10	125.232	6.301	.000
Within Groups	874.431	44	19.873		
Total	2126.746	54			

Table 8
The mean terminal rhyme probabilities in two adjacent verse lines

	N	Mean		N	Mean
JT	5	28.644	XS-4	5	20.676
TW	5	16.654	XS-5	5	23.658
YS	5	17.228	XS-6	5	17.502
XS-1	5	27.798	XS-7	5	16.836

XS-2	5	26.986	XS-8	5	15.698
XS-3	5	24.846			

Table 7
Homogeneity of variance test result of mean terminal rhyme probabilities in two adjacent verse lines

Levene Statistic	df_1	df_2	Sig.
1.707	10	44	.109

Table 8
Intergroup linear polynomial contrasts of the mean terminal rhyme probabilities in two adjacent verse lines

	Contrast	Value of Contrast	Std. Error	t	*df*	Sig. (2-tailed)
Assume equal variances	1	-78.5580	20.9097	-3.757	44	.001
	2	-55.1520	16.9168	-3.260	44	.002
	3	40.7680	16.9168	2.410	44	.020
	4	36.1760	16.9168	2.138	44	.038
Does not assume equal variances	1	-78.5580	22.0491	-3.563	4.728	.018
	2	-55.1520	17.9443	-3.074	5.055	.027
	3	40.7680	13.0207	3.131	6.362	.019
	4	36.1760	10.5954	3.414	8.330	.009

According to Table 8, traditional poems have the highest mean terminal rhyme probabilities (28.64%), while poems in the 21st century (XS-8) have the lowest (15.70%). The homogeneity of variance test in Table 9 (with F=1.707, Sig. = 0.109, $p > 0.05$) shows that there are no significant differences among the five text groups in every corpus. So, the intergroup linear polynomial contrasts result in the column of "Assume equal variances" in Table 10 is accepted. In the table 10, we present several comparisons: the mean terminal rhyme probabilities in two adjacent verse lines of the ancient poetry comparing to the new poetry (numbered as comparison "1"), the ancient poetry comparing to the Mandarin new poetry (numbered as comparison "2"), of the comparison of Taiwanese poetry and the Mandarin new poetry (numbered as comparison "3"), and the comparison of translated poetry and the Mandarin new poetry (numbered as comparison "4").

Our H_0 hypothesis is that

There is no significant difference between every two groups.

According to Table 10, the significant coefficients of the contrasts are 0.001, 0.002, 0.020, and 0.038 ($p < 0.05$), which means that the hypothesis has to be rejected, and there are significant differences among the groups. This intergroup contrast reveals at least two points:

1. The status of the terminal rhyme in two adjacent verse lines is quite different from that in the ancient poetry and all the modern poetry.
2. In all the new poetry, the status of the terminal rhyme in two adjacent verse lines is quite different among the Mandarin new poetry, Taiwanese poetry, and translated poetry.

Based on the results in Table 10, we further hypothesize that there would be some significant inner group differences among the ancient poetry and all the modern poetry in different ages. However, the homogeneity of variance test results in Table 10 point out that variances of the mean values among the corpora are constant, which means that the new poetry texts in some ages may share similar rhyming property with the ancient poetry texts. Accordingly, a further LSD test in Table 11 is made with “SPSS”.

Table 9
LSD test of the mean terminal rhyme probabilities in two adjacent verse lines of the selected corpora

	JT	TW	YS	XS-1	XS-2	XS-3	XS-4	XS-5	XS-6	XS-7	XS-8
JT	-	0.000	0.000	0.766	0.560	0.185	0.007	0.084	0.000	0.000	0.000
TW	0.000	-	0.840	0.000	0.001	0.006	0.161	0.017	0.765	0.949	0.736
YS	0.000	0.840	-	0.001	0.001	0.010	0.228	0.027	0.923	0.890	0.590
XS-1	0.766	0.000	0.010	-	0.775	0.301	0.015	0.149	0.001	0.000	0.000
XS-2	0.560	0.001	0.001	0.775	-	0.452	0.030	0.244	0.002	0.001	0.000
XS-3	0.185	0.006	0.010	0.301	0.452	-	0.146	0.676	0.012	0.007	0.002
XS-4	0.007	0.161	0.228	0.015	0.030	0.146	-	0.296	0.266	0.180	0.084
XS-5	0.084	0.017	0.027	0.149	0.244	0.676	0.296	-	0.034	0.020	0.007
XS-6	0.000	0.765	0.923	0.001	0.002	0.012	0.266	0.034	-	0.814	0.526
XS-7	0.000	0.949	0.890	0.000	0.001	0.007	0.180	0.020	0.814	-	0.688
XS-8	0.000	0.736	0.590	0.000	0.000	0.002	0.084	0.007	0.526	0.688	-

The LSD test results indicate that:

1. There is no significant difference in terminal rhyme probabilities in two adjacent verse lines among the ancient poems, the Mandarin new poems before the 1940s (XS-1, XS-2, XS-3) and the Maoist time poems (XS-5), with the p-values being 0.766, 0.560, 0.185 and 0.084 ($p > 0.05$) respectively.
2. Taiwanese poems (TW), translated poems (YS), poems in the 1940s (XS-4), and all the poems since the 1980s (XS-6, XS-7, XS-8) share other kind of common rhyming patterns.

4.2. ANOVA test in three adjacent verse lines

Table 12 lists the result of the ANOVA test of the mean probabilities in three adjacent verse lines. The values of F, df_1, and df_2 are 13.786, 10, and 44, while $p<0.001$. There is a significant difference in the probabilities among the texts. A post hoc pairwise comparison is conducted to distinguish the groups that are significantly different from others.

Table 10
ANOVA of the mean terminal rhyme probabilities in three adjacent verse

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	6232.610	10	623.261	13.786	.000
Within Groups	1989.216	44	45.209		
Total	8221.827	54			

Table 11
The mean terminal rhyme probabilities in three adjacent verse lines

	N	Mean		N	Mean
JT	5	77.666	XS-4	5	50.148
TW	5	41.780	XS-5	5	58.796
YS	5	54.914	XS-6	5	43.756
XS-1	5	55.442	XS-7	5	40.536
XS-2	5	59.588	XS-8	5	39.542
XS-3	5	53.576			

Table 12
Homogeneity of variance test result of mean terminal rhyme probabilities in three adjacent verse lines

Levene Statistic	df_1	df_2	Sig.
2.428	10	44	.021

Table 13
Intergroup linear polynomial contrasts of the mean terminal rhyme probabilities in three adjacent verse lines

	Contrast	Value of Contrast	Std. Error	t	*df*	Sig. (2-tailed)
Assumes equal variances	1	-278.5820	31.53741	-8.833	44	.000
	2	-219.9440	25.51502	-8.620	44	.000
	3	67.1440	25.51502	2.632	44	.012
	4	-37.9280	25.51502	-1.486	44	.144
Does not assume equal variances	1	-278.5820	29.96581	-9.297	4.948	.000
	2	-219.9440	24.43468	-9.001	5.327	.000
	3	67.1440	24.37260	2.755	5.335	.037
	4	-37.9280	15.55753	-2.438	8.610	.039

According to Table 13, the traditional poems have the highest mean probabilities (77.67%), meanwhile the poems in the 21st century (XS-8) have the lowest (39.54%). This result is very similar to the results presented in 3.1.

The variance homogeneity test in Table 14 (with $F = 2.428$, Sig. = 0.021, $p < 0.05$) reveals that there are significant differences among the groups. So the intergroup linear polynomial contrasts result in the column of "Does not assume equal variances" in Table 15 is accepted. In the table 15, several comparison results are presented: the mean terminal rhyme probabilities in two adjacent verse lines of the ancient poetry comparing to that of all the new poetry (numbered as comparison "1"), the ancient poetry comparing to the Mandarin new poetry (numbered as comparison "2"), the comparison of Taiwanese poetry and the Mandarin new poetry (numbered as comparison "3"), and the comparison of translated poetry and the Mandarin new poetry (numbered as comparison "4"). The H_0 hypotheses of the four contrasts are:

The means of the two groups are the same.

According to the data presented in Table 15, the significance coefficients of the contrasts are 0.000, 0.000, 0.037, and 0.039 ($p < 0.05$), which means that the hypothesis has to be rejected, and that there are significant differences among the groups. This intergroup contrast reveals that:

1. In ancient poetry and in new poetry, the status of terminal rhyme in three adjacent verse lines is quite different.
2. For all the new poetry, the status of the terminal rhyme in three adjacent verse lines is quite different among different groups.

Based on the results in Table 15, we further hypothesized that there would be some significant intra-group differences between ancient poetry and other poetry. However, the variance homogeneity test in Table 14 reveals that variances of the mean values among the corpora are constant, which means that maybe Mandarin new poetry texts in some eras share the same rhyming property with the ancient poetry texts. Hence, a further LSD test is made in Table 16.

Table 14
LSD test of the mean terminal rhyme probabilities in three adjacent verse lines of the selected corpora

	JT	TW	YS	XS-1	XS-2	XS-3	XS-4	XS-5	XS-6	XS-7	XS-8
JT	-	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TW	0.000	-	0.003	0.002	0.000	0.008	0.055	0.000	0.644	0.771	0.601
YS	0.000	0.003	-	0.902	0.278	0.755	0.268	0.366	0.012	0.002	0.001
XS-1	0.000	0.002	0.902	-	0.335	0.663	0.220	0.435	0.009	0.001	0.001
XS-2	0.000	0.278	0.335	0.164	-	0.032	0.853	0.001	0.000	0.000	0.000
XS-3	0.000	0.008	0.755	0.663	0.164	-	0.425	0.226	0.026	0.004	0.002
XS-4	0.000	0.055	0.268	0.220	0.032	0.425	-	0.048	0.140	0.029	0.016
XS-5	0.000	0.000	0.366	0.435	0.853	0.226	0.048	-	0.001	0.000	0.000
XS-6	0.000	0.644	0.012	0.009	0.001	0.026	0.140	0.001	-	0.453	0.327
XS-7	0.000	0.771	0.002	0.001	0.000	0.004	0.029	0.000	0.453	-	0.816
XS-8	0.000	0.601	0.001	0.001	0.000	0.002	0.016	0.000	0.327	0.816	-

The LSD test results as shown in Table 16 indicate that:

1. There are significant differences in terminal rhyme probabilities in three verse lines between ancient poetry and all the new poetry, with $p < 0.001$.
2. Compared with the results in Table 11, the difference between TW and YS is quite significant ($p<0.05$), which may mean that the terminal rhyme probabilities in three adjacent verse lines may be the most significant difference between Taiwanese poetry and translated poetry.
3. The rhyming properties of the poems before the 1980s are very similar to the translated poems, with the values of p being 0.092, 0.278, 0.755, 0.268 and 0.366 ($p > 0.05$) respectively.
4. Poems in the 1940s (XS-4) have similar rhyming properties to the poems in XS-1 and XS-3 (p being0.220 and 0.425 respectively, $p > 0.05$). Meanwhile,

their differences from the poems in XS-5 and XS-6 are not quite significant as well (*p* being 0.048 and 0.140 respectively). It can be inferred that the1940s may be a turning point for the development of Mandarin new poetry, especially the development of poetic rhyming.

5. Discussion

In section 4.2, we investigated the terminal rhyme probabilities within three adjacent verse lines of ancient poetry, Taiwanese poetry, translated poetry and Mandarin new poetry. With the data in Table 16, we can confirm the conclusion of Voigt, Jurafsky (2013) that the terminal rhyme probability in new poetry is considerably lower than the probability in ancient poetry. Meanwhile, according to Table 11, it is very clear that ancient poetry, Mandarin new poetry before the 1940s, and poetry in the Maoist period share almost the same rhyming patterns. The conclusion of Voigt, Jurafsky (2013) could therefore be revised:

1. Terminal rhyme probability in three adjacent verse lines of ancient poetry is much higher than that of all the new poetry. This result is consistent with Voigt, Jurafsky's (2013) observation that cross-sentence rhyming greatly decreases in new poetry. So the main difference in rhyming system between ancient poetry and new poetry mainly lies in rhyming on three adjacent verse lines.
2. Though the terminal rhyme probability in two adjacent verse lines of the new poetry is lower than that of the ancient poetry, in a certain period, the terminal rhyme probabilities of the new poetry and the ancient poetry are still similar with each other. However, this observation is not mentioned in Voigt & Jurafsky (2013).
3. The rhyming patterns of Taiwanese poetry are very close to the poetry of 1940s (XS-4) and the poetry after 1978 (XS-6, XS-7, XS-8). This similarity may be related to the developing history of new poetry in mainland China and Taiwan, especially the immigration history of the representative poets. Many of the earliest modern poets, such as "the modernist school", immigrated to Taiwan from mainland China during the 1940s and1950s. These immigrant poets include the three poetic predecessors, JiXian[7], Qin Zihao[8] and Zhong Dingwen[9], as well as some other famous poets, such as Luofu[10],

[7] Jixian (Jìxián 纪弦, 1913-2013), formally called Lu Yu (LùYú 路逾), born in Hebei province, immigrated to Taiwan in 1948, one of the poetic predecessors of Taiwanese poets.

[8] Qin Zihao (Qín Zǐháo 覃子豪, 1912-1963), born in Sichuan province, immigrated to Taiwan in 1947, one of the poetic predecessors of Taiwanese poets.

[9] ZhongDingwen (Zhōng Dǐngwén 钟鼎文, 1914-2012), born in Anhui province, immigrated to Taiwan in 1949, one of the poetic predecessors of Taiwanese poets.

[10] Luofu (Luòfū 洛夫, 1928--), born in Hunan province, immigrated to Taiwan in 1949, one of the best known modernist poets in Taiwan.

Yu Guangzhong[11], Zheng Chouyu[12], Yang Lingye[13], Zhou Mengdie[14], etc. All of these poets were formerly active poets in mainland China during the 1930s and 1940s before their immigrations. This may indicate to why the rhyming patterns of Taiwanese poetry are quite similar to that of Mandarin new poetry in the 1940s. Secondly, since 1978, thanks to the reform and opening-up policy, the political and social environment in mainland China has been constantly changing and improving. Poets can create their works with much more freedom. The creation and writing of poetry has therefore become more modern and diverse. Also, because of the increasingly close communication and exchange between the poets in mainland China and Taiwan, poems in the two areas therefore have more in common, including the rhyming patterns (Yeh 1993). This would be the reason why the rhyming patterns of Taiwanese are so similar to that of the poems in XS-6, XS-7 and XS-8.

4. Taking both Table 11 and Table 16 into consideration, the mean terminal rhyme probabilities in two adjacent verse lines of translated poetry and Taiwanese poetry are not significantly different, with $p = 0.840$ ($p > 0.05$). However, the probabilities in three adjacent verse lines are quite different, with $p = 0.003$ ($p < 0.05$). These results indicate that translated poetry and Taiwanese poetry attach different degrees of importance to rhyming in three adjacent verse lines. What's more, with the data in Fig. 2, it is easy to see that translated poetry is more likely to adopt this kind of rhyming than Taiwanese poetry. Compared with all the new poetry, the terminal rhyme probabilities in two adjacent verse lines of translated poetry are similar with that of the poetry in TW, XS-4, XS-6, XS-7, and XS-8. Meanwhile, the probabilities in three adjacent verse lines of translated poetry are similar with that of the poetry in XS-1, XS-2, XS-3, and XS-5. Translation cannot ignore the language features of the original texts, which means that the translation of poetry has much to do with the properties of the original language, especially the poetic properties in the language such as the stress of the syllables in the verse lines, etc. Furthermore, translation is widely seen as the re-creation of the original works. For instance, Ji (2013) shows that once a text is translated, the genre of the translated version is definitely different from the original one. Pan et al. (2016) quantitatively study ten Shakespear's sonnets and four

[11] Yu Guangzhong (Yú Guāngzhōng 余光中, 1928--), born in Jiangsu province, immigrated to Taiwan in 1949, a well known poets, essayists, translators in Taiwan.

[12] Zheng Chouyu (Zhèng Chóuyǔ 郑愁予, 1933--), formally called Zheng Wentao (ZhèngWéntāo 郑文韬), born in Shandong province, immigrated to Taiwan in 1949, a well known poet in Taiwan.

[13] Yang Lingye (Yáng Lìngyě 羊令野, 1923-1994), formally called Huang Zhongcong (Huáng Zhòngcóng 黄仲琮), born in Anhui province, immigrated to Taiwan in 1950, a well known poet in Taiwan.

[14] Zhou Mengdie (Zhōu Mèngdié 周梦蝶, 1921-2014), formally called Zhou Qishu (Zhōu Qǐshù 周起述), born in Henan province, immigrated to Taiwan in 1947, a well known poet in Taiwan.

corresponding English-Chinese translated sonnets, and conclude that both the use of words and the distribution of part of speech are significantly different between the original English sonnets and their Chinese translated texts

5. The investigation of the rhyming patterns of the new poetry indicated that it is reasonable to divide the short history of new poetry into several periods or eras. According to the data in section 3, the eight periods of new poetry can be clustered into three groups: the first group consists of the poetry in XS-4, the second group consists of the poetry in XS-1, XS-2, XS-3 and XS-5, and the third group consists of the poetry in XS-6, XS-7 and XS-8.

The 1940s were very special in the history of the new poetry. As Wu (cf. Hong, 2010a: 151) states that the old or traditional poetic things (e.g. old poetry style, old poetic language, old poetic aesthetics criteria, etc.) in this era collapsed, but the new ones had not been built up yet. Thus, the poetry in this era attempted to change the old poetic features and to create brand new poetic categories. The LSD comparison in Table 11 tells us that, the mean terminal rhyme probabilities in two adjacent verse lines of the poetry in XS-4 are neither significantly different from that in XS-3 and XS-5, nor from those in TW, YS, XS-6, XS-7 and XS-8. Furthermore, the LSD comparison in Table 16 yields a very similar result that the mean terminal rhyme probabilities in three adjacent verse lines of the poetry in XS-4 are again neither significantly different from those in XS-1, XS-3 and XS-5, nor from those in TW, YS, and XS-6.

The LSD comparison in Table 11 and Table 16, as well as Fig. 1 and Fig. 2, indicate that the poetry in XS-1, XS-2, XS-3, and XS-4 pay attention on the terminal rhyme more than other new poetry. Although the earliest Mandarin new poetry (poetry in XS-1) is often considered to have broken the relation with the ancient poetry, it is probably not entirely the case. The LSD comparison (see Table 11) indicates no significant differences between the poetry in JT and that in XS-1. It means the rhyming patterns of the earliest poetry are not much influenced by the ancient poetry. Meanwhile, most active poets in 1920s and 1930s formed a poetic school, "new moon[15]", to study the rhymes of the new poetry. They believe that rhyme should be the key property of poetry. Poems in Maoist time (XS-5) are very specific (Yu 1983). Most of the poems are odes, slogans, and folk songs, whose genre rely heavily on rhymes.

Acknowledgement

The present study is funded by "Huaqiao University's Academic Project Supported by the Fundamental Research Funds for the Central Universities (15SKGC-QG12)".

[15] The "new moon" poetic school (新月派) has been founded in 1926. Poets in the school stood against free verse but advocated rhymes. Some of the representative poets are Hsu Chih-mo (徐志摩), Wen Yiduo (闻一多), Zhu Xiang (朱湘), etc.

References

Chao, Y. (2011). *A Grammar of Spoken Chinese*. Beijing: The Commercial Press.

Chen, M. (1979). Metrical Structure: Evidence from Chinese Poetry. *Linguistic Inquiry 10*(3), 371-420.

Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford, New York: Oxford University Press.

Greene, R., Cushman, S., Cavanagh, C., Feinsod, H., Ramazani, J., Marno, D., Slessarev, A. (2012). *The Princeton Encyclopedia of Poetry and Poetics* (*4th edition*). Princeton/Oxford: Princeton University Press.

Hong, Z. (2010a). *The Brief One-hundred-year History of Chinese New Poetry* (Bǎinián Zhōngguó xīnshī shǐlüè "百年中国新诗史略"). Beijing: Beijing University Press.

Hong, Z. (2010b). *Volumes of Chinese New Poetry* (*Vol. 1-10*) (Zhōngguó xīnshī zǒngxì (1-10 juàn) "中国新诗总系(1-10 卷)"). Beijing: Beijing University Press.

Hu, Y. (1995). *Modern Chinese* (Xiàndài hànyǔ "现代汉语"). Shanghai: Shanghai Educational Publishing House.

Huang, B., Liao, X. (2002). *Modern Chinese* (Xiàndài hànyǔ "现代汉语"). Beijing: Higher Education Press.

Ji, M. (2013). *Exploratory Statistical Techniques for the Study of Literary Translation*. Lüdenscheid: RAM-Verlag.

Kao, J., Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In: Elson, D.K., Kazantaseva, A., Mihalcea, R., & Szpakowicz, S. (eds.). *Workshop on Computational Linguistics for Literature -- The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 8-17*. Madison, USA: Omnipress, Inc.

Lee, W.-S., Zee, E. (2014). Chinese Phonetics. In: Huang, C.T. James, Li, Y.H. Audrey, Simpson, A. (eds.). *The Handbook of Chinese Linguistics: 369-399*. West Sussex: John Wiley & Sons Ltd.

Lin, H. (2012). A New Rhyming Scheme of Chinese Classic Poetry. *Asian Journal of Social Sciences & Humanities 1*(2), 107-114.

Masters, E. L. (1915). What is poetry. *Poetry 6*(6), 306-308.

McDougall, B.S. (1994). Modern Chinese Poetry (1900-1937). *Modern Chinese Literature 8*(1/2), 127-170.

Pan, X., Chen, X., Liu, H. (2016). *Harmony in Diversity: The Language Codes in English-Chinese Poetry Translation. A quantitative study on Shakespearean sonnets translation.* Manuscript.

Qi, G. (2009). *Thesis on Poetic Rhythm* (Shīwén shēnglǜ lùngǎo"诗文声律论

稿”). Beijing: Zhonghua Book Company.

Shibles, Warren A. (1994). Chinese Romanization system: IPA transliteration. *Sino-Platonic Papers 52,* 1-17.

Sun, Y. (2008). Simple Review of the New Poetry Developing History in 1920s (1920 niándài Zhōngguó xīnshī shùlüè “1920 年代中国新诗发展述略”). *Journal of Beijing University 45* (2), 89-98.

Voigt, R., Jurafsky, D. (2013). Tradition and Modernity in 20th Century Chinese Poetry. In: *Proceedings of Second Workshop on Computational Linguistics for Literature*. USA: Atlanta, Georgia.

Yeh, M. (1993). Contemporary Chinese Poetry Scenes. *Chicago Review 39*(3/4), 279-283.

Yu, S. (1983). Voice of Protest: Political Poetry in the Post-Mao Era. *The China Quarterly 96,* 703-719.

Lexical Text Compactness with Link Length Taken into Account

Gejza Wimmer[1], *Ján Mačutek*[2]

Abstract. A new, more detailed approach to the analysis of lexical text compactness is suggested. It is based on the same idea as the previous one (sentences are considered linked if they share autosemantic words in common), but here also distances between linked sentences are taken into account. It seems that there are two patterns of properties of lexical text compactness, one for short (50 sentences or less) and one for longer texts. Statistical tests are provided for both cases.

***Keywords**: textology, lexical text compactness, statistical tests.*

1. Introduction

Mačutek and Wimmer (2014) presented a relatively simple measure of lexical text compactness (hereafter *LTC*), according to which the more sentences there are that share autosemantic words in common, the more compact the text is. This measure is defined as

$$LTC = \frac{L}{\binom{n}{2}},$$

where L is the number of linked sentences (i.e., the number of pairs of sentences which share at least one content word - noun, adjective, verb, or adverb) and n is the number of sentences in the analyzed text.

In this contribution we develop a more detailed analysis of the *LTC*; specifically, the lengths of links (i.e., distances between linked sentences) will be considered. If, e.g., the first sentence is linked with the fifth sentence, the length of this link is four. Theoretically, if all sentences were mutually linked, there would be $n - 1$ links of length 1, $n - 2$ links of length 2, etc. In general, the maximum possible number of links of length j is $n - j$.

Let us take into consideration the relative number of links of a given length and denote by $r(j)$ the number of observed links of length j divided by $n - j$ (i.e., by the maximum possible number of links of length j). A text is thus characterized by the numbers

[1] Mathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 81473 Bratislava, Slovakia, and Department of Mathematics, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 97401 Banská Bystrica, Slovakia,
e-mail: wimmer@mat.savba.sk

[2] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com

$$r_1, r_2, \dots, r_{n-1},$$

which can be interpreted as measures of link exploitations. These numbers take their values from the interval $\langle 0,1 \rangle$.

The values of $r_1, r_2, \dots, r_{n-1}$ were evaluated from 59 texts written by the Czech author Karel Čapek[3] (nine fairy tales FT1-FT9, and ten texts from each of the following types: journalistic texts J1-J10, private letters L1-L10, scientific texts on aesthetics AE1-AE10, short stories S1-S10, travel books T1-T10). The lengths of these texts vary from 5 to 314 sentences. Some text characteristics can be found in Mačutek et al. (2016), Table 1 (pp. 827-828). Based on our observations, we tentatively suppose that there are two different patterns of the values of $r_1, r_2, \dots, r_{n-1}$, one for short texts (up to 50 sentences) and one for longer texts (more than 50 sentences). The two patterns can be seen in Fig. 1.

