Editors:
Ivan Obradović, Emmerich Kelih, Reinhard Köhler

Editorial assistance:
Sheila Embleton, Hermann Moisl

IQLA

www.iqla.org

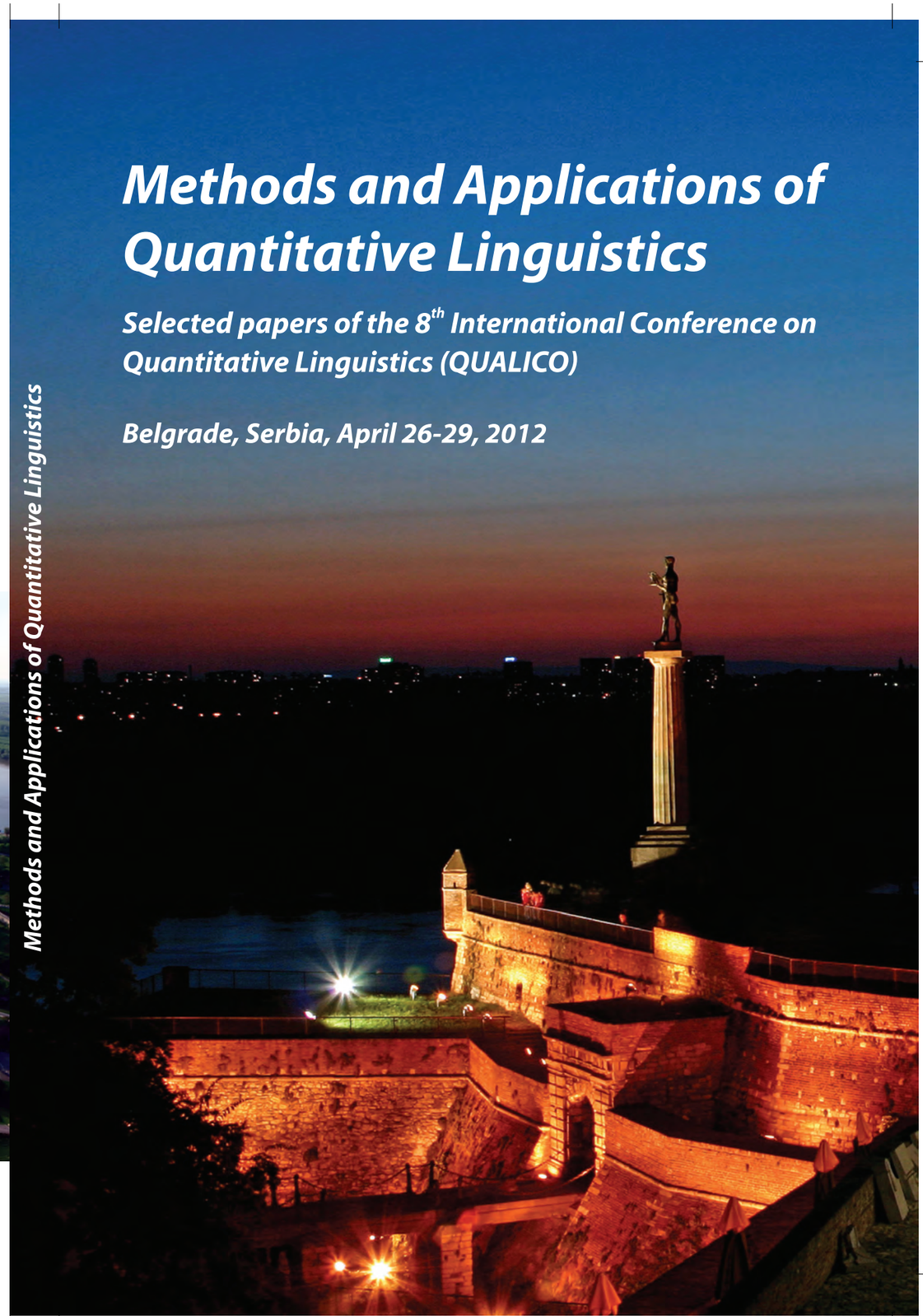
Printed with the support of the Government of the Republic of Serbia
Ministry of Education, Science and Technological Development

Methods and Applications of Quantitative Linguistics

Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)

Belgrade, Serbia, April 26-29, 2012

Methods and Applications of Quantitative Linguistics



Methods and Applications of Quantitative Linguistics

Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012

Editors:
Ivan Obradović, Emmerich Kelih, Reinhard Köhler

Editorial assistance:
Sheila Embleton, Hermann Moisl

Technical editor:
Milan Radonjić

Printed with the support of the Government of the Republic of Serbia Ministry of Education, Science and Technological Development

QUALICO April 26-29, 2012, Belgrade, Serbia

Programme Committee:

Peter Grzybek (Graz), Emmerich Kelih (Graz), Reinhard Köhler (Trier), Cvetana Krstev (Belgrade), Ján Mačutek (Vienna), George Mikros (Athens), Ivan Obradović (Belgrade), Duško Vitas (Belgrade)

Organising Committee:

Cvetana Krstev, Ivan Obradović, Duško Vitas

Publishers

UNIVERSITY OF BELGRADE

ACADEMIC MIND, Belgrade, Serbia

Printed in Serbia by

ACADEMIC MIND, Belgrade

Circulation

80 copies

ISBN 978-86-7466-465-0

Previous QUALICO Proceedings:

Köhler, Reinhard; Rieger; Burghard B. (eds.) (1993): Contributions to Quantitative Linguistics. Proceedings of the First International Conference of Quantitative Linguistics, QUALICO, Trier, 1991. Dordrecht; Boston; London: Kluwer Acad. Publ.

Polikarpov, Anatolij A. (ed.) (1994): Qualico-94. 2-aja Meždunarodnaja konferencija po kvantitativnoj lingvistike 20-24 sentjabrja 1994 g. Moskva: MGU, Filologičeskij fakul'tet.

Köhler, Reinhard (ed.) (2009): Issues in Quantitative Linguistics. Lüdenscheid: Ram-Verlag (Studies in Quantitative Linguistics, 5).

Grzybek, Peter; Kelih, Emmerich; Mačutek, Ján (eds.) (2010): Text and Language: Structures • Functions • Interrelations • Quantitative • Perspectives. Wien: Praesens.

Preface

The contributions to this volume are selected and peer-reviewed papers on the basis of talks which were delivered on the occasion of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia. The first QUALICO was held in Trier, Germany in 1991, the second one was organised in Moscow in the year 1994 by colleagues from the Faculty of Humanities of the Lomonosov University in Moscow and the department of Computational Linguistics at Trier University, during which the International Quantitative Linguistics Association (IQLA) was founded. Follow-up conferences have been organised regularly since 1997 by IQLA (QUALICO 1997 in Helsinki, Finland, QUALICO 2000 in Prague, Czech Republic; QUALICO 2003 Athens, Georgia USA, 5th Trier Symposium on QL 2007 and QUALICO 2009 in Graz, Austria).

QUALICO 2012 was, as usual, open to original papers without any limitations as to thematic or methodological aspects. Therefore, a large variety of topics has been addressed and numerous fields of linguistics and related fields have been touched.

A number of contributions to the conference can be subsumed under the general problem of **stylometrics**, in particular authorship attribution or text classification with respect to stylistic features unveiling the authors' membership in social or other groups, among them gender, age, or literary tradition. In this volume, *George Mikros* presents a stylometric approach to the analysis of the language of blogs, where authorship attribution and gender determination are the main topics. Another paper in this group *is D. Benedetto's, M. Degli Esposti's, and G. Maspero's* report on a specific philological problem of disputed authorship. The approach presented is a mathematical method using an estimation of entropy.

Fundamental linguistic research is focussed on by another group of contributions. *Beliankou, Köhler, and Naumann* introduce us into a new field of quantitative-linguistic research by giving an example of a study on the analysis of the discourse-pragmatic problem of argumentation structures. *Ján Mačutek and Radek Čech* investigate the diversification of the cases of Czech nouns; their results support earlier investigations into comparable matters. Furthermore, interesting facts are found and reported which shed new light on frequency effects in morphology. On the background of the Menzerath-Altmann Law, *Peter Grzybek* offers results of his research on the relations between sentence length and the lengths of other linguistic levels, which are not directly adjacent to the sentence level. A new mathematical approach to Polikarpov's and Krylov's "word life cycle" hypothesis is presented by *Vasiliy Poddubnyy and Anatoliy Polikarpov*. They describe a dynamic model of the evolution of linguistic signs based on dissipative processes and provide the empirical data on which the model was tested. *Matilde Trevisiani's and Arjuna Tuzzi's* paper

presents a mathematical masterpiece as an attempt at unveiling chronological patterns in the frequency of words in a text corpus. The specifity of this study is that the mathematical model tries to capture not only general tendencies, which could be motivated by linguistic or philological considerations, but at the same time the natural fluctuations and statistical deviations as a part of the deterministic model. *Makoto Yamazaki* delivers one of the so far very few quantitative works on text cohesion. His method is based on finding lexically identical elements in sequences of text passages.

The following papers can be grouped under the label **application-oriented**: *Gordana Đuraš, Ernst Stadlober and Emmerich Kelih* focus their study on the analysis of the distribution of word length in Slovenian and Russian texts of different type. Within a text classificatory context, they propose generalisations of the Poisson distribution and compare several commonly used estimators. *Ivan Obradović* presents a method for extracting semantically related word pairs from aligned texts using frequency based-ranking. The results are promising with respect to computational-linguistic applications and may be of importance to automatic translation systems. Japanese texts are analysed by *Haruko Sanada* with respect to their thematic concentration using Altmann's and Popescu's TC measure, which is based on the H index as introduced into linguistics by Ioan-Iovitz Popescu. The results suggest that thematic concentration is heavily influenced by text type.

Another group consists of contributions with **methodological** topics. *Łukasz Dębowski* tests Hilberg's conjecture for a selection of English prose using the Lempel-Ziv algorithm and finds an upper bound for the exponent. This contribution is interesting mainly from a mathematical point of view and may have implications in informatics. String similarity and the comparison of methods to find cognates is dealt with by *Antonella Delmestri and Liviu P. Dinu*. They suggest that orthographic learning methods may accurately detect traces of sound changes left in the orthography and outperform static phonetic systems. *Jiří Milička* delivers a paper on the mathematical relation between TTR and rank-frequency distribution. He describes a method to predict the development of the TTR curve for large corpora. Non-linearity is the focus of *Hermann Moisl's* paper. He demonstrates the nature of data nonlinearity, reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects. *Relja Vulanović* extends the Piotrowski-Altmann Law to several dimensions by generalizing the logistic differential equation to a system of partial differential equations. The application of multi-dimensional variants of the law is exemplified on data from cross-linguistic studies of parts-of-speech systems.

All in all the selected papers provide a broad overview on current research topics in quantitative linguistics and related areas.

The editors, July 2013

CONTENTS

PREFACE **4**

I STYLOMETRICS **9**

Dario Benedetto, Mirko Degli Esposti, Giulio Maspero
Authorship Attribution and Small Scales Analysis Applied to a Real Philological Problem in Greek Patristics 11

George Mikros
Authorship Attribution and Gender Identification in Greek Blogs 21

II FUNDAMENTAL LINGUISTIC RESEARCH **33**

Andrei Beliankou, Reinhard Köhler, Sven Naumann
Quantitative Properties of Argumentation Motifs 35

Peter Grzybek
Close and Distant Relatives of the Sentence: Some Results from Russian 44

Ján Mačutek, Radek Čech
Frequency and Declensional Morphology of Czech Nouns 59

Vasiliy Poddubnyy, Anatoliy Polikarpov
Stochastic Dynamic Model of Evolution of Language Signs Ensembles 69

Matilde Trevisani, Arjuna Tuzzi
Shaping the history of words 84

Makoto Yamazaki
Quantitative Analysis of Text Structure Using Co-occurrence of Words 96

III APPLICATION-ORIENTED RESEARCH **105**

Gordana Đuraš, Ernst Stadlober, Emmerich Kelih
The Generalized Poisson Distributions as Models of Word Length Frequencies 107

Ivan Obradović
A Method for Extracting Translational Equivalents from Aligned Texts 119

Haruko Sanada
Thematic concentration in Japanese prose 130

IV METHODOLOGICAL ISSUES **141**

Łukasz Dębowski
Empirical Evidence for Hilberg's Conjecture in Single-Author Texts 143

Antonella Delmestri, Liviu P. Dinu
An Assessment of String Similarity Methods for Cognate Identification 152

Jiří Milička
Rank-frequency Relation & Type-token Relation: Two Sides of the Same Coin 163

Hermann Moisl
Measurement of nonlinear distance in data derived from linguistic corpora 172

Relja Vulanović
A Multidimensional Generalization of the Piotrowski-Altmann Law 184

PART I

STYLOMETRICS

Authorship Attribution and Small Scales Analysis Applied to a Real Philological Problem in Greek Patristics

Dario Benedetto[1], Mirko Degli Esposti[2], Giulio Maspero[3]

[1]Dipartimento di Matematica, Università Sapienza, Roma, Italy.
benedetto@mat.uniroma1.it
[2]Dipartimento di Matematica, Università di Bologna, Italy.
mirko.degliesposti@unibo.it
[3]Pontifical University of the Holy Cross, Rome, Italy.
maspero@pusc.it

Abstract. We combine the traditional philological approach and recent mathematical techniques to a real case in Authorship Attribution (A. A.): the attribution of a letter written in Greek in the 4th century by Basil of Caesarea or his brother Gregory of Nyssa, two influential Christian theologians. Their extensive work is fully analysed and compared having recourse to the methods discussed in [1] and [2]. Here we refine our analysis introducing a new method that makes possible the attribution and comparison of texts at smaller scales. Our novel method is based on two similarity (pseudo) distances, based, respectively, on the statistics of *n*-grams and on zip-like algorithms [1].

Keywords: authorship attribution, n-grams, entropic methods, zip-like algorithms, Greek

1 Introduction

This contribution is a development of the work presented in the talk given at the Qualico 2012 conference. This original material (see Section 5) is heavily based on references [1] and [2], where the interested reader can find additional details and results.

The main point is that we approach a real problem in Authorship Attribution (A.A.) combining mathematical tools and philology [7]. In this way we have been able to develop our methods taking into account the specific problem and its characteristics. The main contribution of the present article is the introduction of a new method for attribution at small scales. This makes possible the analysis of multi-author contributions *inside* a single work. We have first tested the method on artificial texts and

applied it later to the specific problem. We will extend this kind of analysis to a select group of works of the same corpus, which cannot be correctly attributed, but may be a collection of writings by different authors.

2 Basil, Gregory and the disputed Ep. 38

Basil of Caesarea [3] and his brother Gregory of Nyssa [4] were two main theologians of the 4th century, whose thought and teachings were fundamental for the definition of the Christian doctrine on the Trinity. They wrote a large corpus of works debating with Eunomius of Cyzicus, a heretic who denied that God is one and three as Christian reading of the Gospel implies. Their discussion was almost thirty years long and many works were devoted to it. *Epistula 38* (Ep. 38) is one of them [1]. It is a letter that was transmitted in Basil's epistolary corpus [5], but that has also been attributed to his brother Gregory of Nyssa.

Ep. 38 is in some way a perfect A.A. problem: it is known in advance that the work belongs to either Basil or Gregory, i.e. it is a genuine two-class classification problem. Moreover, the statistical approach is expected to give good results, as the productions of these authors are conspicuous. The key characteristic of the problem that raises our hope of good results is that both Basil and Gregory of Nyssa have produced extensive works against the same Eunomius, in such a way that they discuss the same subjects and use the same quotations and vocabulary.

It seems reasonable to suppose that these works can be effective testing elements to compare the possibility of authorship of Ep. 38 for Basil and Gregory. In fact, the differences according to a statistical distance between Ep. 38 and Basil's or Gregory's works to counter Eunomius should be principally due to their personal styles, because the contents are very similar. And Ep.38 has the same subject as these works: this is another good point that justifies a quantitative approach to author style recognition, as the one presented here and in more detail in [1].

3 The Corpus

The Corpus studied is composed of all the known works by Basil and Gregory of Nyssa. The digitalized texts can be found in *Thesaurus Linguae Graecae* (TLG)[1]. included all the known works, also the spurious and dubious ones (see [1] or [2] for a detailed list). In fact, our main concern is to develop a reliable A.A. method for exploring not only the clearly attributed works, but we try to extend the analysis also to spurious and dubious ones. Because of that, the whole corpus has been divided into 3 sets with specific tasks in mind:

1. A *reference set*: the work by Basil and the sum of the three books by Gregory of Nyssa written against Eunomius, that from now on will be denoted

[1] http://www.tlg.uci.edu/. The digital library has been developed by the University of California, Irvine.

respectively as B0 and G0. We must carefully take into account the fact that the latter is almost 6 times longer than the former. In fact, the (character) lengths of these two texts, after a suitable coding explained in [1] are

$$|\mathrm{B0}| = 172\,342 \text{ and } |\mathrm{G0}| = 1\,017\,314.$$

2. A large *controlled corpus* composed by the works and letters composed both by Basil and by Gregory. The former wrote 43 works and more than 300 letters with undisputed authorship [5]. Due to their numbers and great difference in length, the letters have been divided into three groups, according to their extensions: the first one includes the letters longer than 2500 characters (L); the letters whose length is between 2500 and 1250 form the second group (M) and the letters shorter than 1250 constitute the third one (S). Correspondingly, also 57 works and 25 letters by Gregory have been included in this set [1]. This corpus has been used, as we will soon explain, in order to verify the efficiency and the stability of our method.
3. a set of almost 100 *dubious texts* which contains Ep. 38, i.e. the work under investigation.

The problem is approached by "measuring" the distances between the works in the second corpus and the reference set, in order to be able to finally analyse Ep. 38, together with the other writings with disputed attributions.

4 The methods, numerical indicators and ranking

The two methods implemented here are the development of two methods that have already been used in a project concerning the attribution of Antonio Gramsci's papers [6]. Each method essentially defines a kind of similarity distance between texts: given any pair of texts, we compute a positive number that can be interpreted as the distance between the texts. A small distance means that the two texts are quite similar, whereas a large distance means a high degree of dissimilarity.

The first method is based on n-grams frequency distribution and it is one of the simplest possible measures of the similarity between texts (we refer to [1] for details). Our second method (LZEW) is based on *data compression* and its role in the estimation of the entropy of a source. In short, we estimate the similarity of the two texts measuring the number, the positions and the lengths of the common subsequences of the texts, in a way suggested by compression algorithms like LZ77.

The very basic idea is that each method in principle allows one to compare any given text to be attributed with the texts of known attribution in the reference set. A key point is that both methods are quite sensitive to the length of the reference texts (in particular the entropic one): longer texts are usually richer and contain several different expressions, leading to small similarity distances that collocate them in the neighbourhood of any text, whoever the real author is.

In order to construct a reference corpus with a homogeneous distribution of lengths of the texts, we have split B0 and G0 in pieces of approximately the same dimension. In this way we have transformed the problem of managing texts of different sizes into the problem of managing a different number of texts of the same size.

To be more precise, we analyse an unknown text by calculating its similarity distance (using either one of the two methods) with all the subsets of G0 and of B0, respectively, and we put these values in an ordered list, from the closest fragment to the furthest one, extracting the rank of the authors (G = Gregory and B = Basil). For that we introduce an efficient *voting algorithm* that allows us to summarize all the information present in the ranking though a single number. We might call it a *basileanity index*: a positive value of that index means that the text is closer to the Basilean corpus, a negative value means that the text is closer to the Gregorian corpus. The greater is the distance from zero of the index, the greater is the reliability of the attribution.

The length used for the splitting of the reference texts has been selected through simple empirical considerations. These suggest that we combine analyses at two different scales:

- A large scale where B0 is left untouched. In this way we have cuts of about 170,000 characters long and G0 is divided in 6 equal parts.
- A small scale with cuts of about 11,000 characters long. With this choice we have exactly 16 parts for B0 and 94 parts for G0.

The choice of two very different scales of comparison can be heuristically motivated: confronting any unknown text with pieces that are 170,000 characters long enhances the weight of rare stylistic features, whereas using cuts of 11,000 characters long amplifies frequently used stylistic patterns. In order to correlate the different information arising from the two different scales analysis, both the entropic distance and the *n*-grams distance with n = 11 have been implemented at these scales, yielding four different attributions for each disputed text.

In summary, we have selected four methods: LZWE-170, LZWE-11 (the entropic method with B0 and G0 split in texts of size 170,000 and 11,000 respectively), N-11-170 and N-11-11 (the n-grams methods with $n = 11$ with B0 and G0 divided as before).

We applied them to the whole of Basil's corpus and to Gregory's one. For each given text, a fixed single method returns an attribution either to Basil (B) or Gregory (G). Combining now all the methods, we can have up to 5 different results, namely: all 4 methods return the same attribution (B or G), only three methods coherently give the same attribution or, finally, the draw possibility, when 2 methods return B, while the other two attribute the text to G. Our global attribution strategy was to consider a text attributed if and only if at least three methods attributed it to the same author.

The final results of our computations are summarized through the standard indexes from information retrieval, excluding the short Basil M and S letters that are analysed separately [1]:

	Basil	**Gregor**
recall	0.87	0.90
precision	0.96	0.93
F-measure	0.91	0.91

We remark that if we choose as true attribution only in the case of complete agreement of the four methods, the overall precision clearly increases (up to 97% for

both authors) but with a natural degradation of the recall (around 70-80%) and the F-score (below 90%).

We now turn our attention to Ep. 38. The good results on the controlled corpus of already attributed works suggest that the answer to the attribution problem for the letter under investigation should be meaningful. Our final result is that Ep. 38 is attributed to Gregory by all the methods. The high precision (97%) in the cases of complete agreement of the four methods seems to suggest that the answer can be trusted. Moreover the letter is 18,083 characters long, i.e. of sufficient length to give good statistics.

Nevertheless, the large philological interest for this letter can motivate further analysis. In particular, in the next Section, we try to develop a method capable of analysing the internal structure of the work, in such a way that we can check if the attribution changes within the text.

5 Going to smaller scales

We now introduce a novel method in order to analyse the texts of our corpus at smaller scale, looking for possible pieces attributed to different authors: given a text of length l, we compute the *Basilean index* over a sliding window of length n with each one of the methods described in the previous section. The window slides from the character in the first position $c = 1$ to the character in the position $c = l - n$, and the value of the index is assigned to the central position of the window, namely $c + n/2$. In this way we define the index function $I(c)$, $c \in [n/2,\ l - n/2]$.

It is quite natural to interpret the points of sudden change of $I(c)$ as the points of the text at the intersection between two regions written by different authors, or at least where the author modified his style or subject.

In order to obtain a good method, it is necessary to analyse different values of the scale n, while keeping in mind two opposite requirements: n must be small, if we aim to analyse the fine structure of the text, but n must be large enough to give good statistics and hence yield a reliable attribution of the fragment. After intensive numerical computations, we can empirically conclude that a good compromise is $n = 2,500$. Moreover we decided to use only the methods with cuts of size 11,000: the choice of 170,000 would only give 6 possible values of the *Basilean index*, making its variation along the text too discontinuous and less readable.

We have also to choose between the n-grams methods and the entropic method. An ideal index should have the important additivity property: if A and B are two sections of a given text (or just two different texts), then

$$I(AB) = I(A) \times |A|/(|A| + |B|) + I(B) \times |B|/(|A| + |B|) \tag{1}$$

i.e. $I(AB)$ should be equal to the weighted average of $I(A)$ and $I(B)$. This property will ensure the coherence between the (less accurate) information at smaller scale and the more solid information we gain at higher ones. Unfortunately neither the values of the relative entropy nor the similarities computed through n-grams are

additive. Because of this, it is clear that neither our indexes, obtained from the values through the complex ranking procedure, will be additive.

Nevertheless, we can numerically estimate how much the methods violate the additivity property. We have used 7 texts belonging to *B* and 7 texts belonging to *G*, chosen in such a way to represent different sizes, and we have verified the behaviour of the entropic and the *n*-grams methods: we have sliced the texts in sections of 2,500 characters and we have compared the *Basilean index* ("index" in the table) with the average of the index values over the small sections ("mean" in the table). Finally, with $\triangle$ we indicate the difference. We do this analysis using 12-grams.

text	index	mean	$\triangle$	index	mean	$\triangle$
B078	4.29	0.95	3.34	7.73	6.46	1.27
B085	0.52	0.34	0.18	1.79	2.38	-0.59
B086	9.76	5.59	4.17	5.41	5.7	-0.29
B093	13.61	4.77	8.84	6.81	5.56	1.25
B096	15.04	9.32	5.72	17.58	13.82	3.76
B097	14.66	9.19	5.47	14.39	11.52	2.87
B544	17.64	9.78	7.86	17.57	12.91	4.66
G001	-6.99	-1.78	-5.21	-6.01	-4.06	-1.95
G003	-12.24	-7.79	-4.45	-10.04	-6.07	-3.97
G016	-12.95	-8.2	-4.75	-7.43	-5.27	-2.16
G041	-14.15	-7.74	-6.41	-7.98	-5.9	-2.08
G045	-14.41	-10.55	-3.86	-12.61	-9.11	-3.5
G549	-12.06	-5.23	-6.83	-3.8	-2.32	-1.48
G552	-12.09	-7.1	-4.99	-6.16	-4.49	-1.67

As we can see, the index mean always returns the correct attribution and LZEW-11 produces systematically smaller errors D. For this reason, we use only the method LZEW-11 to perform the small scale analysis. It is worth mentioning that the index mean is consistently smaller, in absolute value, than the index over the whole text, in agreement with the idea that for smaller texts attributions are more uncertain, viz. they have index values closer to 0.

We perform a test of this method using an artificial text, obtained from the concatenation of three texts, the first and the last by Gregory, the intermediate by Basil. The resulting text is 14,553 characters long and author changes are located at the positions 7,010 and 10,384. In Fig. 1. we represent, in black, the function $I(c)$. The horizontal line at 0 distinguishes the positive values of the index (*Basilean regions*) from the negative values (*Gregorian regions*).

How does one find the points where the author changes? It is important to remark that these points may not coincide with the positions where the graph intersects the horizontal line, because $I(c)$ is calculated on a region of 2,500 characters. To overcome this problem, we suggest the following procedure: first, we introduce a prior hypothesis on the number of pieces into which the text must be split with respect to either the authorship or the style. Then we construct a simplified model of the text in which the index of each piece is an unknown constant. This means that

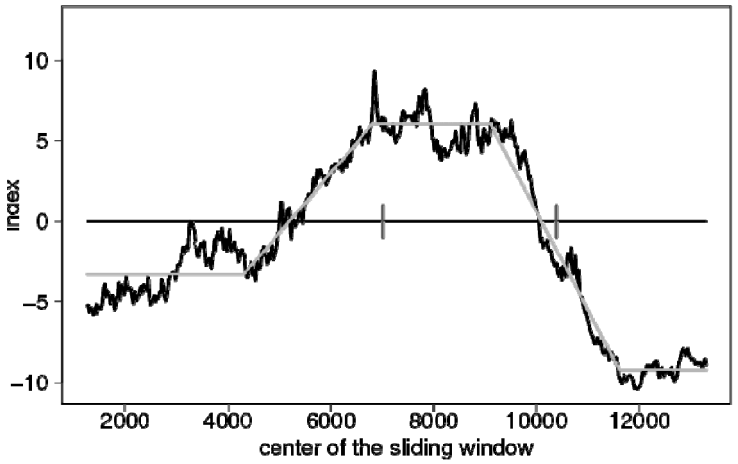


Fig. 1. The index function $I(c)$ for an artificial text (in black). The first and the third section are authored by Gregory, the second by Basil. The two short vertical lines show the points of separation of the three texts. The grey line is the index function $J(c)$ obtained via a 3-texts model.

also the points of separation of the different pieces are unknown. For any choice of the unknown parameters (the constants and the separation points), we can calculate the corresponding index function $J(c)$, and we find the better constants and the better separation points minimizing, over all the possible choices, the quadratic difference

$$\Sigma_c (J(c) - I(c))^2 \tag{2}$$

We performed this optimization procedure for the index function associated to the artificial text in Fig. 1. In Fig. 2 the grey lines represent the 3-texts model of the text, and the short vertical lines indicate the true point of separation of the text. The grey lines in Fig. 1 represent, instead, the graph of the corresponding $J(c)$. Note that the position of the second separation point is not exactly where $I(c) = 0$, but that this simplified model gives an almost correct value. On the contrary, the first value of separation for the first part is not well identified. The reason is that the first part of the artificial text, authored by Gregory, has an internal variation of the index: there is a main section with strong Gregorian features, but a final part classified as Basilean.

In order to take into account this phenomenon, it is useful to study a model with more separation points. In figures 3 and 4 we show the results obtained considering a 6-texts model. A consequence of the increasing number of parts is that we obtain a better agreement of $J(c)$ with $I(c)$ (see Fig. 3), and we are also able to detect the first transition point (*see* Fig. 4).

It is evident that the last part of the first text by Gregory is strongly attributed to Basil and that the central text authored by Basil is divided in two regions with a

Fig. 2. The 3-texts model for the artificial text. Only the second transition point is correctly detected.

different index value. Without this additional analysis, a first graphical investigation of the plot in Fig. 1 would probably lead to superficial considerations.

Let us note that the accuracy of the model obtained is measured by the number

$$\Sigma_c(J(c)-I(c))^2/\Sigma_c(I(c)-I)^2, \tag{3}$$

where I is the average of $I(c)$. This value is just the portion of the variance of $I(c)$ that the simplified model is able to capture. With the 3-texts model we achieve 93.81% of the variance, while with the 6-texts model we achieve 95.76%. This kind

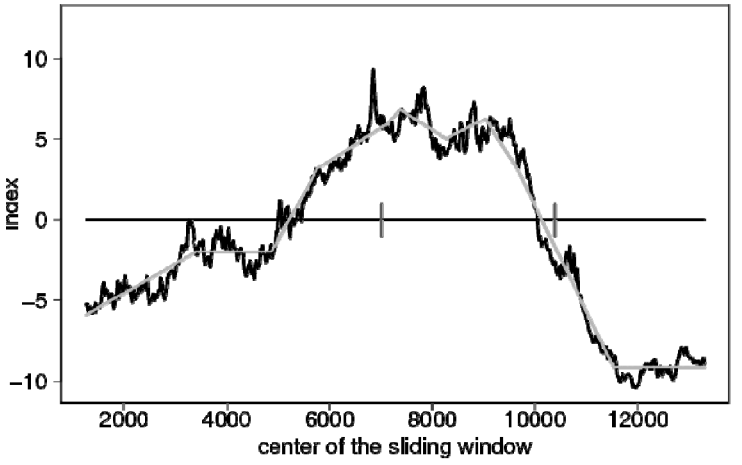


Fig. 3. $J(c)$ for the 6-texts model for the artificial text.

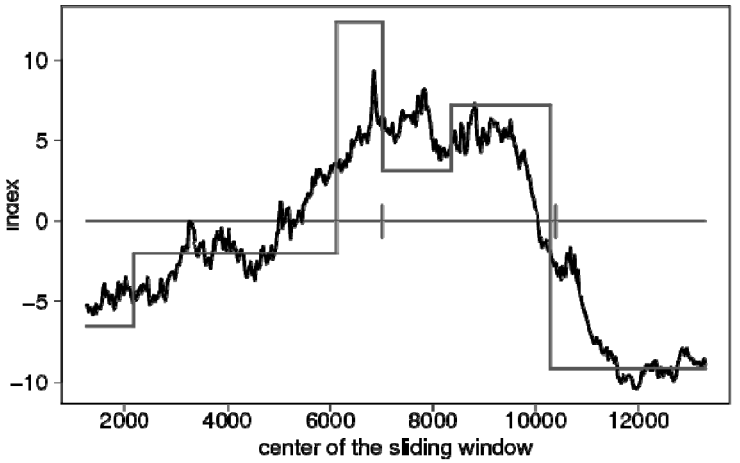


Fig. 4. The 6-texts model for the artificial text. Both the real transition points are correctly detected.

of analysis is quite delicate and the initial choice of the number of separation points of the text is crucial. A possible strategy for choosing this number is to increase the number until the explained variance is bigger than a given fixed value.

In Fig. 5 we show a 5-texts model analysis on Ep. 38 that is able to explain 85.24% of the variance. We note a strongly Gregorian segment of about 20,000 characters, and a little final segment with Basilean attribution. It is very interesting to note the presence of a plateau between 5,000 and 10,000 characterized by an index closer to zero. This can explain the difficulties in the attribution at the philological level. The Basilean shift at the end of the text is a common phenomenon and can be

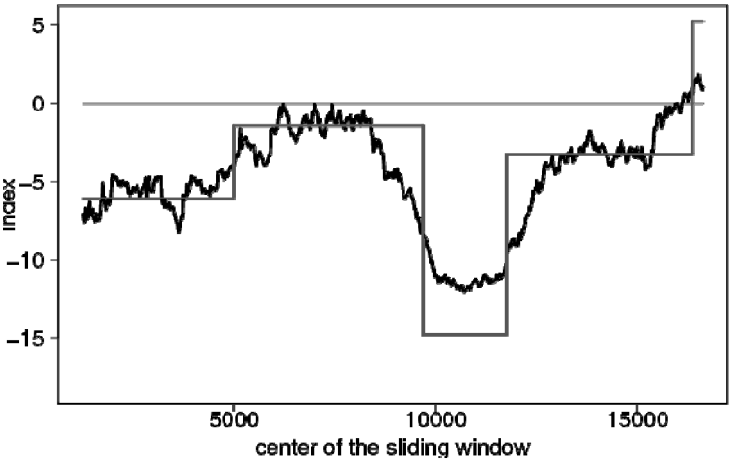


Fig. 5. The 5-texts model for Ep.38.

philologically explained with a change of style due to the consideration that the endings of literary pieces at that time were somehow fixed by rhetorical norms.

Conclusions

We have approached a real philological problem in A.A. designing our computational methods in order to take into account the specificities of Ep. 38 and of Basil of Caesarea's and Gregory of Nyssa's corpora. In particular, we have very strong indications that Ep. 38 was, in fact, written by Gregory of Nyssa.

Furthermore, we have developed a new method to study authorship at small scales, that can be very useful also in general. In fact, one typical cause of disputed authorships is that of multiple-authors who composed different parts of the works analysed.

Moreover, the small scales analysis has showed an internal structure of the work, that can explain the difficulties in the attribution process and that can be connected to a previous source of the work, like the act of a local synod or a circular letter, authored neither by Basil nor by Gregory. This hypothesis is worth being investigated both at the numerical and the philological level.

References

1. Benedetto, D., Degli Esposti. M., Maspero, G.: The puzzle of Basil's Epistula 38: a mathematical approach to a philological problem. To appear in the Journal of Quantitative Linguistics (2012)
2. Benedetto, D., Degli Esposti. M., Maspero, G.: Who wrote Basil's Epistula 38? A Possible Answer through Quantitative Analysis, in J. Leemans – M. Cassin (eds.), Gregory of Nyssa's Contra Eunomium III. Proceedings of the Twelfth International Gregory of Nyssa Colloquium (Leuven, 14-17 September 2010). Brill, Leuven, to appear (2012)
3. Rousseau, P.: Basil of Caesarea. University of California Press, Berkeley (CA), Los Angeles (CA), London (1998)
4. Meredith, A.: Gregory of Nyssa. Routledge, London, New York (1999)
5. Fedwick, P.J.: Bibliotheca Basiliana Universalis I. Brepols, Turnhout, 620-623, 674-678 (1993)
6. Basile C., Benedetto D., Caglioti E., Degli Esposti M.: An example of mathematical authorship attribution. J. Math. Phys. 125211 (2008)
7. Maspero, G., Leal, J.: Revisiting Tertullian's Authorship of the Passio Perpetuae through Quantitative Analysis. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.). Text and Language. Structures Functions – Interrelations – Quantitative Perspectives. Praesens, Wien, 99-108 (2010)

Authorship Attribution and Gender Identification in Greek Blogs

George K. Mikros

Department of Italian Language and Literature, School of Philosophy
National and Kapodistrian University of Athens, Athens, Greece
gmikros@isll.uoa.gr

Abstract. The aim of this study is to obtain authorship attribution and author's gender identification in a corpus of blogs written in Modern Greek language. More specifically, the corpus used contains 20 bloggers equally divided by gender (10 males & 10 females) with 50 blog posts from each author (1,000 posts in total and an overall size of 406,460 words). From this corpus we calculated a number of standard stylometric variables (e.g. word length statistics and various vocabulary "richness" indices) and 300 most frequent word and character *n*-grams (character and word unigrams, bigrams, trigrams). Support Vector Machines (SVM) were trained on this data, and the author's gender prediction accuracy in 10-fold cross-validation experiment reached 82.6% accuracy, a result that is comparable to current state-of-the-art author profiling systems. Authorship attribution accuracy reached 85.4%, an equally satisfying result given the large number of candidate authors $(n = 20)$.

Keywords: Authorship Attribution, Author profiling, Blogs, Machine Learning, Support Vector Machines, Gender Identification, Stylometry

1 Introduction

Over the last two decades Automatic Authorship Identification (AAI) has been evolved in a highly dynamic research strand exploiting recent advances in a number of fields like Artificial Intelligence, Linguistics and Computing. Furthermore, AAI research now is concerned not only with problems of authorship in the broad field of the Humanities (Literature, History, Theology), but also with applications in various law-enforcement tasks such as Intelligence, Forensics e.g. [1-4] etc. The major application areas are described below:

1. Authorship Attribution: This is the most common authorship identification analysis with the study of the *Federalist Papers* by Mosteller & Wallace [5]

being a typical example. In this case we are trying to find who is the author of one or more disputed texts among a closed set of $2, 3 \ldots n$ known authors. This scenario assumes that we are certain that at least one of the possible authors is actually the author of the disputed texts and that an adequate corpus in size and quality for every possible author is available [6].

2. Author verification: In this case we are investigating whether certain text(s) were written by a specific author. We are assuming an open set of authors and each dubious document must be attributed to the specific author without reference to corpora from other authors [7-9].
3. Author profiling: In some applications related to Information Retrieval or Opinion Mining and Sentiment Analysis we are interested in identifying the author's gender [10-12], age [13] or psychological type [14-16].

In this paper we will focus in the first and the third type of research, namely authorship attribution and author profiling. More specifically, we will try to detect automatically the author of a blog and his/her gender training machine learning algorithms using data from a Modern Greek blog corpus.

2 Language Usage in Blogs

During the last decade the Internet has evolved from a static field of simple information provision into a digital carrier of language production characterized by interactivity and dynamic configuration of the online textual content.

Blogs are among the best known Web tools that have transformed Web communication and overcame the unidirectionality of standard online communication. Up to 2011 approximately 181 million blogs have been created worldwide, producing 900,000 posts every day which are being read by 77% of internet users (Source: NM Incite). Since many blogs are important information nodes and attract many more readers than most of the traditional printed media, they can exert influence in language usage and produce linguistic innovations accelerating linguistic change. For this reason, blog language usage has started to attract attention and become a challenging research subject in the linguistic community.

Blogs represent a new text genre with interesting characteristics. They combine personal views, news and reporting on current events [17]. Their structure is a hybrid containing both monologue and dialogue features. At the same time they are both log entries reflecting personal opinions and open calls for public discussion [18]. Mishne [17] studied in detail various properties of linguistic usage in English blogs and showed that they present increased usage of personal pronouns and words relating to personal surroundings emerging from personal experience. Furthermore, he examined the linguistic complexity of the blogs using the perplexity measure [19] and the out-of-vocabulary rate (OOV) and found that their linguistic structure was more complex than most of the similar written genres (e.g. personal correspondence). Increased perplexity, according to Mishne, equates with increased irregularity in linguistic usage (i.e. free-form sentences, decreased compliance with grammatical rules etc.). In addition, blogs presented increased OOV rates, meaning

that blog texts exhibit a topical diffused vocabulary, with many neologisms, possible typographical errors and increased level of references to named entities from the blogger's personal environment.

Another interesting characteristic of the blog's linguistic structure is its equilibrium between spoken and written language. Sentence construction in blogs is highly variable using selectively structures from both spoken and written norms [20]. An equally important effect in language usage in blogs is the age of the bloggers. Half of the them are aged 18-34 (Source: The Social Media Report: Q3 2011, MN Incite, Nielsen). For this reason, formality in language usage is decreased, with shorter that average sentence lengths and lower readability scores in the best-known readability formulas (Gunning-Fog, Flesch-Kincaid, SMOG).

3 Gender Identification in Blogs. A Literature Review

Blogs' textual production is increasing rapidly. At the same time anonymous posting often covers illegal acts ranging from copyright infringement to criminal offences. AAI methods can be effectively employed in the framework of Forensic Linguistics. Due to their special linguistic structure described in the previous section, anonymous blog posts represent a serious challenge for both the stylometric features and the machine learning methods used to reveal a malicious blogger's identity [10, 11, 21, 22].

The detection of the blogger's gender is an equally important research issue with many possible applications including forensics, online audience identification for targeted advertisement and socio(linguistic) analysis on gender identity issues.

Schler et al. [10] used a large blog corpus (37,475 blog posts totaling 300 million words) and tried to predict both the authors' gender and age. The specific study used 1,502 features including specific content words, selected parts-of-speech, function words and blogs specific features such as "blog words" - lol, haha, ur etc. - and hyperlinks. The machine learning algorithm used was Multi-Class Real Winnow and the prediction accuracy for the author's gender reached 80.1%. Interestingly, the authors noted that despite the great diversity found among stereotyped word content usage between men and women, the most important gender distinctive features were semantically neutral (such as frequent functional words and Parts of Speech).

Argamon et al. [11] have also examined how age and gender affect writing style and topic in blog postings. They presented an analysis based on 140 million words mined from 46,747 English language blogs. They extracted the 1,000 most frequent words from this corpus and recorded their frequency in each blog. Using these data they performed a factor analysis in order to find groups of related words that tend to occur in similar documents. Results indicated that women bloggers prefer personal pronouns, conjunctions and auxiliary verbs while male bloggers use more articles and prepositions. Prediction accuracy of the bloggers' gender using the 1,000 most frequent words reached 80.5%. The researchers, however, warn that style and content effects are highly correlated and it may be that the choice of content determined particular style preferences, or both content and style may be influenced by a single underlying variable such as genre preference.

In another study [23], 73 Vietnamese bloggers' gender was predicted using a variety of machine learning algorithms and stylometric features based on character and word units. The classification accuracy for gender reached 83.3% with the word-based features to contribute more to the gender identification than the character-based features.

Mohtasseb and Ahmed [24] studied a large number of demographic characteristics of authors including gender in blog texts. They trained Support Vector Machines (SVMs) using various standard stylometric indices and 88 features from the Linguistic Inquiry Word Count (LIWC) [25], a special psycholinguistic lexical database that groups words into specific psychological categories. Results indicated that men's posts could be recognized more accurately that women's under all experimental conditions.

Mukherjee and Liu [26] also studied author gender classification in blog posts. They proposed a new class of features which are POS sequence patterns that are able to capture complex stylistic regularities of male and female authors. Furthermore, they proposed an ensemble feature selection method which takes advantage of many different types of feature selection criteria. These methods were tested in 3,100 blog posts and compared against known public domain gender detection systems (Gender Genie, Gender Guesser) and relative published algorithms [10, 11, 27]. In all cases their proposed methodology proved considerably more accurate.

Sarawgi et al. [28] studied the effect of text topic and genre in the accuracy of automatic gender identification methods. Using a sophisticated experimental design and multiple datasets (mostly blogs of different topics), they compared multiple machine learning methods controlling for genre and topic bias. They noticed that the most robust approach was based on character-level language models which used morphological patterns, rather than token-level language models that learned shallow lexico-syntactic patterns. In addition, they traced statistical evidence of gender-specific language styles beyond topics and genre, and even in modern scientific papers.

4 Research Methodology

4.1 The Greek Blog Corpus (GBC)

In order to explore authorship attribution and gender identification in Greek blogs we had to develop from scratch a Greek blog corpus (GBC). For this reason we harvested the Greek blogosphere from September 2010 till August 2011 and manually collected 100 Greek blogs equally divided to 50 male and 50 female bloggers. Since topic can induce significant bias into stylometric measurements [29], we decided to explore only a part of the collected corpus, using blogs that share the same topic. In this study we used 20 blogs (10 male and 10 female authors) with a common topic (Personal affairs), with a total of 1,000 blog posts counting 406,460 words. For each author we collected the 50 most recent blog posts.

A close examination of the word length descriptive statistics reveals that male and female bloggers produce texts that vary considerably in size even when the topic is

roughly the same. Female (fm) bloggers produce longer posts with less variation in size (Mfm = 423.4 words, SDfm = 243.6) than male (ma) bloggers (Mma = 389.5, SDma = 351.1).

4.2 Stylometric Features and Classification Algorithm

Authorship attribution has a long history of using a large variety of textual features in order to correlate them with a specific author's style. In the present study we will use a wide set of stylometric features in order observe their association with authorship and author gender. The feature list we mined is extensive and contains both "classic" stylometric features such as lexical "richness" and word length measures, and "modern" features borrowed from Information Retrieval and Language Modeling such as character and word n-grams. The detailed list of the features used in this study is the following:

"Classic" stylometric features

- Vocabulary "richness"
  - Yule's K, [30]
  - Functional Density, [31]
  - Percentage of Hapax and Dis-legomena
  - Ratio of Dis to Hapax-legomena, [32]
  - Lexical Entropy and Redundancy, [33]
- Word Length
  - Average Word Length – AWL (in characters)
  - Standard Deviation of Average Word Length – sd AWL
  - Word Length Spectrum: Normalized frequency of 1, 2, 3 ... 14-letter words.
- Letter frequencies
  - Normalized frequencies of each letter.

"Modern" features

- Character bigrams
- Character trigrams
- Unigrams (words)
- Word bigrams
- Word trigrams

For each character and word n-gram feature group described above we counted the 300 most frequent features and normalized their frequency in 100-word text size. Feature counting was performed using customized PERL scripts and the total vector size produced was 1,356 features.

This vector fed the Sequential Minimal Optimization (SMO) algorithm [34], an optimized version of the Support Vector Machines (SVMs). SVMs represent the state-of-the-art in machine learning methods regarding text classification and have been used extensively in authorship attribution research [35-38]. They are suited for

solving binary classification problems, though there are many extension methods that make them appropriate also for multi-class problems. They project the points of the training sample to a higher dimension area and find a hyperplane that separates with the best possible way the points of the two classes. Points from the testing sample are classified according to the side of the hyperplane in which they are located. Vectors which define the hyperplane are called support vectors.

Evaluation of the classification performance was obtained using accuracy, i.e. percentage of the texts that were attributed correctly to their author, or author's gender. In order to avoid random fluctuations in algorithm performance we used 10-fold cross-validation methodology, i.e. we took the mean accuracy of 10 different complementing training and testing cycles with each cycle to use 90% of the data as training sample and 10% as validation sample.

5 Results

Using the features and the algorithm described in the previous section we had 85.4 accuracy in authorship attribution and 82.6 in gender identification. Both reported accuracies can be considered as excellent regarding the data size and the number of candidate authors $(n = 20)$. This last parameter is very important since two-class authorship attribution problems are less demanding and most stylometric methods can successfully deal with them.

In order to understand better the impact of the number of candidate authors on the evolution of the authorship attribution accuracy we created a controlled experiment. We segmented our data into 4 size groups (2 authors, 4 authors, 8 authors, 16 authors). For each size group we selected 10 different author combinations using stratified random sampling. Reported accuracy measures are based on the mean of these 10 different author combinations in each size group. This method minimizes systematic errors which can intervene due to an unusual stylistic (dis)similarity between specific authors. In total we ran 40 (4×10) classification experiments using SMO in 10-fold cross-validation scheme. The mean accuracies are displayed in figure 1.

Classification performance is directly related to the number of candidate authors. When we examine 2 candidate authors the obtained classification accuracy is very high (97.7%). Accuracies drop as the number of candidate authors increases with the lowest accuracy reported in the 16-author group (86.9%). In order to evaluate further the impact of candidate group size to the classification accuracy, we examined the full experimental data using one-way ANOVA with dependent variable the obtained accuracies and independent variable the group size. Results were statistically significant at the 0.05 level $(F(3,36) = 53.4,\ p < 0.05)$ indicating that overall means in the different group sizes are indeed different. In order to further explore which specific group sizes differentiate, we applied the Tukey post hoc test. Results indicated that all group sizes differ statistically significantly between themselves except the 2 and 4-groups. This means that authorship attribution accuracy using the above mentioned combination of features and algorithm performs its best up to

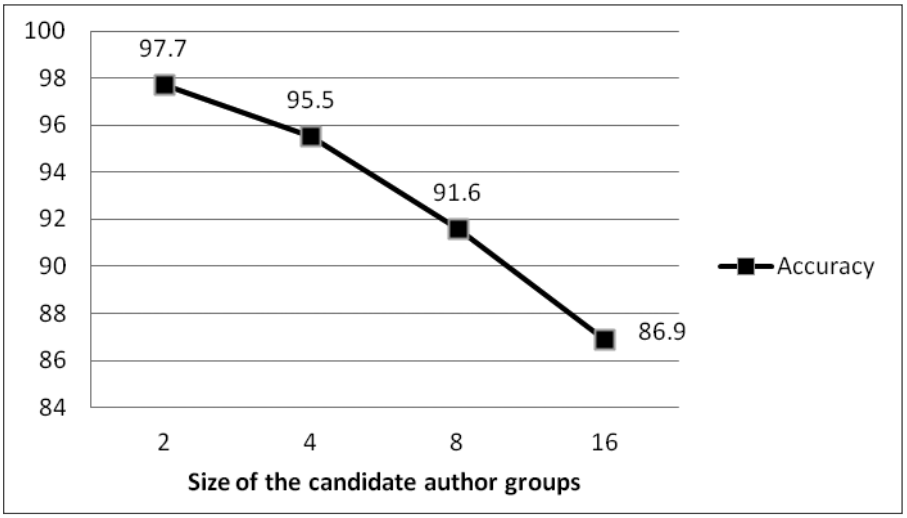


Fig. 1. Influence of the number of candidate authors

4 candidate authors and then its performance drops linearly as the number of the authors is increasing exponentially.

Another important research question we confronted was related to the influence of the stylometric features in each type of attribution, i.e. authorship and gender. We applied Information Gain [39], a well-known feature selection algorithm for text classification tasks and recorded the 10 most influential features in authorship attribution and gender identification task. The relative importance of each feature in the two classifications is displayed in figure 2.

A general conclusion that can be drawn from examining feature importance in the two classification tasks is that specific character *n*-grams carry significant authorship information while specific word n-grams have increased importance in author gender identification. Another finding that deserves comment is that word length measures (AWL, sd AWL) convey both authorship and gender evidence.

In order to explore further the way n-grams reveal authorship and gender patterns we performed authorship and gender classification using only these as features. We recorded the classification accuracy first using all *n*-gram features and in a second step we performed classification without a specific n-gram feature group. We subtracted the new accuracy from the one that was based on all n-grams, resulting in a relative difference that could be explained as the importance of the feature group that was missing, i.e. the larger the difference, the larger the importance of this feature group in the classification. We calculated all these differences by removing sequentially all *n*-gram feature groups one at a time for both classification tasks (authorship and gender). N-gram importance in relation to the classification task is displayed in figure 3.

In the above chart we observe two different tensions regarding word n-grams. As we move from words (ung) to word bigrams (wbg) and trigrams (wtg) gender identification is getting more accurate. The exact opposite trend can be observed in

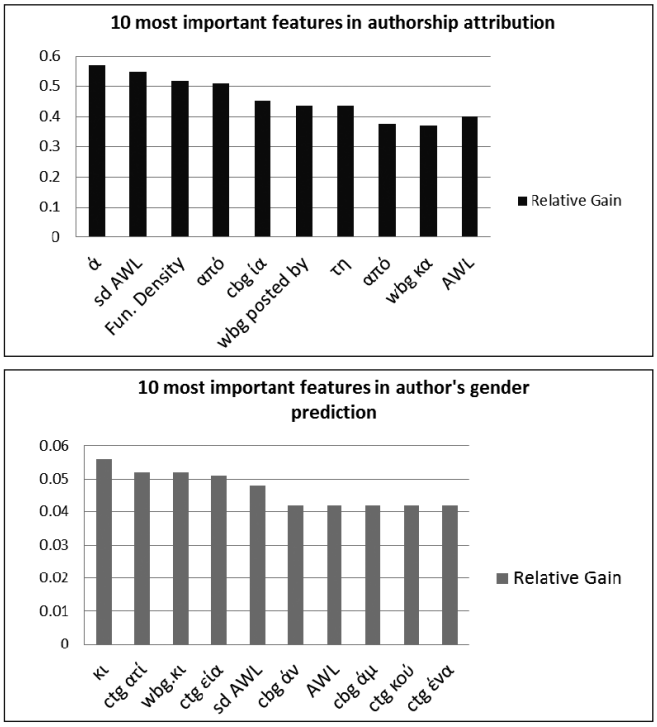


Fig. 2. 10 most important stylometric features for authorship attribution and gender identification.

authorship identification, where increasing the size of word sequences leads to a linear drop in accuracy. This trend could be related to the way the syntax is connected to these tasks. It seems that gender distinctions are associated with specific syntactic patterns while authorship is based more on most frequent words usage.

Character *n*-grams follow a similar trend. As we move towards longer sequences, gender identification becomes more accurate, meaning that morphological information is highly relevant to the way gender is manifested in a text. On the other hand, simple character frequency is the most productive feature group among subword features in authorship attribution. These trends in word and character *n*-grams reveal that authorship and gender classifications are quite different tasks which utilize complementary linguistic means. Author gender finds expression using specific morpho-syntactical patters which differentiate male from female authors. Authorship on the other hand, is based on the selection of high frequency words and their idiosyncratic usage by each author. This phenomenon is partly reflected in the representation of specific characters and their derived usefulness in authorship attribution, since specific very frequent words increase the frequency of their constituent characters. This complementarity, however, is not absolute. N-grams function as markers of both authorship and gender, and their increased discriminatory power in each of these tasks is just an indication that gender and authorship exploits more or less specific elements of the grammatical spectrum.

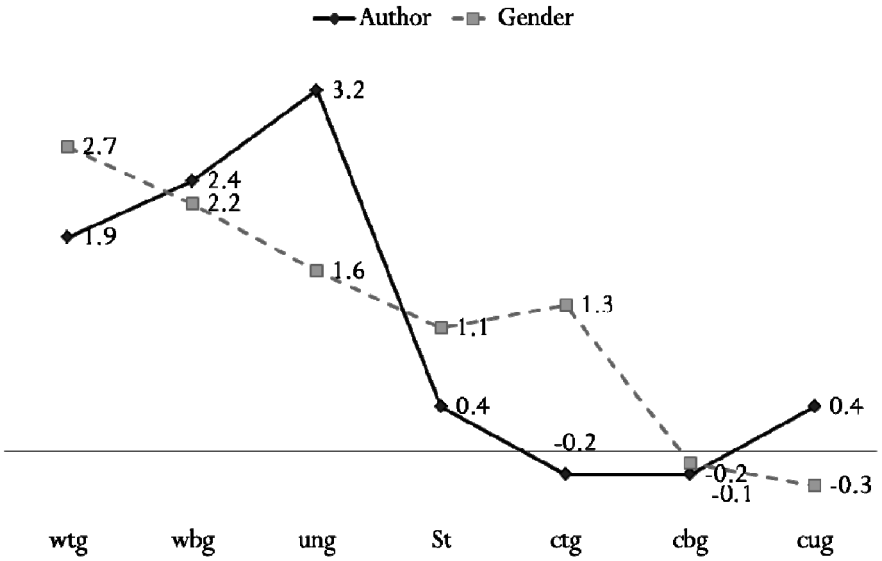


Fig. 3. Relative importance of *n*-gram feature groups in the authorship attribution and the gender identification task.

6 Conclusions

The present study has investigated methods for authorship attribution and gender identification in Greek blogs using state-of-the-art machine learning algorithm (SVM) and a large variety of stylometric features. Authorship attribution and author gender prediction in blog posts reached reasonable accuracy (85.4% & 82.6%) with many candidates $(n = 20)$.

Furthermore, the relation of authorship attribution accuracy to the number of candidate authors was examined. Using a controlled experimental design our methodology performed optimally up to 4 candidate authors. From this point, authorship attribution accuracy dropped linearly as the number of candidate authors was increased exponentially.

Another finding of this study was that author identification and gender detection are two different tasks with distinct patterns of stylometric feature interaction. As we moved towards longer lexical chains (bigger word *n*-grams), we noticed an increase in the author's gender identification accuracy. An opposite trend was spotted in the authorship classification task. As we moved towards single words (unigrams), we noticed an increase in author identification accuracy. The same trend was detected in the character *n*-grams. Longer sequences of character *n*-grams led to a better accuracy rate in gender identification while shorter n-grams and single characters boosted accuracy in authorship attribution. These observations lead us to the conclusion that author gender is conveyed through specific syntactical and morphological patterns while authorship seems to rely on over- or under-representation of specific high frequency words.

References

1. de Vel, O., Anderson, A., Corney, M.W., Mohay, G.: Multi – Topic E-mail Authorship Attribution Forensics. Proceedings of ACM Conference on Computer Security – Workshop on Data Mining for Security Applications. Philadelphia, PA, USA (2001).
2. Chaski, C.E.: Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence 4(1), pp. 1-13 (2005).
3. Iqbal, F., Binsalleeh, H., Fung, B.C.M., Debbabi, M.: Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation 7(1-2), pp. 56-64 (2010).
4. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Communications of the ACM 49(4), pp. 76-82 (2006).
5. Mosteller, F., Wallace, D.L.: Applied bayesian and classical inference. The case of The Federalist Papers. 2nd ed. Springer-Verlag, New York (1984).
6. Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3), pp. 233-334 (2008).
7. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. Proceedings of 21st International Conference on Machine Learning, July 2004, pp. 489-495. Banff, Canada (2004).
8. Van Halteren, H.: Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing (TSLP) 4(1), pp. 1-17 (2007).
9. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: E-mail Authorship Verification for Forensic Investigation. Proceedings of the 2010 ACM Symposium on Applied Computing (SAC '10), March 22-26, 2010, Sierre, Switzerland, pp. 1591-1598. ACM, New York (2010).
10. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 27-29 March 2006, Stanford, California, pp. 199-205. (2006).
11. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self–expression. First Monday, 12(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/ fm/article/view/ 2003/1878 (2007).
12. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), pp. 401-412 (2002).
13. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text 23(3), pp. 321-346 (2003).
14. Luyckx, K., Daelemans, W.: Personae: A corpus for author and personality prediction from text. In: Calzolari, N., et al. (eds.). Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28-30 May 2008. Marrakech, Morocco (2008).
15. Luyckx, K., Daelemans, W.: Using syntactic features to predict author personality from text. Proceedings of Digital Humanities 2008 (DH 2008), pp. 146-149. (2008).
16. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.: Lexical predictors of personality type. Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification, 8-12 Jun 2005. St. Louis, MO (2005).
17. Mishne, G.: Applied Text Analytics for Blogs. University of Amsterdam, Amsterdam (2007).
18. Nilsson, S.: The function of language to facilitate and maintain social networks in research weblogs. Umeå Universitet (2003).

19. Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J., Lai, J.C.: Class-based n-gram models of natural language. Computational Linguistics 18(4), pp. 467-479 (1992).
20. Chafe, W., Danielewicz, J.: Properties of spoken and written language. In: Horowitz, R., Samuels, J.S. (eds.). Comprehending oral and written language, pp. 83-113. Academic Press, New York (1987).
21. Mohtasseb, H., Ahmed, A.: Two-layered Blogger identification model integrating profile and instance-based methods. Knowledge and Information Systems, pp. 1-21 (2011).
22. Mohtasseb, H., Ahmed, A.: More blogging features for author identification. Proceedings of the International Conference on Knowledge Discovery (ICKD'09), 6-8 June 2009, Manila, Philippines pp. 534-539. (2009).
23. Dang Duc, P., Giang Binh, T., Son Bao, P.: Author profiling for vietnamese blogs. Asian Language Processing, 2009 (IALP '09), pp. 190-194. (2009).
24. Mohtasseb, H., Ahmed, A.: The Affects of Demographics Differentiations on Authorship Identification. In: Ao, S.-I., Gelman, L. (eds.). Electronic Engineering and Computing Technology, Vol. 60, pp. 409-417. Springer, Heidelberg (2010).
25. Pennebaker, J.W., Francis, M.E.: Linguistic Inquiry and Word Count: LIWC2001. Erlbaum Publishers Mahwah, NJ (2001).
26. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 207-217. Association for Computational Linguistics, Cambridge, Massachusetts (2010).
27. Yan, X., Yan, L.: Gender classification of weblog authors. Computational Approaches to Analyzing Weblogs, 27-29 March 2006, Stanford University, California, USA, pp. 228-230. American Association of Artificial Intelligence, (2006).
28. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: tracing stylometric evidence beyond topic and genre. Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 78-86. Association for Computational Linguistics, Portland, Oregon (2011).
29. Mikros, G.K., Argiri, E.K.: Investigating topic influence in authorship attribution. In: Stein, B., Koppel, M., Stamatatos, E. (eds.). Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, Vol. 276, pp. 29-35. CEUR, Amsterdam, Netherlands (2007).
30. Tweedie, F.J., Baayen, H.R.: How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities 32(5), pp. 323-352 (1998).
31. García, A.M., Martín, J.C.: Function Words in Authorship Attribution Studies. Literary and Linguistic Computing 22(1), pp. 49-66 (2007).
32. Hoover, D.L.: Another Perspective on Vocabulary Richness. Computers and the Humanities 37(2), pp. 151-178 (2003).
33. Oakes, M.P.: Statistics for corpus linguistics. Edinburgh University Press, Edinburgh (1998).
34. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.). Advances in kernel methods, pp. 185-208. MIT Press, Cambridge (1999).
35. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship Attribution with Support Vector Machines. Applied Intelligence 19(1), pp. 109-123 (2003).
36. Escalante, H.J., Solorio, T., Montes-y-Gómez, M.: Local histograms of character N-grams for authorship attribution. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp. 288-298. Association for Computational Linguistics, Portland, Oregon (2011).
37. Houvardas, J., Stamatatos, E.: N-Gram Feature Selection for Authorship Identification.

In: Euzenat, J., Domingue, J. (eds.). Artificial Intelligence: Methodology, Systems, and Applications, Vol. 4183, pp. 77-86. Springer Heidelberg (2006).

38. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3), pp. 378-393 (2006).
39. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Fisher, D.H. (ed.). Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97), pp. 412-420. Morgan Kaufmann Publishers Inc., San Francisco, CA. (1997).

PART II

FUNDAMENTAL LINGUISTIC RESEARCH

Quantitative Properties of Argumentation Motifs

Andrei Beliankou, Reinhard Köhler, Sven Naumann

CL/DH, Universität Trier, 54286 Trier, Germany
{a.beliankou, koehler, sven.naumann}@uni-trier.de

Abstract. Some aspects of argumentative texts can be investigated by means of 'motifs', a unit which was introduced into linguistics for analyzing syntagmatic properties of text. In the present paper, two problems are presented and discussed which are connected with the definition of motifs on the basis of relations in argumentative structures and with many other aspects of text structure. One of them is the fact that the elements are not always numerical, the other one is the dimensionality of the structures under study: while motifs are sequential units, rhetorical and other structures are represented by two-dimensional trees. The paper presents a method which enables motif-based analyses on multi-dimensional categorical data and gives examples of hypotheses and evaluations using such methods.

Keywords: syntactic motif, polytextuality, RST, syntactic complexity, rhetorical structure, description formalisms.

1 Introduction

Motifs in the sense of [5] were originally defined on the basis of numerical (or at least ordinal) variables such as frequency, length, polysemy, or polytextuality of linguistic units. There are, however, also categorical variables in linguistics which (temporarily) resist metrification. An example of such a variable is the type of relation in argumentation structures, e.g. justification, elaboration, concession, circumstance etc. Motif studies are motivated by the wish to investigate texts with quantitative methods not only with respect to unordered sets of elements (such as vocabularies and other inventories) but also with respect to the sequences of linguistic elements. Therefore, the analysis of the syntagmatic dimension of argumentation elements in texts seems to be worthwhile as well.