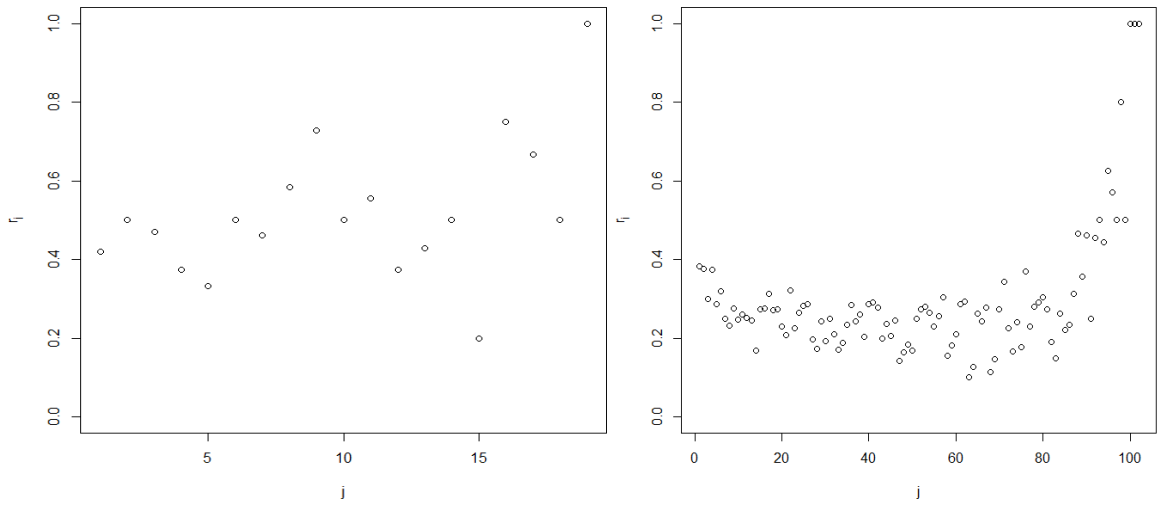


Figure 1. Values of $r_1, r_2, \dots, r_{n-1}$ for a journalistic text J2 (left, 20 sentences) and for a fairytale FT2 (right, 103 sentences).

We suppose that the sequences of the relative numbers of links $r_1, r_2, \dots, r_{n-1}$ can be modelled by a linear function for short texts and by a quadratic function for longer ones[4]. This is a pragmatic decision based on our analyses of the above-mentioned texts, which enables us to apply the apparatus of mathematical statistics described in the following section. Of course, the possibility that other, more appropriate models will be found cannot be excluded.

[3] The analyzed texts can be found at https://www.mlp.cz/cz/projekty/on-line-projekty/karel-capek/ (accessed on 13 July 2016), cf. also Mačutek et al. (2016).

[4] Data and a computer program (written in the statistical environment *R*) which gives the relative numbers of links (i.e., values of $r_1, r_2, \dots, r_{n-1}$) and the text characteristics from this paper as its output can be sent upon request (jmacutek@yahoo.com).

2. Statistical model

Mačutek and Wimmer (2014) showed that the number of observed links of length *j* follows the binomial distribution with the parameters p_j and *n-j* (cf. Wimmer and Altmann 1999, pp. 21-27); the distribution will hereafter be denoted as *bin*(p_j, *n-j*). Thus, every element r_j of the vector *R* from (1) is a realization of $\frac{1}{n-j} bin\left(p_j,\ n-j\right)$, which is a distribution with mean p_j and variance $\frac{p_j(1-p_j)}{n-j}$. As the binomial distribution *bin*(p_j, *n-j*) can be approximated for sufficiently large *n-j* by $N\left(p_j,\ (n-j)p_j\left(1-p_j\right)\right)$, i.e., by the normal distribution with mean p_j and variance $(n-j)p_j\left(1-p_j\right)$, the vector

$$\boldsymbol{r} = (r_1, r_2, \dots, r_{n-1})^T$$

can be considered a realization of the random vector

$$\boldsymbol{R} = (R_1, R_2, \dots, R_{n-1})^T, \tag{1}$$

which follows approximately the $N_{n-1}(\boldsymbol{p}, \boldsymbol{V})$ distribution, i.e., the *n*-1-dimensional normal distribution with mean $\boldsymbol{p} = (p_1, p_2, \dots, p_{n-1})$ and covariance matrix

$$\boldsymbol{V} = \sigma^2 \begin{pmatrix} \frac{1}{n-1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \sigma^2 \boldsymbol{H}. \tag{2}$$

It is assumed that random variables $(R_1, R_2, \dots, R_{n-1})$ are independent (reasons why they can be considered independent are explained in Mačutek and Wimmer 2014), and that $p_1, p_2, \dots, p_{n-1}$ do not differ much from each other. The variance factor σ^2 serves as a correction of the covariance matrix $\boldsymbol{H}$; it enables data to be fitted better (this approach is well-known e.g. in metrology, see e.g. Fan 2010).

The following formulae for estimates, statistical tests and confidence intervals are taken from Kubáček (1988).

2.1 Short texts

The basic assumption about $p_1, p_2, \dots, p_{n-1}$ in short texts (cf. Figure 1 left) is that they lie on a straight line, i.e.,

$$p_i = a_S + b_S i, \qquad i = 1,2, \dots, n-1. \tag{3}$$

The number

$$m_S = a_S + b_S \frac{n}{2} \tag{4}$$

is the mean of the probabilities that two sentences in a text are linked. The values of the parameter b_S show whether the probabilities of links $p_1, p_2, \dots, p_{n-1}$ tend to increase (if $b_S > 0$) or decrease (if $b_S < 0$) with increasing distance between sentences. Furthermore, the coefficient σ^2 indicates how the real (observed) values of $r_1, r_2, \dots, r_{n-1}$ are spread around their theoretical means $p_1, p_2, \dots, p_{n-1}$. For smaller values of σ^2 they are concentrated close to their means, whereas if σ^2 is large, they are more dispersed.

For short texts, the random vector $\boldsymbol{R}$ from (1) follows the linear regression model $\left(\boldsymbol{R};\ \boldsymbol{A}\begin{pmatrix} a_S \\ b_S \end{pmatrix}; \boldsymbol{V}\right)$, i.e., the mean of $\boldsymbol{R}$ is

$$\boldsymbol{A}\begin{pmatrix} a_S \\ b_S \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n-1 \end{pmatrix}\begin{pmatrix} a_S \\ b_S \end{pmatrix}$$

and its covariance matrix is $\boldsymbol{V} = \sigma^2 \boldsymbol{H}$, with the matrix $\boldsymbol{H}$ defined in (2). In addition, it is assumed that $\boldsymbol{R}$ is normally distributed.

It is well known that the optimal estimator of parameters a_S and b_S is

$$\begin{pmatrix} \widehat{a_S} \\ \widehat{b_S} \end{pmatrix} = (\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{R}$$

and the optimal estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-3} \boldsymbol{v}^T \boldsymbol{H}^{-1} \boldsymbol{v},$$

where $\boldsymbol{v} = \boldsymbol{A}\begin{pmatrix} \widehat{a_S} \\ \widehat{b_S} \end{pmatrix} - \boldsymbol{R}$. The unbiased estimator of the covariance matrix of $\begin{pmatrix} \widehat{a_S} \\ \widehat{b_S} \end{pmatrix}$ is

$$\widehat{\boldsymbol{V}} = \widehat{\sigma^2} (\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A})^{-1}.$$

The optimal linear estimator of p_i, $i = 1,2, \dots, n-1$, is

$$\widehat{p_i} = \widehat{a_S} + \widehat{b_S} i, \qquad i = 1,2, \dots, n-1.$$

The $(1-\alpha)$-confidence interval for p_i is

$$\left(\widehat{p_i} - t_{n-3}\left(1 - \frac{\alpha}{2}\right) \sqrt{(1, i)(\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A})^{-1} \begin{pmatrix} 1 \\ i \end{pmatrix}}; \right.$$
$$\left. \widehat{p_i} + t_{n-3}\left(1 - \frac{\alpha}{2}\right) \sqrt{(1, i)(\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A})^{-1} \begin{pmatrix} 1 \\ i \end{pmatrix}} \right),$$

where $t_{n-3}\left(1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of the Student t-distribution with *n*-3 degrees of freedom. The optimal linear estimator of m_S from (4) is

$$\widehat{m_S} = \widehat{a_S} + \widehat{b_S}\frac{n}{2},$$

and the $(1-\alpha)$-confidence interval for m_S is

$$\left(\widehat{m_S} - t_{n-3}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1,\frac{n}{2}\right)(\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1}\left(\frac{1}{n}\right)};\right.$$
$$\left.\widehat{m_S} + \quad t_{\ n-3}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1,\frac{n}{2}\right)(\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1}\left(\frac{1}{n}\right)}\right)$$

The hypothesis $H_0: b = 0$ (i.e., that the distance between two sentence does not have any influence on the probability of the sentences being linked) can be tested using the test statistic

$$T_{b_S} = \frac{\widehat{b_S}}{\widehat{\sigma^2}\sqrt{\{(\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1}\}_{2,2}}}.$$

If a realization of $|T_{b_S}|$ is greater than $t_{n-3}\left(1-\frac{\alpha}{2}\right)$, the null hypothesis is rejected at the significance level α.

When comparing two short texts, one can test the hypothesis $H_0: b_S^{(1)} = b_S^{(2)}$ (i.e., that the influence of distances between two sentences on the probability of links is the same in the two texts), with $b_S^{(1)}$, $b_S^{(2)}$ being the coefficients in the linear regression (3) for the two texts. If the confidence intervals

$$\left(b_S^{(1)} - t_{n_1-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\{((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(1)}\}_{2,2}};\right.$$
$$\left.b_S^{(1)} + t_{n_1-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\{((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(1)}\}_{2,2}}\right)$$

and

$$\left(b_S^{(2)} - t_{n_2-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\{((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(2)}\}_{2,2}};\right.$$
$$\left.b_S^{(2)} + t_{n_2-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\{((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(2)}\}_{2,2}}\right)$$

are disjoint (i.e., their intersection is the empty set), the null hypothesis is rejected at the significance level α (n_1, n_2 are numbers of sentences and $((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(1)}$, $((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(2)}$ matrices pertaining to the two texts under study).

One can also test the hypothesis $H_0: m_S^{(1)} = m_S^{(2)}$, i.e., the hypothesis that the mean probabilities $m_S^{(1)}$, $m_S^{(2)}$ of links in two texts do not differ. If the confidence intervals

$$\left(m_S^{(1)} - t_{n_1-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_1}{2}\right)((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(1)}\left(\frac{1}{n_1}\right)};\right.$$
$$\left. m_S^{(1)} + t_{n_1-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_1}{2}\right)((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(1)}\left(\frac{1}{n_1}\right)}\right)$$

and

$$\left(m_S^{(2)} - t_{n_2-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_2}{2}\right)((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(2)}\left(\frac{1}{n_2}\right)};\right.$$
$$\left. m_S^{(2)} + t_{n_2-3}\left(1-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{2}{2}\right)((\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})^{-1})^{(2)}\left(\frac{1}{n_2}\right)}\right)$$

are disjoint, the null hypothesis is rejected at the significance level α.

The hypothesis $H_0: \sigma_{(1)}^2 = \sigma_{(2)}^2$ regarding the variance factors in two short texts introduced in (2) can also be tested, using the test statistic

$$F = \frac{\widehat{\sigma_{(1)}^2}}{\widehat{\sigma_{(2)}^2}}.$$

The value of this statistic is then compared with the quantiles of the Fisher-Snedecor distribution. If it is greater than $F_{n_1-3,n_2-3}\left(1-\frac{\alpha}{2}\right)$ or less than $F_{n_1-3,n_2-3}\left(\frac{\alpha}{2}\right)$, the null hypothesis is rejected at the significance level α.

2.2 Longer texts

Regarding longer texts (those containing more than 50 sentences, cf. Section 1), we assume that

$$p_i = a_L + b_L i + c_L i^2, \qquad i = 1,2,\dots,n-1, \qquad (5)$$

which means that $p_1, p_2, \dots, p_{n-1}$ are supposed to lie on a parabola. The coefficients a_L, b_L and c_L describe the shape of the parabola (we assume that $c_L > 0$).

In the case of longer texts, the mean of the probabilities that two sentences in a text are linked is

$$m_L = a_L + b_L \frac{n}{2} + c_L \frac{n(2n-1)}{6}.$$

The random vector R from (1) follows the linear regression model $\left(\boldsymbol{R}, \boldsymbol{B}\begin{pmatrix} a_L \\ b_L \\ c_L \end{pmatrix}, \boldsymbol{V}\right)$, i.e., the mean of $\boldsymbol{R}$ is

$$\boldsymbol{B}\begin{pmatrix} a_L \\ b_L \\ c_L \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & n-1 & (n-1)^2 \end{pmatrix}\begin{pmatrix} a_L \\ b_L \\ c_L \end{pmatrix}$$

and its covariance matrix is $V = \sigma^2 H$; see (2). The interpretation of the coefficient σ^2 remains the same as for short texts (cf. Section 2.1). Again, it is assumed that $\boldsymbol{R}$ is a normally distributed random vector.

The optimal estimator for the parameters a_L, b_L and c_L is

$$\begin{pmatrix} \widehat{a_L} \\ \widehat{b_L} \\ \widehat{c_L} \end{pmatrix} = (\boldsymbol{B}^T \boldsymbol{H}^{-1} \boldsymbol{B})^{-1} \boldsymbol{B} \boldsymbol{H}^{-1} \boldsymbol{R}$$

and the optimal estimator for the coefficient σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-4} \boldsymbol{w}^T \boldsymbol{H}^{-1} \boldsymbol{w},$$

with $\boldsymbol{w} = \boldsymbol{B}\begin{pmatrix} \widehat{a_L} \\ \widehat{b_L} \\ \widehat{c_L} \end{pmatrix} - \boldsymbol{R}$. The unbiased estimator of the covariance matrix of the vector $\begin{pmatrix} \widehat{a_L} \\ \widehat{b_L} \\ \widehat{c_L} \end{pmatrix}$ is $\widehat{\boldsymbol{W}} = \widehat{\sigma^2}(\boldsymbol{B}^T \boldsymbol{H}^{-1} \boldsymbol{B})^{-1}$.

The optimal linear estimator of p_i, $i = 1,2, \dots, n-1$, is

$$\widehat{p_i} = \widehat{a_L} + \widehat{b_L} i + \widehat{c_L} i^2, \qquad i = 1,2, \dots, n-1.$$

The $(1-\alpha)$-confidence interval for p_i is

$$\left(\widehat{p_i} - t_{n-4}\left(1 - \frac{\alpha}{2}\right)\sqrt{(1, i, i^2)(\boldsymbol{B}^T \boldsymbol{H}^{-1} \boldsymbol{B})^{-1}\begin{pmatrix} 1 \\ i \\ i^2 \end{pmatrix}};\right.$$

$$\widehat{p_i} + t_{n-4}\left(1-\frac{\alpha}{2}\right)\sqrt{(1,i,i^2)(\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1}\begin{pmatrix}1\\ i\\ i^2\end{pmatrix}}\Bigg).$$

The optimal linear estimator of m_L is

$$\widehat{m_L} = \widehat{a_L} + \widehat{b_L}\frac{n}{2} + \widehat{c_L}\frac{n(2n-1)}{6}$$

and

$$\left(m_L - t_{n-4}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1,\frac{n}{2},\frac{n(2n-1)}{6}\right)(\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1}\begin{pmatrix}1\\ \frac{n}{2}\\ \frac{n(2n-1)}{6}\end{pmatrix}};\right.$$
$$\left.m_L + t_{n-4}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1,\frac{n}{2},\frac{n(2n-1)}{6}\right)(\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1}\begin{pmatrix}1\\ \frac{n}{2}\\ \frac{n(2n-1)}{6}\end{pmatrix}}\right)$$

is the $(1-\alpha)$-confidence interval for m_L.

For two long texts, one can test the hypothesis $H_0: m_S^{(1)} = m_S^{(2)}$. If the confidence intervals

$$\left(m_L^{(1)} - t_{n_1-4}\left(1\right.\right.$$
$$\left.-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_1}{2},\frac{n_1(2n_1-1)}{6}\right)((\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1})^{(1)}\begin{pmatrix}1\\ \frac{n_1}{2}\\ \frac{n_1(2n_1-1)}{6}\end{pmatrix}};$$

$$m_L^{(1)} + t_{n_1-4}\left(1\right.$$
$$\left.\left.-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_1}{2},\frac{n_1(2n_1-1)}{6}\right)((\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1})^{(1)}\begin{pmatrix}1\\ \frac{n_1}{2}\\ \frac{n_1(2n_1-1)}{6}\end{pmatrix}}\right)$$

and

$$\left(m_L^{(2)} - t_{n_2-4}\left(1-\frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_2}{2},\frac{n_2(2n_2-1)}{6}\right)((\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1})^{(2)}\begin{pmatrix}1\\ \frac{n_2}{2}\\ \frac{n_2(2n_2-1)}{6}\end{pmatrix}}\right.;$$

$$m_L^{(2)} + t_{n_2-4}\left(1 - \frac{\alpha}{4}\right)\sqrt{\left(1,\frac{n_2}{2},\frac{n_2(2n_2-1)}{6}\right)((\boldsymbol{B}^T\boldsymbol{H}^{-1}\boldsymbol{B})^{-1})^{(2)}\begin{pmatrix}1\\ \frac{n_2}{2}\\ \frac{n_2(2n_2-1)}{6}\end{pmatrix}}\Bigg)$$

are disjoint, the null hypothesis is rejected at the significance level α.

As with short texts, the hypothesis $\sigma_{(1)}^2 = \sigma_{(2)}^2$ can also be tested using the test statistic

$$F = \frac{\widehat{\sigma_{(1)}^2}}{\widehat{\sigma_{(2)}^2}}.$$

If the value of the test statistic F is greater than $F_{n_1-4,n_2-4}\left(1-\frac{\alpha}{2}\right)$ or less than $F_{n_1-4,n_2-4}\left(\frac{\alpha}{2}\right)$, the null hypothesis is rejected at the significance level α.

3. Examples

We will apply the formulae from Sections 2.1 and 2.2 to two short texts (journalistic texts J1 and J2) and to two longer texts (fairytales FT1 and FT2). The results can be found in Tables 2 and 3 below.

Table 2
Some characteristics of two journalistic texts (J1 and J2).

	J1	**J2**
$\widehat{a_S}$	0.493	0.418
$\widehat{b_S}$	-0.025	-0.003
T_{b_S}	-7.712	-1.043
comparing b_S in the two texts	(-0.060; 0.010)	(-0.019; 0.012)
$\widehat{\sigma^2}$	0.058	0.248
$\widehat{m_S}$	0.245	0.368
comparing m_S in the two texts	(0.031; 0.435)	(0.236; 0.500)

For these two texts, the differences between the values of the parameters b_S and between the mean values of link probabilities m_S are not significant at the 0.05 level (the respective confidence intervals have in both cases non-empty intersections). The 0.95-quantiles of the Student t-distributions for texts J1 and J2 are 2.101 and 2.048, respectively, which means that the hypothesis on the zero value of the parameter b_S is rejected for text J1, but not for text J2.

Table 3
Some characteristics of two fairytales (FT1 and FT2)

	FT1	**FT2**
$\widehat{a_L}$	0.480	0.332
$\widehat{b_L}$	-0.004	-0.005
$\widehat{c_L}$	0.00002	0.00006
$\widehat{\sigma^2}$	0.113	0.149
$\widehat{m_L}$	0.365	0.263
comparing m_L in the two texts	(0.329; 0.400)	(0.221; 0.305)

One can see that the value of m_L is significantly higher for text FT1, as the confidence intervals are disjoint (the interval bounds were again computed with α set at 0.05.

4. Conclusion

Several new approaches for analyzing the *LTC* were presented in this paper. In addition to the models and tests from Mačutek and Wimmer (2014), the statistical apparatus introduced here also takes into account link lengths and exploitation rates of links with different lengths. Thus, not only can the *LTC* itself be tested; the dependence of link probabilities on the link length is, at least tentatively, also described - see formulae (3) and (5) (Sections 2.1 and 2.2). The paper also provides a tool for testing differences between the slopes of the linear dependencies in two short texts (cf. the test for differences between parameters b_S in Section 2.1). In principle, tests for other parameters from formulae (3) and (5) can be derived as well.

Given that the last few r_i's take their values from just a few observations (the extreme case is r_{n-1} which reflects just one pair of sentences, the first and the last one, hence r_{n-1} is either 0 or 1), quite high fluctuations in their values can be expected. Therefore, one could omit, e.g., the last five of them in each text, even at the cost that our methodology is then not applicable to very short texts. Such a treatment of units which occur but rarely, and therefore bring too much instability to the data, is quite common in mathematical modelling of linguistic data (e.g., Kelih 2010, Mačutek and Mikros 2015).

Acknowledgement

Supported by the grant VEGA 2/0047/15 (G. Wimmer, J. Mačutek) and APVV-15-0295 (G. Wimmer).

References

Fan, H. (2010). *Theory of Errors and Least Squares Adjustment*. Stockholm: Royal Institute of Technology (KTH), Division of Geodesy and Geoinformatics.

Kelih, E. (2010). Parameter interpretation of the Menzerath-Altmann law: evidence from Serbian. In: Grzybek, P., Kelih, E., Mačutek, J (eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives: 71-79.* Wien: Praesens.

Kubáček, L. (1988). *Foundations of Estimation Theory.* Amsterdam: Elsevier.

Mačutek, J., Koščová, M., Čech, R. (2016). Lexical compactness across genres in works by Karel Čapek. In: Mayaffre, D., Poudat, C., Vanni, L., Magri, V., Follette, P. (eds.), *Statistical Analysis of Textual Data: 825-832.* Nice: University Nice Sophia Antipolis.

Mačutek, J., Mikros, G.K. (2015). Menzerath-Altmann law for word length motifs. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 125-131*. Berlin, Boston: de Gruyter.

Mačutek, J., Wimmer, G. (2014). A measure of lexical text compactness. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical Approaches to Language and Text Analysis: 132-139.* Lüdenscheid: RAM-Verlag.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm

Continuous Modelling of Verse Lengths in Welsh and Gaelic Metrical Psalmody

Andrew Wilson
Lancaster University

Abstract. Twenty 17th century metrical psalms (ten in Welsh and ten in Scottish Gaelic) were used to investigate the suitability of the Zipf-Alekseev function for modelling verse lengths in these Celtic languages. Good fits were obtained for both languages; however, the parameter *c* was always negligible, hence future work might consider using the simple power function, if this proves to be the case more generally. Future work needs to look at a broader sample of authors, dates, genres, etc. in order to arrive at a full synergetic model for verse lengths in Welsh and Scottish Gaelic.

Keywords: *Welsh, Scottish Gaelic, verse lengths, Zipf-Alekseev function, power function.*

1. Introduction

Poetry can come in many forms but the commonest involves sets of verses with a fixed metrical pattern of long and short, or stressed and unstressed, syllables. Such poetic metres allow very limited scope for variation in the lengths of lines when they are measured in syllables and deviations from the metre occur only rarely. The only exceptions to this limit on variability are those metres where a certain number of syllables of a particular type may be substituted for a different number of another type - e.g., in the classical Latin hexameter and pentameter, where a choice of either two short syllables or one long syllable is allowable in many positions. However, if we measure the lengths of verses in words, rather than in syllables, then there is more scope for variation. Nevertheless, the constraints of the metrical form, and the word-length structure of a language's lexicon, will still have some effect on the number of words in a verse. What we have here, then, is a potential stochastic regularity that enters into a larger synergetic linguistic system (Köhler, 1987).

Verse length, in words, has not yet been extensively researched for any language. Best (2012a) cites only two previous works by Muller (1972) and Grotjahn (1979). However, in a recent series of studies, Best (2012a, 2012b, 2013) has begun to explore verse lengths in German and Old Icelandic using discrete probability distributions. He has found that, so far, the 1-displaced Binomial distribution constitutes a good model for the verse lengths in both languages. However, a problem with discrete modelling, as noted in the context of word-length research by Mačutek, Altmann (2007), is that many different distributions may be required to cover the world's languages, and different distributions may even be required within a single language to account for variations in date, text-type, etc. Whilst most of these distributions may be derivable as special cases from a more general law, their proliferation complicates the picture, especially when one seeks to interpret parameter variation from a func-

tional perspective. But discrete modelling is not the only form of modelling that can be applied to language and texts. Mačutek, Altmann (2007) have noted that discrete and continuous modelling are merely two sides of the same coin. Since all modelling involves an idealized abstraction of entities and properties from concrete reality, it is also possible to consider essentially discrete phenomena from a continuous perspective - which is the case, for example, in the modelling of rank-frequency relations using Zipf's power law. Using a continuous function in place of a discrete distribution opens up the possibility of arriving at a unique model of a phenomenon, where only the parameters of the function are free to vary from text to text, regardless of language or other typological considerations. Having once established the general appropriacy of the model, the nature and grounds of any variation in the parameter values can then be studied more systematically across different data sets.

Popescu, Best, Altmann (2014) thus began to experiment with a single continuous function – the Zipf-Alekseev function – to study the distribution of lengths for different units in texts. This function derives from the differential equation:

$$dy/y = (A + B \ln x /Dx)\, dx$$

where x is the length class and y is the frequency of the length class. A is the language, text-type, or style constant; B is the force of the speaker or writer; and D is the equilibrating force of the community (Koch 2014). Solving and re-paramterizing this equation gives the Zipf-Alekseev formula:

$$y = ax^{b+c \ln x}$$

The aim of this paper is to provide a preliminary test of whether this model can also account for the distribution of verse lengths in Celtic poetry, specifically within the genre of metrical psalmody in Welsh and Scottish Gaelic.

2. Data and method

This study is the first to consider Celtic verse lengths from the theoretical perspective of Popescu, Best, Altmann (2014), hence the focus is on establishing the appropriacy of the model, rather than on investigating patterns of parameter variation across authors, genres, dates, etc. Thus, in order to minimize predictable variation within the data samples for this pilot study, the scope of the experiment was restricted to 17th century metrical psalmody, a poetic genre that exists for both Welsh (a P-Celtic or Brythonic language) and Scottish Gaelic (a Q-Celtic or Goidelic language). Within the individual language samples, the metrical psalms considered here allowed date, metre, genre, and - at least for Welsh – authorship to be held constant, so that any language-internal parameter variation or failures in model fit must be attributable to other, less immediately obvious factors. Also, since the data are free translations of the biblical book of Psalms, some of which

are quite lengthy, it was possible to obtain a reasonably sized corpus of sufficiently long texts that adhere to these criteria, something that is not always possible with poetic genres that favour short texts. If we were to rely on other texts at this stage in the research, we could encounter difficulty in finding sets of sufficiently long texts that do not also exhibit substantial internal variation in authorship, metre, date, etc.