2 Data

In our study we used the data extracted from the Potsdam Commentary Corpus [10, 11]. This resource is an ongoing project combining multilayered linguistic annotations and some interesting aspects of the regional language. Since we wanted to focus on quantitative aspects of discourse structure, our choice of resources was determined by the fact that PCC is the most mature corpus of German with an annotation layer using the formalism called RST [6]. For other languages there are several copora and RST parser available [3, 4, 8, 7, 13]. Furthermore, the tree structure of rhetorical relation was encoded in a flexible xml-based format baptized URML [9]. One of the core features of the proposed format is its ability to allow underspecification, which has both advantages and disadvantages. Annotators may postpone their decision and leave a difficult case unresolved, "underspecified". The obvious disadvantage is an eventual inconsistency of the corpus and the need for reviewers to take normalization steps at a later stage.

The corpus was produced at the Potsdam University. It is based on short commentaries from a regional daily newspaper "Märkische Allgemeine Zeitung". [1] The source for the texts and the genre were chosen guided by the following considerations: (1) the texts are short; (2) the lexical richness is less than in a national daily and (3) every article is rather opinionated and thus can be annotated unambiguously.

The version of the PCC we took for our analysis consists of 172 texts with an average text length of 10.9 sentences per document. All 1876 sentences are split into 2771 units produced by human annotators. On average, a sentence contains 1.5 discourse units per sentence. The average length of a discourse unit is 11.7 tokens.

Due to our focus on RST, in this study we want to pay particular attention to the quantitative evaluation of this annotation level. The corpus annotation scheme contains 23 different relation types (18 mononuclear vs. 5 multinuclear). In the evaluated part of the corpus, we observed the following absolute frequencies, e.g. for some mononuclear relations: elaboration – 730, evaluation – 268, evidence – 264, summary – 1. In case of multinuclear relations these frequencies are: list – 166; contrast – 54; joint – 42; sequence – 27; conjunction – 1.

The whole corpus contains 220 RST Trees spread over 124 single-tree documents and 48 documents with two trees. The majority of the annotated relations (2075 occurrences) are mononuclear, and only a few relations (290 occurrences) have multinuclear nature. The average relation tree contains 32.4 nodes.

Besides rhetorical structures, the core of PCC is annotated on five levels with: (1) POS tags, (2) syntactic constituents, (3) connectors with the scope, (4) co-reference markers and (5) information structure relations. The layers are widely independent and use different annotation schemas. While the POS annotations were done automatically, other layers required manual or semi-automatic (in case of syntactic structures) labor.

[1] The online version of this newspaper can be found under http://maerkischeallgemeine.de.

3 Motifs on the basis of two-dimensional structures

Syntagmatic relations and syntagmatic units can be defined on sequential structures only. Obviously, the linguistic surface structure of a text and its underlying logical structure or *rhetorical structure* as RST coins it should not necessarily be equal. The logical structure of a text can be captured in form of a two-dimensional tree:

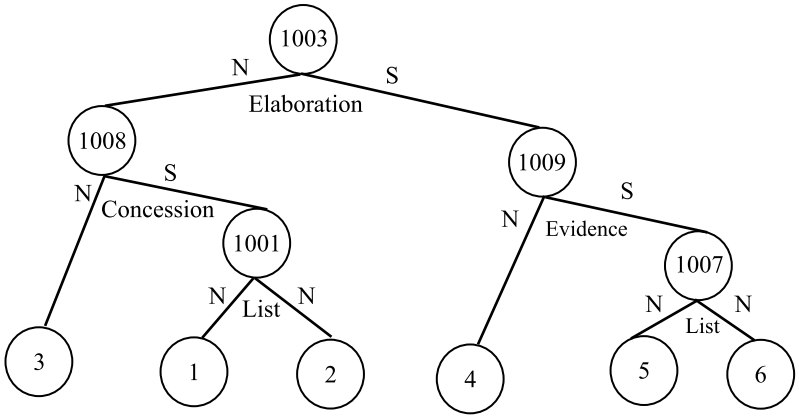


Fig. 1. Example of a text abstract, represented using RST: [**1** *Wie schwierig es ist, in dieser Region einen Ausbildungsplatz zu finden,*] [**2** *haben wir an dieser und anderer Stelle oft und ausführlich bewertet.*] [**3** *Trotzdem bemühen sich Unternehmen sowie die Industrie- und Handelskammer Potsdam den Schulabgängern Wege in die Ausbildung aufzuzeigen.*] [1 How difficult it is to find an apprenticeship in this region] [2 have we mentioned several times and commented on extensively.] [3 Even so are companies and the *Industrie- und Handelskammer Potsdam* trying to show graduates a way to find an apprenticeship.]

Each node represents an *elementary discourse unit* (EDU) which roughly corresponds to a sentence or a clause on the surface level. Each EDU is classified either as a nucleus (N) in case the information it conveys are essential for the logical structure of the text or as satellite (S) when it just provides some additional information. A RST-relation is assigned to each non-leaf node of the tree specifying the rhetorical relation that holds between its child-nodes.

In order to derive motifs from RST-trees, it is necessary to map each tree onto a set of linear objects (such as sequences, strings, etc.). We used a simple algorithm to obtain sequences from a tree: for each tree, a set of paths of non-leaf nodes is computed. A path (in our case) can be represented as a sequence of <label, relation, depth>-triples: e.g. $<<$ 1003, elaboration, 0 $>$, $<$ 1008, concession, 1 $>$, $<$ 1001, list, 2 $>>$.

Depending on the kind of motifs we are interested in different variants of the path-building algorithm are applied: We compute either paths of maximal length (left-to-right, depth-first) or paths where no relation occurs more than once.

4 Length of R-motifs

There are several ways to form motifs from categorical data without scaling them. One of the possibilities is the following definition:

An R-motif is an uninterrupted sequence of unrepeated elements.

An example of the segmentation of a text fragment (represented as a sequence of argumentative relations) into R-motifs is the following:

[”elaboration”], [”elaboration”, ”concession”], [”elaboration”, ”evidence”, ”list”, ”preparation”, ”evaluation”, ”concession”], [”evidence”, ”elaboration”, ”evaluation].

The first R-motif consists of a single element because the following relation is a repetition of the first; the second one ends also where one of its elements occurs again etc.

On this basis, the lengths of these R-motifs in the Potsdam commentary corpus were determined. The distribution of the motif lengths turned out to abide by the hyper-Binomial distribution (cf. Fig. 2):

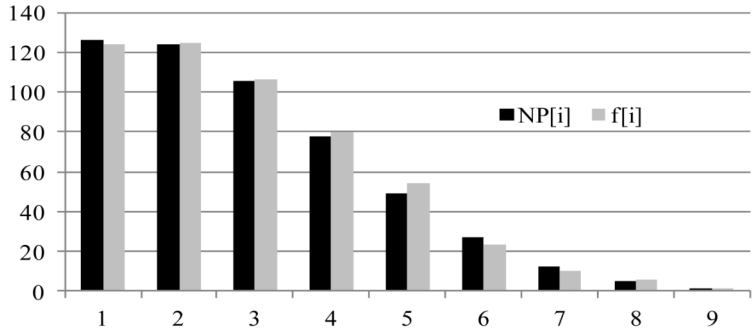


Fig. 2. The distribution of the length of R-motifs in the Potsdam RST Corpus

This finding can be linguistically interpreted if one assumes that the observed length distribution is a consequence of a diversification process [1, 2]. Altmann's approach to modeling such processes is based on a simple consideration: the probability for the process to form a class (in our case a length class) x is proportional to the probability of finding elements of class $x-1$. In this approach, the proportion is not a constant but a function $g(x)$. If this function is a rational dependency of the form

$$g(x) = \frac{d - bx}{a + cx} \tag{1}$$

with the four parameters a, b, c, and d the hyper-Binomial distribution is obtained

Table 1. Result of fitting the hyper-Binomial distribution to the R-motif length data using Altmann Fitter (3.1)

x[i]	f[i]	NP[i]
1	124	126.3
2	125	124.04
3	106	105.56
4	80	77.74
5	54	49.38
6	23	26.93
7	10	12.52
8	6	4.92
9	1	1.61

Table 2. Result of fitting the hyper-Binomial distribution to the R-motif length data using Altmann Fitter (3.1)

Parameters				
n 14	M 13.9696	q 0.98	Ord I 0.9483	Ord S 1.3327
X^2 2.0995	$P(X^2)$ 0.8352	DF 5	C 0.004	R 0.9974
N 529	m1 2.9244	m2 2.7731	m3 3.6957	m4 24.6937
Skewness 0.8003	Excess 11.8287	Entropy 0.8172	Repeat rate 0.4338	

[cf. 1, 2]:

$$P(X=x) = \frac{\binom{n}{x}}{\binom{m+x-1}{x}} q^x P_0 \tag{2}$$

where $d/b-1=n$; $a/c+1=m$; $b/c=q$ and $P_0 = [{}_2F_1(1,-n;m;-q)]^{-1}$, i.e. the inverse of the Hypergeometric function. The four parameters can be interpreted in the following way: d represents a constant cognitive influence toward more complex structures (and hence longer chains of varying relations) in order to shed light on the statement which is argued for from all sides; b is the coefficient of x which takes care of the degree with which the complexity-increasing influence of d is decelerated with increasing complexity and variation of the argumentation. The elements of the denominator decrease complexity and length of the motif; parameter a represents a constant limiting effect, responsible for comprehensibility of the argumentation and the cognitive limitations of the recipient (hearer or reader), supported from the term cx, which adds a limiting force, increasing with increasing

variation of the argument types, and hence promoting repetition and strengthening of an already uttered argument type.

5 Length of D-motifs

Another way of defining motifs on the basis of categorical data is: *A D-motif is an uninterrupted depth-first path of elements in a tree structure*. The length of motifs determined in this way displays a behavior that differs considerably from that of the R-motifs. A linguistically interpretable theoretical probability distribution which can be fitted to the empirical frequency distribution is the mixed negative binomial distribution (cf. Fig. 3 and Tables 3 and 4).

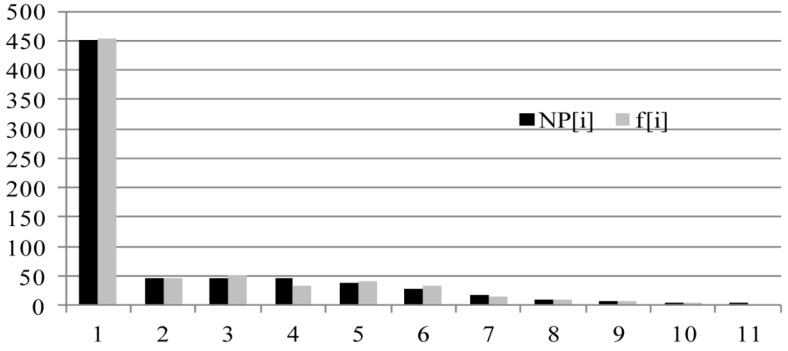


Fig. 3. Fitting the mixed negative binomial distribution to the D-motif data

Table 3. Result of fitting the mixed negative binomial distribution to the D-motif data using Altmann Fitter (3.1)

x[i]	f[i]	NP[i]
1	454	451.32
2	46	45.88
3	52	46.29
4	33	46.03
5	40	38.04
6	33	27.52
7	16	18.03
8	11	10.93
9	6	6.22
10	5	3.36
11	1	3.39

A mixed distribution is the result of a combination of two processes. We are concerned here with the combination of two diversifications, which result both in the

Table 4. Result of fitting the mixed negative binomial distribution to the D-motif data using Altmann Fitter (3.1)

Parameters				
K 7.5963	p1 0.9965	p2 0.688	A 0.6434	Ord S 3.6217
X^2 8.3176	$P(X^2)$ 0.2157	DF 6	C 0.0119	R^2 0.9985
N 697	m1 2.2195	m2 4.2804	m3 15.5021	m4 97.6665
Skewness 1.7505	Excess 44.2069	Entropy 0.5601	Repeat rate 0.2932	Ord I 1.9285

negative binomial distribution but with different parameters. We can assume that one process concerns the diversification of depth in the tree structure, the other one is a consequence of the diversification of the length of the paths in the trees. Therefore, we will test the hypothesis that the distribution of depths is compatible with the negative binomial distribution. Otherwise, the mixed negative binomial distribution could not be justified as a model of D-motif lengths.

Fitting the negative binomial distribution to the depth distribution of the relations in the corpus yielded good Chi-square values except three of the relations (cf. Table 5).

Table 5. Fitting the negative binomial distribution to the depth distributions of the relations in the PCC using Altmann Fitter (3.1.). The three marked relations yield a poor fit according to $P(X^2)$; according to R^2, also these fits are acceptable.

Relation	**X^2**	**$P(X^2)$**	**DF**	**R^2**	**P**	**N**
hline background	1.93	0.748	4	0.9646	0.8724	114
circumstance	1.66	0.645	3	0.9206	0.441	25
nonvolitional-cause	2.68	0.6121	4	0.9309	0.8998	83
volitional-cause	1.89	0.5953	3	0.819	0.9094	38
contrast	3.56	0.4684	4	0.7273	0.6544	30
condition	5.52	0.3562	5	0.6978	0.9935	32
concession	7.76	0.1703	5	0.8068	0.9769	98
nonvolitional-result	8.46	0.1326	5	0.5461	0.8473	36
volitional-result	2.5	0.1138	1	0.4932	0.8939	12
elaboration	12.79	0.0773	7	0.9476	0.9992	413
evaluation	17.34	0.0081	6	0.7307	0.5445	192
preparation	19.4	0.0035	6	0.8487	0.8992	124
evidence	24.64	0.0001	4	0.7882	0.8972	180

6 Conclusion

Our study is an attempt at investigating a pragmatic phenomenon – the argumentation structure of texts – with quantitative means. The method we applied is based on the analysis of motifs as previously introduced for linguistic units and properties such as word length, frequency and polytextuality. As argumentation structures are described (1) in form of two-dimensional trees and (2) the concepts used to describe the relations in these structures are categorical, a way had to be found to (1) transform the corresponding trees into sequential data and (2) two map there sequences on quantitative values.

We showed that RST structures can be analyzed in this way and that theoretical probability distributions can be found as models of these text phenomena. Moreover, linguistically plausible interpretations of the mathematical models can be given.

It goes without saying that this study is only a first step towards a quantitative investigation of argumentation structures: many more texts and corpora must be analyzed, and the results from different text types and from different languages compared. A huge number of properties and units open up new vistas for the quantitative research on argumentation structures and other pragmatic fields.

References

1. Altmann, G. Modeling diversification phenomena in language. In: Rothe, U. (ed.): Diversification Processes in Language: Grammar, pp. 33-46. Rottmann, Hagen (1991)
2. Altmann, G.: Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds): Quantitative Linguistik / Quantitative Linguistics. Ein internationales Handbuch / An International Handbook, pp. 646-658. de Gruyter Berlin, New York (2005)
3. Carlson, L., Marcu. D., Okurowski, M.E.: RST Discourse Treebank. Pennsylvania, Linguistic Data Consortium (2002)
4. Da Cunha, I., Torres-Moreno, J.M. and Sierra, G.: On the Development of an RST Spanish Treebank. In: Proceedings of Fifth Law Workshop, ACL 2011, pp. 1-10 (2011)
5. Köhler, R., Naumann, S. (2008): Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.): Data Analysis, Machine Learning and Applications, p. 637-646. Heidelberg, Berlin (2008)
6. Mann W. C., Thompson S. A.: Rhetorical structure theory: description and construction of text structures. In: Kempen, G. (ed.): Natural Language Generation, pp. 85-96. Nijhoff, Dordrecht (1987)
7. Marcu, D.: The rhetorical parsing of unrestricted text: a surface-based approach. Computational Linguistics, 26/3, 395-448 (2000)
8. Pardo, T.A.S., Nunes, M.G.V.: On the development and evaluation of a Brazilian Protugiese Discourse Parser. In: Journal of Theoretical and Applied Computing, 15/2, 43-64 (2008)
9. Reitter, D., Stede, M.: Step by step: underspecified markup in incremental rhetorical analysis. In: Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), Budapest (2003)

10. Stede. M.: Surfaces and depths in text understanding: the case of newspaper commentary. In: Proceedings of the HLT/NAACL Workshop on Text Meaning, Edmonton/AL, (2003)
11. Stede, M.: The Potsdam commentary corpus. In: Proceedings of the Workshop on Discourse Annotation, 42nd Meeting of the ACL, (2004)
12. Wimmer, G., Altmann, G. Thesaurus of univariate discrete probability distributions. Stamm, Essen (1999)
13. Yue, M., Liu, H.: Probability Distribution of Discourse Relations Based on a Chinese RST-annotated Corpus. In: Journal of Quantitative Linguistics 2, 107-121 (2011)

Close and Distant Relatives of the Sentence:

Some Results from Russian

Peter Grzybek

University of Graz, Department for Slavic Studies
peter.grzybek@uni-graz.at

Abstract. In this contribution, sentence length is studied from a Menzerathian perspective. Whereas the Menzerath-Altmann law models the construct-constituent relation of linguistic entities from two directly neighboring levels, the present study focuses on the relation of the sentence to linguistic entities from 'indirect' neighbors. In detail, the sentence length ↔ word length and the sentence length ↔ chapter length relations are submitted to analyses of Russian texts.

Keywords: Word ↔ Sentence ↔ Chapter length, Menzerath-Altmann law

1 Introduction

Studies of sentence length have repeatedly been related to different kinds of questions, including in a broad research spectrum of linguistics and text analysis: reaching from sentence length assumed to be an author or style specific feature to questions of text difficulty (or comprehensibility), the length of sentences has been regarded a major criterion of text construction. There is no need to offer here an extensive presentation of the history of sentence length research, which usually is considered to start with [22]. With regard to Russian, the important works by [15, 16, 17] deserve mention, which concentrated on the question of frequency distribution of sentence length, i.e., the question with which frequency do sentences of a given length occur in the material under study: after scholars like [26] or [23] had referred to the allegedly author-specific dimension of sentence length, [15, 16, 17] conducted more detailed analyses, paying attention to different text types and functional styles as well as further intralingual factors. Extending this line of research, more recent studies in the field of quantitative linguistics –, e.g., [20, 21], [10, 11] – have predominantly treated the question of sentence length from a theoretical modeling perspective.

Relations between the length of sentences and that of linguistic entities and constructs from other levels have been studied to a much lesser degree. In classical

structural concepts, starting from a sentence perspective, such ”vertical” relations may be assumed to exist in both ”downward” and ”upward” directions: in the first case, sentence length may be assumed to be related to the length of its constituents (like clauses, or phrases), in the second case, to larger textual units (such as paragraphs).[1] Studies on relations in both directions are likely to be interpreted in terms of the well-known Menzerath-Altmann law (Mal), according to which those units, which constitute a given linguistic construct, are the shorter the longer the construct itself is [1, 2, 4].[2]

These assumptions generally hold true, however, only for direct constituents: from a mathematical point of view, the Mal does not necessarily imply transitivity, so that no conclusions may be drawn with regard to indirect constituents coming into play when, in structuralist terms, an intermediate linguistic level is skipped, or leapfrogged. Notwithstanding these theoretical objections, there may well be, however, empirically speaking, systematic cross-level relations. From a linguistic point, this might even turn out to fully plausible: if, for example, there is a decrease of clause length with an increase of sentence length, it seems reasonable to assume that, as a consequence, relatively shorter clauses are in turn characterized by longer words, so that an increase of word length would go along with an increase of sentence length. As a result, we would thus be concerned with direct or indirect constituents, which shall be termed here ‚close' and 'distant' relatives.

The study of such cross-level relations yields important insight into general principles of global text processes across levels. It may also eventually provide valuable empirical corroboration in favor of the Mal in case clear evidence is lacking from the analysis of direct relations; exactly this is the case with regard to Russian sentence relations. Whereas there are generally almost no studies available in the 'upward' direction[3], one might object that analyses in the 'downward' direction have repeatedly proven the Menzerath-Altmann law to be valid for the sentence $\leftrightarrow$ clause relation – for Russian, however, the situation is different, since related studies have not provided consistent results, what has led to the assumption, that the Mal might not hold valid for this language [21].

From these deficits, the major objective of this contribution arises: the overall aim is to point out the need for systematic studies, by providing and theoretically interpreting some promising preliminary results, as a basis for future work. To achieve this goal, we will start with the analysis in the ‚downward' direction (Section 2)

[1] Strictly speaking, it may be highly misleading to juxtapose 'downward' vs. 'upward' directions, in this context: after all, the Menzerath-Altmann law, concerning linguistic constructs and their constituents (necessarily from a 'lower' level, in structuralist terms), should more likely be generally seen as a ”top-down” law – only heuristiclly, i.e. focusing a specific level (here: the sentential level), such a terminology may be justified.

[2] In this form, the Menzerath-Altmann law has been conceived as as law relevant for intratextual relations, i.e., it refers to relations within given linguistic material (as text, a corpus, etc.). It must clearly be set apart from the Arens-Altmann law, which is based on similar assumptions, but refers to intertextual relations, i.e., it is based on averages of texts, which represent a vector of averages [7].

[3] In fact, this holds true not only for Russian, but holds generally true, with the exception of [18] recent study on German (see below).

before then turning to the 'upward' direction (Section 3). In both cases, we will not confine the analyses to the length relations between the entities under study, but, by way of a pre-condition and requirement to be met, will test if the units concerned are regularly organized with regard to their frequencies, on each of the linguistic levels at stake.

2 Sentence length $\leftrightarrow$ word length

With regard to the ‚downward' direction, relations between sentence length and the length of linguistic units or constructs from "lower" (i.e., sub-sentential) levels have previously been studied with regard to Russian data, e.g., by [21] and [5]. Analyzing he relation between sentence length and clause length, [21] found her results not be consistent with the Mal, assuming that it might not hold valid for the sentence $\leftrightarrow$ clause relation in Russian (ibd., 609). Attempting to explain these findings, [21] offered two (not mutually exclusive) options: 1. the Mal might not, at least not in its "standard" form, hold for Russian (i.e., the boundary conditions of a general law would significantly differ for Russian), 2. for Russian, a different definition of either sentence as the construct and/or of clause as relevant measuring unit might be needed as compared to other languages.[4] A third factor may (also) have played a crucial role, due to the fact that [21] analyzed only Chapter XVII of Book IV from L.N. Tolstoj's *Anna Karenina* [*Анна Каренина*], summing up, according to her counting[5], to an overall number of 231 sentences – a data basis, which may well not have been large enough for far-reaching conclusions.

In order to exclude possibly intervening problems of clause definition, [5] have skipped the intermediate level of clauses: concentrating on the sentence $\leftrightarrow$ word relation, the authors' assumption was that, in case of some regular relation, this would be an indirect proof of the Menzerath-Altmann law being valid for Russian, too. [5] indeed found corroborating evidence, concentrating on what they termed the "core data structure" from $4 \leqslant \mathrm{Se}L \leqslant 30$ words per sentences, excluding shorter and longer sentences from analysis. In this contribution, we will therefore maintain the argumentation outlined above, but extend the data basis by including short sentences from 1-4 words per sentence into the model.

2.1 Sentence length frequencies

Based on a sentence definition, according to which a sentence is a closed textual unit ended by a period, a question mark or an exclamation mark followed by a capital letter, sentence length is defined here by the number of words, which in turn follow an orthographic-phonetic definition (see below). According to these definitions, the text consists of 19297 sentences, the shortest consisting of one word only,

[4] Studies with languages other than Russian have tended to define clauses on the basis of finite verb forms, a definition which is likely to be inadequate for Russian with its high number of (adverbial) participles.

[5] According to the above-mentioned definition, the overall number of sentences sums up to an even smaller number of 199 sentences.

the longest of 151 words; average sentence length is $\bar{x} = 13.89$ word per sentence ($s = 11.08$). Table 1 represents the frequency occurrences (fx) of sentences with x words (column *WoL* can be ignored here and will be referred to further below).

Table 1. Sentence length frequencies in *Anna Karenin*a

SeL	*f*	*WoL*	*SeL*	*f*	*WoL*	*SeL*	*f*	*WoL*	*SeL*	*f*	*WoL*
1	340	2.2647	26	199	2.2578	51	21	2.3081	77	3	2.3853
2	725	2.1359	27	194	2.2518	52	17	2.3032	78	3	2.1880
3	1093	2.0494	28	189	2.2132	53	25	2.2936	79	2	2.4810
4	1296	2.0355	29	159	2.272	54	15	2.1432	80	1	2.4500
5	1294	2.1037	30	145	2.2614	55	11	2.3306	81	1	2.1975
6	1145	2.1007	31	130	2.2759	56	12	2.2188	84	1	2.3810
7	1115	2.1363	32	115	2.2508	57	11	2.1722	87	1	2.3448
8	992	2.1569	33	106	2.2573	58	12	2.2716	88	1	1.9318
9	975	2.185	34	91	2.3284	59	7	2.3850	90	1	2.2444
10	874	2.1684	35	81	2.279	60	8	2.2500	91	1	2.4286
11	840	2.1871	36	86	2.2474	61	14	2.2951	92	2	2.1957
12	736	2.2183	37	78	2.2367	62	4	2.2540	98	1	2.6429
13	648	2.2204	38	72	2.2376	63	9	2.3086	99	1	2.5051
14	643	2.2253	39	55	2.1939	64	5	2.3844	100	1	2.2700
15	598	2.2317	40	53	2.2552	65	5	2.3723	106	1	2.1792
16	534	2.2328	41	49	2.2339	66	4	2.1174	116	1	2.1724
17	489	2.2283	42	50	2.3248	67	3	2.4030	125	1	2.3200
18	489	2.2295	43	39	2.2302	68	7	2.2647	138	1	2.8768
19	373	2.2585	44	41	2.2955	69	4	2.1884	151	1	2.9470
20	362	2.2297	45	30	2.3237	70	4	2.3071			
21	320	2.2579	46	39	2.2737	71	6	2.1174			
22	330	2.2625	47	27	2.1875	72	1	2.1528			
23	271	2.2784	48	26	2.2131	73	2	2.1233			
24	244	2.2609	49	24	2.2645	74	4	2.4527			
25	230	2.2610	50	26	2.1677	75	1	2.2267			

As compared to [10] results, who found the negative binomial distribution (in its 1-shifted form) to be a good model for Russian prose of different genres and authors, an extended version of this model is needed for Tolstoj's *Anna Karenina*, which consists of a mixture of two negative binomial distributions, each of them with different parameter values for k and p (and $q = 1 - p$), resulting in a (1-shifted) mixed negative binomial distribution with weights α and $1 - \alpha$:

$$P_x = \alpha \binom{k_1 + x - 2}{x - 1} p_1^{k_1} q_1^{x-1} + (1 - \alpha) \binom{k_2 + x - 2}{x - 1} p_2^{k_2} q_2^{x-1} \quad x = 1, 2, 3, \ldots \tag{1}$$

This may well be due to the fact that the novel contains different sentence regimes with differing length distributions, e.g. for dialogical, narrative or descriptive sequences – no systematic studies as to this point are available, however. Fig. 1 represents the result[6] in graphical form, with parameter values $k = 2.47$, $p_1 = 0.26$, $p_2 = 0.12$ and the weighting factor $\alpha = 0.52$; the goodness-of-fit of this model is excellent, with $C = X/N = 0.0075$.[7]

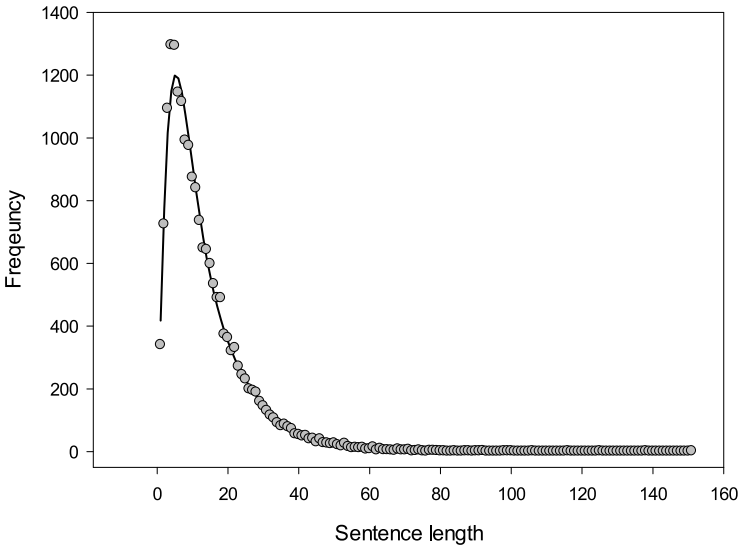


Fig. 1. Sentence length frequencies $(f_x,\ Np_x)$ in Tolstoj's *Анна Каренина*

We can thus summarize that the first requirement according to our postulates is met, namely, that the distribution of sentence length is not chaotic, but is regularly organized and follows a well-known regularity.

2.2 Word length frequencies

Word length frequencies have repeatedly been dealt, and the procedures need no detailed mention here. Thus immediately turning to Tolstoj's Анна Каренина, we see that on the whole, 258384 words[8] occur in the running text. Word length average is $\overline{x} = 2.22$ syllables per word $(s = 1.18)$. Table 2.2 represents the frequencies (f_x) for each individual length (x), ranging from $x_{\min} = 1$ to $x_{\max} = 10$ syllables per word. The values in the third column (Np_x) will be referred to further below.

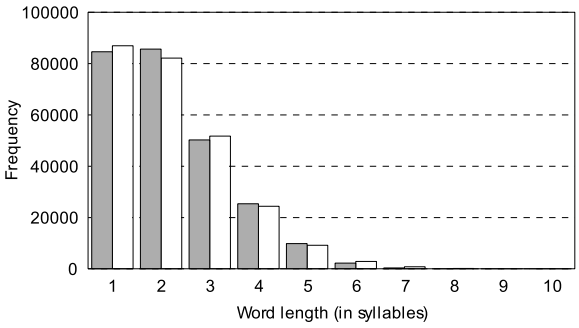
[6] In order to reduce the overall number of parameters, the mixed negative binomial distribution is calculated here with $k_1 = k_2 = k$.

[7] A value of $C < 0.02$ is interpreted to be a good, a value of $C < 0.01$ a very good fit.

[8] A word, or rather word form, is defined here as an orthographic-phonetic unit, so that, for example, zero-syllable words, like the prepositions 'в' [in], 'к' [to], 'с' [with], are treated like clitics.

Table 2 / Fig. 2. Word length frequencies in Tolstoj's *Anna Karenina*

x	f_x	Np_x
1	84629	86960
2	85649	82134
3	50241	51717
4	25338	24423
5	9828	9227
6	2257	2905
7	370	784
8	65	185
9	6	39
10	1	7



As to a theoretical model for the observed word length frequencies, it turns out that, in case of Tolstoj's *Anna Karenina*, the one-parameter Poisson distribution is a sufficiently good model[9] , albeit in its left-truncated form[10], which is also known by the name of positive Poisson distribution:

$$P_x = \frac{e^{-a}a^x}{x!(1-e^{-a})} \quad x = 1,2,3,\ldots \tag{2}$$

With parameter value $a = 1.89$, we obtain the theoretical values (Np_x), presented in the third column of Tab. 2.2 (see above), graphically represented in Fig. 5: the grey bars depict the observed, (f_x), the white ones the theoretical (Np_x) frequencies. As the discrepancy coefficient of $C = 0.003$ (cf. fn. 5) shows, the fit can be considered to be excellent. Since thus our second requirement is met, too, saying that word length is not chaotic, but regularly organized in Анна Каренина, we may next turn to the study of the relation between these two.

2.3 Sentence length ↔ word length in *Anna Karenina*

Table 3 (see above) shows the results for word length (*WoL*) depending on sentence length (*SeL*): for each *SeL* not only their frequencies (f_x) are given, but also average *WoL* for each *SeL*. Fig. 6 represents the results graphically.

A number of observations are clearly visible, at first sight:

– in the central area of *SeL* (i.e., in the interval of ca. $3-4 > SeL > 33$ words per sentences, there is a non-linear increase of *WoL* with an increase of *SeL*;

[9] It goes without saying that models with more parameters yield even better results; in our case, the differences are minimal, however, so that the model with the minimal number of parameters should be preferred.

[10] Whereas in case of a displacement by 1 the whole model is shifted by one position to the right, a left-truncation is based on the assumption that there can be no frequencies for the class $x = 0$, resulting in a theoretical "elimination" of this class.

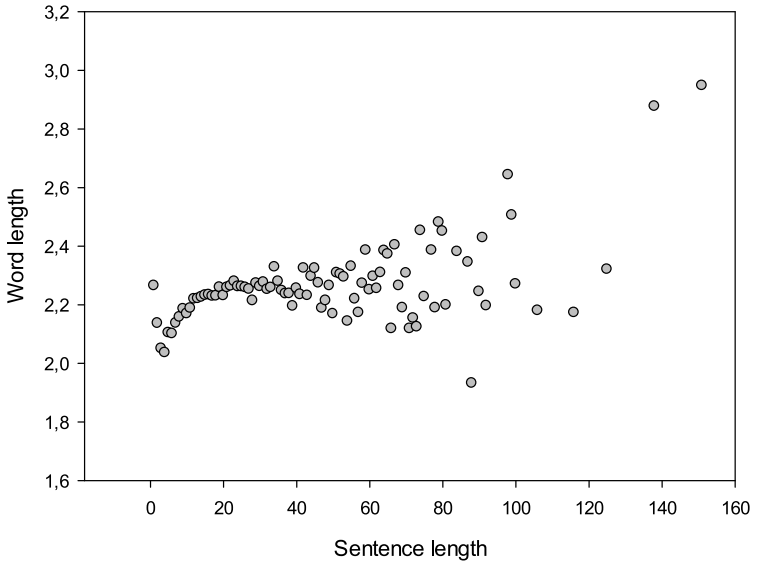


Fig. 2. Word length $\leftrightarrow$ 1 sentence length in *Anna Karenina*

– very short sentences (ca. 1-4 words per sentence) follow a reverse trend, these sentences being characterized by relatively longer words;
– *WoL* variation increases beyond *SeL* of ca. 33 words per sentence.

With regard to an explanation of the last point, two (not mutually exclusive) options are available, a statistical and a linguistic one. According to the statistical option, one might suspect an insufficient number of observances to be responsible for an instable average *WoL*. It has already previously been assumed that a minimal number of $f_{\mathrm{SeL}} > 30$ is necessary to provide sufficient stability. The results found now do not seem to corroborate this suggestion since, as can be seen from Table 3, this requirement is met in the data up to $SeL \leqslant 46$, for $SeL \leqslant 33$ averages are even based on frequencies of $f_{\mathrm{SeL}} = 100$. The linguistic option incorporates the fact that, in calculating *SeL*, the level of clauses is leapfrogged; taking this into account, one may even consider it to be surprising that up to $SeL \approx 30$ there seems to be a relatively stable tendency. With this in mind, it seems to be reasonable, from a linguistic point of view that mechanisms of self-regulation do not operate beyond a specific *SeL*, since they are not accessible to the producer's (intentional of non-intentional) control any more. In this context, it may be worthwhile taking into account human information processing and memory span limits, what refers back to Miller's "magical number" 7 ± 2 and its linguistic-syntactic interpretation by [27, 28], fully in accord with more recent insights into quantitative syntax [12, 13, 14]: assuming clauses in Russian to be constructed of 4-5 words, on the average [21], complex sentences with up to 7 clauses would correspond quite accurately with an upper limit of $SeL \approx 30$ words per sentence, average *WoL* for longer sentences in that case varying around some relatively constant value, the amount of variation depending on the number of observances per data point.

In our context, we are thus faced with the task to find a theoretical model for the relationship of *SeL* and *WoL* for the interval of $1 \leqslant SeL \leqslant 33$ (the upper limit in our case being justified by the minimal frequency of $f_{\mathrm{SeL}} > 100$).

2.4 Word and sentence length in light of the Menzerath-Altmann law (Mal)

The Mal, as it is known today, has been proposed by [1]; it generally postulates a proportionality relation between a linguistic construct and the entities which constitute it, more exactly: between the decrease of the length of a given constituent with an increase of the construct's length. Mathematically expressing this assumption of decrease as $y' = -a$, results in the differential equation

$$\frac{y'}{y} = -a \tag{3}$$

with the solution

$$y = Ke^{-ax}. \tag{4}$$

In order to grasp more complex relations, too, with an initial increase up to a maximum at $x \neq 0$, [1] suggested an extension of differential equation (3) by adding an inverse proportionality component, so that from differential equation

$$\frac{y'}{y} = -a + \frac{b}{x} \tag{5}$$

solution (4) is obtained for $b = 0$, whereas for $b \neq 0$ two options arise, namely, for $a = 0$

$$y = Kx^b, \tag{6}$$

and for $a \neq 0$

$$y = Kx^b e^{-ax}. \tag{7}$$

For linguistic purposes, (6) has often been considered to be the "standard form" of the Mal. However, none of these models is adequate[11] to model the data structure in the interval of $1 \leqslant SeL \leqslant 30$; which obviously asks for a more complex model. Such a model is provided by [24, 25]: as compared to the above-mentioned Menzerathian formulae, it offers some extensions and generalizations which thus far have only sparely been applied to construct-constituent relations. This approach is generally based on the differential equation

$$\frac{y'}{y} = \left(a + \frac{b}{x} + \frac{c}{x^2} + \frac{d}{x^3} + \cdots\right). \tag{8}$$

[11] Concentrating on the "core data structure" $(4 \leqslant SeL \leqslant 30)$ of Anna Karenina only [5], model (6) would result in a determination coefficient of $R^2 = 0.84$, with parameter values $K = 2.00$ and $b = 0.0381$ (Fig. 7a), as compared to $R^2 = 0.92$ for equation (7), with $K = 1.85$, $b = 0.0858$, and $a = 0.0031$ (Fig. 7b).

As can be seen, differential equation (12) is obtained for $c, d, \ldots = 0$ from (8), which for $a, b, c, \ldots \neq 0$ generally has the solution

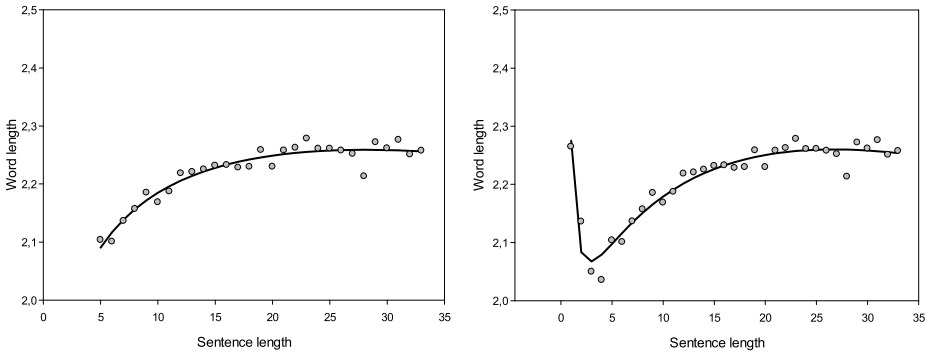
$$y = Ke^{ax}x^{b}e^{-c/x-d/2x^{2}-\cdots}. \tag{9}$$

From (9), also equations (4, 6, 7) can be obtained, but as compared to these, we have an additional (optional) factor $e^{c/x}$, which results from (8) with $d = 0$. As a result, we thus have a system of six functions with maximally four parameters (K, a, b, c), with which we are likely to model more complex relations, too.The most complex is (VI), the remaining five can be interpreted to be its special cases for specific parameter values or constellations. Table 3 presents for each of these six functions the parameter constellation and the resulting number of parameters.

Table 3. Functions of the Mal and its extensions

I	$y = K \cdot e^{ax}$	$a < 0, b, c = 0$	2
II	$y = K \cdot x^{b}$	$b < 0, a, c = 0$	2
III	$y = K \cdot e^{ax} \cdot x^{b}$	$a, b \neq 0, c = 0$	3
IV	$y = K \cdot e^{-c/x}$	$c > 0$	2
V	$y = K \cdot x^{b} \cdot e^{c/x}$	$b, c \neq 0$	3
VI	$y = K \cdot e^{ax-c/x} \cdot x^{b}$	$a, b, c \neq 0$	4

In fact, modeling the complete data structure in the interval $1 \leqslant SeL = 33$, is possible with model (VI) which, with parameter values $K = 1.74$, $a = 0.0038$, $b = 0.1098$ and $c = 0.1526$, results in $R^2 = 0.92$ (Fig. 4b).[12]



a) $4 \leq SeL \leq 30$: $y = K \cdot e^{c/x}$ **b)** $1 \leq SeL \leq 30$: $y = K \cdot e^{ax} \cdot x^{b} \cdot e^{-c/x}$

Fig. 4. Modeling sentence length ↔ word length in *Анна Каренина*

[12] Interestingly enough, the 2-parameter model (IV) yields identically good results ($R^2 = 0.92$) for the "core data structure", with parameter values $K = 2.30$ and $c = 0.49$, as compared to the 3-parameter model (III/11c), depicted in Fig. 4a.

We can thus summarize that the sentence length ↔ word length relation in Russian is not chaotic, but follows regular patterns which can be modeled in the framework of concepts well-known in the field of quantitative linguistics. It remains an open question what this means for the sentence length ↔ clause length relation in Russian: given our results, it seems not unlikely that [21] failure as to Russian is due to inappropriate linguistic definitions and/or to sparse data material; it is not excluded, however, that (further) reasons may be found in specifics of Russian syntax. The findings that (a) for the modeling of the "core data structure" a 2-parameter function (Table 3, IV) is similarly appropriate as compared to a 3-parameter function from the "usual" ones (Table 3, III), and that (b) for the integration of the very short sentences into a common model the addition of two, not only one parameter is needed, may seem surprising at first glance; it might plausibly be explained, however, by a look at differential equations (5) and (8). Function IV depicted in Fig. 8a, covering the "core data structure", is based on the differential equation

$$\frac{y'}{y} = a + \frac{c}{x^2}. \tag{10}$$

Obviously, no "correction" is needed for the reverse (as compared to the general) tendency and the "disturbance" by short sentences (since for $b = 0$ the term b/x is not part of the function); however, it seems necessary to include, in addition to the simple constant $-a$, the squared component c/x^2 from (8), resulting from $c \neq 0$, interpreting it to be an interfering factor, due to skipping the intermediate level of clauses. This would be in line with the fact that, in order to cover the whole data structure (Fig. 8b), both $b \neq 0$ and $c \neq 0$, resulting in the differential equation

$$\frac{y'}{y} = a + \frac{b}{x} + \frac{c}{x^2}. \tag{11}$$

From this perspective, we might be concerned with a plausible and complete interpretation of the whole model, which needs to be tested, of course, with more data, also from other languages, paying due attention to further possibly intervening (modifying) factors.

3 In search of supra-sentential regularities

As has been pointed out above, the lack of studies on sentence length with specific regard to the sentence's "upward' relations with 'higher' (supra-sentential) levels is even more evident, and it applies not only to Russian: rather, such studies represent an absolute desideratum in the whole field of research. Attempts in this direction have been suggested as early as in the second decade of the 20th century, in context of Russian formalism, very much ahead of its time [7, 8]. But such theoretical claims have hardly ever been empirically tested, and if so, then not systematically.

There are but a few attempts to relate the sentence and its length to supra-sentential linguistics structures. According to [9], for example, who attempts to describe texts

in connection with the basic formulation of the Menzerath-Altmann law, it is "evident that texts cannot consist directly of sentences; there must be at least one level between sentences and the entire text [...]." In his own approach, this intermediate level "evidently should contain a structure consisting of semantically based units."

Whereas the resulting textual units – for which the term 'hreb' has been established [3] – are thus semantically defined, [18] has pursued the question, if paragraph length might be systematically related to sentence length. [18] analyzed German texts from different types – journalistic, literary (prose and drama), scientific – and of varying text length (from 60 to 228,939 words). Measuring paragraph length in the number of sentences per paragraph, she found various distribution models to be relevant (Zipf-Alekseev, negative binomial, hyper-Pascal), without finding a clear relation between text type and any one of these models. Likewise, Neumann's results as to the sentence $\leftrightarrow$ paragraph relation yielded good fits for the standard Menzerathian equation (cf. II, Tab. 4) only in some cases, even after data pooling. The assumption that in these cases data originating from individual texts were too sparse, is corroborated by the finding that results were much better for homogeneous corpora, although not equally well across text types: although results were good for corpora of Wikipedia, journalistic and scientific articles, literary texts did not follow this tendency. Summarizingly, it seems likely that, one the one hand, a regular sentence $\leftrightarrow$ paragraph relation is characteristic of specific text types only, and that it is a mass phenomenon, demanding sufficient data, on the other.

Table 4. Chapter length freqeuncies in Tolstoj's *War and Peace*

x	f_x	Np_x	x	f_x	Np_x	x	f_x	Np_x
1-10	1	1.37	101-110	13	18.09	201-210	0	0.53
11-20	8	10.30	111-120	11	13.74	211-220	0	0.35
21-30	19	22.26	121-130	10	10.19	221-230	2	0.23
31-40	36	32.57	131-140	7	7.40	231-240	1	0.15
41-50	38	38.71	141-150	7	5.28	241-250	0	0.10
51-60	30	40.33	151-160	5	3.71	251-260	0	0.06
61-70	47	38.32	161-170	5	2.57	261-270	0	0.04
71-80	39	34.02	171-180	3	1.76	271-280	0	0.03
81-90	27	28.66	181-190	6	1.20	281-290	1	0.04
91-100	24	23.16	191-200	3	0.80			

Given these findings and assumptions, particularly as the literary texts are concerned, it seems reasonable to follow the same path as in case of the "downward' direction, i.e., to skip the level of paragraphs and study the sentence $\leftrightarrow$ chapter relation (which does not, of course, exist in shorter text types).

Analyzing Tolstoj's *War and Peace* [Война и мир] from this perspective, includes the analysis of 336 Chapters [Глава] of this novel, distributed over 15 Parts [Часть] and 4 Books [Книга]. Throughout the text, Chapter length ranges from $x_{\min} = 1$ to $x_{\max} = 284$ sentences per chapter. Table 4 presents the Chapter length frequencies (x), pooled by intervals of 10, and the frequencies (f_x) for each chapter length interval. The values in the third column are the theoretical frequencies

(Np_x), based on the hyper-Pascal distribution

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0 \tag{12}$$

which, in its 1-shifted form, turns out to be a good model ($P[X^2] = 0.09$, with parameter values $k = 4.04$, $m = 0.30$, and $q = 0.56$).[13]

A graphical representation of observed (black) and theoretical (white) frequencies can be found in Fig. 5.

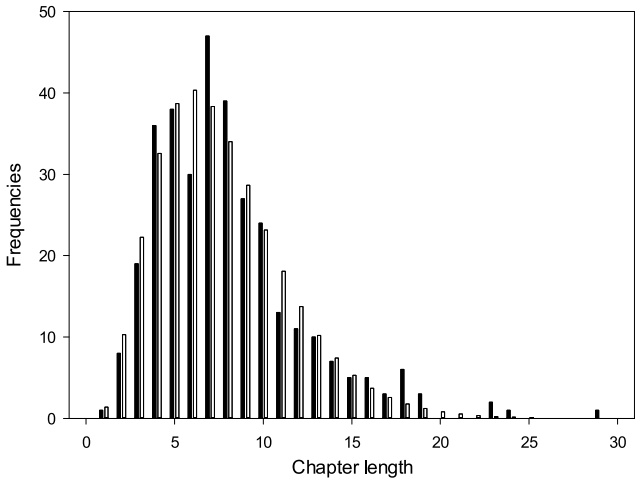


Fig. 5. Chapter length freqeuncies in Tolstoj's *War and Peace*

Since thus our requirements are met again – namely, regular frequency organization of both sentence length and chapter length – we can finally turn to the relation between these two levels, considering sentence length to be the dependent, chapter length the independent variable.

Figures 6a and b present the corresponding results in graphical form: Fig. 6a contains the original data points, with average sentence length for those chapters with identical length. It can clearly be seen that there is a nonlinear decrease of sentence length with an increase of chapter length.

Fig. 6b – based on the same data, pooled, however, in intervals per 30 for which weighted averages are calculated (given in Table 5) – makes this trend even clearer.

As can clearly be seen, under these circumstances, the data follow the standard Menzerathian function $y = a \cdot x^{-b}$ – in our case: $SeL = a \cdot ChL^{-b}$: with parameter values $a = 63.86$ and $b = 0.33$, the fit is excellent ($R^2 = 0.98$).

[13] Other models, like the mixed Poisson or the mixed negative binomial distribution, yield good results, too, but the hyper-Pascal distribution is least vulnerable to pooling procedures and interval size manipulation; of course, more rigid pooling yields even better fitting results.

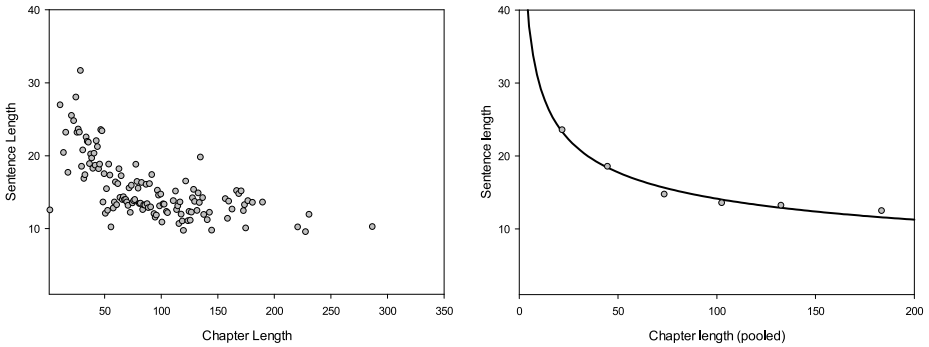


Fig. 6. Modeling sentence length 1 chapter length in Tolstoj's *War and Peace*

Table 5. Chapter length ↔ sentence length

ChL	*SeL*
21.89	23.51
44.83	18.51
73.61	14.68
102.71	13.49
132.73	13.17
183.67	12.43

4 Conclusions

In this contribution, focusing on 'downward' and 'upward' indirect relatives of the sentence in a synergetic textual framework, it could be shown that the well-known Menzerathian principle is at work, even when directly neighboring levels are leapfrogged: both the sentence ↔ word length relation (skipping the intermediate level of clauses) and the chapter ↔ sentence length relation (skipping the level of paragraphs) follow the Mal. Regardless these promising results, there are a number of caveats, however, which should be paid attention to in future more systematic work:

1. The results have been obtained with Russian texts; research must be extended to other languages, too, and it may well be that some kind of "local", or language-specific, modifications will have to be taken into account. In any case, the findings obtained for Russian provide clear (albeit indirect) evidence in favor of the notion that the Mal is fully valid for this language, too – an assumption which has recently been casted doubt upon.
2. The results have been obtained for literary texts; future studies will have to take into account possible text-type specifics. This holds true for both the 'downward' and 'upward' directions; in the latter case, we are faced with the emergence of an almost new spectrum of questions, and it may well turn out that for some text types (e.g., shorter ones) the paragraph 1 sentence relation is more relevant than the chapter ↔ sentence relation.

It goes without saying that the study of distant relatives (in terms of indirect relations) may eventually provide but indirect evidence as to Menzerathian processes primarily regulating direct relations; it may as well turn out, however, that regular indirect relations are more than a textual epiphenomenon; in this case, their study would provide deep insight into processes of dynamic text construction in general.

References

1. Altmann, G.: Prolegomena to Menzerath's Law. In: Glottometrika 2. Brockmeyer, Bochum, 1-10 (1980)
2. Altmann, G., Schwibbe, M. H.: Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Olms, Hildesheim (1989)
3. Altmann, G., Ziegler, A.: Denotative Textanalyse. Praesens, Wien (2002)
4. Cramer, I. M.: Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook, pp. 650-688. de Gruyter, Berlin, New York (2005)
5. Grzybek, P., Kelih, E., Stadlober, E.: The relation between word length and sentence length: an intra-systemic perspective in the core data structure. Glottometrics, 16, 111-121 (2008)
6. Grzybek, P., Stadlober, E., Kelih, E.: The Relationship of Word Length and Sentence Length. The Inter-Textual Perspective. In: Decker, R., Lenz, H.-J. (eds.), Advances in Data Analysis, pp. 611-618. Springer, Berlin, Heidelberg (2007)
7. Grzybek, P.: Michail Lopatto: Attempt at an Introduction into the Theory of Prose (1918). Glottometrics, 23, 70-80 (2012a)
8. Grzybek, P. = Грижбек Петер: "Опыт введения в теорию прозы": Современные изображения к забытому наследию М. Лопатто с точки зрения квантитатвной лингвистики". In: *Антропология культуры*. Москва. [In print] (2012b)
9. Hřebíček, L.: Text levels. Langauge Constructs, Constituents, and the Menzerath-Altmann Law. wvt, Trier (1995)
10. Kelih, E.: Untersuchungen zur Satzlänge in russischen und slowenischen Prosatexten. Band 1 & Band 2. Graz, M.A. Thesis (2002)
11. Kelih, E., Grzybek, P.: Satzlänge: Definitionen, Häufigkeiten, Modelle. (Am Beispiel slowenischer Prosatexte). In: Quantitative Methoden in Computerlinguistik und Sprachtechnologie. [Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie // Journal for Computational Linguistics and Language Technology. 20, 31-51 (2005)
12. Köhler, R.: Syntactic structures: properties and interrelations. Journal of Quantitative Linguistics, 6, 46-57 (1999)
13. Köhler, R.: Quantitative analysis of syntactic structures. In: Mehler, A., Kóhler, R. (eds.), Aspects of Automatic Text Analysis, pp. 191-209. Springer, Berlin, Heidelberg (2007)
14. Köhler, R.: Quantitative Syntax Analysis. De Gruyter Mouton, Berlin, Boston (2012)
15. Leskiss, G.A.: "О размерах предложения в русской научной и художественной прозе 60-х годов XIX в.". Voprosy jazykoznanija, 2; 78–95 (1962)
16. Leskiss, G.A.: "О зависимости можду размером предложения и характером текста", in: Voprosy jazykoznanija, 3; 92–112 (1963)

17. Leskiss, G.A.: "О зависимости между размером предложения и его структурой в разных видах текста", Voprosy jazykoznanija, 3; 92–112 (1964)
18. Neumann, S.: Das Menzerath-Altmann-Gesetz als Textcharakteristikum. Trier, M.A. Thesis (2009)
19. Roukk, M.: Satzlängen in Texten von A. Tschechow. Göttinger Beiträge zur Sprachwissenschaft, 5, 113-120 (2001a)
20. Roukk, M.: Satzlängen im Russischen. In: Best, Karl-Heinz (ed.), Häufigkeitsverteilungen in Texten, pp. 211-218. Peust & Gutschmidt, Göttingen (2001b)
21. Roukk, M.: The Menzerath-Altmann Law in translated texts as compared to the original texts. In: Grzybek, P.; Köhler, R. (eds.), Exact Methods in the Study of Language and Text. Mouton de Gruyter, Berlin, 605-610 (2008)
22. Sherman, L. A.: Some observations upon the sentence-length in English prose, in: University of Nebraska Studies, 1, 119-130 (1888)
23. Williams, C.B.: A note on the statistical analysis of sentence-length as a criterion of literary style, in: Biometrika, 31, 356-361 (1940)
24. Wimmer, G., Altmann, G.: Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), Quantitative Linguistik · Quantitative Linguistics, pp. 791-807. de Gruyter, Berlin, New York (2005)
25. Wimmer, G., Altmann, G.: Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), Contributions to the science of text and language. Word length studies and related issues, pp. 329-337 Springer, Dordrecht, NL, (2006)
26. Yule, U. G.: On sentence length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika, 30, 363-390 (1938/39)
27. Yngve, V. H.: A model and a hypothesis for language structure. Proceedings of the American Philosophical Society, 194, 444-466 (1960)
28. Yngve, V. H.: From Grammar to Science: New Foundations for General Linguistics. Benjamins, Amsterdam, Philadelphia (1996)

Frequency and Declensional Morphology of Czech Nouns

Ján Mačutek[1], Radek Čech[2]

[1]Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia
jmacutek@yahoo.com

[2]Department of General Linguistics, Philosophical Faculty, Palacký University, Křížkovského 10, 771 80 Olomouc, Czech Republic
cechradek@gmail.com

Abstract. The relationship between frequency and declension of nouns in Czech is analyzed. The nominative is taken as the basic form of nouns. We define the measure of morphological change as the number of phonetic changes in the stem plus the number of phonetic changes in the declensional suffix. Two approaches were examined: 1) nominative singular as the basic form of a noun regardless of its grammatical number, 2) nominative singular as the basic form of a noun in the singular and nominative plural as the basic form of a noun in the plural. In both cases, the relation "the lesser the change, the higher the frequency" is observed.

Keywords: declension, morpho-phonetic, change frequency, case, noun.

1 Introduction

The relationship between the frequency and a word form's "behavior" has been well known for almost two centuries. As M. Krug noticed in [18], J. Grimm [8] already observed the correlation between frequency and irregularity: "Auxiliaria, d.h. Verba, welche sehr häufig gebraucht werden und statt ihrer lebendigen Bedeutung abstrakte Begriffe annehmen, tragen gowöhnlich solche Unregelmässigkeiten an sich." [Auxiliaries, i.e. verbs which are used very frequently and which take on abstract notions instead of their vivid meanings, usually display such irregularities] (cited and translated according to [8], p. 8). However, except for quantitative linguistics [15], [16], Zipf's approach [24], [25] and so called usage-based models [1], [2], [3] [4], [10], where the frequency is considered to be one of the central "powers" shaping properties of language, the focus on frequency effects on word form (or phoneme, syllable, sentence, etc.) is rather exceptional among linguists,

cf. "A newcomer to the field of linguistics might be surprised to learn that for most of the twentieth century facts about frequency of use of particular words, phrases, or constructions were considered irrelevant to study of linguistic structure. To the uninitiated, it does not seem unreasonable at all to suppose that high-frequency words and expressions might have one set of properties and low-frequency words and expressions another" ([2], p. 5).

In the present study, the relationship among frequency, morphology and phonetics is observed. Particularly, it is assumed that frequency has an impact on the number of morpho-phonetic changes in a particular word form. We hypothesize that *the greater the magnitude of a morpho-phonetic change, the lower the frequency of word forms with the magnitude* (the number of changes with respect to the nominative is considered, cf. Section 2). We can find, to our knowledge, two theoretical sources for the reasoning of this kind: 1) synergetic linguistics [13], [14] and 2) a usage-based cognitive approach [3]. As for 1), we preliminarily suppose that the relationship could be viewed as a result of a mutual interaction of so-called requirements, namely "Minimisation of producing effort", "Minimisation of encoding", and "Minimisation of memory effort" for a speaker, and "Minimisation of decoding", "Minimisation of memory effort", and "Minimisation of inventories" for a hearer. As for 2), from the cognitive point of view the effects of high token frequency are considered as follows: "because exemplars are strengthened as each new token of use is mapped onto them, high-frequency exemplars will be stronger than low frequency ones, and high-frequency clusters – words, phrases, constructions – will be stronger than lower frequency ones. The effects of this strength (lexical strength [1]) are several: first, stronger exemplars are easier to access, thus accounting for the well-known phenomenon by which high-frequency words are easier to access in lexical decision tasks. Second, *high-frequency, morphologically complex words show increased morphological stability*." ([3], p. 24, italics by the authors of this paper).

2 Language material and methodology

For the testing of a hypothesis concerning morpho-phonetic changes a language with a rich inflexional system should be used, for obvious reasons. Therefore, Czech has been chosen; in particular, we focused on Czech nouns.

The system of Czech noun declensional morphology contains seven cases (nominative, genitive, dative, accusative, vocative, locative, instrumental). Throughout the paper, we use the following abbreviations for the cases: N – nominative, G – genitive, D – dative, A – accusative, V – vocative, L – locative, I – instrumental. The morphology is expressed by inflectional endings (desinences) added to the stem, cf. the word forms of masculine noun *táta (daddy)* (Table 1).

The endings do not merely express information regarding case but also number and gender. Therefore, Czech is typologically ranked among fusional languages (one ending denotes more than one morphological category). Further, morpho-phonetic alternations are typical for Czech, e.g. an elision of "e" in the stem of the word *pes (dog)*

$$\mathrm{pes}(\mathrm{N},\mathrm{singular}) \qquad \mathrm{ps}-\mathrm{a}(\mathrm{G},\mathrm{singular})$$

Table 1. An example of declensional morphology of Czech nouns: word *táta (daddy)*

case	singular	plural
N	tát-a	tát-ové
G	tát-y	tát-
D	tát-ovi	tát-ů
A	tát-u	tát-y
V	tát-o	tát-ové
L	tát-ovi	tát-ech
I	tát-ou	tát-y

or an alternation ”k” to ”c” in word *kluk (boy)*

$$\text{kluk}\,(\text{N}, \text{singular}) \qquad \text{kluc} - \text{i}\,(\text{N}, \text{plural}).$$

According to a traditional view (cf. [11] and [23]) the alternations (and other changes) are governed by rules which have no relation to frequency.

Two texts were used for the analysis, namely the book of travel ”Obrázky z Holandska” (Pictures from the Netherlands) written by Karel Čapek and the short novel ”Krásná Poldi” (Beautiful Poldi) written by Bohumil Hrabal. The language data were taken from *Dictionary of Karel Čapek* [5] and from *Dictionary of Bohumil Hrabal* [6], which are, actually, lemmatized and morphologically tagged authors' corpora. The lemmatization and morphological tagging allow us to process data automatically; for example, all forms of the lemma *kráva (cow)* and their frequencies can be easily obtained as follows.

Table 2. Morphological tagging in [5] and [6]

word form	tag (morphology)	frequency
krávy	NNFP1——A——	4
krávy	NNFS2——A——	3
krávy	NNFP4——A——	2
kráva	NNFS1——A——	2
krav	NNFP2——A——	2
kravami	NNFP7——A—1-	1
krávu	NNFS4——A——	1
kravách	NNFP6——A—1-	1

A letter in the first column depends on a part of speech (we chose only nouns denoted by N) The information regarding grammatical number and case, needed for our analysis, is represented by letters (S = singular, P = plural) in the fourth column and by digits (1 = N, 2 = G, etc.) in the fifth columns of the tags.

The nominative form was taken as the basic form of each word. Then, for each lemma the number of phonetic changes/alternations with respect to the nominative in the stem and the number of phonetic changes/alternations with respect to the nominative in the suffix were determined manually. The total number of changes/alte-

rnations (i.e., those from the stem plus the ones from the suffix) was used as a magnitude of the morphological change. For example, let us consider all singular cases of the lemma *stůl (table)* (Table 3).

Table 3. Magnitudes of change in word *stůl (table)*

case	word form	number of changes
N	stůl	0
G	stolu	2
D	stolu	2
A	stůl	0
V	stole	2
L	stolu	2
I	stolem	3

The nominative is regarded as the basic form, therefore it is assigned zero changes (as well as accusative which is represented by the same word form); genitive, dative, and locative display one change in the stem (alternation *ů-o*) and the inflectional ending is represented by one vowel (*u*), whereas the basic form (i.e., nominative) does not have any suffix (or it has a zero-morpheme suffix); so, for all these cases two changes are assigned; instrumental displays one change in the stem (alternation *ů-o*) and the inflectional ending represented by a vowel (*e*) and a consonant (*m*); therefore, three changes are assigned to it.

We remind readers that making a suffix longer (or adding it if there was none) is taken into account (cf. Table 3, the word *stůl* in the nominative and instrumental case). The same applies to an elimination of phonemes (cf. the word *pes* in Table 4).