The Welsh metrical psalms used here are all by Edmund (in Welsh: Edmwnd) Prys, a noted Welsh academic, clergyman, and poet (Williams 2004, Morgan 2011). They were published in 1621 as an appendix to the Welsh version of the Book of Common Prayer, and all use a metre of alternating 8 and 7 syllable verses.

The Scottish Gaelic texts are of less certain authorship. Their textual history is somewhat complex, but the basic facts are these. According to MacTavish (1934), the first 50 psalms were originally translated into verse by a group of three translators and published in 1659. The rest of the psalms, together with revisions of the first 50, were completed by a different team of five translators. The authors of the individual psalm translations are, unfortunately, not identified in the text. The text, as published, dates from 1694. All of the psalms use a metre of alternating 8 and 6 syllable verses.

The texts chosen for both languages were psalms 18, 22, 37, 68, 78, 89, 104, 106, 107, and 119. These were selected as being the ten longest psalms in the Welsh translation (which was analysed first).

For the purposes of this experiment, a word was considered to be any one or more alphabetic characters with a space, punctuation mark, or line break on either side. Apart from removing blank lines, title lines, verse numbering, etc., no attempt was made to modify the orthography of the texts, and the frequencies of line lengths (in words) were counted for each text using a reliable program written in Python. The Zipf-Alekseev function was then fitted to the frequency data from each text using TableCurves. A fit was considered acceptable if it achieved the usual benchmark R^2 of 0.8.

3. Results

The results for fitting the Zipf-Alekseev function to the Welsh data are shown in Table 1, and the results for the Scottish Gaelic data are shown in Table 2.

Table 1
Welsh data

Welsh Psalm 18			Welsh Psalm 78		
Words	Verses	Z-A	Words	Verses	Z-A
2	1	1.12	3	11	9.09
3	5	9.30	4	50	51.56
4	42	38.36	5	74	71.96
5	49	52.08	6	45	47.39

6	40	38.90	7	22	20.79
7	22	20.97	8	9	7.57
8	9	9.60	9	1	2.87
a = 23.8415, b = -7.4907, c = 2.95399985E-007, R^2 = 0.9811			a = 29.9705, b = -9.4974 c = 3.8600554E-009, R^2 = 0.9946		

Welsh Psalm 106			Welsh Psalm 22		
Words	Verses	Z-A	Words	Verses	Z-A
3	14	11.39	3	3	4.41
4	37	36.16	4	29	28.11
5	37	44.69	5	43	43.87
6	47	33.48	6	31	30.59
7	13	19.14	7	15	13.82
8	4	9.56	8	3	5.22
a = 20.1539, b = -6.4036 c = 5.71773026E-006, R^2 = 0.7848			a = 32.2565, b = -10.0819 c = 2.67570099E-010, R^2 = 0.9925		

Welsh Psalm 89			Welsh Psalm 107		
Words	Verses	Z-A	Words	Verses	Z-A
2	1	1.01	3	14	8.28
3	6	5.09	4	25	29.92
4	35	35.33	5	43	39.06
5	59	58.76	6	27	29.43
6	43	43.49	7	20	16.46
7	22	20.61	8	3	7.98
8	6	7.86			
a = 32.0086, b = -9.9062 c = 3.40804279E-010, R^2 = 0.9977			a = 22.1484, b = -6.9826 c = 9.00508214E-007, R^2 = 0.8720		

Welsh Psalm 37			Welsh Psalm 104		
Words	Verses	Z-A	Words	Verses	Z-A
3	2	5.08	3	8	5.10
4	32	28.31	4	27	25.01
5	38	43.61	5	33	39.87
6	39	32.44	6	43	33.40
7	14	16.23	7	16	19.54
8	3	6.74	8	5	9.48
a = 29.0385 b = -9.0277 c = 3.07743607E-009 R^2 = 0.9206			a = 25.4990 b = -7.7907 c = 3.39436242E-008 R^2 = 0.8336		

Welsh Psalm 119			Welsh Psalm 68		
Words	Verses	Z-A	Words	Verses	Z-A
3	16	11.81	3	6	6.88
4	88	87.50	4	37	35.92
5	138	141.69	5	45	46.79
6	108	101.83	6	31	28.80
7	45	46.57	7	12	12.00
8	9	16.67	8	1	4.39
a = 31.7840202 b = -9.88209848 c = 1.11953174E-009 R^2 = 0.9904			a = 30.3977335 b = -9.74183401 c = 2.35895246E-009 R^2 = 0.9869		

Table 2
Gaelic data

Gaelic Psalm 18			Gaelic Psalm 78		
Words	Verses	Z-A	Words	Verses	Z-A
3	5	1.50	2	1	1.00
4	19	23.15	3	2	2.66
5	74	69.22	4	32	31.72
6	46	53.45	5	79	78.57
7	28	19.33	6	68	70.39
8	4	4.99	7	41	35.34
			8	8	12.89
			9	1	4.30
a = 52.931569 b = -15.9864633 c = 6.61118031E-018 R^2 = 0.9491			a = 39.3325563 b = -11.7439952 c = 4.06266085E-013 R^2 = 0.9894		

Gaelic Psalm 106			Gaelic Psalm 22		
Words	Verses	Z-A	Words	Verses	Z-A
3	7	2.64	4	11	11.44
4	30	31.83	5	43	42.61
5	70	67.81	6	36	36.51
6	45	48.41	7	14	13.57
7	23	18.90	8	4	3.62
8	4	5.64			
9	1	1.95			
a = 42.9799367 b = -13.1898584 c = 4.1792278E-014 R^2 = 0.9844			a = 58.4149 b = -17.4302 c = 2.49867889E-020 R^2 = 0.9992		

Gaelic Psalm 89			Gaelic Psalm 107		
Words	Verses	Z-A	Words	Verses	Z-A
3	3	2.63	3	4	1.28
4	31	31.54	4	16	13.62
5	70	69.13	5	44	46.71
6	50	51.53	6	49	45.27
7	23	21.09	7	17	21.15
8	6	6.50	8	9	6.80
9	1	2.20	9	1	2.25
a = 42.2759444 b = -12.911623 c = 6.43650511E-014 R^2 = 0.9978			a = 49.6690963 b = -14.6549996 c = 2.68463569E-017 R^2 = 0.9734		

Gaelic Psalm 37			Gaelic Psalm 104		
Words	Verses	Z-A	Words	Verses	Z-A
3	3	5.21	3	5	2.05
4	29	26.69	4	22	20.77
5	39	41.91	5	47	49.31
6	36	33.57	6	45	41.85
7	19	18.53	7	18	19.88
8	6	8.48	8	6	7.07
			9	1	2.56
a = 26.7255756 b = -8.22524685 c = 1.52898657E-008 R^2 = 0.9733			a = 40.3866493 b = -12.1446769 c = 1.3095497E-013 R^2 = 0.9848		

Gaelic Psalm 119			Gaelic Psalm 68		
Words	Verses	Z-A	Words	Verses	Z-A
3	18	18.03	3	9	4.67
4	124	121.96	4	27	28.15
5	185	189.15	5	48	48.04
6	140	135.11	6	39	39.47
7	63	62.66	7	26	21.57
8	18	22.87	8	3	9.53
9	2	7.62			
a = 30.3282497 b = -9.46292484 c = 5.26340048E-009 R^2 = 0.9967			a = 28.8103115 b = -8.79483242 c = 2.68327415E-009 R^2 = 0.9438		

Tables 3 and 4 provide a summary of the estimates for parameters *a* and *b*. (Parameter c is negligible in all cases and will be commented on in the next

section.)

Table 3
Parameters a and b in the Welsh data

Psalm	a	b
18	23.84	-7.49
22	32.26	-10.08
37	29.04	-9.03
68	30.40	-9.74
78	29.97	-9.50
89	32.01	-9.91
104	25.50	-7.79
106	20.15	-6.40
107	22.15	-6.98
119	31.78	-9.88

Table 4
Parameters a and b in the Scottish Gaelic data

Psalm	a	b
18	52.93	-15.99
22	58.41	-17.43
37	26.73	-8.23
68	28.81	-8.79
78	39.33	-11.74
89	42.28	-12.91
104	40.39	-12.14
106	42.98	-13.19
107	49.67	-14.65
119	30.33	-9.46

4. Conclusion

In all but one case (namely the Welsh translation of psalm 106), an acceptable model fit was possible using the Zipf-Alekseev function. Even this one exception was only slightly below the R^2 cut-off of 0.8, with a value of 0.78. Most probably this is due to a boundary condition - e.g., corrections by editors. Overall, this experiment suggests that the Zipf-Alekseev function may well prove to be a generalizable model for verse lengths in both Welsh and Scottish Gaelic. It also provides further support for it as a cross-linguistic model for verse length.

Examining Tables 3 and 4, it is clear that parameter *a* is always a positive number here, in the approximate range [20, 60]. The values of parameter *a* tend, on the whole, to be larger for the Scottish Gaelic texts (minimum = 26.73, maximum = 58.41) than for the Welsh texts (minimum = 20.15, maximum =

32.26); seven out of the ten Gaelic texts have parameter *a* larger than the maximum for the Welsh texts. Parameter *b* is always a negative number in the approximate range [-18, -6]. The absolute values of parameter *b* tend again to be larger in the Gaelic texts than in the Welsh texts: the precise ranges, as negative numbers, are [-17.43, -8.23] for Gaelic and [-10.08, -6.4] for Welsh, with seven out of the ten Gaelic texts falling below the lower bound of the range for Welsh.

It is notable that parameter *c* is very small in all cases – indeed, one might say negligible, as it is far smaller than the two or three decimal places to which figures are typically rounded. If this continues to prove the case with other Welsh and Gaelic texts, it might then be more appropriate, in future, to attempt to model verse lengths in these languages using the simple power function: $y = ax^b$.

Modelling the relationship between parameters *a* and *b* for these texts shows that they are also linked by a power function $y = cx^d$ in both Welsh and Gaelic. The Welsh data give estimates of $c = -0.327$ and $d = 0.987$, with $R^2 = 0.992$. The Gaelic data give estimates of $c = -0.356$ and $d = 0.957$, with $R^2 = 0.997$.

Based on these few data alone, little more can be said about the estimated parameters from a functional perspective. Further work should attempt to apply these insights to other verse data from Welsh and Scottish Gaelic, in order to arrive at a full and explanatory synergetic model of verse lengths in these two languages. Patterns of parameter variation according to author, date, genre, etc., will need to be examined.

Acknowledgements and Data Access Statement

The Scottish Gaelic texts are from Text 188 - Sailm Dhaibhidh - courtesy of the Digital Archive of Scottish Gaelic project (www.dasg.ac.uk). The Welsh texts are from Morgan (2011).

References

Best, K.-H.. (2012a). How many words are in a verse? An exploration. In: Naumann, S., Grzybek, P., Vulanovic, R., Altmann, G., (eds.), *Synergetic linguistics. Text and language as dynamic systems:13-22*. Wien: Praesens.

Best, K.-H. (2012b). Zur Verslänge bei G. A. Buerger. *Glottometrics 23, 56-61*.

Best, K.-H. (2013). Zur Verslänge im Altisländischen. *Glottometrics 25, 22-29*.

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum: Brockmeyer.

Koch, V. (2014). *Quantitative film studies: Regularities and interrelations exemplified by shot lengths in Soviet feature films*. 2 vols. Ph.D. Thesis, University of Graz.

Köhler, R. (1987). System theoretical linguistics. *Theoretical Linguistics 14, 241–257.*

MacTavish, D.C. (1934). *The Gaelic Psalms, 1694*. Being a reprint of the edition

issued by the Synod of Argyll in that year. With an historical introduction by Duncan C. MacTavish. Lochgilphead: James M.S. Annan.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Morgan, A. (2011). *Astudiaeth o Salmau Cân (1621) Edmwnd Prys.* 2 vols. Ph.D. thesis, Aberystwyth University.

Muller, C. (1972). *Einführung in die Sprachstatistik*. München: Hueber.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

Williams, G.A. (2004). The poetic debate of Edmwnd Prys and Wiliam Cynwal. *Renaissance Studies 18(1), 33-54.*

German Loanwords in Polish and Remarks on the Piotrowski-Altmann Law

Kamil Stachowski
Jagiellonian University, Cracow, Poland

Abstract. As signalled in the title, the paper has a dual purpose: it discusses the German lexical influence on Polish and also the Piotrowski-Altmann law. The first focus requires a qualitative historical linguistic approach, whereas the second is more quantitative in nature. The paper shows that when these two approaches are combined new insights can be gained.

***Keywords**: Polish, Loanwords, Piotrowski-Altmann law*

1. Introduction

Germans and Poles have been neighbours since before either became what could today be considered a nation. If one combines this fact with the consequences of the cultural and political geography of Europe, and its history, it becomes evident that the lexical influence of German on Polish must be rather substantial, and ancient. The present paper analyzes the sources (section 2) and some of the phonetic adaptations (section 3), using a primarily quantitative approach, before producing not only factual but also methodological conclusions (section 4).

Frequently mentioned is the Piotrowski-Altmann law. This is an equation which describes the progression of a change in language. The most common variant is given in eq. 1, but here a slightly modified version will be used (eq. 2; see Stachowski 2013: 110f for an explanation). The two are equivalent, with $A = ln(a)/b$.

$$P_x = \frac{c}{1 + ae^{-bx}} \tag{1}$$

$$P_x = \frac{c}{1 + e^{-b(x-A)}} \tag{2}$$

where $a, A > 0$, and $b, c \neq 0$.

The advantage of eq. (2) is that in it all the coefficients are linguistically meaningful: A denotes the moment in time when the progression of change stops accelerating and begins to decelerate (the point of inflection), b the overall speed of the change (the slope), and c its intensity (the height).

P_x itself has at least two meanings. One is the absolute count of a feature ($c > 1$), e.g. the absolute number of loanwords, as a cumulative sum. This application is used in section 2 and figs. 1 and 2. The other is the proportion of a feature ($c = 1$), e.g. of a phonetic sequence being rendered in a specific way. This

is the (attempted) application in section 3 and figs. 3–11. See also 3.2 and 4.

2. Data

There can be no doubt that German has exerted a very significant influence over Polish. Yet, to the best of my knowledge, only two comprehensive studies have been written on the subject, both quite recently: Czarnecki (2014), and WDLP. The findings of these studies are inconsistent in several regards. Let us first discuss these studies individually (2.1–2.5), take a brief detour (2.6), and only then make a more direct comparison (2.7).

The Piotrowski-Altmann curve has been fitted to both Czarnecki and WDLP's datasets and also to some of their subsets. The results are shown in Figure 1, and the coefficients are given in Table 1.

Table 1
Coefficients for the fitting of the Piotrowski-Altmann curve to the data from Czarnecki (2014) and WDLP.

Dataset	**A**	**b**	**c**	**R^2**
(a) Czarnecki	1521.28	0.01670	2594.05	0.9900
(b) WDLP	1740.70	0.06671	4048.03	0.9589
(c) ≈ Czarnecki ∩ WDLP	1547.13	0.01909	1255.97	0.9879
(d) ≈ Czarnecki ∪ WDLP	1620.48	0.00840	4404.10	0.9581
(e) WDLP main entries	2334,72	0.00425	11831.79	0.9613
(f) WDLP multiple	A_1= 1545.24 A_2= 1888.98	b_1 = 0.01876 b_2 = 0.02209	c_1= 1933.94 c_2 = 1963.68	0.9963

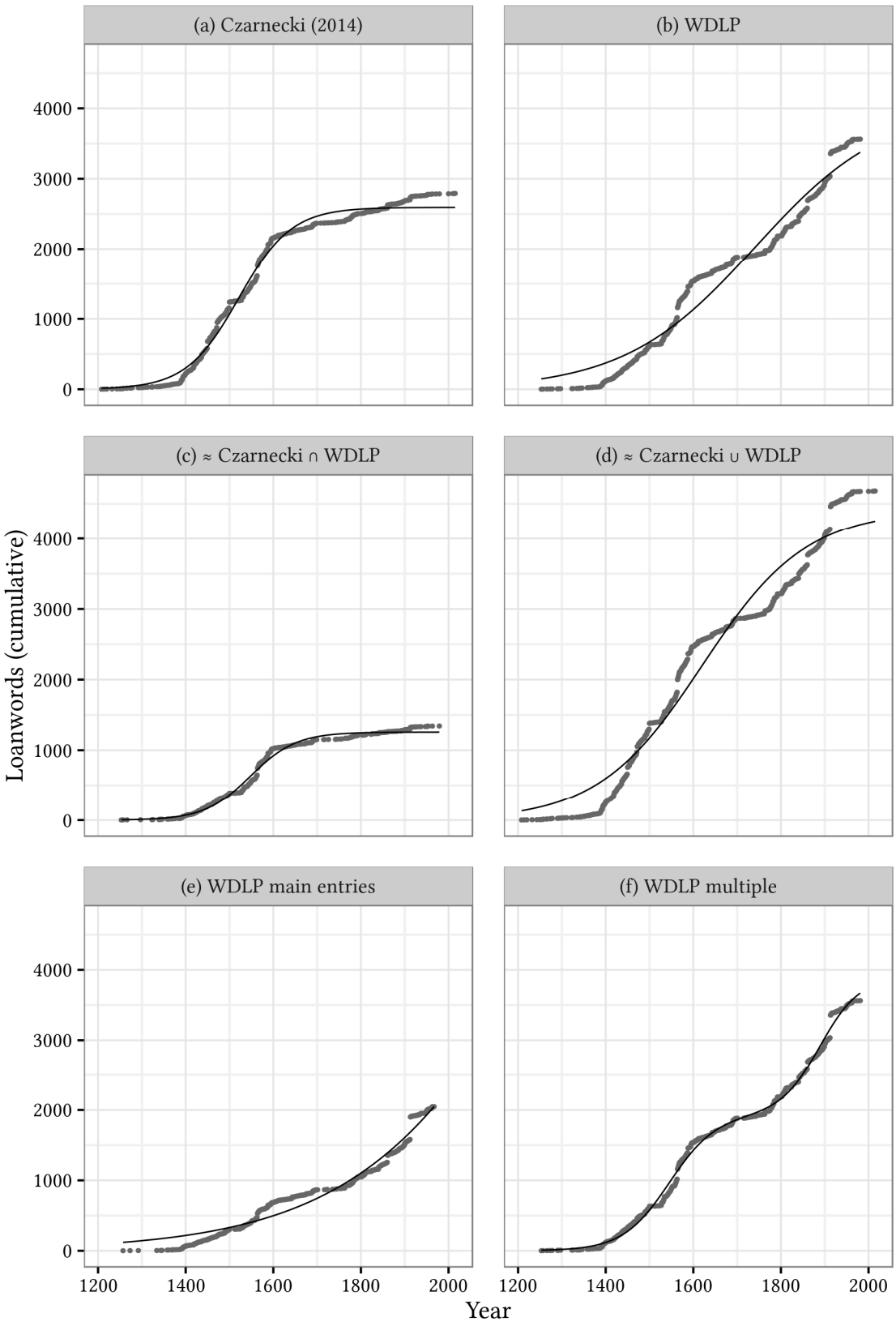


Figure 1. Fitting of the Piotrowski-Altmann curve to the data from Czarnecki (2014) and WDLP. See Tab. 1 and section 2

2.1 (a) Czarnecki (2014)

Let us begin with Czarnecki (2014). His list contains 5800 forms in total: 4936 are actual German borrowings in Polish, and the rest are either words which entered Polish through Czech mediation, Gothic loanwords, or unclear cases which may as well be native formations.

In fact, the list is a little longer because more than thirty words occur twice or even thrice in it, often with different datings and etyma, and in various places in the supposedly alphabetical order. For example, the corpus contains *szlachta*, which is dated 1390, 1378, and 1423, and derived from OHG *slahta*, MHG *slechter herr*, and MHG *slëhte* + OCz. *šlechta*, respectively (p. 267). This inconsistency is just one of the many in Czarnecki's book. Coupled with multiple unclear and unexplained cases (of which there are several examples discussed below), they force me to judge the work as unreliable. This matter will be discussed further in this section but ignored later on.

If we consider direct borrowings, the time frame is 1205–2014. The oldest words are *grenica* (1205, < MHG *grenize* or a native formation), *Roprachtovo* (1208, < MHG *Ruoprëcht*), and *barta* (1213 or 1472, < MHG *barte* rather than < OHG *barta*), whereas the newest are *Aspirin* (2014, < G *Aspirin*), *Autoland* ~ *AutoLand* ~ *Auto Land* (2014, < G **Autoland*), and *Urinal* (2015, < G *Urinal*). In this tiny sample, only *barta* is also listed in WDLP, where it is dated 1472. The form *Roprachtovo* is surprising, not least because of the use of ‹v›, a letter that is essentially completely absent from Polish orthography. In the collective list (p. 252), Czarnecki does not cite any source for *Roprachtovo* (he only cites Taszycki 1974–76 for *Ruprecht* but I could not find *Roprachtovo* in it); on p. 93, however, he cites *Jungandreas* (1928: 173) as the source for *Roprahtovo*, using ‹h› instead of ‹ch›; unfortunately, I was not able to confirm this attestation. The three newest borrowings are even more debatable. Firstly, *Aspirin* is not a common noun, as Czarnecki classifies it, but a proper name trademarked by Bayer; NKJP does not suggest any other use (I am also unaware of it) and, incidentally, dates the first usage to 1997. *Autoland* is likewise a proper name, not used in any other function, dated 1999 by NKJP, and in my opinion not necessarily a German borrowing at all (though it is, contrary to Czarnecki's asterisk, attested in German, as a proper name). Lastly, *Urinal* is also a proper name, of a medicine manufactured by a Czech company, and attested in NKJP since 2005.

Czarnecki (2014) lists German words in Polish, regardless of whether they were borrowed directly or through Czech mediation (though these are marked as such), and of whether they are ultimately of German origin or not. He also does not discern whether the word currently is or ever was included in the literary variety of Polish. As a result, many words are listed multiple times in various renderings. MHG *widerkouf*, for example, is featured no less than thirty times, as *wederkaf*, *wedyrkof*, *widerkuf*, *wyderkaw*, etc. This is not a problem, but it is a design choice that needs to be borne in mind; more on this in 2.4.

Czarnecki provides clear datings for 2792 out of the 4936 words in his list; the rest are given alternate dates, ranges, termini ante or post quos, or are only dated with precision to one century. The clearly dated words, however, conform to the Piotrowski-Altmann law exceptionally well ($R^2 = 0.99$; see fig. 1a and tab.

1). It will be important to note, in light of the coming comparison with the dataset from WDLP, that Czarnecki's data depart from the curve only slightly around year 1800, and generally appear to be a result of essentially a single period of increased influence.

2.2 (b) WDLP

WDLP's dataset consists of 5014 forms grouped into 2446 entries (more on this in 2.4). Unlike Czarnecki, the authors of WDLP chose to only include words that were borrowed directly from German to Polish *and* are native in German – a quite unorthodox decision.

The time frame is 1253–1981. The oldest words are *oberszar* (1253, < MHG *über-schar* or NHG *Überschar* or *Oberschar*), *kram* (1257, < MHG *krâm* or NHG *Kram*), and *lantwójt* (1266, < MHG *lant-voget*); the newest are *branzel* (1975 with a question mark, < NHG *Brandsohle*), *ecie-pecie* (1979, < NHG *Hätschepetsch ~ Hetschepetsch*), and *fakelzug* (1981, < NHG *Fackelzug*).

In this small sample, *oberszar*, *kram*, and *ecie-pecie* are confirmed by Czarnecki (2014), while *lantwójt*, *branzel*, and *fakelzug* are missing. The latter is also missing from NKJP (a Google search for *fakelzug*, however, returns several websites in Polish); *branzel* is found in NKJP since 1965, and *ecie-pecie* since 1997.

Similarly to Czarnecki (2014), WDLP lists both the literary forms of borrowed words, and their alternate adaptations which either remained dialectal or entirely hapax legomena. In a way, it too lists many words multiple times. The above-mentioned MHG *widerkouf* (spelt with a hyphen in WDLP and presented as one of two possible etyma, along with MLG *wedderkop*) has as many as 22 different adaptations. Unlike Czarnecki, however, WDLP presents Polish words using the modern orthography. This is important because, although Polish spelling is fairly historical, it no longer marks vowel length which has been present in Polish for a better part of the history of the language. Its only vestige is today the letter ‹ó› (pronounced [u] = ‹u›) while ‹á› and ‹é› are completely disused. One result of this, for example, is that the difference between *hałda* (dated 19–20th c. in Czarnecki and 1573 in WDLP) and *hałda* (dated 1573 in Czarnecki) is lost entirely. Luckily, such cases are relatively rare: there are 172 examples, out of which only eleven have clear datings.

Out of 5014 words in the WDLP dataset, 3563 are dated with precision to one year. They conform very well to the Piotrowski-Altmann law ($R^2 = 0.9589$; see tab. 1), but it is clear that from fig. 1b that, unlike Czarnecki's data, they represent in fact not one but two periods of increased influence. This fact has not escaped the attention of the authors of WDLP; more on this in 2.5. Interestingly, the *A* coefficient which indicates the point in time when the influx of new words stops to accelerate and then begins to slowly decelerate, falls here on the year 1741 – almost precisely on the midpoint between the two sigmoids, when the influx had actually nearly stalled in reality, only to regain momentum about half a century later.

2.3 (c)–(d) Czarnecki (2014) and WDLP

There are at least two ways in which Czarnecki's and WDLP's sets can be sensibly unified: intersection and union.

For the intersection, I chose those words which are attested both in Czarnecki (2014) and in WDLP, and have the same clear dating in both. There are only 1343 such words, i.e. less than a half of Czarnecki's clearly dated words, and not much more than a third of WDLP's. These words fall into the period 1253–1979, which is a range that is very close to either of the sources' own. They follow the Piotrowski-Altmann curve nearly as closely as Czarnecki's dataset alone, resulting in R^2 of 0.9879 (see fig. 1c and tab. 1). The division into two periods of influence, which is only lightly marked in Czarnecki's data but very clear in WDLP's, is all but invisible here.

For the union, I took the intersection as described above, plus all those words that are only listed in one of the sources, and have clear datings. There are 4683 words in total, 1343 from the intersection, 1291 from Czarnecki (2014), and 2049 from WDLP. Their dates range from the year 1208 to 2015 (the 1205 *grenica* mentioned in 2.1 above was not included because its etymology is uncertain). The more numerous contribution from WDLP appears to have dominated this set, resulting in a curve similar to that of WDLP alone (see 2.2), composed of two, clearly separate periods of influence, and R^2 of 0.9581 (see fig. 1d and tab. 1). See also 2.6 for the results of fitting the Piotrowski-Altmann curve to erroneously unified datasets.

A is the most readily readable of the coefficients in Piotrowski-Altmann law (see 1 above). I would like to point to the discrepancy between the four datasets so far discussed. According to Czarnecki's data, the turning point of German influence on Polish, the point when the influx of new words began to slow down, was in 1521. According to WDLP's data, if one ignores the double-sigmoid shape of the curve, this point occurred considerably later, in 1741. If one considers only those words on which both sources agree, the first view prevails (A = 1547), but if one takes them both at face value and examines their union, the result falls in between, in year 1620.

2.4 (e) WDLP main entries

It was mentioned in 2.2 above that WDLP groups various phonetic variants of a single etymon into main entries. This is a perfectly reasonable practice from the editorial point of view, but let us now consider what implications it has for the linguistic interpretation of a fitting to the Piotrowski-Altmann law.

Out of 2446 main entries, 2047 are dated to within one-year. The dates given by WDLP are those of the oldest variant, not necessarily the one chosen as the main entry. Let us, however, consider the reality of the borrowing. Polish literary language forms in the 16th–17th century; standard German in the 17th–18th; both began to dominate only in the 18th century; in Poland, the marginalization of dialects was nearly complete by the outbreak of World War II, whereas in Germany, the process is ongoing even today. What this means is that for a better part of the history, and therefore about two thirds of the variants listed in

WDLP, the borrowing occurred not so much from German to Polish as from one of German dialects to one of Polish dialects. Later, when those dialects produced between them the so-called literary variety, this new dialect may have or may not have inherited one specific rendering. The others remained generally restricted to dialects. Singular cases may have entered the literary language whilst it was not yet quite as rigid as it is today, and co-existed with other renderings for some time, but ultimately they too were ousted. Therefore, for those pre-literary words it is more correct to consider not the oldest variant, but only the one which eventually won out. For newer borrowings, the situation is in fact very similar.

Let us apply the Piotrowski-Altmann equation to the earliest attestations of those variants which the authors of WDLP chose for main entries. The fit is very good, $R^2 = 0.9613$ (see fig. 1e and tab. 1), but interestingly, it is actually only the initial part of the curve that covers the entire dataset. According to this prognosis, German influence on Polish is only gaining speed at the moment; the influx of new words will keep intensifying until the year 2335, and it will not slow down noticeably until around the 34th century, by which time Polish will have borrowed more than 11,000 words in total.

See 2.7 for a little commentary on the methodological question posed by this result.

2.5 (f) WDLP multiple

Let us return to considering all the attested variants, not just the ones chosen for the main entries. It was mentioned in 2.2 above that the dates given by WDLP seem actually to form not on one but two sigmoid curves, reflecting two separate periods of increased influence. The authors of WDLP have also noticed this characteristic; see especially Hentschel (2001, 2009). Indeed, having read those papers, I expected that a single sigmoid would not fit unbinned data very well. I was wrong, as is attested by $R^2 = 0.9589$. But it is also true that a sum of two sigmoids fits them exceptionally well; with $R^2 = 0.9963$ (see fig. 1 and tab. 1). A similar result was obtained with Turkic glosses in Polish (Stachowski K. 2013: 113f) and, I expect, can be obtained with the frequency with which the word *terrorismo* has appeared in end-of-year speeches of the presidents of Italy (Köhler/Tuzzi 2015: 117). All of these datasets pose a methodological question which we shall return to in 2.7.

2.6 Excursus: Garbage in, gospel out

While preparing the union of Czarnecki's and WDLP's datasets, I made a mistake and counted their intersection twice. The resulting dataset contained 6355 elements and proved to follow the Piotrowski-Altmann curve quite closely; in fact marginally closer than the result of the correctly performed union: $A = 1583.02$, $b = 0.01075$, $c = 5702.31$, and $R^2 = 0.9663$ (as compared to 0.9581, see Tab. 1).

Intrigued, I experimented with a few other datasets unified in a somewhat irrational manner, and found that the Piotrowski-Altmann law described them all with an astounding accuracy, R^2 not dropping below 0.96. An example is presented in fig. 2a and tab. 2; it is the union of three datasets: Russian borrowings

in Aleut (based on Bergsland 1994), Ottoman borrowings in Hungarian (based on Kakuk 1973; see also Stachowski 2013), and WDLP. For comparison, in fig. 2b, is a fitting of the curve to an entirely random sample of 200 numbers from the interval [1, 100]. The fit is in fact a little better, $R^2 = 0.9888$; see tab. 2. I was able to further improve it to about 0.993 (based on one hundred trials) by binning the data into 10-year intervals, and to about 0.9964 by binning into 20-year intervals.

Conclusions from this little experiment are in 2.7.

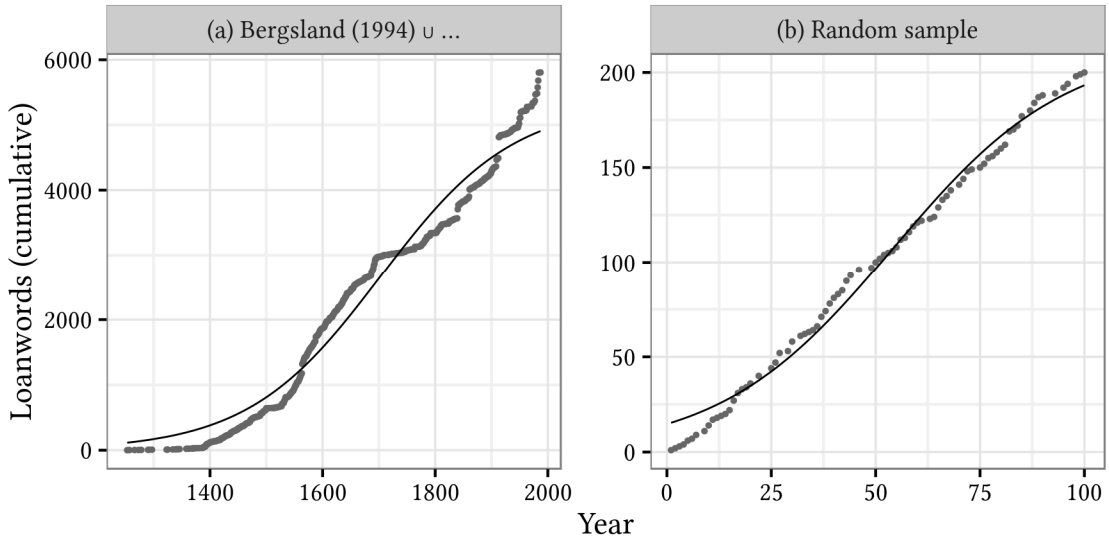


Figure 2. Fitting of the Piotrowski-Altmann curve to (a) the union of datasets from Bergsland (1994), Kakuk (1973), and WDLP, and (b) a random sample. See tab. 2 and subsection 2.6.

Table 2
Coefficients for the fitting of the Piotrowski-Altmann curve to (a) the union of datasets from Bergsland (1994), Kakuk (1973), and WDLP, and (b) a random sample. See fig. 2 and subsection 2.6.

Dataset	***A***	***b***	***c***	R^2
(a) Bergsland ∪ …	1703.57	0.00846	5351.52	0.9698
(b) Random sample	54.14	0.04818	214.48	0.9888

2.7 Observations

The first observation is about the reliability of sources (see 2.1–2.3). Czarnecki (2014), unfortunately, must be thoroughly reviewed before it can be used. Figures 1a–d serve to show the discrepancy between Czarnecki and WDLP. An interesting point is the lack of the second, (19th-century) sigmoid in the intersection of the two datasets (fig. 1c). It raises the question of why this intersection is consistent with Czarnecki's data but inconsistent with WDLP's. The two comple-

ments would need to be very carefully examined in order to formulate an answer, but, in light of Czarnecki's shortcomings, I think this may not be necessary. Nonetheless, I should like to repeat here that I find WDLP's two decisions: to present all words in the modern orthography with vowel length marking omitted, and to exclude words which are not native to German, to be quite unfortunate.

Other than that, it may be said that the two datasets contain a large number of examples (2792 in Czarnecki (2014), 3563 in WDLP) and they cover a very similar period from the 13th to the 20th/21st century, but that the pictures that they paint of the chronology of German influence on Polish are rather different. They agree about the first wave of borrowings, the one that peaked in mid-16th century, but WDLP then shows another wave culminating in the mid-19th century, which Czarnecki (2014) almost entirely omits. Indeed, to see the error in the latter view, one needs to merely remember that since late 18th century all of western and a part of central Poland was annexed by Prussia and has remained under German control till the end of World War I.

The second observation pertains more to the application and interpretation of the Piotrowski-Altmann equation than to the datasets themselves (see 2.4). It seems that one should be able to freely choose whether to include all the various phonetic renderings of an etymon, ephemeral as they may be, or to limit oneself to just the more 'successful' variants. In fact, WDLP makes a note if a specific rendering is a hapax legomenon or only attested in a single source. But the two subsets, of words with and without such annotations, both appear to be quite representative of the entire dataset in that they clearly fall on a double sigmoid, and both yield sensible results when the Piotrowski-Altmann curve is fitted to them (words not marked (2274 examples): A = 1749.76, b = 0.00633, c = 2608.28, R^2 = 0.9495; words marked (1289 examples): A = 1721.93, b = 0.00731, c = 1429.97, R^2 = 0.9657). This is not the case when only the main entries are taken into account. Visually, they may appear to fall on a curve fairly similar to the one drawn by the entire dataset, but the results of the fitting are quite different (fig. 1e, tab. 1). Only time, specifically the next millennium, can tell whether they will prove more accurate.

The third observation returns to the question of fitting multiple curves to a single dataset (see 2.5). This possibility was also mentioned in Stachowski (2013: 113f) with reference to Turkic loanwords in Polish, but the case discussed here is considerably clearer. The influx of German words into Polish can be approximated by a single sigmoid to a high level of accuracy, but it is clear from figures 1b and f that the data actually fall on two sigmoids, and this raises the question whether it should. Perhaps both approaches are valuable in their own ways. The multi-curve can pinpoint the exact moments in time when the influx of new words peaked, adding precision to a more historically-oriented perspective, while a single curve shows that even an influence which is clearly composed of two separate waves does overall follow the epidemic curve, bringing a minor new insight to the more quantitatively-oriented outlook.

The fourth and last observation will be perhaps more helpful to those who do not deal with quantitative analyses of linguistic data on an everyday basis. It was shown in 2.6 that the Piotrowski-Altmann law can describe linguistically

nonsensical data, or indeed random data, as accurately as an honest, sound dataset such as that of WDLP. One may be tempted to say 'Well, this just fits anything, then!' and disregard the idea altogether as being too broad to be able to produce an actual insight, but this would be a mistake. Firstly, the Piotrowski-Altmann curve does not in fact fit any odd bit of linguistic data, as will be demonstrated in section 3. It would also not be much use if it only fitted some of the appropriately prepared datasets but not others. Secondly, a ruler is not broken just because it can measure a broken chair. The Piotrowski-Altmann law describes how change progresses in a language, the coefficients of the regression capture numerically the time, the rate, and the intensity of the change, but a single equation cannot be expected to validate the linguistic sense of the data it was fed.

3. Method

The area of application of the Piotrowski-Altmann law is not limited to the absolute number of loanwords. Among others, it has been used with good or very good results to describe morphological changes in German verbs (Best 1983), the shift from *ward* to *wurde* in German (Best/Kohlhase 1983, Kohlhase 1983), *e*-epithesis in German verbs (Imsiepen 1983), the shift from *vi* to *ci* in Italian (Köhler/Tuzzi 2015), or the relation between grammatical markers (including fixed word order) and the number of word classes (Vulanović 2013). What these analyses have in common is that in all of them the measured value is a proportion, as opposed to a cumulative sum of absolute counts.

When reading through WDLP, it is quite evident that certain elements, sequences of sounds or parts of compounds, occur more frequently than others, and are not always rendered in the same way. I selected a dozen such elements. Let us first take a closer look at them (see 3.1), and then make some observations about the Piotrowski-Altmann law that they inspire (see 3.2).

3.1 Adaptations

Out of the dozen features selected for examination, five are elements of compounds (*-eisen*, *-haus*, *-holz*, *-meister*, *-stein*), four are sounds or sequences of sounds (‹ei›, *ke*, *l*, *VNC*), and three are affixes (*-er*, *ge-*, *-ung*). Let us look at them in the alphabetical order.

3.1.1 *‹ei›*

The graphic sequence ‹ei› appears in the etyma of 720 words in WDLP's dataset. I discarded words with significantly unclear etymology, and those in which the ‹ei› was entirely omitted (e.g. *blumistyka* < NHG *Blumisterei*, *ząząel* < NHG *Zaumseil*), or it was inside *-meister*, *-stein*, or *-eisen* as these, owing to their frequency, appear to have received special treatment, and will be discussed separately. This left 503 words containing 508 instances of ‹ei›. The adaptations attested in them are: *aj* (150 examples), *ej* (138), *y* (80), *e* (64), *a* (36), *i* (29), *u* (8), *ę* (1), *ij* (1), and *yj* (1). Of these, 333 instances are clearly dated.

Because the number of different renderings is quite high, fig. 3 shows only one aspect: whether the final *-j* of the original diphthong has been preserved. An evolution in time is clearly visible but, so far as the proportions are concerned, it hardly resembles the sigmoid of the Piotrowski-Altmann law.

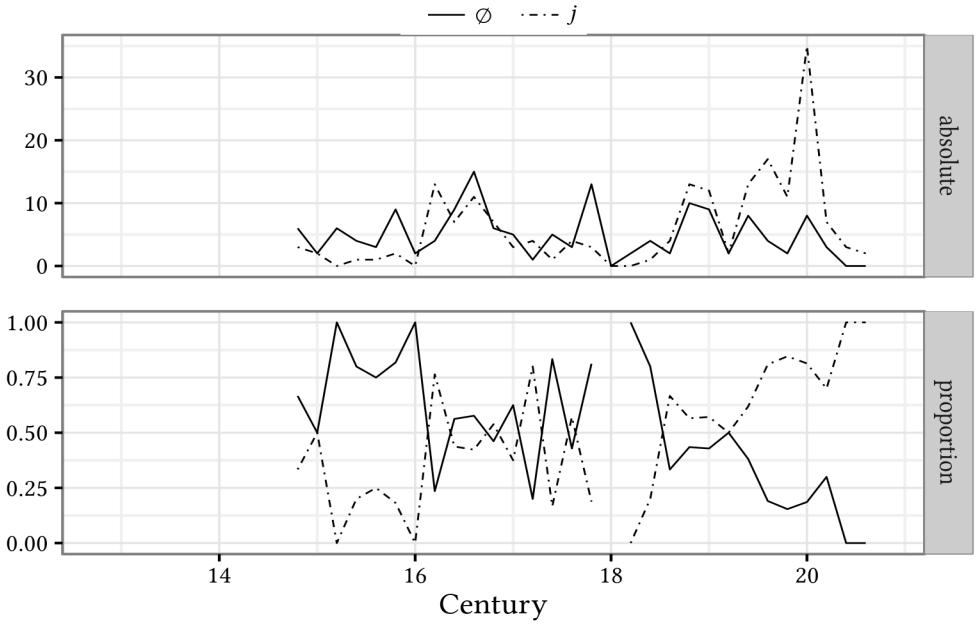


Figure 3. Rendering of *-j* in ‹ei›. Data binned into 20-year intervals. See 3.1.1.

3.1.2 *Eisen*

There are borrowings of 47 compounds with *Eisen*. I discarded three words: *arzenagiel* < NHG *Eisennagel*, *bygla* < NHG *Bügeleisen*, and *hulejza*/*hulejsa* < NHG *Hohleisen*. The adaptations in the remaining 44 are: *ajza* (18 examples), *ajz* (10), *eza* (6), *ajzen* (2), *ajzy* (2), *ejza* (1), *ejzen* (1), *ejze* (1), *es* (1), *ezyja* (1), and *yza* (1). All are clearly dated.

The set is nevertheless quite small and a graph does not reveal much more than that *ajz* is concentrated mainly around the late 18th century, while *ajza* is concentrated around the late 19th with occasional appearances in the 17th and 18th centuries. The whole dataset does not seem to follow any specific pattern. For most intervals, it is even pointless to calculate proportions because they have just one borrowing in them – or none at all.

3.1.3 *-er*

There are 845 words in WDLP whose etyma end in *-er*. I discarded those where the rendering could not be clearly established or was altogether missing (as in *demfrować* < NHG *Dämpfer*, *fercel* < NHG *Feldscher*, or *sztabstrębacz* < NHG *Stabstrompeter*), and where it was inside *meister* which appears to have been treated differently and is discussed separately. This left me with 704 examples. The adaptations in them were: *er* (377 examples), *arz* (97), *ar* (57), *ra* (37), *erz* (37), ∅ (28), *r* (20), *ir* (11), *el* (6), *irz* (6), *or* (6), *ry* (4), *y* (4), *yr* (4), *a* (2), *ro* (2), *ur* (2), *yrz* (2), *ery* (1), and *usz* (1). Of these words, 504 are clearly dated.

The primary opposition is *-er* against all the other renderings. The absolute

number of borrowings is perhaps less obvious, being distributed between the two spikes in the influx of German loanwords into Polish, but the proportion reveals the trend very clearly, as can be seen in fig. 4. It is, however, not at all similar to the curve of the Piotrowski-Altmann law.

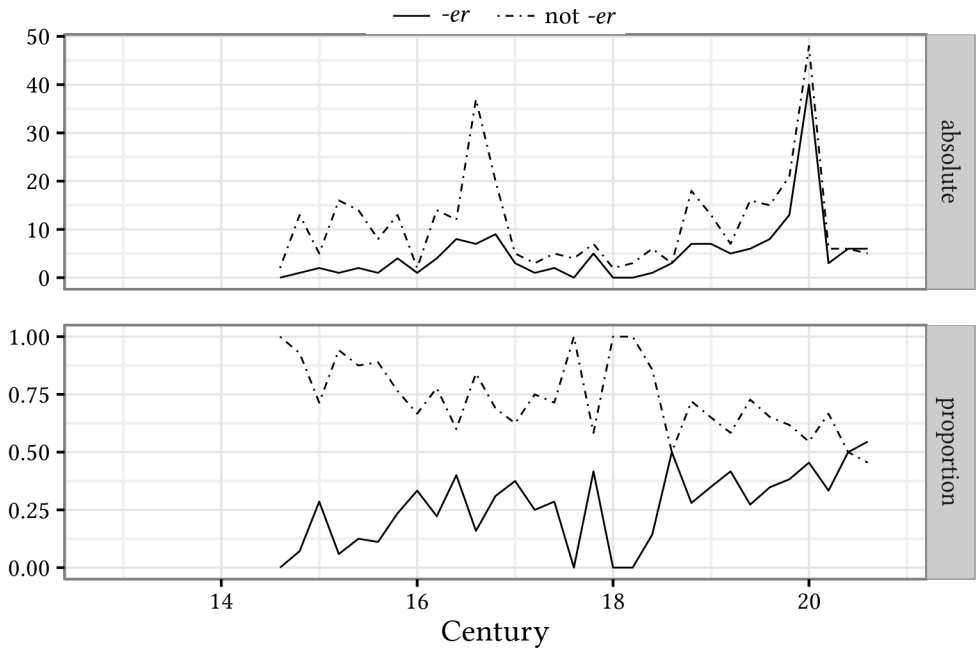


Figure 4: Rendering of *-er*. Data binned into 20-year intervals. See 3.1.3.

3.1.4 *ge-*

There are 40 words in WDLP whose etyma begin with *ge-*. I discarded those where the beginning was not the *ge-* prefix (e.g. *gierować* < NHG *gehren*, *gilować* < NHG *geilen*), and was left with 31. In 24 cases, the *ge-* was rendered as just *g-*, and in seven as *gV-* (*ge-* four times, *gie-* twice, and *ga-* once). 28 words are clearly dated.

The dataset is quite small but it has nonetheless proved sufficient for a graph: fig. 5. Assuming that the drop in *g-* at the beginning of the 16th century is accidental and meaningless, no more can be deduced from the graph than that *gV-* has effectively always been the less popular choice.

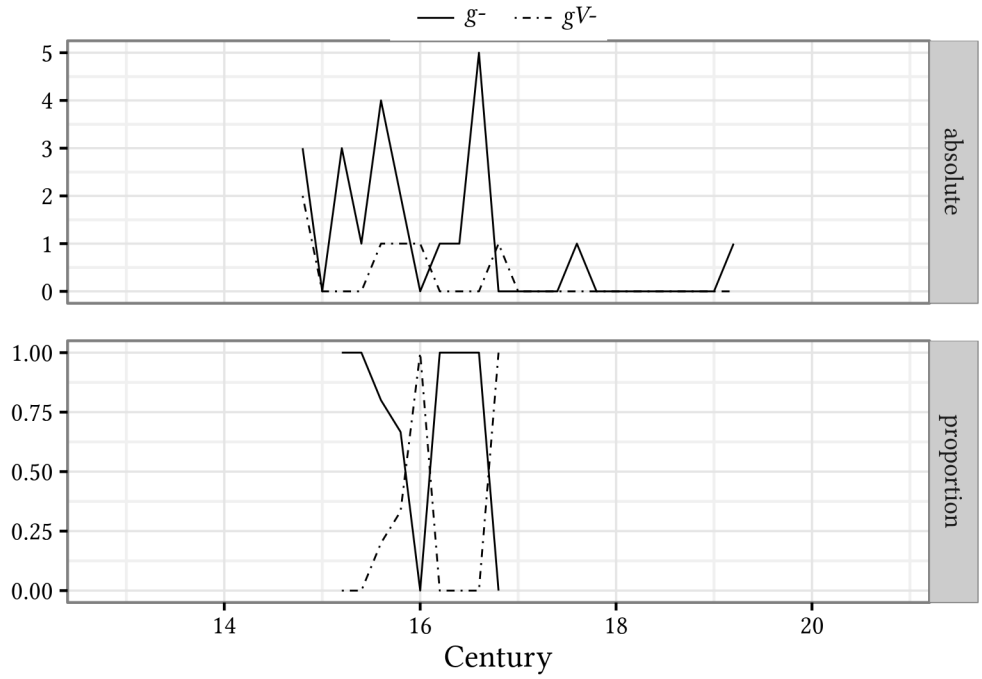


Figure 5: Rendering of *ge-*. Data binned into 20-year intervals. See 3.1.4.

3.1.5 *-ge-, -ke-*

There are 839 words with etyma containing the *g*/*k* + *ae*/*æ*/*e*/*ë* sequence. I discarded those where the sequence was entirely omitted (e.g. *bigajza* < NHG *Bügeleisen*, *ćwikać* < NHG *zwicken*, on account of *-ać* being the Polish infinitive suffix), or where it was: in auslaut, in particular as a part of the *-unge* suffix, followed by the *-er* suffix, or in the *ge-* prefix (the last three appear to have been treated differently, and are discussed separately). This left me with 322 instances in 315 words. The adaptations are: *kie*/*gie* (148 examples; labelled *K´E* in fig. 6), *k*/*g*/*h* (91; labelled *K*), *ke*/*ge* (55; labelled *KE*), *ka*/*ga* (16), *ki*/*gi* (not followed by *e*; 7 examples), *ko* (2), *go* (1), *he* (1), and *że* (1). Of these, 206 are clearly dated.

In the interest of readability, both data and adaptations are binned in fig. 6 (see the paragraph above), and only the top three groups of renderings are included, which comprise 92.7% of the clearly dated examples. In truth, the graph does not seem to clarify much. During the first wave, *K* and *KE* appear to have been almost equally popular while during the second, *KE* dominated all the other renderings. None of the adaptations even approaches the curve of the Piotrowski-Altmann law.

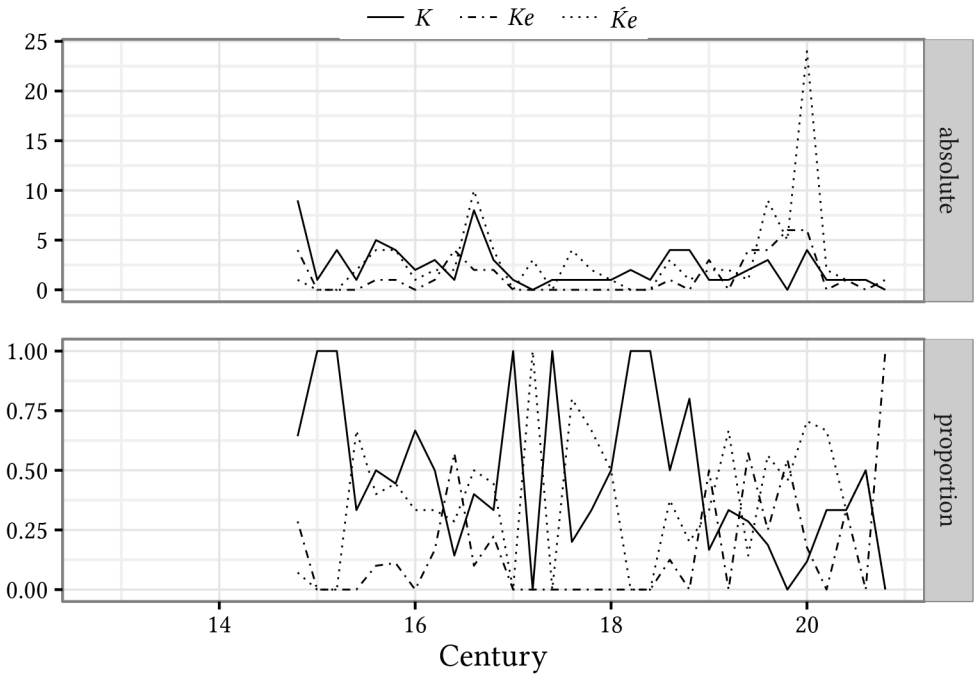


Figure 6: Rendering of *-ke-*. Data binned into 20-year intervals. See 3.1.5.