It should be noticed that we followed the morpho-phonetic approach and phonetic (not phonemic or graphemic) changes were taken into account; for example, in the case of the word *dub (oak)*, pronounced [dup], we counted also change of voice (p-b), cf. Table 4:

Table 4. Magnitudes of change in words *dub (oak)* and *pes (dog)*, word forms in singular are presented for both nouns. Pronunciation is given in square brackets, elimination of a phoneme is marked by Ø.

case	*dub*		*pes*	
	word form	number of changes	word form	number of changes
N	dub [dup]	0	pes	0
G	dubu [dubu]	2	pØsa	2
D	dubu [dubu]	2	pØsovi	4
A	dub [dup]	0	pØse	2
V	dube [dube]	2	pØsa	2
L	dubu [dubu]	2	pØsovi	4
I	dubem [dubem]	3	pØsem	3

Two approaches were applied: first, the nominative singular was taken as the basic form of a noun regardless of its grammatical number; second, the nominative

singular was taken as the basic form of a noun in the singular and the nominative plural as the basic form of a noun in the plural. We thus analyze six datasets (three for each author).

3 Results

In the first step, we modeled the relation between the magnitude of change and the frequency of word forms. We fit all datasets with the function

$$y = a(x+1)^b e^{-cx}, \tag{1}$$

where x is the magnitude, y is the frequency of word forms displaying magnitude x, and a, b and c are parameters. Function (1) is the basic form of the Wimmer-Altmann model [22]. The parameter a is often interpreted as the value which y takes for the smallest x, i.e., in this case a is the frequency of word forms which are the same as the nominative form and they are therefore assigned zero changes. Thus we have

$$y = y(0)(x+1)^b e^{-cx}. \tag{2}$$

The data obtained, together with parameter values b and c from function (2) and values of the determination coefficient R^2, can be found in Table 5. Throughout the paper we use the following notation:

$(S+P)/S$ – nominative singular is the basic form for all nouns regardless of their grammatical number,
S/S – nominative singular is the basic form for nouns in the singular,
P/P – nominative plural is the basic form for nouns in the plural.

Table 5. Magnitude of change and frequency – fitting function (2) to data form [5] and [6]

	Čapek			Hrabal		
	(S+P)/S	S/S	P/P	(S+P)/S	S/S	P/P
magnitude of change	frequency					
0	738	711	426	821	778	295
1	849	456	181	639	395	120
2	566	235	140	382	183	86
3	70	33	44	72	34	25
4	10	3	2	7	1	4
parameters and determination coefficient						
b	2.953	1.572	0.258	1.836	0.981	-0.353
c	1.857	1.506	0.579	1.484	1.337	0.575
R^2	0.959	0.991	0.976	0.983	0.992	0.985

As can be seen, function (2) yields a very good fit in all cases.

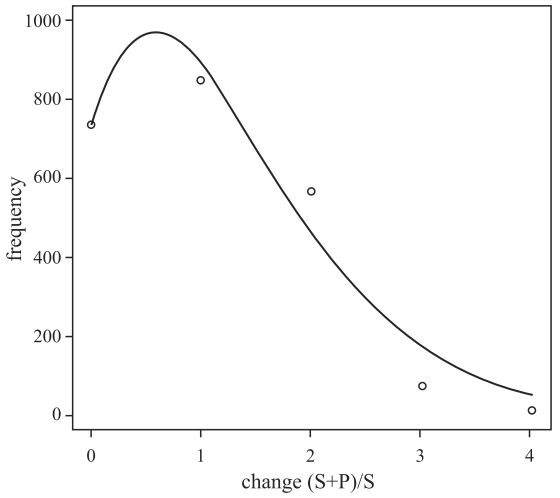


Fig. 1. Fitting function (2) to data from [5], method (S+P)/S

The data and the respective functions for Čapek's book from Table 5 are presented also in Figures 1, 2 and 3.

The data and respective curves for Hrabal's short story are quite similar (cf. Table 5).

We also paid attention to a similar hypothesis introduced by Fenk-Oczlon in [7] (cf. also a collection of open problems in quantitative linguistics [21]): "the more frequent a case in a particular language, the more it tends toward zero coding". Frequencies of cases in the texts by apek and Hrabal can be easily determined

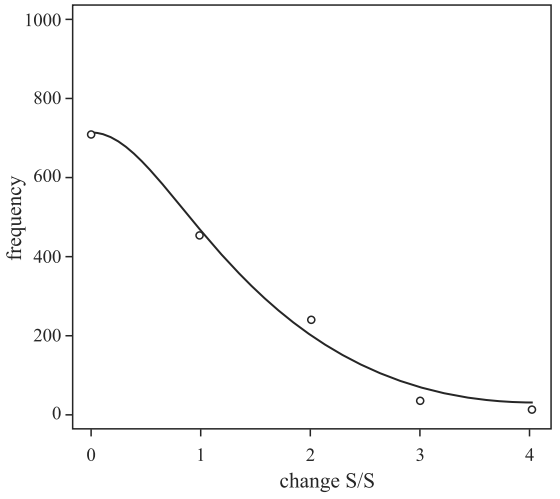


Fig. 2. Fitting function (2) to data from [5], method S/S

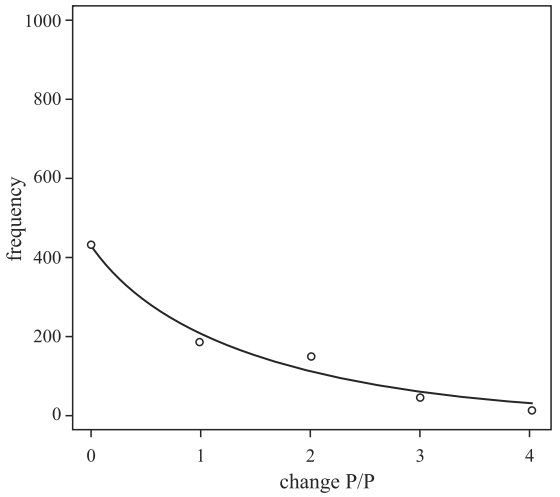


Fig. 3. Fitting function (2) to data from [5], method P/P

from the morphological tags (cf. Table 2). The magnitude of coding was defined as the mean number of changes with respect to the basic form (cf. Section 2). We applied two approaches from Section 2 again, i.e., first, the nominative singular was taken as the basic form of a noun regardless of its grammatical number; second, the nominative singular was taken as the basic form of a noun in the singular and the nominative plural as the basic form of a noun in the plural. The mean number of changes per case was then computed as the total number of changes per case (i.e., the sum of all changes per case) divided by the frequency of the case. Results are presented in Tables 6 and 7.

Table 6. Case frequency and magnitude of coding in the text by Čapek [5] (f – frequency, *mc* – magnitude of coding)

	S+P/S		S/S		P/P	
case	f	*mc*	f	*mc*	f	*mc*
N	707	0.45	462	0.00	245	0.00
G	570	1.21	352	1.01	218	1.31
D	62	1.63	40	1.23	22	2.23
A	382	0.68	237	0.34	145	0.14
V	3	1.67	2	1.50	1	0.00
L	239	1.61	183	1.43	56	2.00
I	270	1.76	162	1.77	106	1.25

The tendency from Fenk-Oczlon's hypothesis can be seen (it can be tested, e.g., by the Kendall correlation coefficient, the p-values are higher than 0.05 in all cases). Nevertheless, the data are not "smooth" and therefore they are difficult to model. Function (2), e.g., does not fit them well (which is true especially for the text by Čapek). Figure 4 presents the data from Čapek's text, with nouns in the singular

Table 7. Case frequency and magnitude of coding in the text by Hrabal [6] (f – frequency, *mc* – magnitude of coding)

	S+P/S		S/S		P/P	
case	*f*	*mc*	*f*	*mc*	*f*	*mc*
N	707	0.45	462	0.00	245	0.00
G	570	1.21	352	1.01	218	1.31
D	62	1.63	40	1.23	22	2.23
A	382	0.68	237	0.34	145	0.14
V	3	1.67	2	1.50	1	0.00
L	239	1.61	183	1.43	56	2.00
I	270	1.76	162	1.77	106	1.25

being considered, i.e., S/S. One obtains a very low value of the determination coefficient $(R^2 = 0.486)$.

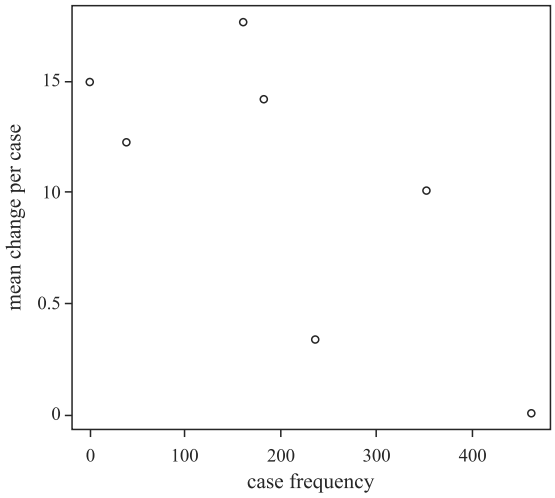


Fig. 4. Case frequency and magnitude of coding, Čapek S/S (cf. Table 6)

This is the most irregular behavior among the datasets from Tables 6 and 7, but only two of the datasets achieve a determination coefficient higher than 0.8 (for apek's text one obtains $R^2 < 0.7$ for all three approaches). However, applying a different methodology (i.e., a different definition of the magnitude of change/case coding, e.g., the length of the suffix only) could lead to different results.

4 Conclusion and ideas for further research

The paper shows that declensional morphology of Czech nouns is clearly related to frequency – if the nominative is taken as the basic form, it holds that "the lesser the

magnitude of change, the higher the frequency" (or, in other words, the more similar the word forms are to their nominative case, the more frequently they occur).

This line of investigation should be broadened in two directions. First, similar studies must be undertaken also for other (and not only Slavic) languages with a rich inflexional system. Second, other parts of speech, especially adjectives, should be studied. Results obtained should be in future interpreted within the synergetic linguistics framework [13], [14]. We suppose that especially connections and interrelations with other properties of morphology and syntax will be established.

The data presented in this paper, especially Tables 6 and 7, can serve as material for building a model for case frequencies (a study on case diversification [20] contains rank-frequency distributions of cases from many texts in German, Russian, Slovak and Slovene).

Relations among the case frequency, the frequency of inflexional suffixes and the length of inflexional suffixes (which can be understood as a measure of inflexion/change) are exploited in psycholinguistics (we mention works [17] for Polish, [12] and [19] for Serbian, and [9] for Slovak). A theoretically based model would therefore also be useful in this research area.

Acknowledgment. The authors were supported by research grants VEGA 2/0038/12 (J. Mačutek), and ESF OPVK 2.3 – Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology (CZ.1.07/2.3.00/20.0161) (R. Čech).

References

1. Bybee, J.: Morphology: A Study of the Relation Between Meaning and Form. Benjamins, Amsterdam (1985)
2. Bybee, J.: Frequency of Use and the Organization of Language. Oxford University Press, Oxford (2007)
3. Bybee, J.: Language, Usage and Cognition. Cambridge University Press, Cambridge (2010)
4. Bybee, J., Hopper, P. (eds.): Frequency and the Emergence of Linguistic Structure. Benjamins, Amsterdam/Philadelphia (2001)
5. Čermák, F. (ed.): Slovník Karla Čapka. Nakladetelství Lidové noviny, Praha (2007)
6. Čermák, F., Cvrček, V. (eds.): Slovník Bohumila Hrabala. Nakladetelství Lidové noviny, Praha (2009)
7. Fenk-Oczlon, G.: Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.) Frequency and the Emergence of Linguistic Structure, pp. 431–448. Benjamins, Amsterdam/Philadelphia (2001)
8. Grimm, J.: Deutsche Grammatik. Dieterich'sche Buchhandlung, Göttingen (1822)
9. Hanulíková, A., Davidson, D.J.: Inflexional entropy in Slovak. In: Levická, J., Garabík, R. (eds.) NLP, Corpus Linguistics, Corpus Based Grammar Research, pp. 145–151. Tribun, Brno (2009)
10. Hopper, P.: Emergent grammar. In: Aske, J., Beery, N., Michaelis, L., Filip, H. (eds.) Proceedings of the 13th Annual Meeting of the Berkeley Linguistic Society, pp. 139–157. Berkeley Linguistic Society, Berkeley (1987)

11. Karlík, P., Nekula, M., Rusínová, Z: Příruční mluvnice češtiny. Nakladetelství Lidové noviny, Praha (1995)
12. Kostić, A., Mirković, J.: Processing of inflected nouns and levels of cognitive sensitivity. Psihologija 35, 287—297 (2002)
13. Köhler, R.: Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum (1986)
14. Köhler, R.: Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) Quantitative Linguistics. An International Handbook, pp. 760–774. de Gruyter, Berlin/New York (2005)
15. Köhler, R., Altmann, G.: Aims and methods of quantitative linguistics. In: Altmann, G., Levickij, V., Perebyinis, V. (eds.) Problems of Quantitative Linguistics, pp. 13–43. Ruta, Chernivtsi (2005)
16. Köhler, R., Altmann, G.: Quantitative linguistics. In Hogan, P.C. (ed.) The Cambridge Encyclopedia of the Language Sciences, pp. 695–697. Cambridge University Press, New York (2011)
17. Krajewski, G., Lieven, E.V.M., Theakston, A.L.: Productivity of a Polish child's inflexional noun morphology: a naturalistic study. Morphology 22, 9–34 (2012)
18. Krug. M.: Frequency as a determinant in grammatical variation and change. In: Rohdenburg, G., Mondorf, B. (eds.) Determinants of Grammatical Variation in English, pp. 7–67. de Gruyter, StateBerlin (2003)
19. Milin, P., Filipović Đurđević, D., Moscoso del Prado Martín, F.: The simultaneous effects of inflexional paradigms and classes of lexical recognition: Evidence from Serbian. Journal of Memory and Language 60, 50–64 (2009)
20. Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G.: Diversification of the case. Glottometrics 18, 32—39 (2009)
21. Strauss, U., Fan, F., Altmann, G.: Problems in Quantitative Linguistics 1. RAM-Verlag, Lüdenscheid (2008)
22. Wimmer, G., Altmann, G.: Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) Quantitative Linguistics. An International Handbook, pp. 791–807. de Gruyter, Berlin/New York (2005)
23. Ziková, M.: Alternace vokálů s nulou v současné češtině – laterální autosegmentální analýza. Dissertation thesis. Filozofická fakulta Masarykovy university, Brno (2008)
24. Zipf, G.K.: The Psycho-Biology of Language. MIT Press, Cambridge (MA) (1935)
25. Zipf, G.K.: Human Behavior and the Principle of the Least Effort: An Introduction to Human Ecology. Addison-Wesley, Cambridge (MA) (1949)

Stochastic Dynamic Model of Evolution of Language Signs Ensembles

Vasiliy Poddubnyy[1], Anatoliy Polikarpov[2]

[1]Computer Science Faculty, Tomsk State University,
Tomsk, 634050, Russian Federation,
vvpoddubny@gmail.com

[2]Laboratory for General and Computational Lexicology and Lexicography,
Faculty of Philology, Lomonosov Moscow State University,
Moscow, 119991, Russian Federation,
anatpoli@mail.ru

Abstract. On the basis of assumptions present in the linguistic model of the life cycle of signs, an attempt is undertaken to build a mathematical model for revealing some mechanisms of evolutionary processes in language. For this purpose a dissipative stochastic dynamic model of language sign-ensemble evolution is proposed. The model satisfies the principle of least action – one of the fundamental variational principles of Nature. The model assumes a Poisson birth flow for language signs and an exponential distribution for their associative-semantic potential (ASP). The model makes it possible to deduce equations for synchronous (momentary) lexical polysemy and rank-frequency distributions. It is important to note that the model parameter values do not differ significantly in terms of the Kolmogorov-Smirnov test from empirical values of polysemy and frequency of use distributions for lexical units chosen from representative Russian and English explanatory and frequency dictionaries.

Keywords: Language sign, evolution, associative-semantic potential, sign meanings, polysemy, dissipative stochastic dynamic model.

1 Introduction

Our previous work developed and tested *a linguistic evolutionary model* based on regularities of a typical *sign life cycle* from language sign inception up to its falling out of linguistic use [1-14]. On the basis of this approach an attempt at mathematical modeling of the language sign life cycle was undertaken by applying *splitting process theory* [4]. The present paper is a further step in the series of studies undertaken by the authors in creation and experimental verification of a novel mathematical model of language evolution, namely, a *stochastic dynamic model of evolution of language sign ensembles* [15] with a special emphasis on evolutionary formation of polysemy and frequency features of lexical units.

2 Major linguistic preliminaries for building the model

Certain dissipative microprocesses constantly occurring with respect to signs in each communicative act constitute the core of *basic evolutionary microdynamics* of any natural language. The most fundamental point is that each communicative act is *an act of hinting by a sender and guessing by a recipient of a certain target sense* by use of some appropriate sign in one of its most appropriate meanings. The most appropriate meaning among all meanings of a linguistic sign for this is that which is the *most similar* to the target sense. Hinting/guessing uncertainty, that is, the ability of any meaning in any act of communication to cover some *new* target sense leads to the *possibility* of *continuous expansion of the sense area for each of the sign's meanings*. Accumulation of such *directed micro-changes* in the sense area of each meaning leads in the course of time to some even more significant *micro-evolutionary tendencies*:

- *increase of an overall diversity of senses* in the sense area of each meaning;
- *increase of an overall diversity of components* of those senses in the sense area of each meaning;
- *narrowing of an intersection of a number of components common for all senses in the sense area of some meaning*;
- *increase of load for a more limited number of components of any meaning* as a result of a narrowing base for association with a reduced number of components common to all meanings' senses;
- *strengthening of overloaded meaning components and weakening of underloaded ones*;
- *abstractivization*, gradual loss of the weakest components of meanings over time;
- *gradual despecification, subjectification and grammaticalization of remaining meaning components* resulting from the loss of more specific, objective and substantive meanings' components during abstractivization process and from the associative attraction to meanings of those sense components from usual contexts of meanings' use which are less specific, subjective and relational than contextual components on the previous stages of meanings' existence; this is because contexts of relatively more abstract meanings become relatively wider with greater repetition in them of the most usual sense components present in any act of communication – components of *subjective relations between communicants, their negative and positive judgments, estimations, etc.;*
- *emergence of new meanings* as a result of unusually strong, "illogical" widening of the sense sphere of the original meanings;
- *gradual retardation of the process of new meaning generation from any existing meaning*, because of exhaustion of gradually occupied valences of it for association with new senses;
- *retardation of the process of new meanings generation at any next step of meanings breed* as a result of *relatively greater degree of abstraction of meanings at more remote steps of the process*;

– *relative increase in stability of meanings at each successive step of their generation* as a result of the increase in their degree of abstractness and *less specific, less objective and more grammatical nature of remaining components.*

The difference between the quantity of meanings *acquired* by a sign and the quantity of meanings which *have fallen out of use* at any given moment of sign's life is the size of its *polysemy*. The curve of polysemy development over time for each sign is, in general, *a unimodal curve with a maximum polysemy achieved very quickly near the beginning of the process, and a gradual diminution over a very long time until disappearance of the last meaning*. This is because, firstly, the loss of any previously acquired meaning happens not immediately, but in some certain time, and, secondly, because the stability of later acquired meanings, on the average, increases.

The same basis can be also used for prediction of the development in time, for each stage of realization in its life cycle, of an *overall semantic (sense) volume of a sign*, measured by the sum of semantic volumes of all meanings, as well of the development in time of an *overall frequency of sign's use*. The main point is that *contribution of each successive meaning to the overall sum of frequencies of sign meanings gradually increases, because of the increase of the degree of abstractness of later born sign's meanings*.

Abstraction of each meaning in its history, in combination with acquisition of new, relatively more abstract meanings and gradual loss of earlier acquired, more concrete meanings form the basis for various directed changes in a sign within its life cycle – *changes in lexical, morphemic and phraseological polysemy, homonymy, synonymy, antonymy, degree of idiomaticity of meanings, relative frequency of sign use, sign length*, and so on [9-10].

3 Initial formalization of some main points of the model

3.1 Associative-semantic potential of a sign

The ability of a sign to generate new meanings is called in the model *"associative-semantic potential (ASP) of a sign"*. ASP is *measured by the maximum number of meanings which a sign is capable of generating (acquiring) during its life cycle. ASP is initially present in the original meaning of a sign, from which its semantic history begins*. Generation of new meanings redistributes to a certain extent the amount of ASP among successive meanings. Abstraction and loss of previously acquired meanings leads to diminution *of remaining amount of ASP, leads to its step-by-step dissipation*.

3.2 Length of a sign life cycle

In our model, the duration of a sign's life cycle is defined by three major factors:

– the *size of ASP*, measured by the *number of meanings which a certain sign is capable of generating during its life span*,

- the *degree of activity of a sign*, measured by the *rate of new meaning generation*, or, in other words, by the *speed of ASP realization*,
- the *degree of sign stability*, measured by the *degree of longevity of each generation of meanings* and, accordingly, *of the sign as a whole*.

4 Mathematical modeling of sign life cycle and that of the ensemble behavior of signs in language

4.1 Basic assumptions

The proposed *dissipative stochastic dynamic model of the development of language signs* assumes a *Poisson character for the stream of language signs' births and an exponential distribution in an ensemble of signs according to their ASP*, and is governed by differential stochastic equations of a special kind: $dx/dt = (G-x)/\tau$ or difference equations of the same kind derived from the principle of the least action for dissipative processes $\triangle k/\triangle t_k = (G-x)/\tau$, where G is *ASP*, $\triangle k = 1$, and $\triangle t_k = t_{k+1} - t_k$.

4.2 Acquisition and loss of sign meanings

Let us write out these equations for i-th sign in an ensemble, having supplied a process of appearance of new meanings by index 1, and a process of loss of meanings by index 2:

$$\begin{gathered} t_{i,k+1}^{(1,2)} = t_{i,k}^{(1,2)} + \tau_i^{(1,2)}/(G_i - k) + \xi^{(1,2)}, \quad k = \overline{1, G_i - 1}, \\ t_{i,1}^{(1)} = t_i, \quad t_{i,1}^{(2)} = t_i + \tau_{0i}, \quad \tau_i^{(2)} > \tau_i^{(1)}, \quad i = \overline{1, N}. \end{gathered} \tag{1}$$

Here $t_{i,k}^{(1,2)}$ is the birth moment, that is, index 1, or the moment of falling out of use (index 2), of a k-th meaning of the i-th sign, G_i. Random ASP of the i-th sign is distributed according to the exponential law with an average value $\langle G \rangle$, and t_i is the random moment of occurrence of sign in a language when $t_{i+1} = t_i + \tau$, where τ is some random interval of time between occurrences of neighboring signs distributed presumably according to the exponential law with an average value $\langle \tau \rangle$. τ_{0i} is some random delay in the beginning of the falling out of use process for meanings of the i-th sign in relation to the moment of the sign's origin in a language, which is distributed according to the exponential law with an average $\langle \tau_0 \rangle$. $\tau_i^{(1,2)} = c^{(1,2)} \langle G \rangle / G_i$ are constants in time of birth processes for new meanings and of time of falling out of use for the meanings, their values being inversely proportional to ASP values so they appear at random. Their distributions are defined by the distribution of their ASP. $\xi^{(1,2)}$ are random fluctuations of the moments of birth and moments of falling out of use of the meanings, which are distributed according to the uniform law with zero averages and semiwidth of intervals of the distribution $\tau_i^{(1,2)}/(G_i - k)$; N is a number of signs (words) in a language.

4.3 Parameters of the dissipative stochastic dynamic model of the polysemy evolution of language sign ensembles

The model (1) is a dissipative stochastic dynamic model of evolution of language sign ensembles. There are 5 parameters in this model:

- *intensity of a stream of new signs* $1/\langle\tau\rangle$, or an average interval of time $\langle\tau\rangle$ between occurrences of the neighboring signs in a stream,
- *average delay of the beginning of the process of sign meanings falling out of use* in relation to the moment of the origin of a sign in a stream $\langle\tau_0\rangle$,
- *average value of the sign's ASP* $\langle G\rangle$,
- *factors* $c^{(1)}$ *and* $c^{(2)}$ $(c^{(1)} \ll c^{(2)})$ *of inversely proportional dependence of constants of time of birth and death processes of sign meanings on a sign's ASP* (these factors define average values of time constants $\langle\tau^{(1,2)}\rangle = c^{(1,2)}Ei(1/\langle G\rangle)$, where $Ei(1/\langle G\rangle)$ is a certain integral exponential function, which increases monotonously and is convex upwards on $\langle G\rangle$, so that $\langle\tau^{(1)}\rangle \ll \langle\tau^{(2)}\rangle$).

4.4 Dissipative stochastic dynamic model of the frequency of use evolution of language signs ensembles

We construct evolutionary models for the prediction of synchronous distributions of lexical sign sense volume, frequency of use, length, age, and so on in a fashion similar to the above model, as intermediate states within some specific process. This is achieved by introducing some additional parameters into the model.

5 Testing of the model for polysemy distributions of lexical signs

5.1 Sources for empirical verification of predictions for synchronous polysemy distribution of words

For testing of the model adequacy, lexical data from 5 representative Russian and English dictionaries of different types were used – 3 Russian and 2 English [16]:

17-volume Russian Dictionary – "Dictionary of the Modern Russian Literary Language" (1948-1965), a "large dictionary" type;
4-volume Dictionary of Russian – "Dictionary of Russian Language" under the editorship of A.P. Evgenieva, 1957–1961, a "medium dictionary" type;
Russian Ozhegov's Dictionary – "Dictionary of the Russian Language" by S.I. Ozhegov (1972, 9-th edition), a "concise dictionary" type;
Shorter – "Shorter Oxford English Dictionary" (1962), a "medium type dictionary" type;
Hornby – A.S. Hornby. "Oxford Advanced Learner's Dictionary of Current English" (1982), a "concise dictionary" type.

Absolute frequencies of words of varying polysemy from these dictionaries are presented in table 1.

Table 1. Absolute frequencies of word polysemy for Russian and English dictionaries

Polysemy	17-volume Russian Dictionary	4-volume Dictionary of Russian	Russian Ozhegov's Dictionary	Shorter	Hornby
1	76382	59920	44166	45965	36210
2	26102	13236	8390	16397	4688
3	9316	4680	2546	7771	1853
4	3989	2060	960	3825	707
5	1928	996	426	2078	366
6	1057	491	231	1240	180
7	587	282	116	792	129
8	342	191	42	489	66
9	222	97	51	308	57
10	151	79	30	237	34
11	118	40	14	130	16
12	83	25	13	131	16
13	37	17	1	79	15
14	38	11	6	84	10
15	35	9	3	55	7
16	22	11	2	43	5
17	7	3	0	35	2
18	12	2	3	21	4
19	14	1	0	21	1
20	14	2	1	19	1
21	9	0	0	10	0
22	4	0	0	11	0
23	1	1	0	9	1
24	2	1	0	7	1
25	1	3	0	4	1
26	3	0	2	4	0
27	1	0	0	2	0
28	1	0	0	1	0
29	0	1	0	6	1
30	1	0	0	3	1
31	1	0	0	4	0
32	0	0	0	2	0
33	1	0	0	2	0
34	0	0	0	2	0
35	0	0	0	2	0
36	0	0	0	1	0
37	0	0	0	3	0
38	0	0	0	0	0
39	0	0	0	1	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	1	0
43	0	0	0	0	0
44	0	0	0	1	0
45	0	0	0	2	0
46	0	0	0	0	0
47	0	0	0	0	0
48	0	0	0	0	0
49	0	0	0	1	0
50	0	0	0	1	0
51	0	0	0	1	0
Total:	120481	82159	57003	79801	44372

5.2 Family of empirical and theoretically derived curves for the polysemy distributions of Russian and English languages

Fig. 1 presents the family of empirical curves for the polysemy distributions in the above dictionaries.

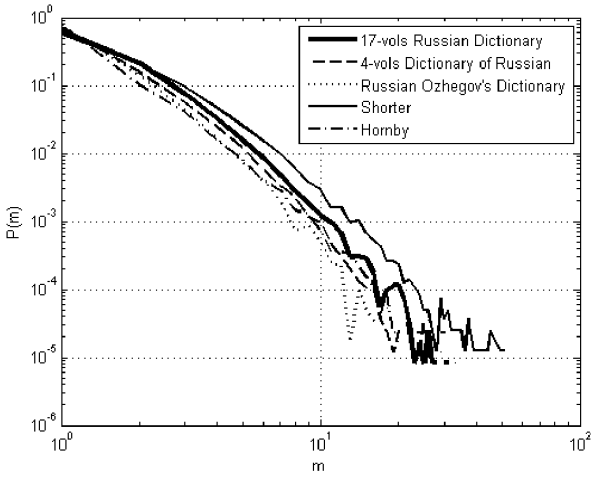


Fig. 1. Family of empirical distributions $P(m)$ for lexical polysemy m of five Russian and English dictionaries

Figs. 2–6 present polysemy distribution curves (shorter curves), *deduced from a stochastic model for predicting some synchronous state (synchronous polysemy dis-*

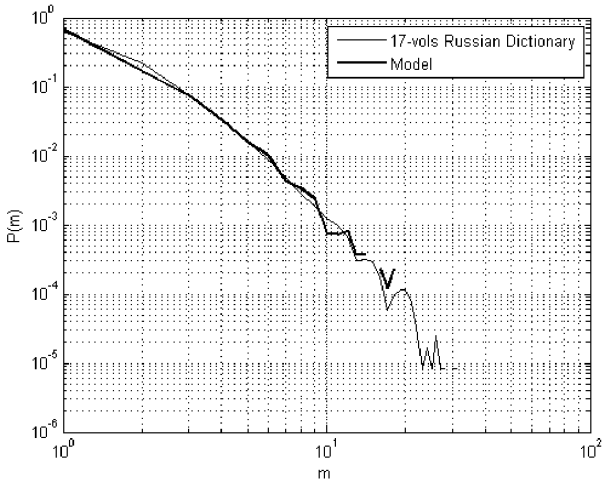


Fig. 2. Distribution for lexical polysemy of the model and of the 17-volume Russian Dictionary

tribution) for each of these empirical polysemy curves (longer curves) for several tens of thousands of lexical units.

The model for each dictionary was identified by the corresponding selection of values for the above-mentioned five parameters.