3.1.6 *Haus*

There are 88 borrowings of compounds with NHG/SilG *Haus* or MHG/MLG *hûs*. None needed to be discarded (including the nine where *Haus* was the first part of the compound). The renderings are: *hauz* (24 examples), *uz* (15), *auz* (14), *haus* (9), *us* (7), *aus* (5), *usz* (4), *hus* (2), *huz* (2), *ans* (1), *ausz* (1), *has* (1), *os* (1), *uza* (1), and *uż* (1). 57 words are clearly dated.

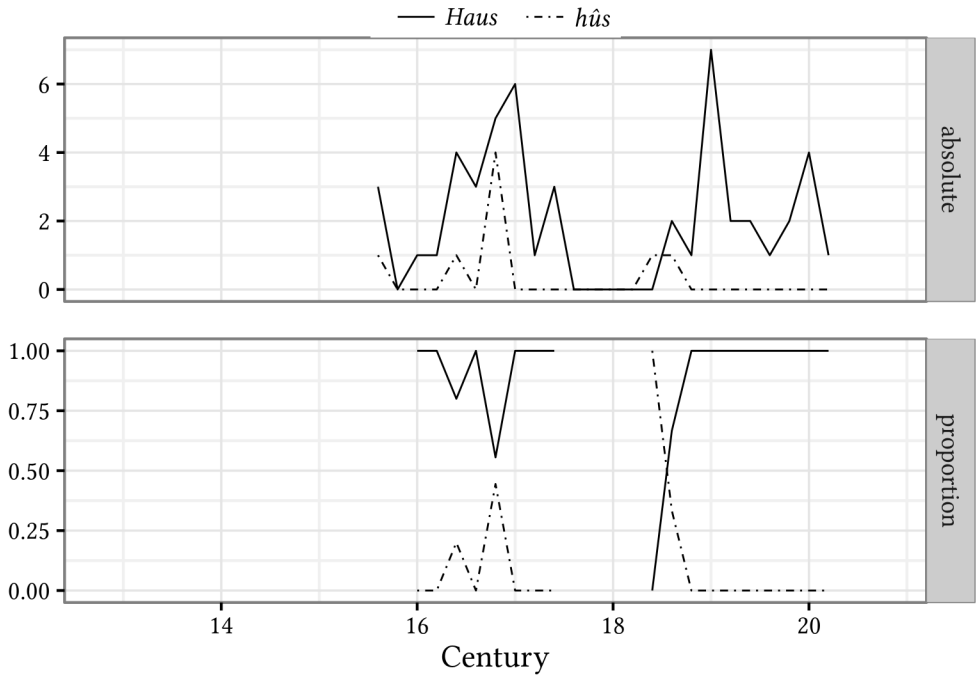
This is not a large dataset but the adaptations are quite diverse and there are too many to be legible in a single graph. It is clear that they are concentrated in two periods, the first ranging from the mid-15th to the mid-17th century, and the other from the early 18th century to the early 20th. This is consistent with the general picture painted by WDLP's data. In both periods, the different renderings are distributed approximately evenly, except for *hauz* becoming a little more prominent in late 19th and early 20th century.

Graphs of single features, such as the preservation of *h-*, rendering of the vowel, or of the final consonant, do not appear to be any more informative. Perhaps the vowel is the most interesting. WDLP reduces all the loanwords to just two etyma, MHG/MLG *hûs* and NHG/SilG *Haus*. One might expect that the first would be most prominent in the earlier wave, the second in the later one, and that the vowel in the Polish rendering would reflect that. But this is not quite the case, as can be seen from figs. 7a and b (for legibility, the latter shows only *au* and *u*, i.e. 96.5% of the clearly dated examples, and disregards the singular cases of *a*, *an* and *o*).

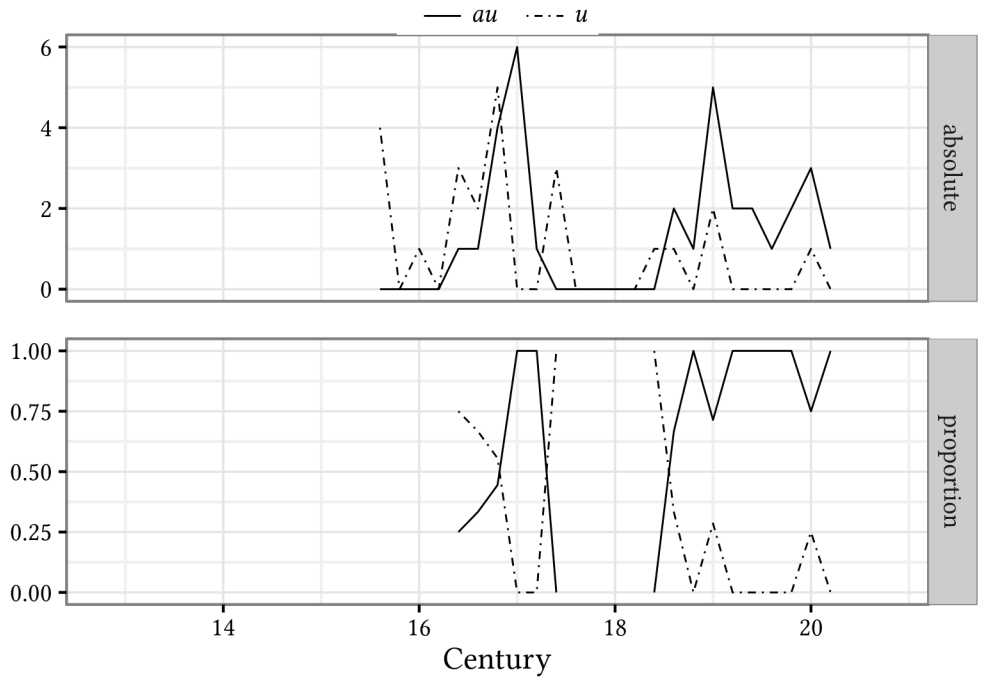
As was mentioned in 2.2, WDLP groups the various phonetic variants of a single word into entries. An etymology is only given for the entry as a whole. In

particular, an etymon with *hûs* is suggested for ten variants, nine of which are rendered with *u*. The exception is *rathaus*, dated 1767 and listed s.v. *ratusz* < MHG *rât-hûs*. It is perhaps most reasonable to view this one case as an omission on the part of WDLP, and conclude that MHG *hûs* was rendered in Polish with *u*.

With *Haus*, however, the situation is less clear. There are 78 phonetic variants derived from *Haus* (including both those with and without a clear dating); 52 are rendered with *au*, 23 with *u*, and one each with *a*, *an* and *o*. The twenty-three rendered with *u* are grouped into six entries. It does not seem very likely that they are all omissions such as *rathaus*. Perhaps they were influenced by the knowledge of previous borrowings of *Haus* that had been rendered with *u*? See also 3.1.9.



(a) The etyma according to WDLP.



(b) Rendering of the vowel

Figure 7. Adaptation of *Haus*/*hûs*. Data binned into 20-year intervals. See 3.1.6.

3.1.7 *Holz*

There are 49 borrowings of compounds with *Holz*. I discarded those where it is unclear exactly which part of the Polish word is the rendering of *Holz*, or it is absent altogether; e.g. *strycholec* < MHG *strich-holz*, NHG *Strichholz*, *Streichholz* (the *ch* may be from *stri<u>ch</u>* or *<u>h</u>olz*), or *wałek* < NHG *Walkholz*, *Walkenmangel*. The five in which *Holz* was the first part of the compound, I preserved. This left me with 38 examples, in which the following renderings are attested: *ulec* (20 examples), *holc* (11), *olc* (3), *holec* (1), *hulc* (1), *olec* (1), and *ólc* (1). Of these, only 27 are clearly dated.

Very little can be said based on this small dataset. *holc* is the most frequent during the first wave in the 15th century, while the second wave, from mid-18th to early 20th century, is dominated by *ulec*. No real patterns can be seen.

3.1.8 *l*

There are 1992 words containing 2002 instances of *l* in their etymon. I did not discard any. In 1895 examples, the rendering was *l*, and in 107 it was *ł*. 1406 instances are clearly dated.

The fairly high proportion of *ł* that can be observed in the initial phase in fig. 8 is probably accidental. Overall, the number of borrowings with *ł* remains fairly constant, and it seems that it is only due to the generally low number of loanwords in this period that *ł* happened to come to relative prominence. A similar spike in the proportion can be observed in early 18th century. If any pattern is to be seen here, it is close to constant.

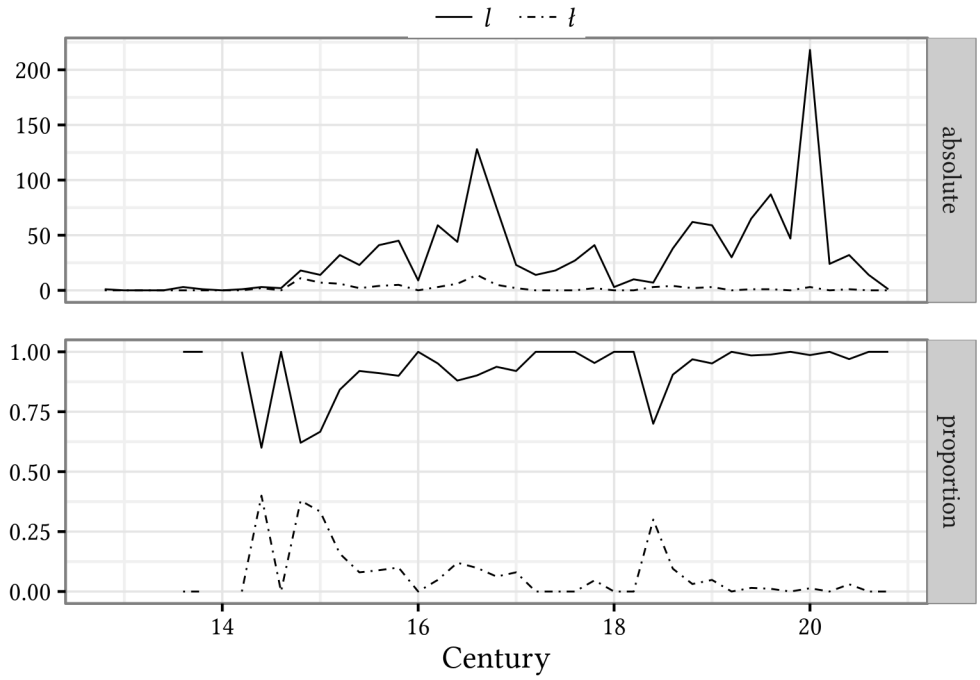


Figure 8. Rendering of *l*. Data binned into 20-year intervals. See 3.1.8.

3.1.9 *Meister*

There are 90 borrowings of compounds with *Meister*. I discarded one where this part was entirely omitted (*firtelnik* < NHG *Viertelsmeister*), and three where the spelling was unclear (*maister* and twice *meister*), but preserved seven in which *Meister* was the initial part of the compound, leaving in total 86 words. The adaptations are: *mistrz* (36 examples), *majster* (29), *mejster* (10), *magister* (6), and *mistr* (5). Of these, 57 are clearly dated.

We will start by looking at the adaptations themselves. For simplicity, I binned them into three groups: *magister*, *mAjster* (= *majster*, *mejster*), and *mistRZ* (= *mistr*, *mistrz*). Before we begin, let it be reminded that WDLP by design excludes words which are not native in German (see 2.2), and therefore all of its data and etymologies pertain to compounds containing *Meister*, but not to *Meister* as an independent word as this is a borrowing of Lat. *magister* (Kluge).

Let us then first look at *magister*. The word is missing from Boryś (2005) and Brückner (1927), but according to Bańkowski (2000), it is attested since the 15th century and originally meant ‘head; manager; commander’. This is consistent with the earliest attestation in WDLP which is from year 1405, inside *och-magister* < MHG *hove-meister* or NHG *Hofmeister*. It is nonetheless quite clear that *magister* cannot be merely an adaptation of G *meister* to Polish phonetics. It must either be a calque, or borrowed from an earlier German form, one more similar to the original *magister*. The former seems to be a more likely explanation.

As for *mAjster*, etymological dictionaries quite unanimously derive it from G *Meister* (except for Brückner 1927 who is somewhat unclear), but Bańkowski (2000) and Boryś (2005) date it to the 18th century, Czarnecki (2014) to the 16th, while the oldest attestation in WDLP is from 1334, but it is inside *berkmejster* < MHG *bërcmeister*.

There is a little more uncertainty about the origin of *mistRZ*. Bańkowski (2000) derives it from OCz. *mistr* or MHG *meister* / MLG *mēster*, Boryś (2005) from OCz. *mistr*/*místr*/*mistř*, and Czarnecki (2014), together with OCz. *mistř*, from OSilG *mistr*/*mestr* (Brückner, again, is unclear). The datings, however, are more unanimous: Bańkowski to the 13th/14th century, Boryś to the 15th, Czarnecki to 1390, and WDLP to 1368, inside *barkmistrz* < MHG *bërcmeister*.

All in all, it seems that it is in fact only *mAjster* that came to Polish directly from German. Those loanwords in which *meister* was rendered either as *magister* or as *mistRZ* can be treated as borrowings of entire compounds, or maybe as borrowings of only the other part of the compound, i.e. of *och*-, *berk*- and *bark*- in the examples above, but not quite as borrowings or renderings of *meister* itself.

Nevertheless, one could hope for a pattern to emerge when the data are graphed. Fig. 9 shows the distribution, from which unfortunately no patterns seem to emerge. The overall tendencies are quite clear but the curves do not in any way resemble that of the Piotrowski-Altmann law.

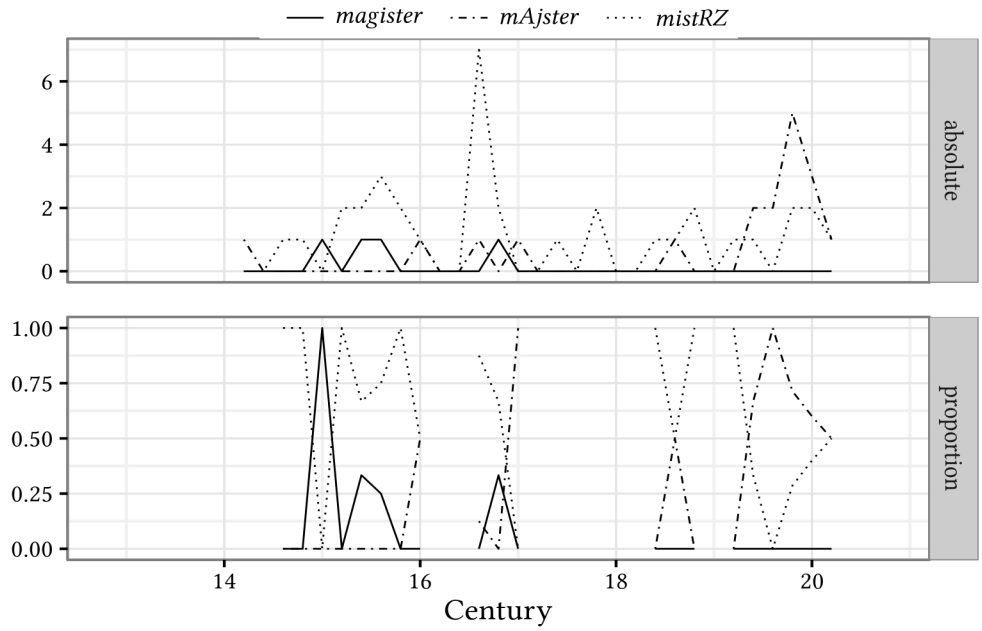


Figure 9. Rendering of *Meister*. Data binned into 20-year intervals. See 3.1.9.

3.1.10 *Stein*

There are 69 borrowings of compounds with *Stein*. I discarded those in which this part was omitted or unclear (e.g. *burztyn* < Eastern MG *bornstein*, *burnstein*, or *strab* &c. < NHG *Strebstein*) but included five cases with *Stein* as the initial part of the compound. This left me with 59 words with surprisingly diversified adaptations: *sztyn* (21 examples), *sztejn* (11), *sztajn* (10), *styn* (5), *sztyna* (3), *stejn* (2), *stin* (2), *sztan* (2), *stan* (1), *styna* (1), and *szten* (1). Of these, 32 are clearly dated.

There are too many different renderings for them to be legible in a graph, and I cannot think of a way to reasonably bin them. Graphs of single features (*s* vs. *sz* in anlaut, the middle vowel, or the gender) do not seem to be any more in-

formative. No meaningful tendencies or patterns can be observed.

3.1.11 *-ung*

There 123 words with etyma ending in *-ung* or *-unge*. I discarded those in which this part was omitted, where it was not the *-ung(e)* suffix, or the rendering was unclear; e.g. *cuhalt* < NHG *Zuhaltung*; *filun(e)k* < NHG *Füllung*; *jung* < NHG *Schiffsjunge*. This left 107 words. The adaptations are: *unek* (46 examples), *unk* (19), *ung* (13), *uga* (12), *ąg* (5), *ynek* (3), *ęg* (2), *ynk* (2), *anek* (1), *ang* (1), *onk* (1), *ónek* (1), and *unga* (1). Of these, 74 are clearly dated.

For simplicity, I binned the different renderings as follows: *UGA = uga*, *UNEK = anek, unek, ónek, ynek, UNG = ang, ąg, ęg, onk, ung, unk, yng, UNGA = unga*. Now, a part of the etymologies point to NHG *-unge*, a part to MHG *-ung*, and a part to both. It seems, however, that the distinction has had little to no impact on the Polish rendering: out of 19 variants with *-unge*, only one yielded *UNGA*, while the remaining eighteen are divided equally between *UNG* and *UNEK*. Out of 15 where WDLP allowed both possibilites, ten resulted in *UNEK* and five in *UNG*. The time of borrowing, it would appear, was also not decisive in any way. Fig. 10 shows the two largest groups (comprising 91.2% of the clearly dated examples) whose distribution does not follow any particular pattern.

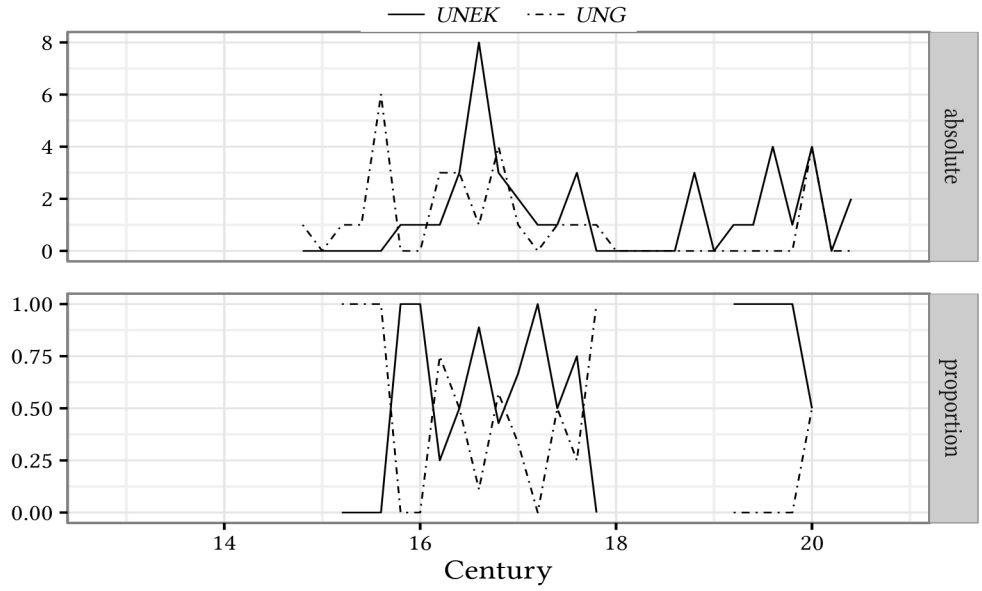


Figure 10: Rendering of *-ung*. Data binned into 20-year intervals..

3.1.12 *VNC*

There are 1268 words with the *VmC*/*VnC* sequence. I discarded those which had the sequence in an- or auslaut (based on the Polish rendering, i.e. including e.g. *antaba* < NHG *Handhabe*), where it was entirely omitted or unclear (e.g. *baleder* < NHG *Ballenleder*, or *biusthalter* < NHG *Büstenhalter*), where it was inside *Zange* used as the final element of a compound (due to their special treatment; 17 cases) or inside the *-ung(e)* suffix (likewise, discussed separately), and also 13 out of 14 different variants of *krużganek* (< MHG *kriuzganc*, NHG *Kreuzgang* or MG *krūzeganc*; the ending was always rendered as *-ganek*). This left 922 in-

stances in 898 words; in 785, the sequence was rendered more or less accurately as *Vm* or *Vn* (labelled *N* in fig. 11), in 98 it yielded a nasal vowel (labelled *Ą*), in 38 the nasality was omitted (labelled Ø), and in one case it was preserved as both a nasal vowel and a nasal consonant at the same time (*mąnsztuk* < NHG *Mundstück*).

Fig. 11 shows the top three groups comprising 99.9% of the clearly dated examples. The bulge in the proportion of ∅ in mid-15th century, and likewise the spike in early 18th century, are probably accidental owing to the generally fairly low number of examples in those periods. Apart from the general domination of *N*, hardly any tendencies or patterns are to be seen.

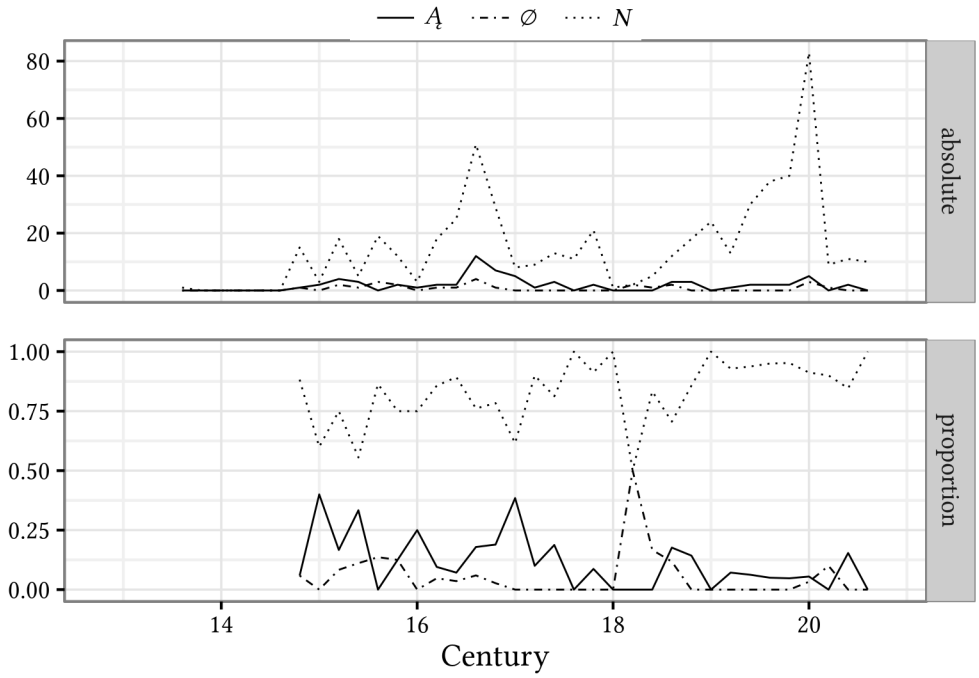


Figure 11. Rendering of *VmC*/*VnC*. Data binned into 20-year intervals.

3.2 Observations

The easiest to observe in the above analyses is the fact that the proportions of various renderings simply do not follow the Piotrowski-Altmann law. Admittedly, the situation here is not quite the same as in the several papers mentioned at the beginning of this section; what they described is the process of one form completely replacing another, while in our examples such shifts do not occur – except perhaps in 3.1.3 (fig. 4) where *-er* appears to be gradually ousting all the other renderings of G *-er*. This dataset can be approximated by the Piotrowski-Altmann equation (A = 21.51, b = 0.27888, c = 1) but only poorly, with R^2 = 0.3455. In all the other cases, the proportions do sometimes remain relatively constant, within a certain range, but mostly, they fluctuate and swap places – but without following the Piotrowski-Altmann curve. It seems that perhaps a stricter definition of the applicability of the law is required.

Naturally, if one considers the cumulative sums of absolute counts of variants with various adaptations, and limits oneself to relatively frequent renderings

(ten or more examples), the data follow the Piotrowski-Altmann curve quite closely, with R^2 ranging between 0.9076 and 0.9855. The only exceptions appear to be *holc* (see 3.1.7) and *Ke* (3.1.5) which insist on being treated as two separate sigmoids, and *mAjster* (3.1.9) which is all but missing until the mid-19th century, at which point the cumulative sum starts to grow almost exponentially. Overall, this is hardly a surprising result, since all of these datasets are more or less random subsets of a dataset which follows the Piotrowski-Altmann law with R^2 of 0.9589 (see 2.2).

The last observation is that the frequencies of various renderings appear to follow a power law distribution. Since I cannot produce a theoretical justification for this phenomenon, I limited myself to just one test using what is perhaps the most popular, the Zipf-Mandelbrot distribution. The results for adaptations with five or more renderings are promising, yielding on average R^2 of 0.9753. I expect this could be improved even further if the theoretical background allowed for a more sophisticated approach, such as in Popescu/Altmann/Köhler (2010). A report containing also examples from other languages will be published (Stachowski [forthcoming]).

4. Conclusions

German vocabulary has permeated the Polish language for more than a thousand years now, and the total number of borrowings exceeds five thousand. Despite these impressive numbers, I am aware of only two attempts at a comprehensive analysis: Czarnecki (2014), and WDLP. Both have their shortcomings, and only the latter can be considered reliable (see 2.1, 2.2, and 2.7). According to this, there have been two distinctly separate periods of increased influence. The first peaked in mid-16th century; the second, progressing slightly faster but resulting eventually in approximately the same number of borrowings, peaked in late 19th century (2.5, and 2.7).

The Piotrowski-Altmann law can be used to quantitatively describe both the whole of German influence on Polish, and the two waves in separation. It can also be used to approximate various other datasets, regardless of their linguistic significance, but one should not be discouraged by this (2.6, and 2.7). The law cannot, however, describe how the Polish rendering of various phonetic and lexical parts of German words changed over time (3.1). This demonstrates that its field of relevance may need to be defined more precisely, especially in light of the already existing dichotomy in its interpretation (1, and 3.2). On the other hand, the frequencies of these various renderings appear to be following a different law; a test fitting to the Zipf-Mandelbrot distribution yields promising results (3.2).

Abbreviations

Cz. = Czech; **G** = German; **H-** = High; **L-** = Low; **Lat.** = Latin; **M-** = Middle; **N-** = New; **O-** = Old; **Sil-** = Silesian

References

Bańkowski A. (2000). *Etymologiczny słownik języka polskiego*. Warszawa.

Bergsland K. (1994). *Aleut Dictionary. Unangam Tunudgusii. An unabridged lexicon of the Aleutian, Pribilof, and Commander Islands Aleut language*. Fairbanks.

Best K.-H. (1983). Zum morphologischen Wandel einiger deutscher Verben. In: Best/Kohlhase 1983b. 107–118.

Best K.-H., Kohlhase J. (1983a). Der Wandel von *ward* zu *wurde*. In: Best/Kohlhase 1983b. 91–102.

Best K.-H., Kohlhase J. (eds.) (1983b). *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte*. Göttingen.

Boryś W. (2005). *Słownik etymologiczny języka polskiego*. Kraków.

Brückner A. (1927). *Słownik etymologiczny języka polskiego*. Kraków.

Czarnecki T. (2014). *Die deutschen Lehnwörter im Polnischen. Untersuchungen zur Chronologie und Geographie der Entlehnungen*. Warszawa.

Hentschel G. (2001). Das deutsche Lehnwort in der Geschichte der polnischen Sprache: Quantitäten in chronologisch qualitativer Perspektive. – Sauerland K. (ed.). *Kulturtransfer Polen – Deutschland. Wechselbeziehungen in Sprache, Kultur und Gesellschaft*. (vol. 2). Bonn. 153–169.

Hentschel G. (2009). Intensität und Extensität deutsch-polnischer Sprachkontakte von den mittelalterlichen Anfängen bis ins 20. Jahrhundert am Beispiel deutscher Lehnwörter im Polnischen. – Stolz C. (ed.). *Unsere sprachlichen Nachbarn in Europa. Die Kontaktbeziehungen zwischen Deutsch und seinen Grenznachbarn*. Bochum. 155–171.

Imsiepen U. (1983). Die *e*-Epithese bei starken Verben im Deutschen. – Best/Kohlhase 1983b. 119–141.

Jungandreas W. (1928). *Beiträge zur Besiedlung Schlesiens und zur Entwicklungsgeschichte der schlesischen Mundart im Mittelalter*. Breslau.

Kakuk S. (1973). *Recherches sur l'histoire de la langue osmanlie des XVI*^{e} *et XVII*^{e} *siècles. Les éléments osmanlis de la langue hongroise*. Budapest.

Kluge = Seebold E. (ed.). (^{25}2011). *Kluge. Etymologisches Wörterbuch der deutschen Sprache*. Berlin, Boston.

Köhler R., Tuzzi A. (2015). Linguistic Modelling of Sequential Phenomena. The role of laws. In: Mikros G..K., Mačutek J. (eds.). *Sequences in Language and Text*. Berlin, Boston. 109–123.

Kohlhase J. (1983). Die Entwicklung von *ward* zu *wurde* beim Nürnberger Chronisten Heinrich Deichsler. Als ein Nachtrag zum vorigen. In: Best/Kohlhase 1983b. 103–106.

NKJP = *Narodowy Korpus Języka Polskiego*. 2016.06.23. http://nkjp.pl.

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

Stachowski K. (2013). The influx rate of Turkic glosses in Hungarian and Polish post-mediaeval texts. In: Köhler R., Altmann G. (eds.). *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the occasion of his 70*th *birthday*. Lüdenscheid. 100–116.

Stachowski K. [in preparation]. A report on the distribution of phonetic rendering of loanwords.

Taszycki W. (ed.). 1974–1976. *Słownik staropolskich nazw osobowych*. (Vol. 4). Ossolineum.

Vulanović R. (2013). A multidimensional generalization of the Piotrowski-Altmann law. In: Obradović I., Kelih E., Köhler R. (eds.). *Methods and Applications of Quantitative Linguistics. Selected Papers of the 8th International Conference on Quantitative Linguistics (QUALICO). Belgrade, Serbia, April 26–29, 2012*. Belgrade. 184–193.

WDLP = de Vincenz A., Hentschel G. (2010). *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache*. http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp.

Probleme der Modellierung von Lehnbeziehungen
(am Beispiel von Serbokroatismen im Slowenischen)

Emmerich Kelih (Universität Wien)
Ján Mačutek (Comenius University)

Abstract. The presented paper deals with the modelling of Serbo-Croatian loanwords in the Slovene language, in particular in journalistic texts. The empirically obtained decrease of loanwords can be modelled by the Altmann-Piotrowski law. In contrary to other approaches non-cumulative data are used, since two subgroups (old and newly appearing) of loanwords are available. Furthermore, it is shown that particular parameters of the Altmann-Piotrowski law can be motivated linguistically and determined empirically. For Slovene a basic stable stock of Serbo-Croatian loanwords is predicted.

Keywords: Altmann-Piotrowski law, Serbo-Croatian loanwords in Slovene, modelling

0. Einleitung

Der vorliegende Beitrag widmet sich der Frage, ob die Häufigkeit des Vorkommens von Serbokroatismen[1] im Slowenischen im Zeitraum von 1945 bis 2005 dem Altmann-Piotrowski Gesetz folgt. Diskutiert wird, ob neu dazu kommende Lehnwörter bzw. im System verbleibende Lehnwörter einen modellierbaren Verlauf einnehmen. Die Grundidee ist, dass die Inkorporation aber auch Verdrängung von fremdem lexikalischen Material in eine Nehmersprache mit einer beschreibbaren Regelmäßigkeit einhergeht und diese statistisch modellierbar ist. In dieser Arbeit wird auch die Diskussion aufgegriffen in der insbesondere die Frage der Kumulierung, d. h. die Aufsummierung von Frequenzen aus linguistischer Sicht kritisch hinterfragt wird. Im vorliegenden Fall werden Daten untersucht, die die absolute Häufigkeit sowohl der neuen Lehnwörter als auch der im Sprachsystem verbleibenden (bereits vorhandenen) Lehnwörter umfassen.

1. Das Altmann-Piotrowski Gesetz und Lehnwörter

Die Übernahme von Entlehnungen innerhalb einer Sprache hängt in erster Linie von sprachexternen Faktoren (Art der Normierungsprozesse, Intensität des

[1] Die Verwendung dieses Begriffes scheint durchaus gerechtfertigt zu sein, da es tatsächlich um „serbokroatische“ Entlehnungen in die slowenische Zeitungssprache seit dem Zweiten Weltkrieg geht. Eine Differenzierung zwischen Serbisch und Kroatisch – die an sich aufgrund der neueren Entwicklungen durchaus gerechtfertigt wäre – kann im vorliegenden Fall (siehe Datenanalyse in Kapitel 2) nicht erfolgen.

Sprachkontaktes, Sprach- und Kulturpolitik, Ausmaß des Purismus usw.) ab. Die Frage, die sich die quantitative Linguistik seit Langem stellt ist, ob die Häufigkeit und die Form der Übernahme von Entlehnungen einem bestimmten Prinzip oder Regularität folgt oder nicht. Es stellt sich insbesondere die Frage, welches Muster bei der Übernahme von Lehngut auf der zeitlichen Achse zu beobachten ist.

Für die Lösung dieser Fragestellung wird in der Regel ein Bezug zu Annahmen und Überlegungen der historischen Sprachwissenschaft zum Sprachwandel im Allgemeinen hergeleitet. Eine der grundsätzlichen Annahmen ist, dass eine Änderung in einem sprachlichen System zuerst langsam in Gang gesetzt wird und sodann – mit zunehmender Akzeptanz – an Geschwindigkeit gewinnt und nachdem ein bestimmter „Sättigungsgrad" erreicht wurde, wiederum langsamer wird. Danach bleibt die Möglichkeit offen, ob sich eine bestimmte Änderung durchsetzt, erhalten bleibt oder, und auch dies ist durchaus plausibel anzunehmen, möglicherweise wiederum verschwindet. Oder in anderen Worten ausgedrückt, es wird angenommen, dass dem Sprachwandel ein S-förmiger Verlauf zugrunde liegt. Diese Vermutung wird auch in der historischen Sprachwissenschaft an unterschiedlichen Stellen immer wieder ins Spiel gebracht (vgl. dazu Blythe/Croft 2012, McMahon 1994), wenngleich es ein Verdienst der quantitativen Linguistik ist, diesen Sachverhalt auch linguistisch begründet und mathematisch ausformuliert zu haben.

Altmann (1983: 106) hat ein mathematisches Modell entwickelt, welches drei unterschiedliche Einflussgrößen für den Zuwachs von neuen Formen in einem sprachlichen System berücksichtigt:

1. Der Anteil[2] neuer, bereits im Gebrauch stehender Formen.
2. Der Anteil alter, noch im Gebrauch stehender Formen und
3. einen Proportionalitätsfaktor, der Auskunft über den jeweiligen Anstieg des Zuwachses gibt und jeweils aus den vorliegenden empirischen Daten zu gewinnen ist.

Diese Größen lassen sich mathematisch in Form einer Differentialgleichung folgendermaßen ausdrücken:

$$(1)\quad p'(t) = rp(t)(1 - p(t)).$$

Hier steht r für den Proportionalitätsfaktor, $p(t)$ für die absolute Häufigkeit neuer Formen zum Zeitpunkt t, $(1–p(t))$ für die absolute Häufigkeit alter Formen zum Zeitpunkt t und $p'(t)$ für den relativen Zuwachs an neuen Formen. Die Differentialgleichung in (1) besitzt die Lösung

$$(2)\quad p(t) = \frac{1}{1 + ae^{-rt}}.$$

[2] Im Folgenden wird von absoluten Häufigkeiten ausgegangen und nicht von relativen Anteilen.

Dabei steht *t* für den Zeitpunkt und *p*(*t*) für die (berechnete) absolute Häufigkeit neuer Formen zu einem bestimmten Zeitpunkt. Bei *a* und *r* handelt es sich um Parameter, die den Verlauf der Kurve bestimmen und die aus den untersuchten Daten geschätzt werden. Der Parameter *a* gibt Auskunft über die Verschiebung auf der Zeitachse, während der Parameter *r* Auskunft über die Steilheit des Anstieges des Verlaufs gibt (je höher *r*, desto steiler der Anstieg der Kurve). Wietere Details dazu in Leopold (2005: 628). Die obige Formel (2) gilt allerdings nur für den Fall, dass ein vollständiger Wandel durchgeführt wird und sich eine sprachliche Innovation vollständig durchsetzt. Derartiges wird aber in der linguistischen Realität im Falle von Lehnwörtern kaum zu beobachten sein. Als Alternative dazu kommt folgende Formel

$$(3)\quad p(t) = \frac{c}{1 + ae^{-rt}}$$

ins Spiel, die zum Zwecke eines unvollständigen Sprachwandels zu verwenden ist. Der zusätzliche Parameter *c* gibt im Falle von c > 1 eine theoretisch mögliche absolute Häufigkeit von sprachlichen Neuerungen wieder; dieser Parameter kann ebenfalls aus den jeweils untersuchten Daten geschätzt werden.

Diese beiden Funktionsgleichungen, vollständiger und unvollständiger Sprachwandel, sind in einer Vielzahl von Studien auf unterschiedlichste Sprachen und Phänomene angewandt worden, u. a. auch bezüglich der Übernahme von Lehnwörtern in das lexikalische System einer Sprache. Dies ist mittlerweile an einer Vielzahl von Sprachen nachgewiesen worden (vgl. dazu Best 2016 mit einer ausführlichen Bibliographie).

Das ursprüngliche „Anwendungsgebiet“ des Altmann-Piotrowski Gesetzes ist der Sprachwandel. Sukzessive ist auch eine Übertragung dieses Gesetzes auf andere Anwendungsgebiete erfolgt. Gemeinsamer Ausgangspunkt ist die Idee eines Wachstumsprozesses. Aus dieser Perspektive kann auch die Inkorporation von Lehnwörtern tatsächlich als ein Wachstumsprozess verstanden werden, der selbstverständlich nicht nur die Zu- sondern auch Abnahme einer Einheit zu berücksichtigen hat. Die Übertragung von Mechanismen und Mustern des Sprachwandels auf die Entwicklung der Vorkommenshäufigkeit von Lehnwörtern auf der chronologischen Achse ist allerdings nicht ohne Kritik bzw. kritische Stellungnahme geblieben. Insbesondere hat sich damit Kempgen (1990) auseinandergesetzt. Es ist an dieser Stelle nicht möglich auf alle relevanten Kritikpunkte einzugehen, sondern es wird auf einen ausgewählten, aber zentralen Punkt einzugehen sein: Dieser betrifft eine ungenügende Ausformulierung der linguistischen Randbedingungen, aber auch der linguistischen Voraussetzungen der Modellierungen. Insbesondere wird die übliche Verwendung von kumulierten Daten bei der Modellierung der Häufigkeit von Lehnwörtern auf der zeitlichen Achse kritisch hinterfragt. Ausgehend von einem bestimmten Zeitpunkt *x* wird diese Häufigkeit mit der Häufigkeit der nächstfolgenden Periode summiert. Somit ergibt sich als letzter Datenpunkt die Summe aller in einer Sprache „belegten“ Lehnwörter. D. h. es wird ein durchgehend vorhandener Integrations- und Wachstumsprozess ins Auge gefasst. Dies entspricht aber nicht immer den linguistischen Gegebenheiten, da gerade aus dem Bereich von Entlehnungen

bekannt ist, dass es nicht immer zu einer erfolgreichen Integration kommt, sondern oft das Phänomen zu beobachten ist, dass bestimmte Lehnwörter eine kurze Zeit vorhanden sind, dann aber wiederum aus unterschiedlichen Gründen aus einem Sprachsystem verschwinden. Es ist daher durchaus plausibel, dass in diesem Zusammenhang von einer sogenannten Überlebenswahrscheinlichkeit von Entlehnungen gesprochen wird, wobei aus linguistischer Sicht vor allem die Faktoren von Interesse sind, die das eine oder andere bedingen.

Gerade in diesem Punkt ist aber auch auf die Komplexität der Datenaufnahme für die empirische Untersuchung zu verweisen, zumal reflektiert werden muss, mit welchen Häufigkeiten operiert wird und auf welche Art und Weise diese bestimmt werden. Es müssen Kriterien gefunden werden, die Auskunft darüber geben, ab wann ein Lehnwort in einer Sprache als „präsent" bzw. „übernommen" gilt. Hier bieten sich natürlich unterschiedliche Möglichkeiten an, wobei zu klären ist, ob die einmalige Verwendung durch einen Sprecher als Kriterium genügt, oder aber eine oftmalige Verwendung durch mehrere Sprecher vorliegen muss. Und wenn Letzteres zutrifft, wie wird diese Verwendungshäufigkeit eruiert? Ist die Registrierung eines Lehnwortes in einem Wörterbuch ausreichend, oder kann dies nur als operationales Hilfskonstrukt angesehen werden?

Vor diesem Hintergrund einer Vielzahl von Fragen würde sich eventuell ein Kompromiss in Form eines gemischten Wörterbuch- und Textansatzes anbieten. Aber auch dies ist bei auf sehr lange Zeiträume bezogene Untersuchungen aus vielerlei Hinsicht problematisch (fehlende historische Textbelege, keine Daten über die entsprechenden Verwendungshäufigkeiten). Sieht man von der grundlegenden Problematik der Datenaufnahme ab, so ist eine weitere Schwierigkeit darin zu sehen, dass man es bei der Modellierung eines lexikalischen Systems mit einem prinzipiell offenen System zu tun hat und gerade die Entlehnung hier eine zentrale Rolle der Ausgestaltung hat. Aber in diesem Fall muss es nicht immer zu einer Kumulation kommen, sondern – sofern man das ganze lexikalische System vor Augen hat – es werden tatsächlich neue „semantische" Felder mit Lehnwörtern besetzt, oder aber alte ererbte Formen durch Entlehnungen ersetzt. Darüber hinaus werden insbesondere bei sehr oberflächlichen Formen des Sprachkontaktes – und das ist aus der Literatur durchaus gut bekannt – insbesondere Substantive entlehnt, die bestimmte technische Neuerungen, konkrete Gegenstände usw. bezeichnen, die wiederum u. a. durch die Dynamik und Vielzahl von Änderungen der außersprachlichen Realität bedingt sind. Somit hat man es mit einem fließenden Kommen und Gehen von Lehnwörtern zu tun. Darin ist auch der Grund zu sehen, dass die Frage der Kumulierung der Häufigkeiten von Lehnwörtern aus dieser Perspektive tatsächlich nicht den linguistischen Gegebenheiten entspricht. Die Modellierung von kumulierten Daten ist somit in erster Linie ein heuristisches Mittel, um grundsätzliche Mechanismen der Inkorporation und Dynamik lexikalischer Strukturen aufzuzeigen.

Dies ändert aber nichts an der Tatsache, dass außer Frage ist, dass die Ab- und Zunahme von Lehnwörter in einer Sprache modellierbar ist. Es ist durchaus legitim anzunehmen, dass entsprechend des Interaktionsmodells zuerst die Verwendung eines Lehnwortes auf wenige Sprecher beschränkt ist und dann –

sofern eben das Lehnwort bestimmte Bedürfnisse in einer Sprechergemeinschaft „erfolgreich“ erfüllt – eine schnelle und massenhafte Verbreitung erfahren kann. Es kann aber auch durchaus sein, dass ein bestimmtes Lehnwort die gestellten Bedürfnisse nicht erfüllt und somit wieder aus dem Sprachsystem fällt bzw. – und dies ist ein oft vorkommender Fall – aus puristischen Gründen abgelehnt wird bzw. sich die äußeren Umstände ändern und der Einfluss einer Gebersprache aus unterschiedlichsten Gründen zurückgeht. Somit hat man es mit einem komplexen Prozess zwischen Integration, Verdrängung und Rückgang zu tun. Die empirische Dimension dieser Fragestellung wird im nächsten Kapitel zu besprechen sein.

2. Untersuchte Daten: Serbokroatismen im Slowenischen

Die im vorangehenden Kapitel angeführte Problematik der Verwendung von kumulierten und nicht kumulierten Daten soll im Folgenden anhand von Lehnwortdaten aus dem Slowenischen durchgespielt werden. Es handelt sich dabei um die Häufigkeit von Serbokroatismen in der slowenischen Zeitungssprache aus den Jahren 1945 bis 2005 (vgl. dazu Jelovšek 2009). Die Autorin bringt zur Eruierung der Daten ein Stichprobenverfahren zur Anwendung, indem sie im Rhythmus von 10 Jahren (1945, 1955, … 2005) jeweils Ende November aus einer Ausgabe der Zeitung Delo die genuin journalistischen Texte (d. h. nicht aus der Werbung, den Wetterberichten usw.) in Hinblick auf die Häufigkeit von darin vorkommenden Entlehnungen aus dem Serbokroatischen auswertet. Das von ihr erhaltene Material beinhaltet (Jelovšek 2009: 70) – und das ist ein Vorteil dieser Art von Datenaufnahme – zwei unterschiedliche Arten von Daten. Ausgehend von der im Jahr 1945 bestimmten Anzahl von Entlehnungen wird in 10-Jahres-Schritten die Anzahl von neu hinzugekommenen Entlehnungen und die Anzahl von alten, d. h. bereits in der Vorperiode vorhandenen Lehnwörtern eruiert. Aus diesen Daten lässt sich die Gesamtanzahl von Serbokroatismen in diesem Zeitungskorpus eruieren. D. h. hier kann genau zwischen neuen, alten und tatsächlich nicht mehr belegbaren Lehnwörtern unterschieden werden, was auch jegliche Form der Kumulierung nicht mehr notwendig erscheinen lässt.

Die Häufigkeit dieser Entlehnungen ist in Tabelle 1 reproduziert (Jelovšek 2009: 71). In der untenstehenden Tabelle sind die Häufigkeiten von belegten Entlehnungen aus dem Serbokroatischen in den untersuchten Zeitungsausgaben zusammengefaßt dargestellt.

Tabelle1
Serbokroatismen im Slowenischen

Jahr	Intervall	Entlehungen		
		alt	neu	Summe
1945	1	178		178
1955	2	37	61	98

1965	3	40	71	111
1975	4	61	52	113
1985	5	31	30	61
1995	6	53	16	69
2005	7	20	13	33

Betrachtet man nunmehr Abb. 1, wird sofort deutlich, dass im Grunde genommen seit dem Jahr 1945 generell ein Rückgang der Entlehnungen zu beobachten ist.

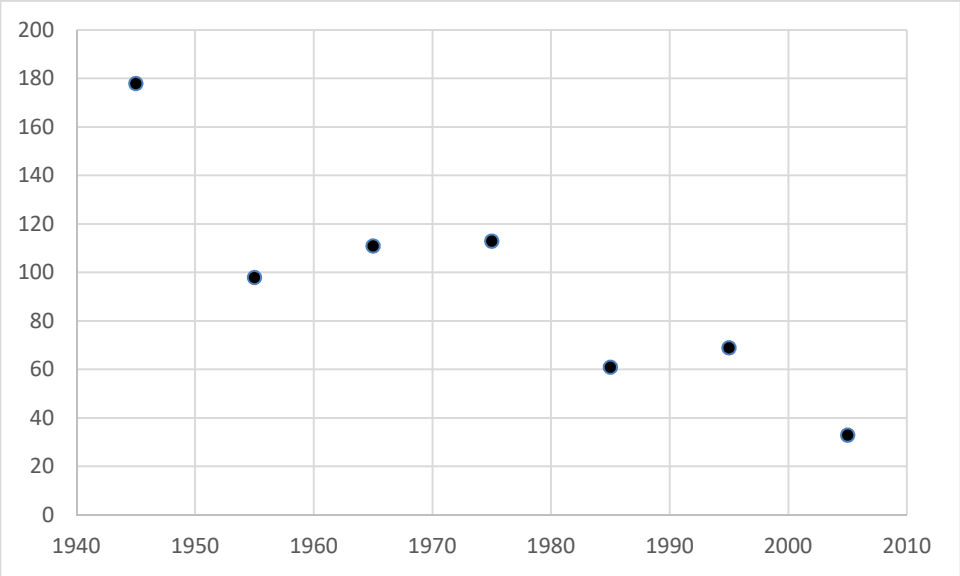


Abb. 1: Absolute Häufigkeit von Entlehnungen aus dem Serbokroatischen seit 1945 (Zeitungskorpus)

Zu beginnen ist mit der Anzahl von neu dazugekommen Lehnwörtern, die einen fast stetigen Abfall zeigt, aber um das Jahr 1965 wieder einen leichten Anstieg nach oben aufweist und dann tatsächlich kontinuierlich zurückgeht. Dennoch ist eine Auf- und Ab-Bewegung zu registrieren. Genau dieser Aspekt ist aber auch von vorrangiger Bedeutung aus der Perspektive des Altmann-Piotrowski Gesetzes. Es wird davon ausgegangen, dass hinsichtlich der Dynamik von Änderungen im Sprachsystem zwei unterschiedliche Sprechergruppen anzunehmen sind: Die eine Gruppe forciert eine bestimmte Innovation (und eine Entlehnung kann ja durchaus als Innovation verstanden werden), während eine andere Gruppe diese nicht übernehmen will und sich erst im Laufe der Zeit eine Form entweder durchsetzt oder nicht. Aus dieser Perspektive erklärt sich diese wellenförmige Bewegung bei den Lehnwörtern, die, sofern einmal im Slowenischen vorhanden, offenbar in gewisser Weise immer wieder „aktiviert“ werden kann. Letztlich hat sich doch – und dieser Befund deckt sich wunderbar mit dem für das Slowenische oftmals beschriebenen Phänomen des Purismus, der sich insbesondere in der zweiten Hälfte des 20. Jh. vor allem gegen den Einfluss anderer südslawischer Sprachen richtete und sich im Slowenischen, zumindest in Bezug auf Manifestationen in der Zeitungssprache, durchgesetzt hatte. Der sprachliche Purismus richtete sich in dieser Zeit vor allem gegen einen allzu starken südsla-

wischen Einfluss und ist auch in der Literatur vielfach beschrieben worden (vgl. Kalin Golob 2009, Požgaj Hadži et al. 2009, Thomas 1997).