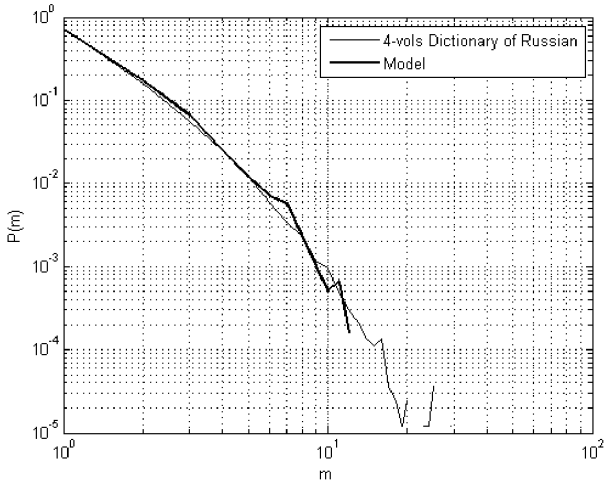


Fig. 3. Distribution for lexical polysemy of the model and of the 4-volume Dictionary of Russian

Model curves, presented in figs. 2–6, correspond to table 2. Model relative frequencies of polysemy, presented in table 2, are derived from modeling data with

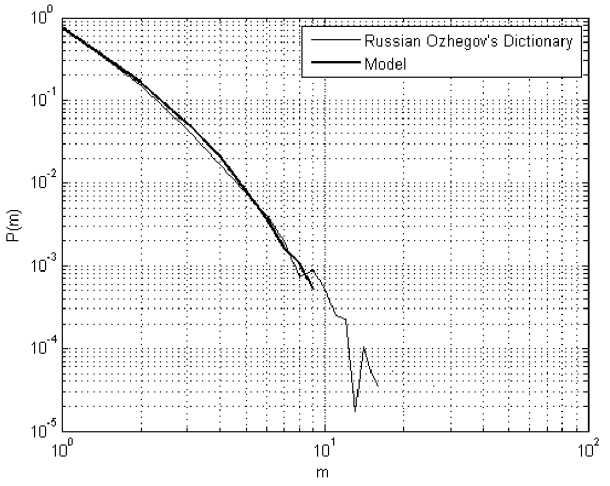


Fig. 4. Distribution for lexical polysemy of the model and of the Russian Ozhegov's Dictionary

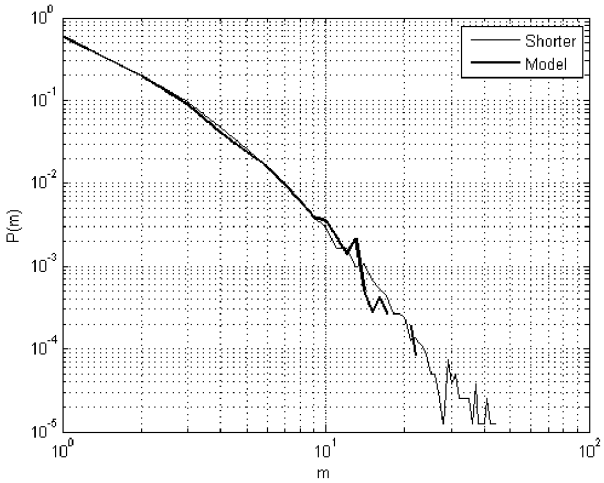


Fig. 5. Distribution for lexical polysemy of the model and of the Shorter dictionary

Table 2. Model absolute frequencies of the polysemy for Russian and English dictionaries

Polysemy	Fig. 2	Fig. 3	Fig. 4	Fig 5	Fig 6
1	5768	4258	1445	7189	5097
2	1398	1027	316	2319	1107
3	619	398	100	1055	331
4	283	150	40	480	107
5	133	73	16	280	55
6	85	43	7	189	22
7	35	34	3	113	16
8	29	14	2	71	11
9	20	6	1	46	10
10	6	3	0	42	2
11	6	4	0	26	3
12	7	1	0	17	0
13	3	0	0	25	0
14	3	0	0	6	0
15	0	0	0	3	0
16	2	0	0	5	0
17	1	0	0	3	0
18	2	0	0	0	0
19	0	0	0	0	0
20	0	0	0	0	0
21	0	0	0	2	0
22	0	0	0	1	0
Total:	8400	6011	1930	11872	6761

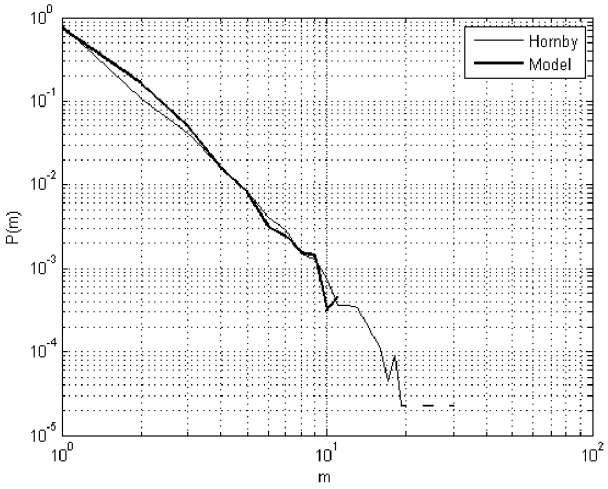


Fig. 6. Distribution for lexical polysemy of the model and of the Hornby dictionary

parameters which provide the maximum vicinity of the polysemy distribution for the dictionaries.

6 Interpretation of polysemy data

6.1 General interpretation of the relation between family of empirical and family of theoretically derived curves for the polysemy distributions

As can be seen, *curves of polysemy distributions deduced from our model (shorter curves on figs. 2–6) do not differ significantly (by the Kolmogorov-Smirnov criterion) from the curves of empirical polysemy distributions (longer curves)* drawn on the basis of the data from the Russian and English dictionaries.

In the table 3, where test results are presented,

- H_0 is the null hypothesis on the identity of empirical and model distributions $P(m)$.
- $K-S$-value is Kolmogorov-Smirnov's criterion for non-zero relative frequencies.
- n is the number of non-zero samples pairs.

6.2 Structural interpretation of parameters of the model

Parameter $\langle G \rangle$ directly influences the degree of inclination of the curve of the polysemy distribution in a double logarithmic scale: the greater the parameter, the smaller the inclination of the curve.

Parameters $c^{(1)}$, usually equal to several conditional units, and $c^{(2)}$, usually equal to about several hundreds of conditional units, influence the curve form.

Table 3. Results of the test for the Kolmogorov-Smirnov criterion

	Fig. 2	Fig. 3	Fig. 4	Fig 5	Fig 6
n	17	12	9	19	11
$K-S$ value	0.1176	0.0833	0.1111	0.1579	0.0909
p-level	0.9994	0.9999	0.9999	0.9563	0.9999
Hypothesis	H_0	H_0	H_0	H_0	H_0

Parameter $\langle\tau\rangle$ in our model is associated with a certain conditional time scale, let us say with "linguistic time".

Parameters $\langle\tau^{(1)}\rangle$ and $\langle\tau^{(2)}\rangle$ through parameter-factors $c^{(1)}$ and $c^{(2)}$ are also correlated with the parameter $\langle G\rangle$ by way of an integrated exponential function $Ei(1/\langle G\rangle)$.

Parameter $\langle\tau_0\rangle$ is the delay of the beginning of the meaning loss process in relation to the beginning of the meaning emergence process.

Values of parameters of the model are presented in the table 4.

Table 4. Parameters of the model for Russian and English dictionaries

Parameters	Fig. 2	Fig. 3	Fig. 4	Fig 5	Fig 6
$\langle G\rangle$	2.4	1.8	1.5	3.5	1.4
$c^{(1)}$	10	1	10	0.1	1
$c^{(2)}$	900	900	900	300	300
$\langle\tau^{(1)}\rangle$	6.752	0.4980	3.984	0.0942	0.3638
$\langle\tau^{(2)}\rangle$	608	448	359	283	109
$\langle\tau\rangle$	0.05	0.05	0.05	0.05	0.05
$\langle\tau_0\rangle$	70	70	70	30	30

6.3 Linguistic-typological interpretation of the difference between empirical curves for the polysemy distributions of Russian and English

We notice that polysemy distribution curves of Russian and English dictionaries, even where values of the parameter $\langle G\rangle$ are close, differ strongly in the values of parameters $\langle\tau_0\rangle$, $\langle\tau^{(1)}\rangle$, $\langle\tau^{(2)}\rangle$. In other words, curves of the polysemy distribution for *The Russian Ozhegov Dictionary* (fig. 4) and for *The English Hornby Dictionary* (fig. 6) have close inclinations, that is, close values of $\langle G\rangle$ (1.5 and 1.4 respectively), but differ in other parameters (in $\langle\tau_0\rangle$ – approximately 2 times, in $\langle\tau^{(1)}\rangle$ – approximately 10 times and in $\langle\tau^{(2)}\rangle$ – approximately 3.5 times in favour of the Russian dictionary).

Based on these results it is possible to come to the conclusion that *the evolutionary parameters* $\langle\tau\rangle$, $\langle\tau_0\rangle$, $\langle\tau^{(1)}\rangle$, $\langle\tau^{(2)}\rangle$, $\langle G\rangle$, that is, the parameters included in the model and responsible for the speed and time length of change of meanings and words in a language *appear to have also synchronous, linguo-typological status*. Namely, the vocabulary of the English language, a language obviously having a

more analytical nature than Russian, is more limited in the set of lexical units but, at the same time, is proportionally richer in phraseological ones. That is why *the synchronous maximum and also average lexical polysemy as well as the maximum and average frequency of use for English lexical units are expected to be greater than for Russian lexical ones. Obviously, this is the case for comparable textual and dictionary materials, at least for parallel corpora and for dictionaries compiled on the same principles. In many cases this is problematic in practice because, for instance, dictionaries of any language may have some unsystematic offset in zones of the most and the least polysemic words*.

The assignment of quantitative values to the above-mentioned evolutionary parameters having the same time-synchronous, linguo-typological status needs special consideration, but this goes beyond the scope of the present paper.

6.4 Typologically determined difference in speed of sign metabolism for English and Russian

Based on the foregoing data, *a "wearing process" of each lexical unit in English is supposed to occur noticeably faster than in Russian. In other words, lexical "metabolism"* (birth, realization of the associative-semantic potential, aging, falling out of use and replacement of lexical meanings and of lexical signs as a whole by some new lexical units [14]) *in English is more intensive than in Russian*. In particular, it can be manifested by a relatively lower average stability of the lexical material of English in comparison with Russian for the same period of time, for example, for a millennium.

The *direct data on stability in time of lexical units of Russian and English* for the last millennium (for instance, according to a glottochronological list), which is at our disposal, *confirms this conclusion*.

For a more detailed analysis of metabolic problems of language life see in works [7; 11; 13; 14].

7 Interpretation of frequency data

7.1 Sources for empirical verification of predictions for regularities of word synchronous frequency-of-use distributions

To test our predictions concerning synchronous frequency distributions of English and Russian words the following frequency wordbooks were used:

NP-RusFrqDict – the frequency wordbook based on 1,350 thousand words from Russian newspaper texts of the last decade of the 20th century.

Pushkin Dictionary – DB for Pushkin's lexicon including frequencies of word usage based on text counts from "Complete Collection of Pushkin's works" (about 500,000 words).

BNC-EngFrqDict – the large frequency wordbook based on 100,350 thousand words from English texts of the second half of the 20th century (British National Corpus).

Both Russian dictionaries were compiled at the Laboratory for General and Computational Lexicology and Lexicography of Moscow University.

7.2 Close correlation between empirical and theoretically derived curves for the frequency-of-use distributions of Russian and English languages

Figs. 7–8 present empirical and theoretically-computationally derived curves for frequency distributions for several tens of thousands of words. For each of the above dictionaries the model was identified by corresponding computational selection of parameters.

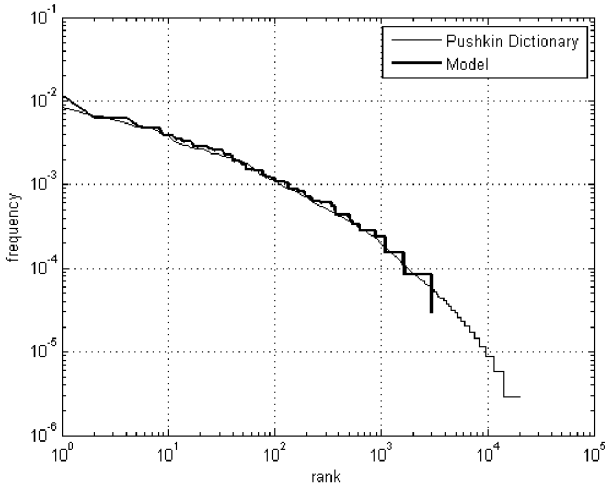


Fig. 7. Rank-frequency distributions for Pushkin dictionary and the model

As can be seen from figs. 7 and 8, the word distributions derived from our model (shorter curves) statistically do not significantly differ from the empirical distributions derived on the basis of the data from Russian and English frequency wordbooks (middle and long curves).

Acknowledgements. We express our gratitude for the partial support of this work by the Russian Foundation for Basic Research grants No. 08-07-00435-a and No. 11-07-00776-a.

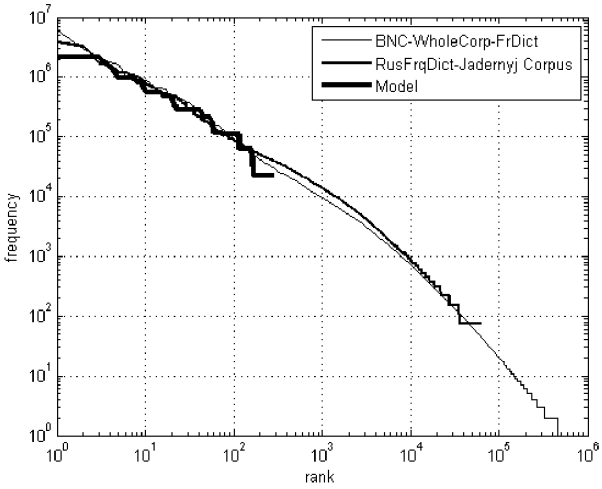


Fig. 8. Rank-frequency distributions for NP-RusFrqDict, BNC-EngFrqDict and the model

References

1. Polikarpov, A.A.: To the Theory of Life Cycle for Lexical Units. In: Applied Linguistics and Automatic Analysis of Texts: Proc. of the Conference (January 28–30th, 1988), p. 66. Tartu University press, Tartu (1988) (in Russian).
2. Polikarpov, A.A.: A Model of the Word Life Cycle. In: Koehler, R., Rieger, B.B. (eds.) Contributions to Quantitative Linguistics, pp. 53–66. Kluwer, Dordrecht (1993).
3. Polikarpov, A.A.: Regularies of New Words Formation: Modeling of the Process and its Experimental Study. In: Language. Verb. Sentence. For 70th Birthday of Professor G.G. Silnitskiy, pp. 211–227. Smolensk (2000) (in Russian), http://www.philol.msu.ru/lex/ articles/words_ex.htm
4. Khmelev, D.V., Polikarpov, A.A.: Basic Assumptions about a Sign's Life Cycle for Mathematical Modelling of Language System Evolution. In: Baayen, R.H. (ed.) Proceedings of the fourth Conference of the International Quantitative Linguistics Association (Prague, August 24–26, 2000). Prague (2000), http://www.philol.msu.ru/lex/khmelev/proceedings/qualico2000.html
5. Polikarpov, A.A.: Cognitive Modeling of Cyclic Processes in Becoming of Lexical System of Language. In: Works of Kazan School in Computer and Cognitive Linguistics, pp. 57-99. Kazan (2001) (in Russian).
6. Polikarpov, A.A.: Cognitive Mechanisms for the Emergence of Some Regularities of Human Language Evolution. In: Text Processing and Cognitive Technologies, No. 9, pp.10–20. Moscow-Varna (2004).
7. Polikarpov, A.A., Selezniova, L.A.: Degree of Abstractness and Subjectivity of Sense – Factors of Variation of the Degree of Safety for Lexical Nominations in Time. In: Problems of Modern Indo-European Linguistics, pp. 327–376. Moscow (2004) (in Russian).

8. Polikarpov, A.A., Filimonova, T.V.: On the System Dependence of Negatively Coloured Objective Denotative and Subjective-Evaluative Features of Phraseological Units on Time. In: Cognitive odeling in Linguistics (CLM-2005). Text Processing and Cognitive Technologies, No. 10, pp. 187–197. Moscow-Varna (2005) (in Russian).
9. Polikarpov, A.A.: Towards the Foundations of Menzerath's Law (On the Functional Dependence of Affixes Length on their Positional Number within Words). In: Peter Grzybek (ed.). Contributions to the Science of Text and Language. Word Length Studies and Related Issues, pp. 215–240. Springer, Dordrecht (2006).
10. Polikarpov, A.A.: System-quantitative Approach in Linguistics. In: Philological Schools and their Role in System Organization of Scientific Studies, pp 35–59. Magenta Publishers, Smolensk (2007) (in Russian).
11. Polikarpov, A.A.: System Dependence of the Degree of Replacement of Old Russian Words in Modern Russian Language on their Age, Categorial Status, Frequency of Use and Polysemy. In: Kochergina, V.A. (ed.) Linguistic Comparativistics in Cultural and Historical Aspects, pp. 232–260. Moscow University Publishing House, Moscow (2007) (in Russian).
12. Polikarpov, A.A., Kukushkina, O.A, Toktonov, A.G.: Check of the Theoretically Predicted Neoderivatological Regularities by the Neological Data From Russian Newspaper Corpus. In: Chernyshova, M.I. (ed.) Theory and History of Slavic Lexicography, pp. 392–427. Moscow (2008) (in Russian).
13. Polikarpov, A.A.: Universal, Typological, Local-Semantic and Specific Linguistic Factors of Historical Replacement of Language Sign Units. In: 8th International Conference on the Languages of the Far East, South-East Asia and West Africa (Moscow, September 22–24, 2009). Theses and Presentations, pp. 125–141. Moscow (2009) (in Russian).
14. Polikarpov, A.A.: On Linguistic Metabolism. In: Festschrift on the Occasion of Jadranka Gvozdanovich's Jubilee. (2013) (in Russian) – in press.
15. Poddubny, V.V., Polikarpov, A.A.: Dissipative Stochastic Dynamic Model of the Development of Language Signs. Computer Research and Modeling, Vol. 3, No. 2, pp. 103–124 (2011) (in Russian), http://www.philol.msu.ru/lex/pdfs/podd-pol_crm_2011_2.pdf
16. Polikarpov, A.A.: Polysemy: System-Quantitative Aspects. In: Acta et Commentationes Universitatis Tartuensis, No. 774, pp. 135–154. Tartu University Press, Tartu (1987) (in Russian).

Shaping the history of words

Matilde Trevisani[1], Arjuna Tuzzi[2]

[1]University of Trieste, Italy
matilde.trevisani@econ.units.it
[2]University of Padua, Italy
arjuna.tuzzi@unipd.it

Abstract. In textual analysis, many corpora include texts in chronological order and in many cases this temporal connotation is crucial to an understanding of their inner structure. In a typical bag-of-words approach, data are organized in contingency tables, the rows reporting the frequency of each word over time-points (shown in columns). These discrete data (temporal patterns for frequencies) may be viewed as continuous objects represented by functional relationships. This study aimed at identifying a specific sequential pattern for each word as a functional object and at grouping these word patterns in clusters. A model-based clustering procedure is proposed, with specific reference to a corpus of end-of-year messages delivered by the ten Presidents of the Italian Republic covering the period from 1949 to 2011.

Keywords: chronological corpora, curve clustering, functional data analysis, functional clustering mixed model, wavelets

1 Chronological corpora

In many applications, the temporal evolution of words, topics and concepts in terms of their occurrence over time is important in order to highlight the distinctive features of texts in various periods of time. In textual analysis applications, many corpora include texts in chronological order, and this temporal connotation is crucial to an understanding of the inner structure of the corpus. Common examples are messages delivered by political and institutional representatives in different years, articles retrieved from newspaper archives, literary works by authors during their active lives, essays by students at different steps of their educational experience, etc. The temporal evolution of a word may be expressed by the sequence of its frequencies over time (tab. 1). In a typical bag-of-words approach, data are organized in word-type time-point contingency tables, which show the occurrence of each word-type at each time-point. It is therefore important to identify the specific

Table 1. Data taken from corpus of end-of-year messages of Italian Presidents (1949-2011)

Lemma_CAT[1]	**1949**	**1950**	**1951**	···	**2009**	**2010**	**2011**
di_PREP	15	14	12	...	134	143	122
il_ART	12	13	17	...	130	82	104
e_CONJ	8	4	5	...	125	101	75
essere_V	12	3	3	...	61	46	46
a_PREP	3	6	7	...	68	57	58
in_PREP	7	9	6	...	63	65	53
uno_ART	2	2	1	...	39	41	31
che_PRON	4	3	5	...	46	34	45
avere_V	2	0	1	...	28	20	14
anno_N	3	2	3	...	6	5	12
potere_V	2	1	1	...	18	13	10
dovere_V	0	0	1	...	8	10	7
fare_V	0	0	0	...	7	8	14
Italia_NM	0	1	1	...	9	15	15
popolo_N	1	1	1	...	0	0	0
io_PRON	0	0	0	...	0	0	0
paese_N	0	2	0	...	12	6	6
pace_N	1	1	0	...	4	1	2
italiano_N	1	1	2	...	5	3	5
mondo_N	0	1	0	...	3	4	2
grande_A	0	0	0	...	6	3	3
voi_PRON	0	0	1	...	2	3	4
nuovo_ADJ	1	1	3	...	2	5	5
augurio_N	1	0	0	...	3	3	3
cittadino_N	0	0	1	...	3	1	2
italiano_ADJ	1	0	0	...	3	2	1
libertá_N	0	1	1	...	0	3	1
vita_N	0	0	0	...	4	4	3
giovane_N	0	0	0	...	3	1	1

sequential pattern of a word-type frequency, and also prototype patterns suitable for clustering word-types portraying similar evolutions.

The aim of the study is twofold: (1) representing each sequence of frequencies as a functional object [14], *i.e.*, as a curve; (2) clustering these curves in order to group words with similar temporal patterns. We also discuss how these aims can be achieved, and highlight some critical points in representing the temporal pattern of words by their frequencies.

[1]CAT = grammar category, A = adverb, ADJ = adjective, ART = article, CONJ = conjunction, N = noun, NM = proper noun, PREP = preposition, PRON = pronoun, V = verb.

In any analysis of time-dependent phenomena (fig. 1a) we expect (or hope) to find easy-to-read patterns such as simple increasing (fig. 1b), decreasing (fig. 1c), meteor (fig. 1d) or constant trends (fig. 1e). Of course, peak-like data (fig. 1f) are less inspiring. In real-life data, most words show irregular peak-like temporal patterns (fig. 2).

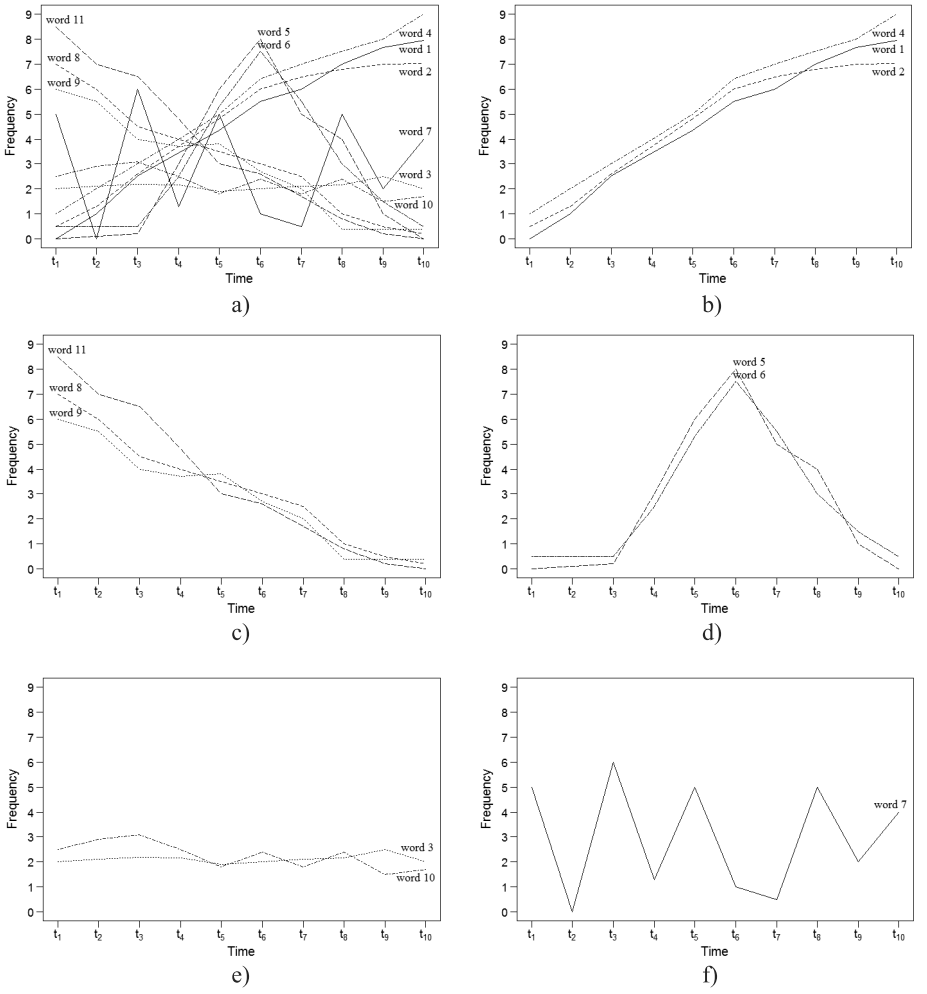


Fig. 1. Examples of temporal patterns

Many approaches aim at identifying the importance of the temporal variable, *e.g.*, cluster analysis and correspondence analysis, but they can only recognize a specific temporal evolution ex-post (*i.e.*, as part of the interpretation of results), because temporal dependency is not part of the method, and most of them can only identify simple shapes in the temporal evolution (*e.g.*, when the data show monotonic increasing or decreasing trends).

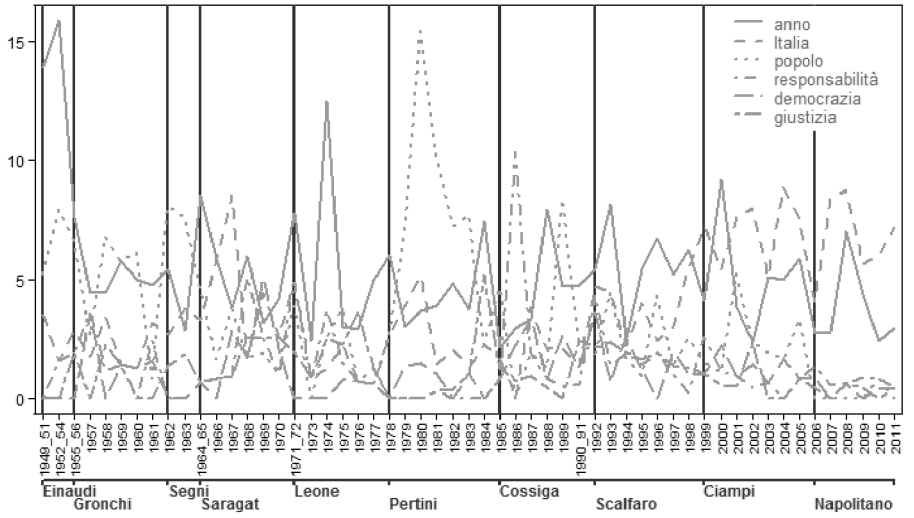


Fig. 2. Temporal pattern of six words taken from corpus of end-of-year messages

2 The corpus of this study

The chronological corpus examined in this study includes 63 end-of-year messages delivered by ten Presidents of the Italian Republic (Luigi Einaudi, Giovanni Gronchi, Antonio Segni, Giuseppe Saragat, Giovanni Leone, Sandro Pertini, Francesco Cossiga, Oscar Luigi Scalfaro, Carlo Azeglio Ciampi and Giorgio Napolitano) over the period from 1949 to 2011. This tradition of giving end-of-year speeches originated in 1949, when Einaudi's first message was broadcast on the radio during the second year of his presidential office. Since then, all Presidents have delivered end-of-year messages and even Cossiga, when he decided to resign, appeared on television and delivered a very short message.

The whole corpus includes 104,152 word-tokens and 10,583 word-types. In terms of word-tokens, the size of the messages (fig. 3) shows an increasing trend over time. Starting with the approximately 200 words of Einaudi's speeches, there was a gradual tendency toward greater length, until the more than 3,500 words of Scalfaro. This was followed by a moderate return to sobriety by Ciampi and Napolitano. Size analysis also shows that the first speech in a President's mandate is generally quite short, and the longest ones are among the last.

To overcome some of the limitations of an analysis based on simple form-types, we chose an analysis based on lemmas. In this study, the lemmatization process associated each form-type with a pair which included a lemma-type and a grammatical category, and involved manual expert correction of preliminary automatic screening of texts. The number of lemma types (6,353) is smaller than the number of word-types.

For these first analyses, we decided to aggregate consecutive messages when their size did not exceed 600 word-tokens (Einaudi 1949 1950 1951, 1952 1953 1954;

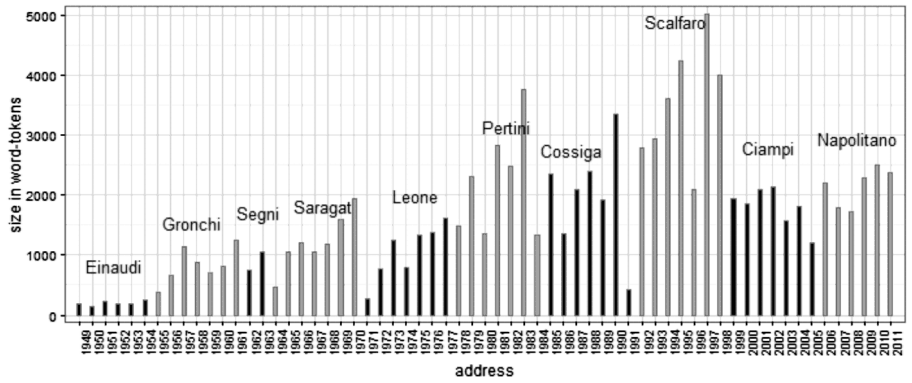


Fig. 3. Size in word-tokens of end-of-year messages

Gronchi 1955 1956; Saragat 1964 1965; Leone 1971 1972; Cossiga 1990 1991). We then selected the first 535 high-frequency nouns (frequencies equal to or higher than 10) and worked with normalized frequencies ($\times 1,000$ tokens).

3 Curve clustering

In functional data analysis, interest focuses on a set of curves, shapes, objects or, more generally, a set of functional observations. In this study the temporal pattern of each selected noun represents a functional observation, then we have a set of 535 peak-like curves to cluster (fig. 4). This is difficult, because we are working with a large number of words, characterized by a relatively low number of observed time occurrences and showing irregular behavior. We need a method capable of dealing with irregular curves (peak-like data), the presence of inter-individual (inter-word) variability, and high dimensional curve clustering. This environment poses many challenges from both statistical and computational perspectives.

In this study, textual data were processed by means of Taltac2 dedicated software [2], [3] and statistical analyses were carried out with R [12]. Developed by a research team from the University of Rome "La Sapienza", Taltac2 is an Italian software program which targets statistical and linguistic resources for computer-assisted statistical analysis of textual data. It consists of a series of tools for studying any kind of corpus, following the logic of both text analysis and text mining. R is a well-known language and environment for statistical computing and graphics, and is available as free software under the terms of the Free Software Foundation's GNU General Public License in source code form.

The main reference for this study is an article by Giacofci et al. [8] who made their procedure available through the R package curvclust. The authors developed these methods for studying signals of mass spectrometry and comparative genomic hybridization data.

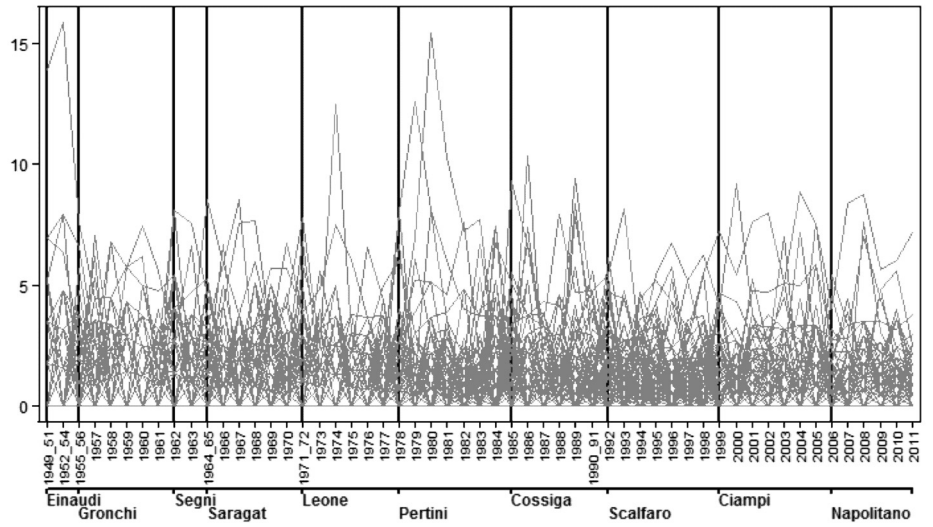


Fig. 4. Temporal patterns of high-frequency 535 nouns

3.1 Wavelet-based clustering for mixed-effects functional models

The temporal evolution of each word i $(i = 1,\ldots,n)$ is represented by a curve $y_i(t)$ observed on M equally spaced time-points $(t = t_1,\ldots,t_M)$. We assume that the model which generated the data involves a functional effect $\mu_i(t)$ and a random measurement error term $E_i(t)$ for each curve $y_i(t)$:

$$\begin{aligned} y_i(t) &= \mu_i(t) + E_i(t) \\ E_i(t) &= N(0,\sigma_E^2). \end{aligned} \tag{1}$$

As we suppose the existence of clusters, each word functional effect should include a principal functional fixed effect $\mu_l(t)$, $l = 1,\ldots,L$ where L is the (unknown) number of clusters. Moreover, in order to handle inter-word variability, $\mu_i(t)$ should include an individual functional random effect $U_i(t)$ as well. Then, model (1) becomes a mixed model:

$$\begin{aligned} y_i(t) &= \mu_l(t) + U_i(t) + E_i(t) \\ E_i(t) = N(0,\sigma_E^2), &\quad U_i(t) = N(0,K_l(s,t)) \end{aligned} \tag{2}$$

where $U_i(t)$ is modeled as a centered Gaussian process independent from $E_i(t)$.