Aus dem Zusammenspiel dieser unterschiedlichen Kräfte ergibt sich somit in Summe die in Abb. 1 ersichtliche Abwärtsbewegung der Häufigkeit in „treppenförmiger" Form. Es lässt sich aber durchaus vermuten, dass langfristig doch ein – wenn auch kleiner – Teil an Lehnwörtern aus dem Serbokroatischen sich im Slowenischen festsetzen und zu einem integralen Bestandteil werden wird. Dies lässt sich auch daran zeigen, dass auch noch im Zeitraum seit 1990 jeweils neue Entlehnungen im Zeitungskorpus registriert werden. Zu klären bleibt nun die Frage einer adäquaten Modellierung dieser Daten, wobei insbesondere auch die Frage der Kumulation zu diskutieren sein wird.

3. Modellierung: Unvollständiger Sprachwandel

In der in Kap. 1 bereits angeführten Kritik von Kempgen (1990) geht es vor allem um das Infragestellen der bei der Modellierung von Lehnwortbeziehungen durchgeführten Kumulierung von Daten. Daraus ergibt sich auf Modellebene bei einem unvollständigen Sprachwandel die Frage einer Ober- bzw. Untergrenze an Entlehnungen, die aber nicht immer plausibel erscheint, da bestimmte Lehnwörter aus dem Gebrauch wieder ausscheiden bzw. sich tatsächlich in ein Sprachsystem eingliedern können. In den in Kap. 2 verwendeten Daten konnte gerade dieses Phänomen auch graphisch gezeigt werden und nunmehr haben wir auch die Gelegenheit dieses Zusammenspiel exakter empirisch zu erfassen, indem konkret eine Überlebensrate der Lehnwörter angesetzt werden kann. Als empirischer Ausgangspunkt kann die Anzahl von 178 Lehnwörtern genommen werden, die durch die Anzahl von jeweils belegten alten Lehnwörtern pro Periode dividiert wird. So ergibt sich für die Periode von 1945 bis 1955 eine Aktivierungsrate von ca. 21%[3]; diese ändert sich im Laufe der Zeit, wie der Tabelle 2 zu entnehmen ist. Wie zu sehen, ist die Aktivierungsrate in der Mitte der 70er Jahre am höchsten; auch Mitte der 90er Jahre scheint sie sehr hoch zu sein, was aber vor allem offenbar eine Art von verzögertem Prozess darstellt. Zu betonen ist, dass man damit auch den Zustand vor 1945 extrapolieren kann, denn es ist nicht anzunehmen, dass die Inkorporation auf 1945 beschränkt ist, sondern auch für den Zeitraum davor ein sukzessives Anwachsen anzusetzen ist.

[3] Diese Aktivierungsrate darf nicht mit einer Überlebensrate, die ebenfalls in der Kontaktlinguistik eine Rolle spielen kann, verwechselt werden. Im Falle der Aktivierungsrate geht es darum, dass Lehnwörter aus dem Inventar des Jahres 1945 in unterschiedlichen späteren Perioden auftreten können. Eine Überlebensrate impliziert, dass Lehnwörter tatsächlich aus dem Sprachsystem fallen und danach nicht mehr verwendet werden und daher kein absoluter Anstieg von Lehnwörtern mehr möglich ist. Dieser Aspekt müsste in Zukunft systematisch untersucht werden.

Tabelle 2
Aktivierungsrate von serbokroatischen Lehnwörtern im Slowenischen

Zeitraum	**Aktivierungsrate**
1955	0,21
1965	0,22
1975	0,34
1985	0,17
1995	0,30
2005	0,11

Die errechnete Aktivierungsrate wird auch für die Modellbildung relevant sein, zumal man hier aus den vorliegenden Daten einen Mittelwert bilden kann. Dieser beträgt 0,22 und besagt, dass jede fünfte Entlehnung aus dem Serbokroatischen aus dem Jahr 1945 in der Zeitungssprache weiterhin aktiv verwendet wird. Dieser Wert ist natürlich nichts anderes als eine Art Richtwert, der aber für die Bestimmung von *c* bei der Modellierung eine Rolle spielen kann.

Der Beobachtungszeitraum von 1945 bis 2005 ist in erster Linie ein Ausschnitt, der insbesondere den Rückgang dokumentiert, nicht aber eine gleichzeitige Zunahme, die ebenfalls als kontinuierlicher Prozess anzusehen ist. Es ist auch tatsächlich sehr plausibel, dies in Kenntnis der slowenischen Sprachgeschichte anzunehmen, da im Grunde seit dem Beginn des 19. Jh. und der eigentlichen Standardisierung des Slowenischen ein stetiger Zuwachs an generell innerslawischen, insbesondere aber südslawischen Entlehnungen zu beobachten ist, der aber in dem von uns verwendeten Datensatz nicht berücksichtigt ist.

Insofern ist es gerechtfertigt, nicht das Modell für eine Zu- und Abnahme heranzuziehen, sondern die entsprechende Formel für einen unvollständigen Sprachwandel. Dies ist auch insofern durchaus plausibel, da auch im gegenwärtigen Kontext weiterhin Serbokroatismen im Slowenischen eine Rolle spielen und natürlich nicht vollständig verdrängt wurden/werden, sondern immer auch noch neue Belege zu finden sind. Insofern wird als geeignetes Modell

$$(4)\quad p(t) = \frac{c}{1 + ae^{-rt}}$$

verwendet.

Da es sich hierbei in unserem Modell um eine Untergrenze handelt (bei kumulierten Daten hätte man es hier mit einer Obergrenze zu tun, die tatsächlich schwer zu begründen ist), kann diese aufgrund der vorliegenden Daten geschätzt werden. Betrachtet man die Häufigkeit von Lehnwörtern aus dem Jahr 2005, so kann man darauf aufbauend annehmen, dass davon – in Anlehnung an den Mittelwert, der sich aus der Vergangenheit ergibt – nur ca. jedes fünfte Lehnwort (ca. 20%) aktiviert wird, sodass hier ein empirisches *c* angesetzt werden kann. Dieses ergibt sich auf der Basis der Teilgruppe von alten Lehnwörtern. Für diese wird angenommen, dass auf der Basis des berechneten Mittelwertes der Aktivierungsrate (0,22) von 178 Lehnwörtern ein $C_1 = 39$ angesetzt wird; darüber hinaus ist anzunehmen, dass von den neuen Lehnwörtern auch in Zukunft einige überleben werden (auf der Basis des empirischen Wertes aus dem Jahr 2005 ergibt

sich ein $C_2 = 3$), was ein $C = C_1+C_2 = 42$ ergibt. Damit geht nicht der Anspruch einher, einen absolut „richtigen“ Grenzwert gefunden zu haben, sondern vielmehr geht es darum zu zeigen, dass es möglich ist, empirisch begründet Untergrenzen zu setzen. Ohne Zweifel muss gerade in diesem Punkt eine sprachspezifische Bestimmung und Diskussion erfolgen. Die Ergebnisse der Anpassungen unter Verwendung von Formel (4) sind in Tabelle 3 zu finden. Die Werte für die Parameter *a* und *r* werden iterativ gewonnen, während C – wie oben begründet – von uns empirisch festgesetzt wurde. Die Ergebnisse ($C = 42$, $a = -0{,}88$, $r = 0{,}15$, $R^2 = 0{,}80$) deuten auf eine akzeptable Modellierung hin; dies kann aber auch durch die geringe Anzahl von Datenpunkten bedingt sein.

Tabelle 3
Anpassungsergebnisse

Jahre	**Intervall**	**LW alt**	**LW neu**	**LW gesamt**	
		emp.	emp.	emp.	theo.
1945	1	178		178	173,14
1955	2	37	61	98	120,66
1965	3	40	71	111	95,70
1975	4	61	52	113	81,23
1985	5	31	30	61	71,88
1995	6	53	16	69	65,40
2005	7	20	13	33	60,69

Zu bedenken ist auch, dass sich darin die Heterogenität der Daten widerspiegelt, die sich aufgrund der Mischung von alten und neuen Lehnwörtern ergibt. Insofern ist das Gesamtergebnis vor diesem Hintergrund dennoch bemerkenswert. Vgl. dazu Abb. 2 für eine graphische Darstellung der empirischen Daten und theoretischen Werte.

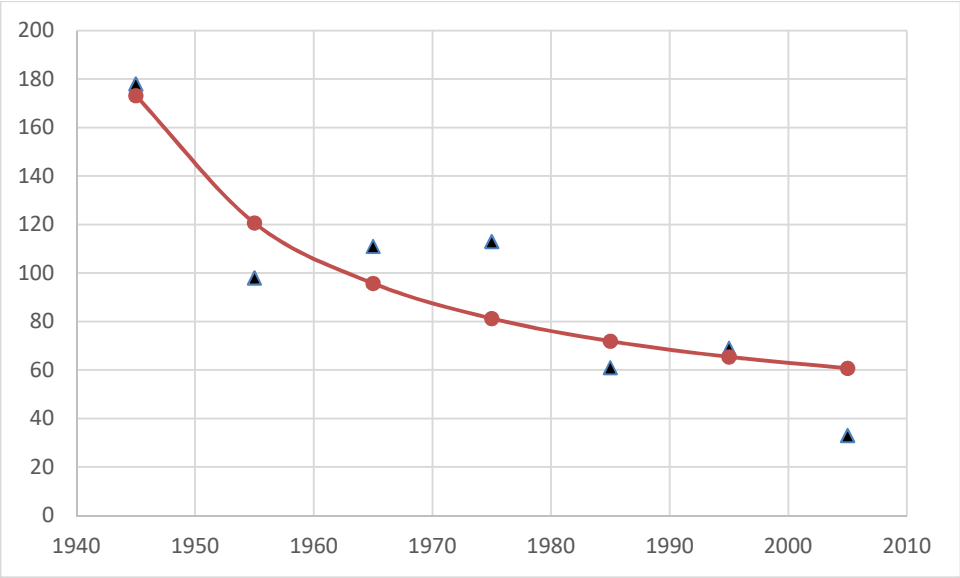


Abb. 2.: Empirische Daten und theoretische Werte für den Rückgang von Serbokroatismen (abs. Häufigkeit) im Slowenischen seit 1945

Insgesamt ergibt sich ein Gesamttrend eines „flachen“ Rückganges von Serbokroatismen im Slowenischen. Wie nun gezeigt werden konnte, kann dies mit der Formel für einen unvollständigen Sprachwandel – wie linguistisch auch völlig einleuchtend – modelliert werden. Zum Abschluss wird nun die Entwicklung des jeweiligen relativen Anteils von alten bzw. neuen Lehnwörtern diskutiert.

4. Relativer Anteil von neuen bzw. alten Lehnwörtern

Die uns vorliegenden Daten geben uns die Möglichkeit, einen weiteren Aspekt der Serbokroatismen im Slowenischen zu beleuchten. Es lässt sich der jeweilige Anteil von jeweils neuen Lehnwörtern berechnen (neue Lehnwörter/ Gesamtanzahl von Lehnwörtern in einer Periode). Diese Daten (auf der Basis von Tab. 1 bzw. 3) sind in Tabelle 4 dargestellt.

Tabelle 4
Serbokroatismen im Slowenischen (Jelovšek 2009: 71)

Jahr	Intervall	alt (abs.)	alt (rel.)	neu (abs.)	neu (rel.)	Summe
1955	2	37	0,38	61	0,62	98
1965	3	40	0,36	71	0,64	111
1975	4	61	0,54	52	0,46	113
1985	5	31	0,51	30	0,49	61
1995	6	53	0,77	16	0,23	69
2005	7	20	0,61	13	0,39	33

Für den relativen Anteil neuer Lehnwörter ergibt sich eine fallende Tendenz. Linguistisch bemerkenswert erscheint der relative Anteil an alten Lehnwörtern. Wie der Abb. 3 zu entnehmen ist, zeigt sich eine wachsende bzw. steigende Tendenz des relativen Anteils von alten Lehnwörtern.

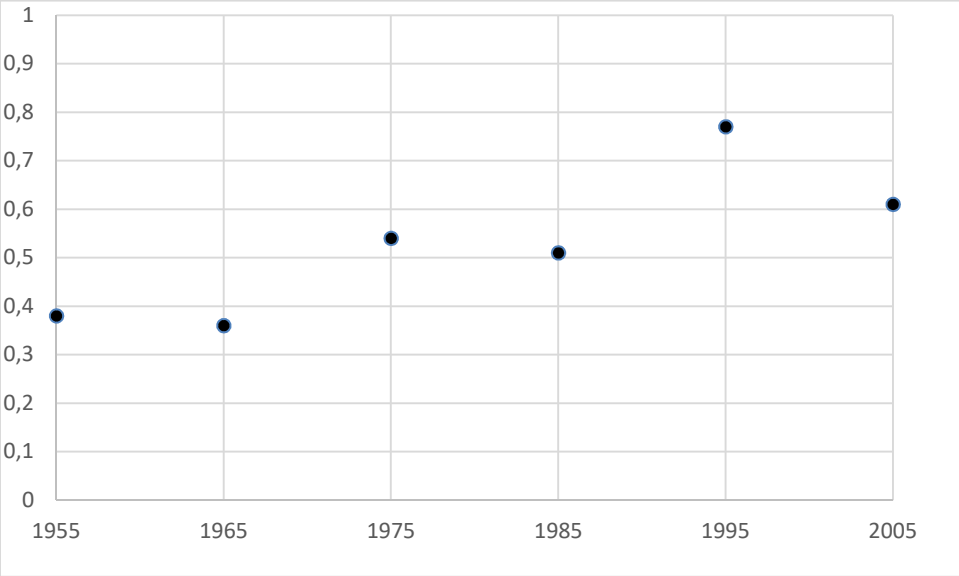


Abb. 3. Relativer Anteil von alten Lehnwörtern am Gesamtbestand (1955-2005)

Dies bedeutet im gegebenen Kontext, dass zwar offenbar die absolute Anzahl von Lehnwörtern aus dem Serbokroatischem im Slowenischen sinkt, es aber offenbar einen „Kernbestand“ an serbokroatischen Lehnwörtern im Slowenischen gibt. Dieser ist mehr oder weniger fest im Sprachsystem verankert und zeigt auch nicht die Tendenz, verdrängt zu werden. Damit scheint tatsächlich eine Stabilisierung zu beobachten zu sein, die darauf hinausläuft, dass immer weniger neue serbokroatische Lehnwörter in das Slowenische eindringen, aber vorhandene dennoch weiterhin verwendet werden. Genau dieser Aspekt konnte bei der Modellierung in Kap. 3 auch in entsprechender Weise berücksichtigt werden, indem der Parameter C = 42 empirisch fixiert werden konnte. Diese Größe hat insbesondere heuristischen Wert, da außerhalb der journalistischen Sprache sicherlich andere Werte anzusetzen sind (man denke in diesem Zusammenhang an die mündliche Umgangssprache bzw. Soziolekte des Slowenischen, die einen weitaus höheren Anteil an Serbokroatismen aufweisen können).

5. Zusammenfassung

Der vorliegende Beitrag ist vor allem der Frage gewidmet, inwiefern für die Modellierung der Übernahme von Lehnwörtern nicht nur die üblich verwendeten kumulierten Daten herangezogen werden können. Für Serbokroatismen im Slowenischen ergibt sich die Häufigkeit von Entlehnungen aus der Wechselwirkung

von der Aufnahme jeweils neuer Lehnwörter und der Verdrängung bzw. dem Nichtverwenden bereits vorhandener Lehnwörter, die bereits zu einem früheren Zeitpunkt in das Slowenische gelangt sind. Auf diese Weise wird man der Dynamik lexikalischer Entlehnungsprozesse durchaus gerecht. Für diese Art der Modellierung kann ebenfalls das bekannte Altmann-Piotrowski Gesetz für einen unvollständigen Wandel erfolgreich verwendet werden. Darüber hinaus kann linguistisch begründet ein empirischer Wert für den Parameter C festgesetzt werden, der allerdings als ein hypothetischer Grenzwert in Erscheinung tritt. Es ist plausibel anzunehmen, dass ein Teil von Lehnwörtern – besonders sofern Perioden einer intensiven Übernahme vorhanden sind – in einem Sprachsystem langsam adaptiert und nachhaltig integriert werden können. Sobald diese Integration abgeschlossen ist, erhöht sich die „Überlebenswahrscheinlichkeit" der Lehnwörter und aufgrund der zwischenzeitlich erfolgten Anpassung auch deren „Resistenz" gegenüber puristischen Eingriffen. Ob und in welcher Form die anhand von Serbokroatismen im Slowenischen beobachtbaren Tendenzen auch für andere Sprachen gelten, können aber nur weitere empirische Studien zeigen.

Danksagung
J. Mačutek was supported by the grant VEGA 2/0047/15.

Verwendete Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Karl-Heinz Best und Jörg Kohlhase (Hg.): *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte:* 54–90. Göttingen: Herodot (Göttinger Schriften zur Sprach- und Literaturwissenschaft, 2),

Best, Karl-Heinz (2016). Bibliography – Piotrowski's law. *Glottotheory* 7 (1), 89–94.

Blythe, Richard A.; Croft, William (2012). S-curves and the mechanisms of propagation in language change. *Language* 88 (2), 269–304.

Kalin Golob, Monika (2009). Linguistic purism in Slovene language: From Trubar to the present. In: Granić, Jagoda (Hg.), *Jezična politika i jezična stvarnost. Language Policy and Language Reality:* 137–146. Zagreb: Hrvatsko društvo za primijenjenu lingvistiku – HDPL.

Kempgen, Sebastian (1990). Zur Modellierung von Lehnbeziehungen. In: Walter Breu (Hg.): *Slavistische Linguistik 1989:* 99–116. München: Sagner.

Leopold, Edda (2005). Das Piotrowski-Gesetz. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Hg.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook:* 627–633. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Jelovšek, Alenka (2009). *Srbohrvatizmi nekoč in danes: Analiza tiska med letoma 1945 in 2005.* Ljubljana: Diplomsko delo.

McMahon, April M. S. (1994): *Understanding language change.* Cambridge: Cambridge Univ. Press.

Požgaj Hadži, Vesna; Balažic Bulc, Tatjana; Miheljak, Vlado (2009). Srbohrvaščina v Sloveniji: Nekoč in danes. In: Vesna Požgaj Hadži, Tatjana Balažic Bulc und Vojko Gorjanc (Hg.): *Med politiko in stvarnostjo. Jezikovna situacija v novonastalih državah bivše Jugoslavije*: 27–39 Ljubljana: Filozofska Fakulteta.

Thomas, George (1997). The impact of purism on the development of the Slovene standard language, in: *Slovenski jezik - Slovene linguistic studies* 1; 133–152.

Fast Calculation of Entropy with Zhang's Estimator

Antoni Lozano[1]
Bernardino Casas[2]
Chris Bentz[3]
Ramon Ferrer-i-Cancho[2,*]

(1) COMBGRAF Research Group. Departament de Ciències de la Computació. Universitat Politècnica de Catalunya. Barcelona, Catalonia, Spain.
(2) Complexity and Quantitative Linguistics Lab. LARCA Research Group. Departament de Ciències de la Computació. Universitat Politècnica de Catalunya. Barcelona, Catalonia, Spain.
(3) Department of General Linguistics, University of Tübingen, Tübingen, Germany.
(*) Corresponding author, rferrericancho@cs.upc.edu

Abstract. Entropy is a fundamental property of a repertoire. Here, we present an efficient algorithm to estimate the entropy of types with the help of Zhang's estimator. The algorithm takes advantage of the fact that the number of different frequencies in a text is in general much smaller than the number of types. We justify the convenience of the algorithm by means of an analysis of the statistical properties of texts from more than 1000 languages. Our work opens up various possibilities for future research.

Keywords: *entropy estimation*, *lexical diversity*, *parallel corpora*.

1. Introduction

In quantitative linguistics, the Shannon entropy of types is a fundamental property of a repertoire. From a Zipfian perspective, vocabularies are shaped by a tension between unification and diversification forces (Zipf 1949). The entropy of types is a measure of the degree of diversification of word use or, equivalently, a measure of lexical diversity: it takes a value of 0 when only one type is used, while it takes its maximum value when all types are equally likely (Ferrer-i-Cancho 2005; Bentz et al. 2015). Theoretically, entropy is a better measure of vocabulary size than the raw number of different types: the entropy of types measures the effective size of the vocabulary, which is related to the concept of typical set in information theory (Ferrer-i-Cancho 2014). In practice, the problem is that the true number of types is unknown but it is possible to estimate entropy by taking into account that not all types have been observed (Grabchak et al 2013). For similar reasons, entropy is also used as an index of the diversity of species in biology (Chao, Shen 2003, Jost 2006).

From an information theoretic perspective, the entropy of words yields a lower bound to the mean length of words assuming uniquely decipherable coding (Cover, Thomas 2006). Besides, word entropy minimization is a core assumption of information theoretic models of Zipf's law for word frequencies (Ferrer-i-

Cancho 2005, Ferrer-i-Cacho, Solé 2003). From a psychological perspective, entropy is hypothesized to be a measure of cognitive cost (Ferrer-i-Cancho 2014).

The Shannon entropy of a text with a vocabulary of V types is defined as (Shannon 1951)

$$H = \sum_{i=1}^{V} p_i \log p_i, \tag{1}$$

where p_i is the probability of the i-th type.

If f_i is the absolute frequency of the i-th type in a text of T tokens, the probability of the i-th type can be estimated by the relative frequency of the type as

$$\hat{p}_i = \frac{f_i}{T} \tag{2}$$

with

$$T = \sum_{i=1}^{V} f_i. \tag{3}$$

The entropy of a vocabulary can be estimated naively by replacing the probabilities in Eq. 1 with the relative frequencies in Eq. (2). This is the so-called plugin or maximum likelihood entropy estimator that is known to underestimate the true entropy in practical applications (Hausser, Strimmer 2009). Here we focus on an alternative method that reduces the original downwards bias of the naïve estimator: Zhang's entropy estimator (Zhang 2012). Our goal is to provide an efficient algorithm for that estimator.

In the design of an entropy estimator there is often a trade-off between its computational cost and the bias of the estimator. The NSB estimator, which is considered one of the best entropy estimators, is computationally expensive (Hausser, Strimmer 2009), and for this reason had to be discarded in Vu et al's (2007) study.

In general, we consider algorithms to estimate entropy whose input consists of the absolute frequencies of the V types. Those algorithms use at least $\theta(V)$ memory for the input frequency of each type. We will refer to the extra memory used by these algorithms as additional memory. It is easy to see that the plugin estimator and the Chao-Shen estimator run in time $\theta(V)$ with $O(1)$ additional memory. The time cost of Zhang's estimator has not been analyzed in detail to our knowledge. For computational purposes, the recommended definetion of Zhang's estimator is (Grabchak et al. 2013)

$$H_z = \sum_{i=1}^{V} \hat{p}_i \sum_{\nu=1}^{T-f_i} \frac{1}{\nu} \prod_{j=0}^{\nu-1} \left(1 + \frac{1-f_i}{T-1-j}\right), \tag{4}$$

where V is hereafter the number of types in a text. Noting that $1 \leq i \leq V$ and $1 \leq \nu$, $j \leq T$ it is easy to derive an algorithm that runs in $O(VT^2)$ time. Here, we present

an algorithm that estimates entropy in $O(WT)$ time, where W is the number of occupied frequencies. Suppose that $n(f)$ is the number of word types that have frequency f in the sample; $n(1),\ldots,n(f),\ldots\ n(T)$ defines the frequency spectrum of the sample (Tuldava 1996). By definition

$$V = \sum_{f=1}^{T} n(f). \tag{5}$$

W is the number of values of f such that $n(f) > 0$.

The remainder of the article is organized as follows. Section 2 presents a couple of algorithms to estimate entropy: one that runs in $\theta(VT)$ time and uses $O(1)$ additional memory and another that runs in $O(WT)$ time and uses $\theta(f_{max})$ additional memory, where f_{max} is the maximum f such that $n(f)>0$. The second algorithm is potentially faster as $W \leq V$. Section 3 investigates some statistical properties of real texts that are needed to justify the convenience of the $O(WT)$ time algorithm. On the one hand, we study the ratio W/V in a couple of parallel corpora, concluding that the $O(WT)$ algorithm is at least 5 times faster than the $\theta(VT)$ algorithm when used to estimate word entropies in real texts. On the other hand, we study the ratio f_{max}/V, concluding that the additional memory required by the $O(WT)$ time algorithm can be neglected compared to V, the cost of storing the table of type frequencies. Section 4 discusses the results.

2. A Faster Algorithm

In general, entropy estimation algorithms take the V frequencies of every type, i.e. $f_1, \ldots, f_i,\ldots,f_V$, as the input and return an entropy estimate (Hausser, Strimmer 2009, Zhang 2012). Here we consider the particular case of algorithms based on Zhang's estimator.

It is convenient to define Zhang's estimator equivalently as

$$H_z = \frac{1}{T}\sum_{i=1}^{V} f_i Q(f_i), \tag{6}$$

where

$$Q(f) = \sum_{v=1}^{T-f} \frac{1}{v} \prod_{j=0}^{v-1} \left(1 + \frac{1-f}{T-1-j}\right). \tag{7}$$

This decomposition leads to Algorithm A:

1. Set K to 0
2. For each i such that $1 \leq i \leq V$ do
    a. Calculate $Q(f_i)$
    b. Sum $f_i Q(f_i)$ to K
3. Return K/T

It is easy to see that the running time of Algorithm A is $O(V \text{ time}(Q))$ where time(Q) is the worst running time of the calculation of Q (Step 2.a. of Algorithm A) for a given type. Since $1 \leq \nu, j \leq T$, an algorithm for the calculation of $Q(f)$ following Eq. 7 naively runs in $O(T^2)$ time and then the running time of Algorithm A becomes $O(VT^2)$. The space cost is given by the size of the input, i.e. $\theta(V)$.

We will show that the running time of the calculation of $Q(f)$ can be reduced to $O(T)$. To see it, it is convenient to define $Q(f)$ as

$$Q(f) = \sum_{\nu=1}^{T-f} \frac{1}{\nu} R(\nu, f), \tag{8}$$

where

$$R(\nu, f) = \prod_{j=0}^{\nu-1} \left(1 + \frac{1-f}{T-1-j}\right). \tag{9}$$

A close observation of Eq. 9 allows one to realize that $R(\nu,f)$ can be defined recursively, i.e. $R(\nu,f) = 1$ if $\nu = 0$ and

$$R(\nu, f) = \left(1 + \frac{1-f}{T-\nu}\right) R(\nu - 1, f) \tag{10}$$

if $\nu > 0$. The recursive definition of $R(\nu,f)$ allows one to calculate $Q(f)$ faster with Algorithm B:

1. Set Q to 0 and R to 1
2. For $\nu = 1$ to $T - f$ do
    a. Multiply R by $\left(1 + \frac{1-f}{T-\nu}\right)$
    b. Sum R/ν to Q
3. Return Q

Algorithm B performs of the order of $T - f + 1$ operations (1 for Step 1 and $T - f$ for Step 2). Thus it runs in time $\theta(T-f)$.

The kind of time optimization that we have performed to calculate $Q(f)$ in Algorithm B, namely the addition of extra local variables to recycle computations from previous iterations, is reminiscent of the one that was applied in an efficient algorithm for the moving average type-token ratio (Covington, McFall 2010).

We obtain Algorithm A' by integrating Algorithm B into Algorithm A. It turns out that the number of operations performed by Algorithm A' is of the order of

$$T(A') = \sum_{i=1}^{V} (T - f_i + 1) = V(T+1) - \sum_{i=1}^{V} f_i. \tag{11}$$

Recalling the definition of T in Eq. 3 one obtains finally

$$T(A') = T(V-1) + V. \tag{12}$$

Since $V \leq T$, Algorithm A' estimates entropy in time $\theta(VT)$, which is a significant improvement with respect to the naïve Algorithm A that runs in $O(VT^2)$. Algorithm A' also needs $\theta(V)$ space.

The running time of the entropy estimation can be reduced further. To see it, it is convenient to restate Eq. 6 equivalently as

$$H_z = \frac{1}{T}\sum_{f=1}^{f_{\max}} n(f) f \mathrm{Q}(f). \tag{13}$$

Notice that one only has to calculate $Q(f)$ when $n(f) > 0$. This formulation leads to Algorithm C:

1. Calculate f_{max} and the frequency spectrum $n(1),\ldots,n(f),\ldots n(f_{max})$ from f_1, $\ldots$, $f_i,\ldots,f_V$
2. Set K to 0
3. For f = 1 to f_{max} do
    a. If $n(f) > 0$ then
        i. Calculate $Q(f)$ with Algorithm B
        ii. Sum $n(f)fQ(f)$ to K
4. Return K/T

Algorithm C calculates $Q(f)$ only for W values of f. Assuming the worst case for the calculation of $Q(f)$ with Algorithm B, namely a running time of $O(T)$, one sees that the running time of Algorithm C is $O(\mathrm{W}T)$ (the cost of the 1st step, the calculation of $n(f)$, is $O(T)$). This suggests that Algorithm C is faster than Algorithm A' (which runs in $\theta(\mathrm{V}T)$ time) because $W \leq V$, with equality if and only if $n(f) \in \{0,1\}$. To see it, notice that W can be defined as

$$W = \sum_{f=1}^{T} b(f), \tag{14}$$

where $b(f)$ is a binary variable that indicates if $n(f) > 0$ ($b(f) = 1$ if $n(f) > 0$ and $b(f) = 0$ if $n(f) = 0$). Noting that $b(f) \leq n(f)$ and recalling the definition of V in Eq. 5, $W \leq V$ follows easily, with equality if and only if $b(f) = n(f)$, i.e. $n(f) \in \{0,1\}$ for all f. The case $n(f) \in \{0,1\}$ never happens in a sufficiently large real text: it only happens for high frequencies as noted by Balasubrahmanyan, Naranan (1996).

Stronger evidence for the superiority of Algorithm C in respect of time efficiency can be obtained easily. On the one hand, the bulk of the time cost of Algorithm C is determined by step 3 and is of the order of

$$T(C) = \sum_{f=1}^{f_{\max}} b(f)(T - f + 1). \tag{15}$$

Step 1 can be carried out in $\theta(\max(V,f_{max}))$ time if one uses a table of size f_{max} to store $n(1),\ldots,n(f),\ldots\ n(f_{max})$. Recalling the definition of V in Eq. 5, it is easy to show that $T(C) \geq V$ since $f \leq T$.

On the other hand, the cost of Algorithm A' given in Eq. 11 can be expressed equivalently as

$$T(A') = \sum_{f=1}^{f_{\max}} n(f)(T - f + 1). \tag{16}$$

Thus, the number of elementary operations saved by Algorithm C (excluding Step 1) is of the order of

$$T(A') - T(C) = \sum_{f=1}^{f_{\max}} (n(f) - b(f))(T - f + 1). \tag{17}$$

It is easy to see that $T(A')$-$T(C) \geq 0$ since $b(f) \leq n(f)$ and $f \leq T$.

The cost of Algorithm C can be calculated with more precision. Expressing Eq. 15 as

$$T(C) = (T + 1)\sum_{f=1}^{f_{\max}} b(f) - \sum_{f=1}^{f_{\max}} b(f)f \tag{18}$$

and recalling the definition of W in Eq. 14, one obtains finally

$$T(C) = W(T + 1) - \sum_{f=1}^{f_{\max}} b(f)f. \tag{19}$$

Combining Eq. 12 and Eq. 19, the cost saved by Algorithm C with respect to A' becomes

$$T(A') - T(C) = (V - W)(T + 1) - T + \sum_{f=1}^{f_{\max}} b(f)f. \tag{20}$$

This shows that the extra time cost of calculating $n(f)$ in Step 1 of Algorithm C (which has cost $\theta(\max(V,f_{max})) \subseteq O(T)$) is balanced by a time saving in the remainder of the algorithm provided that $V - W$ is sufficiently large. In the next section, we will see that V is indeed much larger than W in real texts.

Table 1
Time and space cost of the algorithms

Algorithm	Time	Space
A	$O(VT^2)$	$\theta(V)$
A'	$\theta(VT)$	$\theta(V)$
C	$O(WT)$	$\theta(\max(V, f_{max}))$

T is the number of tokens, V is the number of types, W is the number of different frequencies and f_{max} is the largest frequency where $n(f) > 0$

Algorithm C needs $O(f_{max})$ additional memory if the values $n(1),\ldots,n(f),\ldots n(f_{max})$ are stored in a simple table and thus the overall memory cost of Algorithm C is $\theta(\max(V, f_{max}))$. In contrast, Algorithm A' needs only additional $O(1)$ memory and its overall memory cost is $\theta(V)$. It is crucial to know how big the cost of storing that table is compared to the memory cost of storing the table of type frequencies, that is $\theta(V)$. Table 1 summarizes the time and memory cost of each of the algorithms discussed so far.

In general, further space can be saved if the input is defined differently. If the input is defined by $n(1),\ldots,n(f),\ldots n(f_{max})$, then the space cost of Algorithm C reduces to $\theta(f_{max})$. If the input is defined by a $W \times 2$ matrix whose columns are f and $n(f)$ and whose rows contain only the values of f where $n(f) > 0$, then the space cost of that algorithm becomes $\theta(W)$. These kinds of improvements and their implications for other algorithms are left for future work.

Table 2
Summary of the statistical properties of the texts of the UDHR and the PBC.

		UDHR	PBC
T	Min	105	2836
	Mean	1801.5	290392
	Standard deviation	536.7	215641
	Max	4010	1257218
W/V	Min	0.016	0.0014
	Mean	0.057	0.037
	Standard deviation	0.027	0.031
	Max	0.17	0.21
f_{max}/V	Min	0.035	0.015
	Mean	0.25	1.9
	Standard deviation	0.19	2.4
	Max	1.3	33

T is the length in tokens, W/V is the ratio between the number of different frequencies and the number of types, f_{max}/V is the ratio between maximum frequency and the text length. The statistics on T were rounded to leave only one decimal digit. The statistics on W/V and f_{max}/V were rounded to leave only two significant digits.

3. Statistical Analyses

3.1. The Datasets

We use the Parallel Bible Corpus (PBC; Mayer, Cysouw 2014) and the Universal Declaration of Human Rights (UDHR) (http://www.unicode.org/udhr/) to obtain the frequency of word types across languages.

The version of the PBC we use here comprises 1491 texts that have been assigned 1118 unique ISO 639-3 codes, i.e. unique languages. Some languages are represented by several different translations. The PBC text files are semi-automatically processed to delimit word tokens from punctuation marks by white spaces. Note that this is a non-trivial task for some characters such as apostrophes and hyphens. Depending on the script used to write a language, they have to be either interpreted as part of word tokens or as punctuation. Based on the decisions made in the PBC, word tokens are here defined as strings of alphanumeric characters delimited by white spaces.

The UDHR comprises more than 400 texts with unique ISO codes. However, only 376 of these are fully converted into Unicode. UDHR files do not come with manually checked white spaces between word tokens and punctuation, and hence can bear more noise. We created frequency lists by splitting Unicode strings according to non-word characters, i.e. punctuation and space symbols. For some of the most widespread writing systems (e.g. Latin, Cyrillic, Devanagari and Arabic) the resulting lists of word types were checked by native speakers for misclassified items. In principle, this matches the method described for the PBC. Some of the UDHR texts had to be excluded due to incompleteness, or due to a script that does not support splitting word tokens by white space characters (e.g. Chinese or Mon-Khmer). This yields a sample of 356 UDHR texts with 333 unique ISO codes.

A summary of statistical properties of the lengths of the texts in each collection is provided in Table 2. The minimum T in the PBC (i.e. 2836 tokens) is too low for the whole Bible. It comes from a version of the Bible in Baruya (a language of Papua New Guinea) that has only a few verses translated.

3.2. Results

Table 2 summarizes the statistical properties of W/V in real texts. Interestingly, $W/V \approx 0.057$ on average for the UDHR and $W/V \approx 0.037$ on average for the PBC. This allows one to conclude that Algorithm C is at least 17 times faster than Algorithm A' on average (approximately). Figure 1 suggests a tendency of W to decrease as a function of V in the UDHR and the PBC. Such a tendency is supported by a Kendall τ correlation test: $\tau = -0.35$ and p-value $< 10^{-20}$ for the UDHR, $\tau = -0.084$ for the PBC and p-value $< 10^{-5}$. Our findings indicate that the number of main iterations of Algorithm C will tend to decrease as V increases. The fact that V is much larger than W in general (Fig 1) also indicates that the extra cost of Step 1 in Algorithm C can be neglected.

Table 2 also summarizes the statistical properties of f_{max}/V in real texts. Interestingly, $f_{max}/V \approx 0.25$ on average for the UDHR and $f_{max}/V \approx 1.9$ on average for the PBC. These findings indicate that the extra memory cost of Algorithm C is easy to tolerate in general. For the UDHR this is obvious because f_{max}/V does not exceed 1. Concerning the PBC, f_{max} is about two times V, but it is possible to store the input as a $W \times 2$ matrix employing $\theta(W)$ space as explained in Section 2 and we have already shown that W is on average 17 times smaller than V.

Figure 1 suggests a tendency of f_{max} to decrease as a function of T in the UDHR and the PBC. Such a tendency is supported by a Kendall τ correlation test: $\tau = -0.24$ and p-value $< 10^{-10}$ for the UDHR and $\tau = -0.073$ and p-value $< 10^{-4}$ for the PBC.

4. Discussion

We have provided an efficient algorithm to estimate entropy with less error than the naïve entropy estimator. Our algorithm takes advantage of the fact that W is much smaller than V to save computation time.

Our work opens up new possibilities for future research. First, our algorithm allows us to estimate entropy with Zhang's method in large corpora. Second, our investigation of the relationship between W and V or between f_{max} and V can be seen as emerging topics for research in quantitative linguistics. We have shown that W and f_{max} tend to decrease as V increases, especially for the UDHR. Figure 1 suggests that W decreases linearly with V for the UDHR but the actual functional dependency between W and V should be investigated further. Another issue for future research is the origins of the (at least) two clusters that can be seen in Fig. 1 for the PBC. We suspect that these clusters originate when merging translations of the Bible of different coverage (e.g., translations of the Old and the New Testament versus translations of the New Testament only).

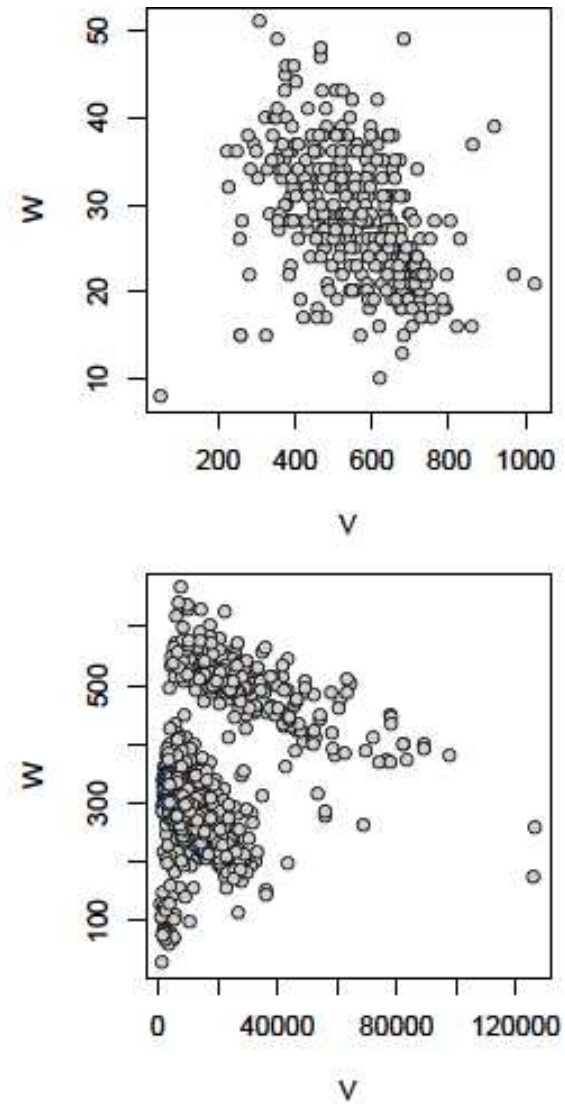


Figure 1. The number of different frequencies (*W*) versus the number of types (*V*) in parallel corpora. Every point corresponds to an individual text. Top: UDHR. Bottom: PBC.

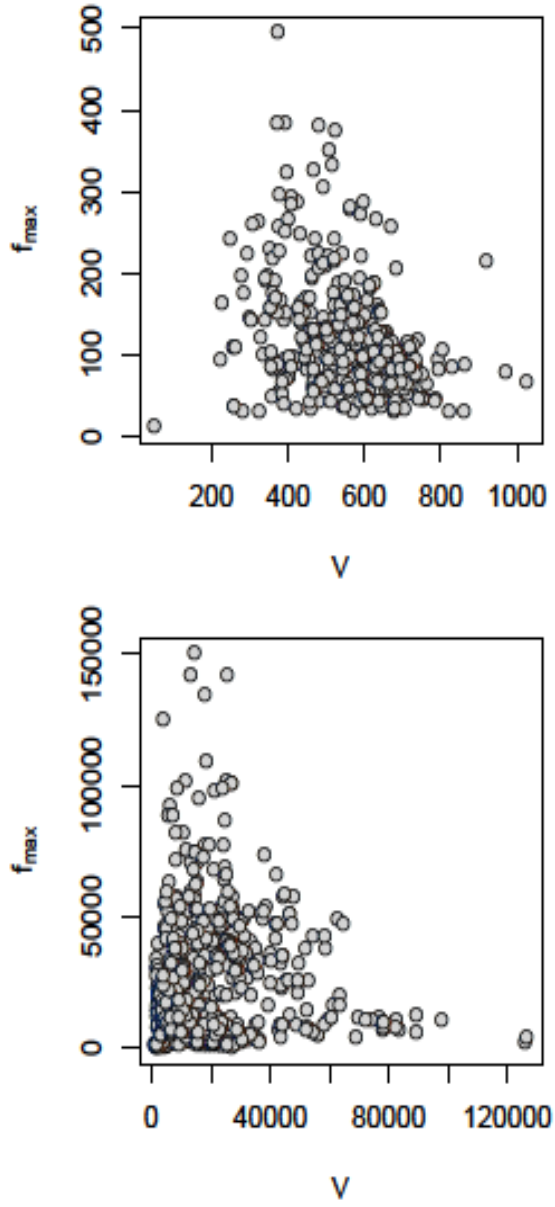


Figure 2. The maximum frequency (f_{max}) versus the number of types (V) in parallel corpora. Every point corresponds to an individual text. Top: UDHR. Bottom: PBC.

Acknowledgements

BC and RFC are funded by the grant 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya). AL, BC and RFC are also funded by the APCOM project (TIN2014-57226-P) from MINECO (Ministerio de Economia y Competitividad). CB is funded by the DFG Center for Advanced Studies "Words, Bones, Genes, Tools" and the ERC grant EVOLAEMP at the University of Tübingen.

References

Balasubrahmanyan, V.K., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics 3, 177-228.*

Bentz, C., Verkerk, A., Kiela, D., Hill, F., Buttery, P. (2015) Adaptive communication: languages with more non-native speakers tend to have fewer word forms. *PLoS ONE 10(6), e0128254.*

Chao, A., Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecolological Statistics 10, 429-443.*

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

Covington, M.A., McFall, J.D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics 17(2), 94-100.*

Ferrer-i-Cancho, R. (2005). Zipf's law from a communicative phase transition. *European Physical Journal B 47, 449-457.*

Ferrer-i-Cancho, R. (2014). Optimization models of natural communication. http://arxiv.org/abs/1412.2486

Ferrer-i-Cancho, R., Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences USA 100, 788-791.*

Grabchak, M., Zhang, Z., Zhang, D. T. (2013). Authorship attribution using entropy. *Journal of Quantitative Linguistics 20(4), 301-313.*

Hausser, J. & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research 10, 1469-1484.*

Jost, L. (2006). Entropy and diversity. *OIKOS 113(2), 363-375.*

Mayer, T., Cysouw, M (2014). Creating a massively parallel bible corpus. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., 3158-3163.* European Language Resources Association (ELRA).

Shannon, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal 30, 50-65.*

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3, 38-50.*

Vu, V.Q., Yu, B., Kass, R.E. (2007). Statistics in Medicine. Coverage-adjusted entropy estimation. *Statistics in Medicine 26, 4039-4060.*

Zhang, Z. (2012). Entropy estimation in Turing's perspective. *Neural Computation, 24(5), 1368-1389.*

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

Der Emeritus

Nachdem man an allen Klippen des Universitätslebens erfolgreich gestolpert ist, wird man zum Emeritus. Im offiziellen Jargon heißt er dann Em-exitus, was in sich ein Stück trauriger Wahrheit trägt. Für die Stiftungen bekommt er dann das Epitheton „Komposty“, was die höhere Evolutionstufe von „Grufty“ ist. Ein Komposty ist kein Wissenschaftler mehr, er ist nur noch lästig. Daher nennen ihn einige Stiftungen auch „Komplästy“ und seine Anträge auf Projekte nehmen nicht den Umweg über den Papierkorb, sondern gehen direkt in die Tonne für Papierentsorgung.

Für den Emeritus macht man öffentlich eine Abschiedsfeier, bei der fast alle Anwesenden weinen (vor Freude, aber dies ist nicht ganz sicher). Die Laudatio, die üblicherweise der Dekan vorliest, ähnelt eher einem Nekrolog, aber einem sehr positiven. „Wie schön war es, als er noch unter uns weilte, den Verlauf der Fakultätssitzungen mit unnötigen Anmerkungen aufhielt, obwohl er wusste, dass wir alle nach Hause gehen wollen, um uns Fußball anzuschauen. Wie schön war es, dass wir wussten, dass er jegliche Meldungen an den Rektor als letzter abschickt, so dass wir uns alle Zeit lassen konnten. Wie schön war es, dass er uns ständig mit originellen Einfällen unterhielt, bei denen es ausschließlich ums Geld für seine zahlreichen Hilfskräfte ging. Sic transit gloria mundi, wie mit ihrem schönen Mund seine Frau Gloria sagte. Jetzt stehen wir ohne ihn da, wir werden ihn eine zeitlang vermissen, dann kommt hoffentlich ein anderer Professor, mit dem man auch reden kann.“

„Was seine wissenschaftliche Tätigkeit betrifft, so frage ich Sie alle: Wann hätte er Wissenschaft machen sollen, wenn er ständig mit Lehrplänen, Berichten, Sitzungen, Streitigkeiten um Assistentenstellen, Ausstattung des Instituts, Bachelorprogrammen, Prüfungen und Unterricht beschäftigt war? Schon das Aufzählen seiner Pflichten lässt uns wundern, dass er überhaupt etwas geschrieben hat. Und das, was er geschrieben hat, haben bösartige Kritiker sowieso durch den dicksten Kakao gezogen. Wir sind die einzigen, die seine Arbeit schätzen und sie in Ehre halten werden, mindestens bis zur nächsten Woche. Dann wird uns wieder die Administrative in ihre Klauen fangen und wir alle werden vergessen, was wir sind.“

„Lobend muss man auch seine Beziehung zu den Studenten erwähnen. Alle bekamen von ihm gute Noten und besonders die Studentinnen schätzten seine Gesellschaft - sogar in Seminarräumen. Wir alle wünschten uns so zu sein wie er es immer war, denn seine Humanität war sprichwörtlich. Er hielt seine Seminare oft in Cafeterias oder auch zuhause und scheute keine Mühe, es den Studenten bequem zu machen.“

„Seine Kenntnisse der Materie waren märchenhaft. Er wusste einfach alles, sogar das, was andere nicht wussten. Und sie haben es gar nicht gewusst, dass sie es nicht wissen, erst unser Emexitus hat es ihnen klar vor die Augen gehalten. Seine Diskussionsbeiträge bei wissenschaftlichen Konferenzen waren oft derartig durchschlagend, dass aus ihnen ganz neue Forschungsrichtungen entstanden sind, besonders in der klinischen Psychiatrie. Die Vortragenden waren oft der Ohnmacht nahe oder erwachten aus ihr erst nach dem Kongress. Man

konnte seinem Wissen nicht gleichgültig gegenüberstehen, man musste es bewundern. Er lernte spielend zehn Sprachen und sprach sie so perfekt, dass nur geborene Franzosen erkannten, dass er höchstwahrscheinlich ein Russe ist, und die Finnen hielten ihn für einen Italiener. Jedem ist bekannt, dass er das absichtlich gemacht hat, um seine Internationalität zur Schau zu stellen. Und nicht an der letzten Stelle, um sich über seine Gesprächspartner lustig zu machen. Er war ein Witzbold und am liebsten erzählte er Witze über Papst, Könige, Präsidenten, Religion und Ähnliches, zum Beispiel über Professoren. Das tat er oft auch im Unterricht, so dass seine Veranstaltungen immer voll besetzt waren. Bezeichnend ist auch, dass seine Hörer nicht dasselbe Fach studierten, sondern zu ihm in der Annahme kamen, dass es sich bei seinen Vorlesungen um ein Kabarett handelt. Er war allgemein beliebt und vielleicht war er auch schwul, aber wir liebten ihn nicht nur deswegen."

Nach dem Dekan macht ein Mitglied seines Instituts eine kurze aber intelligente Rede. Die Anwesenden packen wieder ihre belegten Brötchen ein, weil sie wissen, dass nach dem Schluss ein kleiner Empfang kommt, bei dem sie sich satt essen können. Und wenn schon ein Dozent redet, dann kommt der Schluss sehr schnell, weil kein Dozent bereit ist über seinen ehemaligen Vorgesetzten etwas Böses zu sagen.

Aber man kann sich auch irren. Nach dem Dozenten kommt noch der Rektor, hebt die beiden Hände, um die Entrüstung zu beruhigen und sagt mit bewegter Stimme:

„Meine Damen und Herren, es ist nie so gewesen, dass es nicht irgendwie gewesen wäre. Das haben schon die griechischen Philosophen festgestellt und es hat sich bis heute nicht geändert. Denn wenn es nicht so wäre wie es ist, dann wäre es irgendwie anders, aber auch in dem Falle wäre es irgendwie. Jedoch, auch wenn wir wissen, dass sich mit der Zeit alles ändert, wissen wir auch, dass die Zeit eine subjektive Täuschung darstellt, so dass zum Schluss wieder alles irgendwie ist oder auch anders. Auf jeden Fall bleibt es irgendwie. In diesem Sinne wünsche ich unserem Emexitus noch eine schöne lange Zeit, denn in der Rente kann man viel besser faulenzen als an der Uni."

Nachdem aus dem Publikum das erste laute Gähnen ertönt, schließt der Dekan eilig die Sitzung und lädt alle zu einem kleinen Umtrunk in der Institutsbibliothek ein. Die meisten sind nur deswegen gekommen, und bevor der Emexitus die Bibliothek betritt, ist alles schon aufgegessen.

Man stellt sich in eine Schlange, schüttelt ihm die Hand und verschwindet, weil, na ja, man hat ja Veranstaltungen …

„Junge, Junge," sagt der Emexitus laut. „War das eine schöne Feier! Schade, dass ich die Uni verlassen muss."

„Das ist richtig," sagt der Bibliothekar, der noch am letzten Brötchen kaut, „denn ich muss ja hier alles abschließen und es ist ziemlich spät."

Und so geht der Emexitus und in zwei Wochen weiß keiner, wer er war und was er unterrichtete. Aber auf jeden Fall kann er jetzt Wissenschaft nach Lust und Laune betreiben – wenn er bloß wüsste, was er schon immer sagen wollte …

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics* –** up to now 35 issues – containing articles treating similar themes. The abstracts can be found in http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

The contents of the last issue (Glottometrics 35, 2016) is as follows:

Herausgeber – Editors of Glottometrics