The intrinsic infinite dimension of functions complicates statistical analysis of functional data. Hence, once defined in the functional domain, a classical approach is to convert the original problem into a finite-dimensional one by means of a functional basis representation of the model. Following Giacofci et al. [8] we used a wavelet-based representation of the model (2) and the Discrete Wavelet Transform (DWT) to consider continuous functions on the sole set of M sampled points.

Wavelet representation is based on a father wavelet (or scaling) ϕ and a mother wavelet (or simply wavelet) ψ. Curve $y_i(t)$ has the following decomposition:

$$y_i(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{i,j_0k}\phi_{j_0k}(t) + \sum_{j\geqslant j_0}\sum_{k=0}^{2^j-1} d^*_{i,jk}\psi_{jk}(t)$$

and for the full mixed-effects model (2) the DWT coefficients $\vec{c}_i = [c_{i,j_0k}]_{(k)}$ and $\vec{d}_i = [d_{i,jk}]_{(jk)}$ are:

$$\vec{c}_i = \alpha_i + \nu_i + \varepsilon_i$$
$$\vec{d}_i = \beta_i + \theta_i + \varepsilon_i$$

where the pair (α_l, β_l) represents the scaling and wavelet coefficients of fixed effects $\mu_l(t)$ and (ν_i, θ_i) the scaling and wavelet random coefficients of random effects $U_i(t)$. The number of coefficients depends on the number of clusters and the presence of random effects (number of words in question).

The R package curvclust is dedicated to model-based curve clustering. A set of different models can be estimated (tab. 2). For mixed models (FMM and FCMM) we also have several ways of modeling the variance of random effects: constant and scale-location variance for FMM; constant, scale-location, group-dependent and group-scale-location variance for FCMM. The R package curvclust includes methods for estimating coefficients from the maximum likelihood perspective based on EM-algorithm and for selecting the best model using the framework of penalized likelihoods, *i.e.*, Bayesian Information Criterion (BIC) and Integrated Classification Likelihood criterion (ICL).

Table 2. R package curvclust: list of available models

		group-specific fixed effect	random effect	model
CM	Constant Model	no	no	$y_i(t) = \mu_0(t) + E_i(t)$
FCM	Functional Clustering Model	yes	no	$y_i(t) = \mu_l(t) + E_i(t)$
FMM	Functional Mixed Model	no	yes	$y_i(t) = \mu_0(t) + U_i(t) + E_i(t)$
FCMM	Funct. Clust. Mixed Model	yes	yes	$y_i(t) = \mu_l(t) + U_i(t) + E_i(t)$

4 Preliminary results and discussion

As we do not know a priori which model is the most suitable for our data and how many clusters we need, we have to test and choose among four basic models (tab. 2) and, for models including random effects, among six variants; then we have to decide the number of clusters, comparing all combinations by BIC (fig. 5) and ICL measures. Once we have selected the model and number of clusters, we can estimate all the coefficients and reconstruct the estimated curves for each word.

At present, the FCMM (clustering mixed model) with 15 clusters and constant variance for functional random effects shows the best performance under statistical and

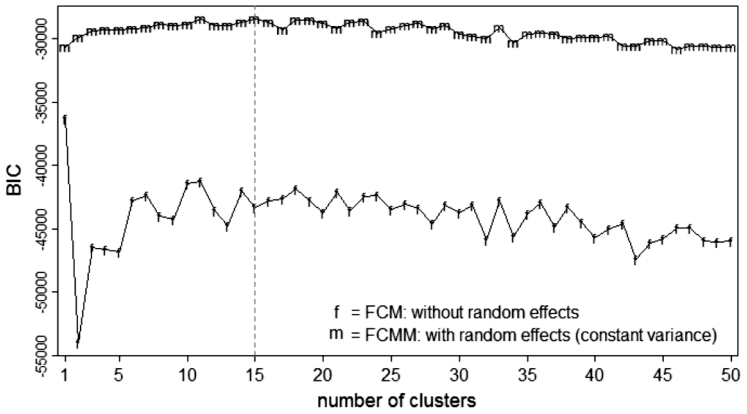


Fig. 5. Comparing FCMs and FCMMs with different number of clusters by BIC

subject-matter considerations [9]. This means that the selected 535 nouns have a cluster structure (group-specific functional fixed effects are important) and show inter-word variability (functional random effects are also important).

For every cluster (out of a total of 15), it is interesting to note which and how many words belong to that cluster and what its temporal pattern is like (in the figures, the mean shape of the curves that belong to the cluster, that is, the group-specific principal fixed effect, is shown in bold). If we exclude some clusters with somewhat flat, undifferentiated trends, observing the most interesting clusters allows us to pass from quantitative analysis to qualitative reading of results.

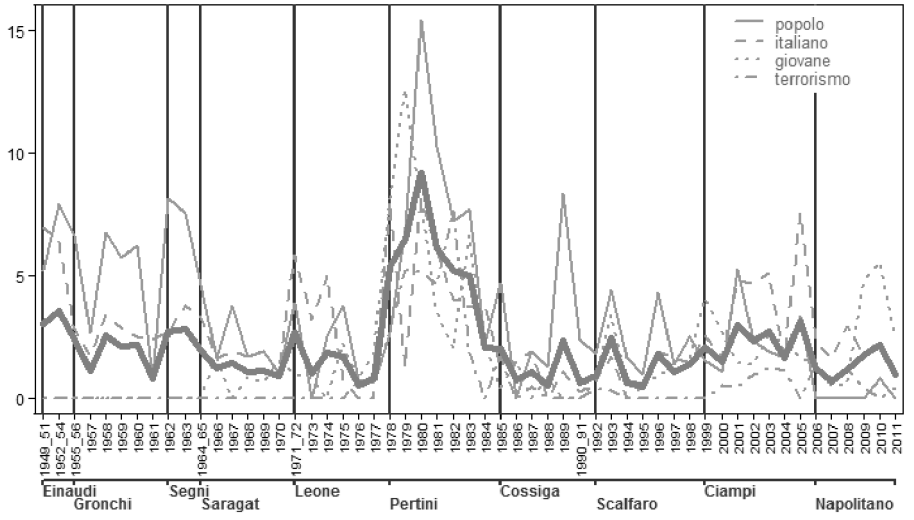


Fig. 6. Pertini's words (lighter lines) and principal fixed effect (bold line) in cluster 14

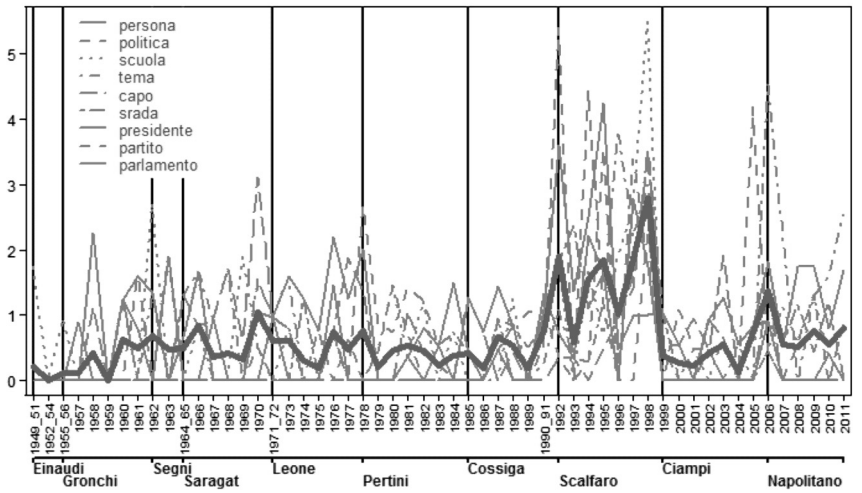


Fig. 7. Scalfaro's words (lighter lines) and principal fixed effect (bold line) in cluster 5

First of all, we note that many clusters are representative of a certain President, i.e., the ones which contain words showing time trends with a peak near the years of the beginning of that President's mandate. For example, one of the clusters (fig. 6) contains all the typical themes dear to President Pertini (popolo italiano [the Italian people], giovane [young people], terrorismo [terrorism]) and another (fig. 7) President Scalfaro's favorites (politica [politics], persona [the individual], capo dello Stato [head of state], strada [way], tema [topic], scuola [education], presidente [president], partito [party], parlamento [parliament]).

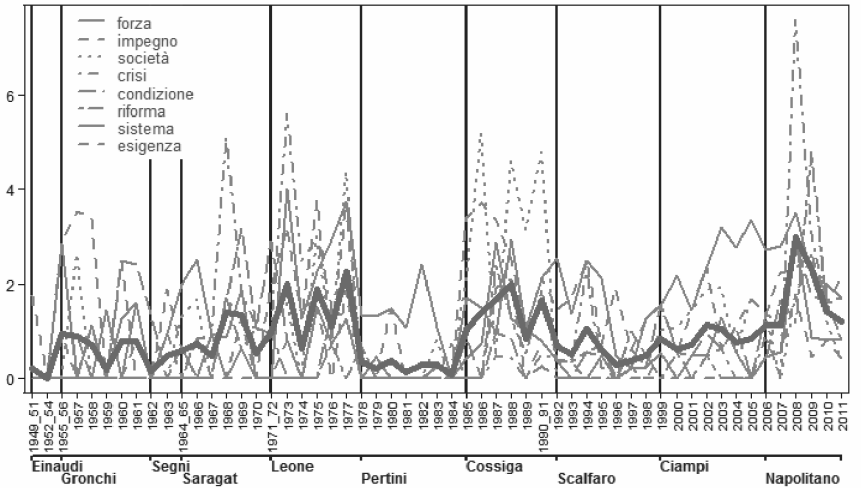


Fig. 8. The crisis words (lighter lines) and principal fixed effect (bold line) in cluster 11

Other clusters also show trends associated with events and history – like that shown in fig. 8, expressing a state of crisis (forza [strength], impegno [commitment], società [society], crisi [crisis], condizione [conditions], riforma [reform], sistema [system], esigenza [requirements]) and showing an increase between the late 1960s and throughout the 1970s (a period of economic crisis and social agitation), between the late 1980s and the early 1990s (a period of political and institutional crisis and the period of enquiries into corruption on the part of politicians called "Mani pulite" [Clean fingers]), and in recent years (worldwide economic crisis). Cluster 9 represents a set of words which have fallen into disuse, appearing in the speeches of the earlier Presidents and gradually disappearing from them (Dio [God], fiducia [trust], solidarietá [solidarity], progresso [progress], spirito [spirit]).

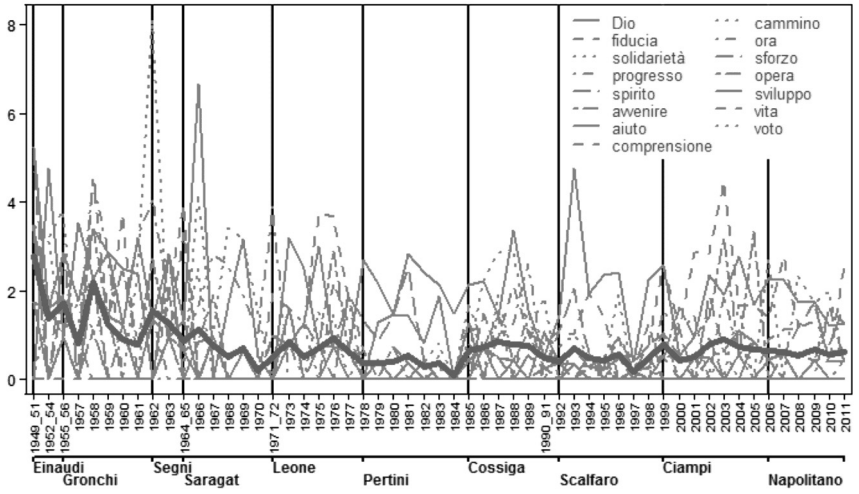


Fig. 9. Old-fashion words (lighter lines) and principal fixed effect (bold line) in cluster 9

5 Open issues and future work

Functional data analysis is a growing area in statistics and may be useful in this case, although at present our analysis should be considered purely explorative. We tried other methods for curve clustering but they did not work, whereas the functional model-based clustering approach proved promising. However, we still cannot justify the use of such a complex, extremely sophisticated model sustained by linguistic theory, and this is clearly one weakness of our approach. As the functional mixed-model is flexible, one of our research aims was to simplify it, to avoid overelaboration, and find a theoretical justification.

The choice to use wavelets is debatable, mainly because splines are preferred in many other contexts (*e.g.*, [15]). However, splines are not appropriate when dealing with high-dimensional data and for modeling irregular curves such as peak-like data.

During the pre-processing phase, it was also necessary to make decisions which may be viewed as debatable. To reduce the effect of the size of the Presidential messages, we decided to adopt a standardized version of frequencies (number of occurrences of nouns per 1,000 tokens) and to aggregate short messages, even though this type of normalization is known to be insufficient to purge frequencies of the size effect, and other methods may be taken into consideration (for example, basing calculations on the concept of politextuality, *i.e.*, number of texts including the word).

One important question underlying this work is whether the corpus analyzed is or is not a chronological corpus. Perhaps it is not chronological in the sense we expected, because many clusters show that the President effect is the main factor shaping the cluster curves and the pure time effect (on higher temporal scales) is not as important as Presidents' individual choices. The institutional nature of this text genre imposes greater constraints, but previous studies ([4], [11], [16]) have already shown that personal traits and individual choices are important, and what a President decides to say (or not to say) is a personal choice and remains unpredictable. If the corpus including the end-of-year messages is not a genuine example of a chronological corpus for testing functional model-based clustering, perhaps we should test the same method on a more intrinsic chronological corpus.

In revealing the effect of the Presidential Office, the model highlights some peculiar properties of these data which may be very interesting from a multi-scale perspective [5], because we have both a time-scale (years) and a President effect (a period of seven years or, sometimes, less). An interesting issue for future research work in this field consists of differentiating lower-scale patterns from higher-scale ones (and vice versa) in order to reveal the importance of a possible President factor and to use it to improve estimation procedures.

References

1. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE PAMI 22(7), 719–725 (2000)
2. Bolasco, S., Baiocchi, F., Morrone, A.: Taltac2: Trattamento Automatico Lessicale e Testuale per l'Analisi del Contenuto di un Corpus (ver.: 2.10), http://www.taltac.it
3. Bolasco, S., Morrone, A., Baiocchi, F.: A Paradigmatic Path for Statistical Content Analysis Using an Integrated Package of Textual Data Treatment. In: Vichi M., Opitz, O. (eds.) Classification and Data Analysis. Theory and Application, pp. 237—246. Springer, Heidelberg (1999)
4. Cortelazzo, M.A., Tuzzi, A. (eds): Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica. Marsilio, Venice (2007)
5. Ferreira, M.A.R., Lee, H.K.H.: Multiscale Modeling: A Bayesian Perspective. Springer, New York (2007)
6. Gareth, M.J., Sugar, C.A.: Clustering for Sparsely Sampled Functional Data. J. Am. Stat. Ass. 98, 397–408 (2003)
7. Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: curvclust. R package, http://cran.r-project.org/web/packages/curvclust/index.html

8. Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: Wavelet-based clustering for mixed-effects functional models in high dimension. Biometrics 69(1), 31-40 (2013)
9. Hennig, C.: A Constructivist View of the Statistical Quantification of Evidence. Constructivist Foundations 5(1), 39–54 (2009)
10. Morris, J.S., Carroll, R.J.: Wavelet-based functional mixed models. J. Roy. Stat. Soc. B Met. 199, 68–179 (2006)
11. Pauli, F., Tuzzi, A.: The End of Year Addresses of the Presidents of the Italian Republic (1948-2006): discoursal similarities and differences. Glottometrics 18, 40–51 (2009)
12. R development core team, R: a language and environment for statistical computing (ver.2.13), http://www.r-project.org
13. Ramsay, J.O., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB. Springer, New-York (2009)
14. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Second edition. Springer, New-York (2005)
15. Reithinger, F., Jank, W., Tutz, G., Shmuell, G.: Modelling price paths in on-line auctions: smoothing sparse and unevenly sampled curves by using semiparametric mixed models. Appl. Statist. 57, 127–148 (2008)
16. Tuzzi, A., Popescu, I.-I., Altman, G.: Quantitative analysis of Italian texts. RAM, Lüdenscheid (2010)

Quantitative Analysis of Text Structure Using Co-occurrence of Words

Makoto Yamazaki

National Institute for Japanese Language and Linguistics, Department of Corpus Studies
10-2, Midori-cho, Tachikawa City, Tokyo, Japan
yamazaki@ninjal.ac.jp

Abstract. This paper investigates the characteristics and structure of text through a quantitative analysis of cohesion. We use the co-occurrence rate – the number of identical words in adjacent paragraphs – to calculate the degree of cohesion. Results show that laws, white papers, and minutes from the National Diet exhibit high co-occurrence rates, whereas newspapers, bestselling books, and magazines exhibit low co-occurrence rates. We also suggest that changes in the co-occurrence rate within a text could be used to study the segmentation of that text.

Keywords: cohesion, text structure, co-occurrence rate, similarity, Japanese corpus

1 Introduction

This paper is intended as an investigation of the analysis of text using quantitative lexicology. Conventional research on vocabulary has focused on non-contextual analysis, because it has dealt with fragments of text. Such analyses have not used the concept of a time axis, i.e., the flow of a text from beginning to end. Only a few attempts such as Yasue [8], Yamazaki [7], and Youmans [9] [10] have tried to capture the structure of text using quantitative lexicology. With the help of a large corpus, we will illustrate the relation between quantitative characteristics of text and the structure of text.

2 Cohesion in Text

Cohesion is an essential property in assembling text as a unified body of writing. Halliday and Hasan [1] writing in the first detailed study of cohesion, describe it as follows:

Cohesion occurs where the ”interpretation” of some element in the discourse is dependent on that of another. The one ”presupposes” the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into a text. (Halliday and Hasan, [1]:4)

There is a similar term called ”coherence,” which is often confused with cohesion. In this paper, we assume that cohesion is a subordinate concept of coherence as Iori ([4]:12) describes. He subordinates cohesion to coherence as an association based on deduction.

Cohesion comprises grammatical and lexical cohesion, with ”reference,” ”substitution,” ”ellipses,” and ”conjunctions” being means to achieve the former, and ”reiteration” and ”collocation” being means to achieve the latter.

For example, Hirabayashi ([2]:72) examined English compositions written by Japanese high school students (average 100 words, 10 sentences) and found 12.87 grammatical cohesions. The breakdown was 4.9 references, 6.32 conjunctions, 1 substitution, and 0 ellipses.

The four types of reiteration that achieve lexical cohesion are (a) the same word, (b) a synonym or near-synonym, (c) a superordinate, and (d) a general word. (Halliday and Hasan, [1]:279)

According to Károly ([5]:162), English texts use ”repetition of different words,” i.e., combinations of (b), (c), and (d), far more often than repetition of the same word (a). At present, it is difficult to determine synonyms and superordinates automatically, because there is no reliable data for synonyms. Even if there were, the results of any studies on this subject would differ according to how broadly they defined ”synonym.” Consequently, this study focuses on the repetition of the same word, but excludes functional words.The same word often is repeated in common text. Among the excerpts from 10,369 library books examined for this study, only 17 instances featured no repetition of the same word, likely because excerpts were brief (22 words or fewer). Repetition of the same word is more likely to occur in longer texts, as our examination shows. Analyses involving repetition of the same word have merits and demerits. Repetition of the same word is easier to discern than other cohesion phenomena. For example, repetitions can be easily observed because they occur more frequently. On the other hand, results depend upon a standard for determining a ”word,” and differ between short-unit words (used in this study) and long-unit words. Also, an analysis of words in repetitions observes only whether words are the same or different, shedding no light on semantic relationships.

3 Data

This study uses the DVD edition of the *Balanced Corpus of Contemporary Written Japanese*[1] (BCCWJ) released in December 2011. It includes an xml file of each

[1]BCCWJ is the first balanced corpus of written Japanese. It was compiled during 2006–2011 by the National Institute for Japanese Language and Linguistics.

individual text that comprises the corpus. This xml file tags morphological information such as word segmentation as well as text structure tags.

This paper focuses on portions of text tagged with the <paragraph> tag in the xml file and analyzes morphological information contained within these portions of text. Although sentences also are valid units for observing cohesion, the BCCWJ <sentence> tag includes text in headlines and captions. Since these would need to be separated from regular text, this study focuses on the <paragraph> tag, which specifically represents all body text. Table 1 shows the number of samples that include <paragraph> tags.

Table 1. Texts with paragraph tags

Registers[2]	**Total no. of text**	**No. of text with paragraph tags**
Publication Books (PB)	10,117	9,742
Magazines (PM)	1,996	1,767
Newspapers (PN)	1,473	1,457
Library Books (LB)	10,551	10,369
White Papers (OW)	1,500	1,496
School Textbooks (OT)	412	0
Public Relationship Reports (OP)	354	354
Bestselling Books (OB)	1,390	1,374
Q&A Bulletin Boards (OC)	91,445	0
Blogs (OY)	52,680	0
Verses (OV)	252	0
Laws (OL)	346	56
Minutes from the National Diet (OM)	159	159
Total	172,675	26,774

Four of the 13 registers – textbooks, Q&A bulletin boards, blogs, and poetry verses - have no text with <paragraph> tags because they do not use the <paragraph> tag. Accordingly, these four registers have been excluded from analysis. Although <paragraph> tags are checked by people, they are essentially assigned automatically. Therefore, figures in Table 1 may contain errors.

Table 2 shows the average number of words in each text, average number of paragraphs, average number of words per paragraph, and average number of different words in each paragraph. Minutes from the National Diet have the highest average number of words per paragraph due to how paragraphs are recognized in the minutes (each speech is transcribed as a single paragraph). The word count excludes auxiliary symbols, spaces, particles, and auxiliary verbs.

Fig. 1 is a boxplot showing the distribution of the number of paragraphs per text for each register. There is an overall rightward trend, that is, to the higher end (you can see if we rotate this figure 90° clockwise). While each distribution has individually distinguishing characteristics, Fig. 1 shows that only library books (LB) and publication books (PB) present similar patterns of distribution.

[2]We use the term ”registe” to represent each sub-corpus within BCCWJ.

Table 2. Numbers of tokens, paragraphs, and types

Registers	Average no. of tokens	Average no. of paragraphs	Average no. of tokens per paragraph	Average no. of types per paragraph
Publication Books (PB)	1,384.61	43.76	50.51	37.06
Magazines (PM)	891.17	29.81	40.05	33.27
Newspapers (PN)	334.33	9.28	38.78	33.33
Library Books (LB)	1,450.16	54.53	45.76	34.70
White Papers (OW)	1,793.10	29.32	64.74	44.33
Public Relationship Reports (OP)	2,903.53	103.14	28.14	23.39
Bestselling Books (OB)	1,404.46	69.30	29.52	24.28
Laws (OL)	219.50	6.93	24.04	15.03
Minutes from the National Diet (OM)	17,885.87	144.06	151.30	76.21

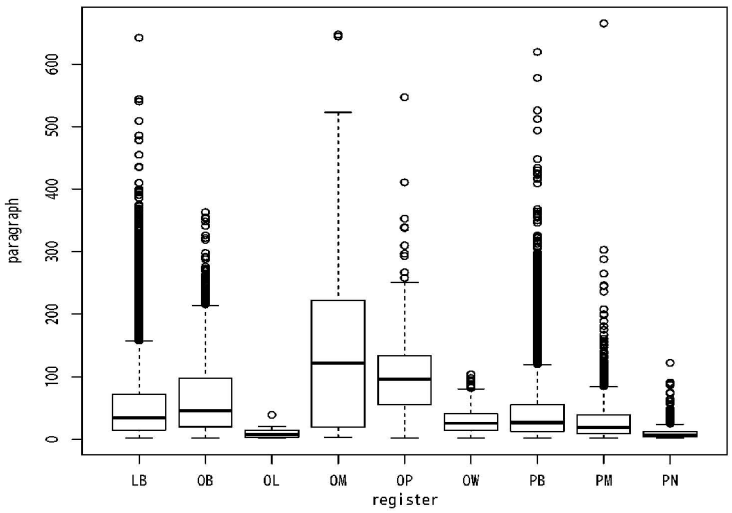


Fig. 1. Distribution of the number of paragraphs per text

4 Cohesion Calculation Method

Calculations include only repetitions of the same word within consecutive paragraphs, as repetitions occurring among widely separated paragraphs may happen by chance. Thus, the calculation was achieved by counting the number of words that appear in a given paragraph and in adjacent paragraphs.

Represented in a more concrete manner, the text would look as follows. Word A appears four times in four consecutive paragraphs. But we count only one co-occurrence between paragraph 1 and paragraph 2 since they are consecutive paragraphs.

Paragraph 1: ****A*****

Paragraph 2: **A**********

Paragraph 3: *******

Paragraph 4: **A***A****

Hoey [3] and Károly [5] calculate not only words in common word forms but even in synonyms. However, these studies do not deal with large volumes of data. Repetition of specific words is the most suitable method for automatic calculations of voluminous data.This study calculates the degree of cohesion - the *co-occurrence rate* – using the following formula:

$$C(a,b) = \frac{F(a,b)}{N_a}$$

a, b: paragraph number (1 to *n*)

$C(a,b)$: co-occurrence rate between paragraph *a* and paragraph *b*

$F(a,b)$: number of tokens in paragraph *a* that appear in paragraph *b*

N_a: number of tokens in paragraph *a*.

The co-occurrence rate is an index that uses Mizutani's [6] non-symmetric similarity. Therefore, there are two co-occurrence rate values within two consecutive paragraphs: the forward co-occurrence rate (FCR) and the backward co-occurrence rate (BCR). When $b = a+1$ in the formula above, the rate given is the FCR; when $b = a-1$, the rate given is the BCR. However, since the first paragraph in a text has no preceding paragraphs and the final paragraph has no following paragraphs, these paragraphs exhibit a co-occurrence rate of zero for the sake of convenience. One limitation arises in using this method to calculate the co-occurrence rate: the text must contain at least two paragraphs. Therefore, 340 one-paragraph texts were excluded from calculations used to derive Table 1. Also excluded from the calculation were auxiliary symbols and spaces, which are not recognized as linguistic expressions, and particles and auxiliary verbs not affecting textual cohesion.

5 Results

Table 3 shows the average number of occurrences per paragraph and average co-occurrence rates. The FCR and BCR are almost identical. We interpret this finding to mean that each register is connected by the same degree of cohesion. Examining results by register, we see that laws, white papers, and minutes from the National Diet have high co-occurrence rates, whereas newspapers, bestselling books, and magazines have low co-occurrence rates.

Table 4 shows the number of co-occurrences and co-occurrence rates by NDC (Japanese library classification) in the register of library books. Although the co-occurrence rate for the register as a whole is 0.19, "Literature" and "Uncategorized" have lower values than other categories. Since NDC data offer no details about the "Uncategorized" category, we cannot suggest reasons for this finding. However, we

Table 3. Number of Co-occurrences and Co-occurrence rate

Registers	Average no. of FC	Average of FCR	Average no. of BC	Average of BCR
Publication Books (PB)	12.98	0.22	12.74	0.22
Magazines (PM)	6.89	0.16	6.82	0.16
Newspapers (PN)	5.99	0.15	5.84	0.16
Library Books (LB)	10.49	0.19	10.36	0.19
White Paper (OW)	20.00	0.31	19.84	0.31
Public Relationship Reports (OP)	5.19	0.18	5.13	0.17
Bestselling Books (OB)	5.49	0.15	5.47	0.15
Laws (OL)	12.16	0.48	12.31	0.47
Minutes from the National Diet (OM)	40.45	0.30	39.01	0.30

FC: forward co-occurrences BC: backward co-occurrences
FCR: forward co-occurrence rate BCR: backward co-occurrence rate

can surmise that ”Literature” has a low co-occurrence rate because it contains many short paragraphs, such as dialogue. This may also explain the low co-occurrence rate for ”Bestselling Books,” many of which are novels, in Table 3. To validate this possibility, let us examine the correlation between average number of words per paragraph and the average co-occurrence rate. Fig. 2 shows there is a positive correlation with a high coefficient of determination of 0.8661.

Table 4. Co-occurrence and library classification

NDC(Japanese Library Classification)	Average no. of FC	Average of FCR	Average no. of BC	Average of BCR
0 General Works	12.97	0.22	12.95	0.22
1 Philosophy	17.55	0.25	17.73	0.24
2 General History	14.80	0.21	14.60	0.21
3 Social Sciences	15.02	0.24	14.84	0.24
4 Natural Sciences	14.32	0.24	13.96	0.24
5 Technology	10.72	0.22	10.56	0.21
6 Industry & Commerce	11.03	0.21	10.82	0.21
7 Arts	12.02	0.20	11.98	0.20
8 Language	10.40	0.21	10.17	0.20
9 Literature	5.07	0.12	4.97	0.12
Uncategorized	3.46	0.13	3.45	0.13

6 Changes in Co-occurrence Rate within a Text

What changes appear in the co-occurrence rate within a single text? Let us consider white papers. Fig. 3 shows the parsing of a text from OW1X_00000 (White Paper on the Japanese Economy, 1979).

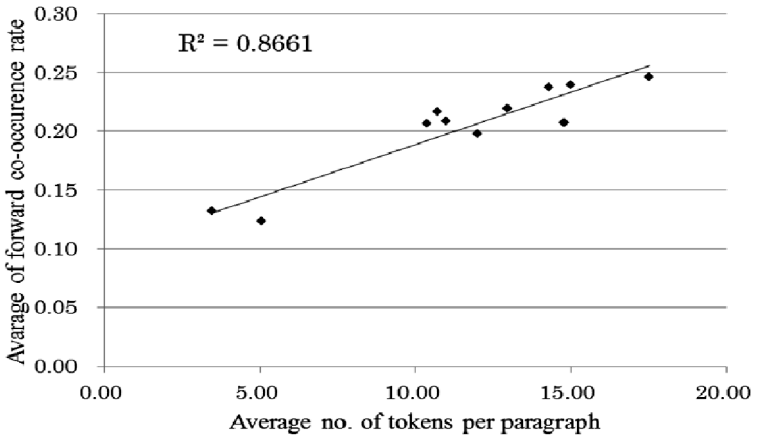


Fig. 2. Correlation between number of tokens per paragraph and average FCR

In Fig. 3 the horizontal axis represents the flow of the text by paragraph numbered from the beginning of the text. There are two bars for each paragraph number. The left represents the BCR and right the FCR.

Fig. 3 shows that 34 of 56 paragraphs have a high co-occurrence rate with a paragraph immediately following. The White Paper on the Japanese Economy has a high FCR around 60%. The three starred points in Fig. 3 indicate the beginning points of new major sections (paragraphs 2, 22, and 36). The nine points designated by downward arrows indicate subheadings within sections (paragraphs 6, 10, 19, 23, 27, 30, 37, 44, and 50). If we compare the FCR (right bar) and the BCR (left bar) in the segments designated by arrows, the FCR is higher eight times out of nine. In the ninth instance, the values are identical. This shows that the first paragraph of a new subject has a high cohesion with the next paragraph, presumably because it introduces a new topic.

Conversely, paragraphs before the arrows signify final summarizing paragraphs. If we check how FCRs and BCRs of these segments compare, we see that they have a higher BCR rate six times out of nine. In other words, they share more words with the preceding paragraph than with the following paragraph. Perhaps they function as a minor conclusion within a text.

Although only one example is presented here, it is possible that changes in the co-occurrence rate within a text could be used to automatically surmise the unit of text.

7 Conclusion and Future Issues

This study has used the co-occurrence rate to examine the degree of cohesion in texts. Results indicated high degrees of cohesion in laws, white papers, and minutes from the National Diet and low degrees of cohesion in newspapers, bestselling

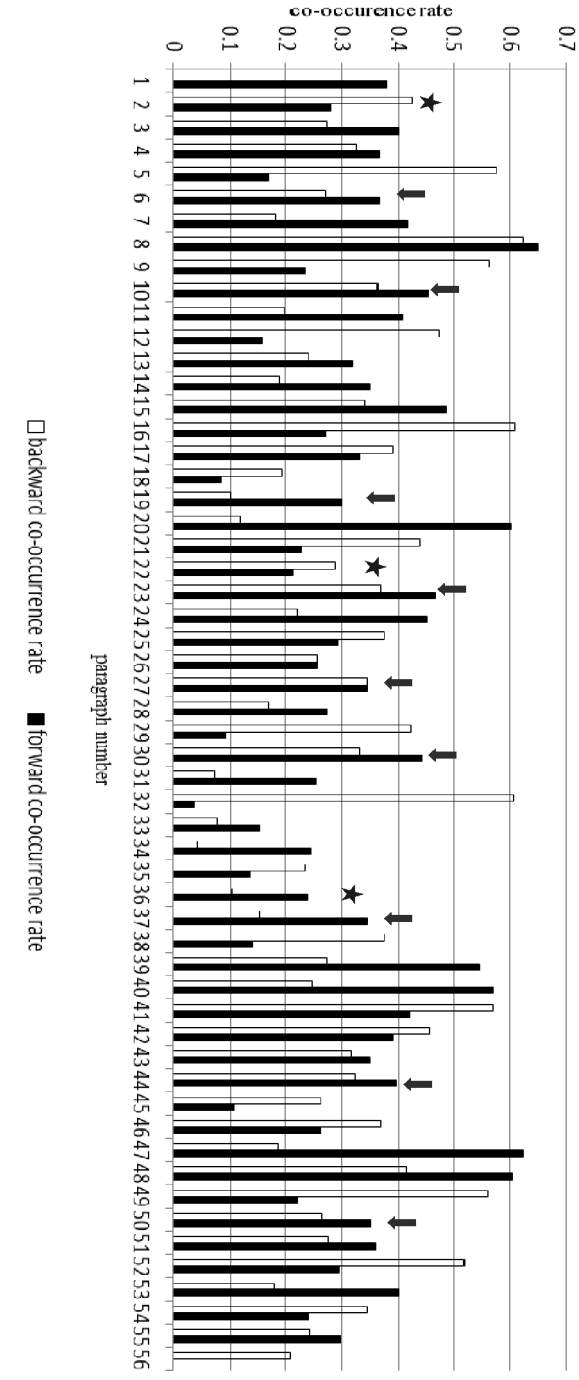


Fig. 3. Changes in Co-occurrence Rate within a Text

books, and magazines. Data by the NDC category showed that literature has a low degree of cohesion, presumably because the data recognizes dialogue as paragraphs. We also have shown that noting changes in the co-occurrence rate could be a way to examine the segmentation of texts.

Three issues merit future study as attempts to describe cohesion objectively.

(1) Examining problems with <paragraph> tags and checking for issues analyzing them.
(2) Exploring ways to measure cohesion using both paragraphs and sentences.
(3) Studying the correlation between methods for grammatical cohesion such as demonstratives and conjunctions.

Acknowledgments. This paper is one outcome of the collaborative research project "Distribution of Vocabulary and Sentence Structures in Texts" conducted from 2011 to 2012 at the National Institute for Japanese Language and Linguistics.

Texts included in the Book registry within the BCCWJ were compiled by MEXT KAKENHI Grant Number: 18061007

References

1. Halliday, M.A.K., and Hasan, R.: Cohesion in English. Longman, London (1976).
2. Hirabayashi, Kenji: Analysis of English Writing by Pre-intermediate Level Japanese Learners of English: An In-depth Study of "Cohesion" in Writing. Bulletin of Aichi Shinshiro Otani College. 2(4), pp. 67–76 (2003)
3. Hoey, Michael: Patterns of Lexis in Text. Oxford University Press, Oxford (1991)
4. Iori, Isao: Nihongo ni okeru tekisuto no kessokusei no kenkyuu. Kuroshio Publishers, Tokyo (2007)
5. Károly, Krisztina: Lexical Repetition in Text. Peter Lang, Frankfurt am Main (2002)
6. Mizutani, Shizuo: Classification of Popular Songs by Lexical Similarity: "Elegy of a Hotspring Town," "Lilu Returned from Shanghai," and Others. Mathematical Linguistics, 12(4), pp. 145–161 (1980)
7. Yamazaki, Makoto: A New Index for Topical Change in the Context. Mathematical Linguistics, 13(8), pp. 346–360 (1982)
8. Yasue, Sawako: Koubun wo otta kotonari gosuu no ugoki. Tokyo Joshi Daigaku Nihon Bungaku, 56, pp. 32–45 (1981)
9. Youmans, Gilbert: A new tool for discourse analysis: the vocabulary management profile. Language, 67(4), pp. 763–789 (1991)
10. Youmans, Gilbert: The vocabulary management-profile, two stories by William Faulkner. Empirical Studies of the Arts, 12(2), pp. 113–130 (1994)

PART III

APPLICATION-ORIENTED RESEARCH

The Generalized Poisson Distributions as Models of Word Length Frequencies

Gordana Đuraš[1], Ernst Stadlober[2], Emmerich Kelih[3]

[1]POLICIES – Statistical Applications, JOANNEUM RESEARCH, Graz, Austria
gordana.djuras@joanneum.at

[2]Institute of Statistics, Graz University of Technology, Austria
e.stadlober@tugraz.at

[3]Institute for Slavic Studies, University of Vienna, Austria
emmerich.kelih@univie.ac.at

Abstract. This study focuses on the analysis of the distribution of word length in Slovenian and Russian texts of different types. Here, word length is measured by the number of syllables per word. Zero-syllable words have not been taken into account in the analysis. The statistical investigation is based on 120 Slovenian and 120 Russian texts. The index of dispersion $d = \frac{s^2}{\bar{x}-1}$ is used as a classifier of texts as follows. In the case of $d < 1$ there is under-dispersion, for $d = 1$ equi-dispersion (standard Poisson) and for $d > 1$ overdispersion of word length distributions. As an appropriate alternative for modeling such kinds of count data we suggest two-parametric generalizations of the Poisson distribution which allow us to model both over- and under-dispersion. Different estimation procedures suggested in the literature, such as the method of moments, the maximum likelihood method and an approach based on the sample mean and on the first frequency class, are compared by applying them to our Slovenian and Russian texts under study. In addition, the performance of the parameter estimators is investigated by a simulation study.

Keywords: Count data, Poisson under/over-dispersion, Parameter estimation techniques

1 Introduction

An essential question when modeling word length frequencies is how to choose an appropriate probability model to describe the observed values. The final choice depends, however, on the interaction of diverse *extra-textual* factors, such as the concrete language under study, individual authorship or text type, influencing both

word length and word length frequencies. Also *theory-driven* factors, in particular the definition of the word and the choice of a measuring unit for its length have to be defined in advance [1]. Taking into account all these boundary conditions allows us to reduce the possible set of distributions. Here, word length is measured by the number of syllables where zero-syllable words, typical for Slavic languages, are considered to be a part of the subsequent word and are excluded from the analysis as such.

The simplest and the most widely used distribution for analyzing count data is the Poisson distribution. The equality of mean and variance, known as *equi-dispersion*, is an important property of the Poisson distribution. But this requirement is sometimes too restrictive for count data. In many practical situations empirical data sets exhibit departures from equi-dispersion which can be either *over-dispersion* (the variance of the count variable exceeds its mean) or *under-dispersion* (the variance is smaller than the mean) with respect to the Poisson model and thus cannot be considered to be fitted by the Poisson distribution. To find out if the count data come from the Poisson distribution we can apply the index of dispersion of a count variable X being defined as the variance to mean ratio [10]. Due to the fact that the texts under study contain no zero-syllable words, 1-displaced versions of the proposed models became relevant for our further research. For this reason, we propose to use the following index of dispersion $\delta = \mathrm{var}(X)/(E(X)-1)$ and estimate it by its empirical value $d = s^2/(\bar{x}-1)$. The 1-displaced Poisson distribution has $\delta = 1$, so it provides an adequate fit only for empirical samples with $d \approx 1$, i.e. for count data where the sample mean diminished by one is near to the sample variance. For this reason, δ can be used as a measure for detecting departures from the 1-displaced Poisson model.

In order to find a general model of word length frequency distributions we construct two-parametric generalizations of the Poisson distribution that cover the whole δ range and are applicable to all dispersion situations.

2 Data Base of the Study

The 120 Slovenian and the 120 Russian texts, serving as data basis for this study include four different text types, namely *journalism*, *poems*, *private letters* and *prose*, thirty texts of each text type being analyzed. These texts were chosen systematically in order to generate a balanced experimental design with an equal number of observations for each text group. The particular selection of the four text types aims to cover the broad textual spectrum and is based on findings from recent word length studies [1, 6]. Table 1 displays the characteristic statistical measures of our text sample: mean word length $(\bar{x})$, sample variance (s^2), text length (TL) and index of dispersion (d).

A closer look at the distributions of word length in Slovenian and Russian texts shows that texts within one single language, but also between languages, differ significantly with regard to the above statistical measures. Figure 1 visualizes the differences in the mean word length between the two languages and the four text types given. Obviously, the average word lengths in Russian texts are significantly longer

Table 1. Statistical measures of Slovenian and Russian texts under study

			$\bar{x}$		s^2		TL		d	
	Text Type	N	min	max	min	max	min	max	min	max
SLO	Journalism	30	2.05	2.46	1.22	1.96	328	1166	1.09	1.35
	Poems	30	1.48	1.90	0.37	0.84	58	626	0.60	1.14
	Private Letters	30	1.72	1.98	0.78	0.98	401	1979	0.97	1.15
	Prose	30	1.73	1.98	0.70	1.04	288	4401	0.95	1.16
RUS	Journalism	30	2.40	2.83	1.46	2.17	320	901	1.03	1.34
	Poems	30	1.76	2.40	0.73	1.60	77	1014	0.76	1.41
	Private Letters	30	1.83	2.52	0.90	2.11	48	488	0.79	1.53
	Prose	30	2.02	2.52	1.15	1.83	236	3154	0.93	1.24

than those in Slovenian texts (Wilcoxon rank-sum test with $p < 0.01$). The linguistic explanation for the apparent differences is that Russian compared to Slovenian has a more complex syllable structure, and tends to have longer word forms on the morphological level, in particular compound words. For further cross linguistic studies of word length in Slavic languages see [7, 8].

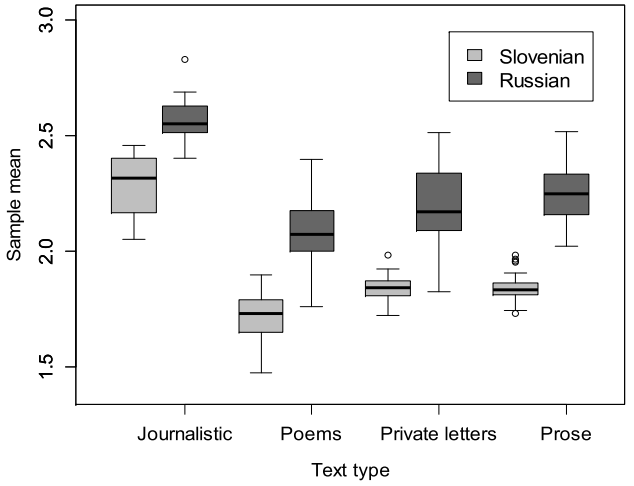


Fig. 1. Differences in word length between languages for different text types

3 Construction of Distributions

In this section we discuss two different approaches for constructing two-parametric generalizations of the Poisson distribution applicable to all three dispersion situations: over-, equi- and under-dispersion.

3.1 The Generalized Poisson Model

Consider the random summation of N mutually independent Borel random variables Y_i with the common probability mass function (pmf) defined by [2, p. 158]

$$P(Y=y)=\frac{(\lambda y)^{y-1}e^{-\lambda y}}{y!},\quad y=1,2,\ldots \tag{1}$$

and zero otherwise, where $0<\lambda<1$. Assuming the number of components N to be a Poisson distributed random variable independent of each Y_i, we obtain the generalized Poisson (GP) distribution as a result. Additionally to the Poisson parameter θ we also have the parameter λ, which when negative causes some kind of truncation. The pmf corresponding to the 1-displaced distribution is defined only over positive integers and given by the formula

$$\pi_{x|\lambda,\theta}=P(X=x)=\begin{cases}\dfrac{\theta(\theta+x\lambda-\lambda)^{x-2}e^{-(\theta+x\lambda-\lambda)}}{(x-1)!}, & x=1,\ldots m+1\\ 0, & x>(m+1),\ \text{for } \lambda<0\end{cases} \tag{2}$$

where $\theta>0$, $\max(-1,-\theta/m)\leqslant\lambda<1$ and m is the largest positive integer such that $\theta+m\lambda>0$ when λ is negative. Additionally, the condition $m\geqslant 4$ is proposed in order to ensure that there are at least five non-zero probability classes in the truncated model when $\lambda<0$ [2, p. 4]. Notice that in the case when $0\leqslant\lambda<1$ the support of the above model does not have to be truncated, hence we have $m=\infty$.

The mean and the variance of the distribution (2) are $E(X)=\theta/(1-\lambda)+1$ and $\mathrm{var}(X)=\theta/(1-\lambda)^3$, hence the index of dispersion is $\delta=1/(1-\lambda)^2$. The parameter λ provides information about the type of the distribution, whereas the parameter θ indicates the intensity of the Poisson process. Obviously, for $\lambda=0$ the GP model (2) simplifies to the common Poisson model. For $0<\lambda<1$ we have $\delta>1$ (over-dispersion), i.e. the model allows for modeling counts where $s^2>\bar{x}-1$ is fulfilled. When $\lambda<0$ we have $\delta<1$ (under-dispersion), hence enabling us to describe $s^2<\bar{x}-1$ data cases.

Furthermore, the GP distribution, suitable for over- and under-dispersion, approximates both the negative binomial (NB) and the binomial (B) model. Since the comparison of any two distributions is reasonable only if some common characteristics are fixed, and all distributions investigated are two-parametric, we fix the first two moments or explicitly mean and index of dispersion and express the 1-displaced model parameters in terms of μ and δ as follows:

- GP(λ,θ): $\mu=\dfrac{\theta}{1-\lambda}+1$, $\sigma^2=\dfrac{\theta}{(1-\lambda)^3}$, thus $\lambda=1-\sqrt{1/\delta}$ and $\theta=(\mu-1)\sqrt{1/\delta}$
- NB(r,p): $\mu=\dfrac{(1-\mathrm{p})\mathrm{r}}{\mathrm{p}}+1$, $\sigma^2=\dfrac{(1-\mathrm{p})\mathrm{r}}{\mathrm{p}^2}$, thus $\mathrm{p}=\dfrac{1}{\delta}$ and $\mathrm{r}=\dfrac{\mu-1}{\delta-1}$
- B(n,p*): $\mu=\mathrm{np}^*+1$, $\sigma^2=np^*(1-\mathrm{p}^*)$, thus $\mathrm{p}^*=1-\delta$ and $\mathrm{n}=\dfrac{\mu-1}{1-\delta}$

Figures 2 and 3 illustrate differences in the probability distributions among GP, NB and B models for different values of μ and δ. The rows indicate if there is any difference that results from increase in dispersion, whereas the columns describe modifications according to the change in mean. The pmfs of the above distributions are computed by using the ratio of two successive probabilities [5]. Figure 2 shows that for low over-dispersion (when δ is close to 1) there is a negligible difference in the pmf's. When μ increases, the pmf's differ as δ increases. But the difference becomes significant only for very large over-dispersion, not relevant for our study (cf. Table 1).

The difference between the binomial and the corresponding GP distribution (cf. Figure 3) is so small that the two lines overlap in most cases. The slight disagreement is only obvious for $\delta = 0.5$, but vanishes as μ increases. However, such low under-dispersion cases are not of interest here (cf. Table 1).

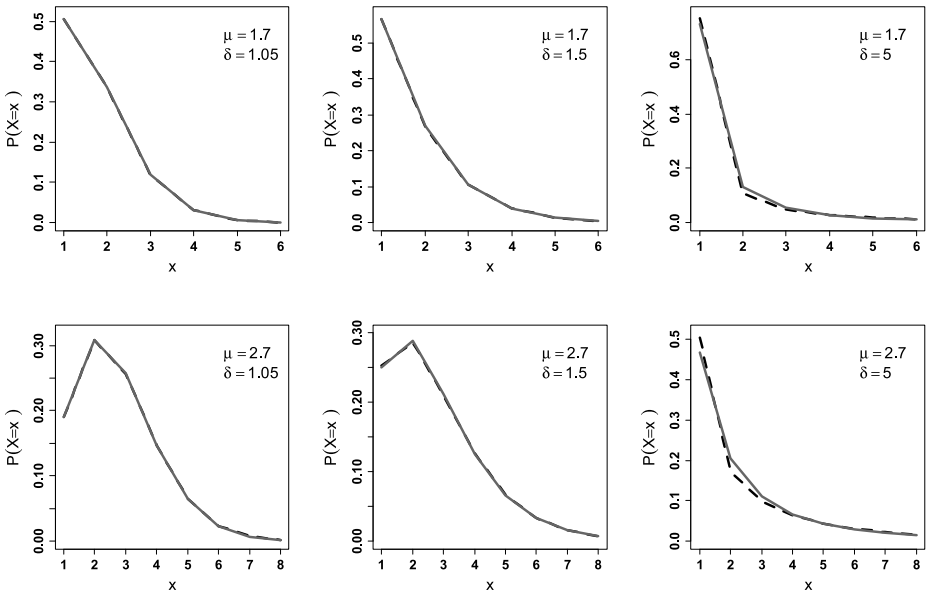


Fig. 2. Comparison of pmf's: GP(λ,θ) (solid line) and NB(r,p) (dashed line) for over-dispersed cases with $\delta = 1.05, 1.5, 5$ and mean $\mu = 1.7, 2.7$

3.2 The Singh-Poisson Model

Another possibility to construct a two-parametric generalized Poisson distribution is to combine the Poisson with the degenerate (one-point) distribution where the probability mass is cumulated at zero-point [4]. In its 1-displaced form, the pmf of a discrete random variable X having SP distribution is given by

$$\pi_{x|\alpha,\theta} = P(X = x) = \begin{cases} 1-\alpha+\alpha e^{-\theta}, & x = 1 \\ \alpha\theta^{x-1}e^{-\theta}/(x-1)!, & x = 2,3,\ldots \end{cases} \tag{3}$$

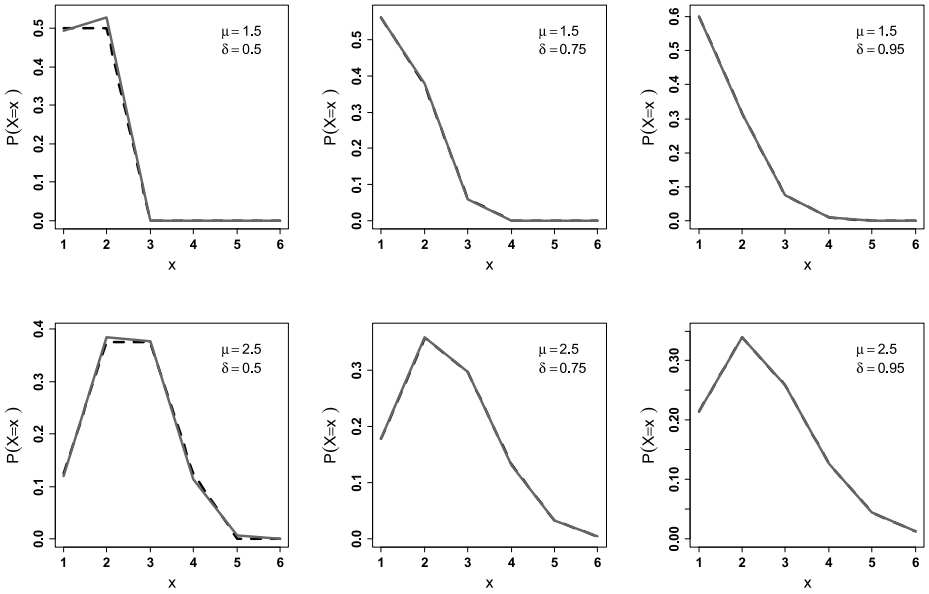


Fig. 3. Comparison of pmf's: $GP(\lambda,\theta)$ (solid line) and $B(n,p^*)$ (dashed line) for under-dispersed cases with $\delta = 0.5, 0.75, 0.95$ and mean $\mu = 1.5, 2.5$

where $\theta > 0$ and $0 < \alpha \leqslant \alpha_{\max} = 1/(1-e^{-\theta})$. Here, $\alpha_{\max}$ denotes the maximal possible value of α for a given θ and results from the constraint $1-\alpha+\alpha e^{-\theta} \geqslant 0$. The first two moments are given by $E(X) = 1+\alpha\theta$ and $\mathrm{var}(X) = \alpha\theta(1+\theta-\alpha\theta)$, hence the index of dispersion is $\delta = 1+\theta(1-\alpha)$. Clearly, under- or over-dispersion is governed only by parameter α, as θ is positive. For $\alpha = 1$ we have equi-dispersion, when $0 < \alpha < 1$ over-dispersion, whereas in case of $1 < \alpha < \alpha_{\max}$ under-dispersion with respect to Poisson variation. For further details see [5].

4 Parameter Estimation

To estimate unknown parameters of the models introduced we consider the three most common estimation procedures: method of moments (MM), maximum likelihood method (ML) and estimation based on sample mean and first frequency class (FF).

The moment estimators are obtained by equating the sample moments to the corresponding theoretical counterparts, and are given by:

- for GP: $\hat{\theta}_{\mathrm{MM}} = \sqrt{\dfrac{(\bar{x}-1)^3}{m_2}}$ and $\hat{\lambda}_{\mathrm{MM}} = 1-\sqrt{\dfrac{\bar{x}-1}{m_2}}$
- for SP: $\hat{\theta}_{\mathrm{MM}} = \dfrac{m_{(2)}}{\bar{x}-1} - 2$ and $\hat{\alpha}_{\mathrm{MM}} = \dfrac{\bar{x}-1}{\hat{\theta}_{\mathrm{MM}}}$.

The maximum likelihood estimator is the value that maximizes the log-likelihood function which by solving score equations results in

- for GP: $\hat{\theta}_{\mathrm{ML}} = (\bar{x}-1)(1-\lambda)$, where $\hat{\lambda}_{\mathrm{ML}}$ is solution of

$$\sum_{i=1}^{k} \frac{f_i(i-2)(i-1)}{(\bar{x}-1)+(i-\bar{x})\lambda} - n(\bar{x}-1) = 0$$

- for SP: $\hat{\alpha}_{\mathrm{ML}} = \dfrac{n-f_1}{n(1-e^{-\hat{\theta}_{\mathrm{ML}}})}$, where $\hat{\theta}_{\mathrm{ML}}$ is solution of

$$\frac{\theta(n-f_1)}{n(\bar{x}-1)} + e^{-\theta} - 1 = 0.$$

To obtain estimators based on mean and first frequency class we equate the sample mean $\bar{x}$ and the relative frequency of the first class f_1/n to the theoretical mean μ and the probability of the first class π_1 and obtain $\hat{\theta}_{\mathrm{FF}} = \log\left(\dfrac{n}{f_1}\right)$ and $\hat{\lambda}_{\mathrm{FF}} = 1 - \dfrac{1}{\bar{x}-1}\log\left(\dfrac{n}{f_1}\right)$ for the GP model. For the SP model it can be shown that $\hat{\alpha}_{\mathrm{FF}} = \hat{\alpha}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{FF}} = \hat{\theta}_{\mathrm{ML}}$.

5 Monte Carlo Simulation Study

We investigate whether the GP model or the SP model performs better for different estimation techniques mentioned above by carrying out a simulation study where all three dispersion situations are considered. Again, we fix the degree of dispersion and the first theoretical moment to obtain the parameterization of the SP model parameters as: $\theta = \delta + \mu - 2$ and $\alpha = (\mu-1)/\theta$. For the parameterization of the GP model parameters see Section 3.1.

This approach enables us to find out the consequence of applying the wrong model whenever the true parameter values, specified in terms of μ and δ, are known. Additionally, we can receive an impression of how effective the proposed estimation techniques are.

With μ and δ known, the model settings result from the above expressions as: (i) for $(\delta,\mu) = (1.34, 2.45)$ we have $(\lambda, \theta_{\mathrm{GP}}) = (0.14, 1.25)$ and $(\alpha, \theta_{\mathrm{SP}}) = (0.81, 1.79)$, (ii) for $(\delta,\mu) = (1.02, 1.85)$ we have $(\lambda,\ \theta_{\mathrm{GP}}) = (0.01, 0.84)$ and $(\alpha,\ \theta_{\mathrm{SP}}) = (0.98, 0.87)$, whereas (iii) for $(\delta,\mu) = (0.78, 1.63)$ we have $(\lambda,\ \theta_{\mathrm{GP}}) = (-0.13, 0.71)$ and $(\alpha,\ \theta_{\mathrm{SP}}) = (1.54, 0.41)$. The sampling experiments are carried out to produce $M = 500$ Monte Carlo samples from both models, and diverse dispersion situations, each of size $n = 1000$. The whole procedure is implemented in the statistical software R. The ML estimation is solved by the Newton-Raphson algorithm for GP model, and for SP by using function **optim()**, available in the software R. For each of the data situations considered here we calculated the mean values of M

parameter estimates, as well as the estimated standard errors. To generate GP and SP random variables we use the inversion method [11]. Table 2 summarizes the simulation results obtained. We encountered no difficulty when performing these simulation experiments.

Table 2. Estimation results for data simulated from the GP and SP models

$GP(\lambda,\theta)$	λ		θ	
	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$\delta > 1$	0.140 (0.022)	0.140 (0.022)	1.251 (0.044)	1.251 (0.040)
$\delta \approx 1$	0.011 (0.022)	0.010 (0.022)	0.840 (0.033)	0.840 (0.033)
$\delta < 1$	–0.130 (0.023)	–0.131 (0.022)	0.710 (0.031)	0.711 (0.030)
$SP(\alpha,\theta)$	α		θ	
	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$\delta > 1$	0.812 (0.024)	0.811 (0.019)	1.788 (0.067)	1.789 (0.057)
$\delta \approx 1$	0.984 (0.049)	0.982 (0.043)	0.869 (0.052)	0.870 (0.048)
$\delta < 1$	1.554 (0.133)	1.551 (0.124)	0.410 (0.039)	0.410(0.037)

In the GP case, all three estimation methods yield similar results, although FF (not shown here) compared to MM has bigger standard errors. Nevertheless, the relative standard errors of θ, calculated as $\mathrm{RSE}_{\bar{\theta}} = \mathrm{se}_{\bar{\theta}}/\theta$, are much smaller than those of $\bar{\lambda}$, regardless of the estimation method and the dispersion case. The parameter estimates of θ in the SP case are of similar precision as those obtained for the GP model in the over- and equi-dispersed data case, but become more imprecise as soon as the under-dispersed data case is at hand. As to the RSE of the parameter α, we found out that they are more precise compared to those of the parameter λ. For $\delta > 1$ we obtain 2-4% accuracy, $\mathrm{se}_{\bar{\alpha}}$ is 2-3% of α, for $\delta \approx 1$ approximately 4-5%, whereas for $\delta < 1$ around 8-8.5% estimation precision for α.

6 Application to Slovenian and Russian Texts

The results of fitting the Singh-Poisson model to 120 Slovenian texts are displayed in Figure 4, left panel. The solid black line in the graph is the reference bound $\mathrm{C} = 0.02$, whereas the dashed line refers to $\mathrm{C} = 0.05$. Obviously, the SP model provides a good fit for the majority of the texts. It seems to be unsuitable, however, for the three poems of Gregorčić, namely "Kesanje" (text no. 66), "Na sveti večer" (text no. 69) and "Pri zibelki" (text no. 76). A closer look at their structure shows that all of them are indeed short texts, having length of 181, 117 and 114 words respectively[1].

[1]Text length is a general problem in quantitative linguistics and cannot be discussed in detail here. Using texts which are "*complete*" from a lexical and paradigmatical point of view is in some

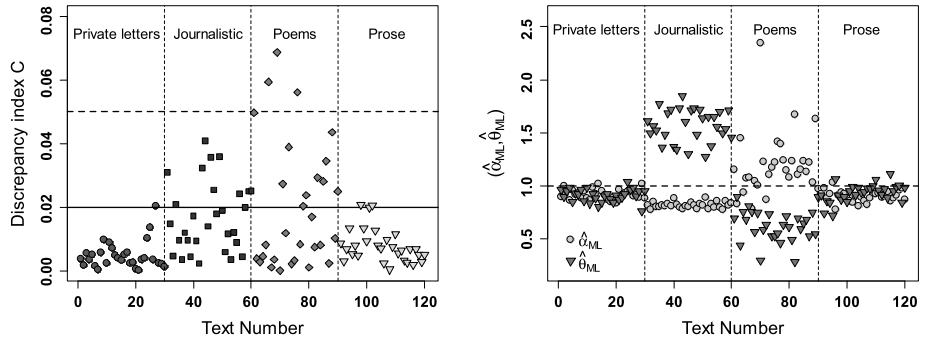


Fig. 4. Results of fitting the Singh-Poisson model to 120 Slovenian texts: discrepancy index (left) and estimated parameter regions (right)

For all 120 Slovenian texts maximum likelihood (ML) estimates of both parameters are computed and each pair of parameters $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ is plotted versus the corresponding text, as shown in Figure 4, right panel. It is evident that each group of texts leads to a different pattern of parameters. In the case of private letters both parameters are very close to each other; the same holds for prose texts, although reversed with respect to the order. Contrary to this, in journalistic texts and poems parameters are quite distant from each other. The $\hat{\alpha}_{\mathrm{ML}}$ outlier in Figure 4 (right panel) refers to Gregorčić's poem "Njega ni!" (text no. 70). This text has only 106 words, $\hat{\alpha}_{\mathrm{ML}} = 2.35$ and $\hat{\theta}_{\mathrm{ML}} = 0.3$ $(\alpha_{\max} = 3.88)$. Surprisingly, the value of $\mathrm{C} = 0.0002$ indicates an extremely good fit here. The results of fitting the Singh-Poisson model to 120 Russian texts are shown in Figure 5. Notice that for five texts the values of C are beyond the reference line. These texts include again four extremely short pri-

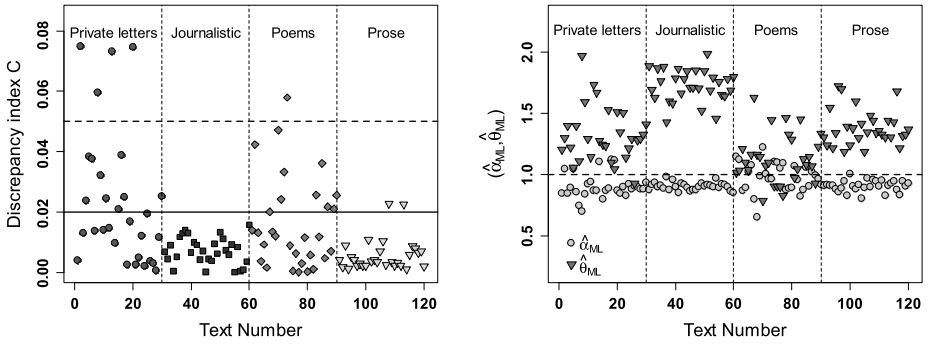


Fig. 5. Results of fitting the Singh-Poisson model to 120 Russian texts: discrepancy index (left) and estimated parameter regions (right)

cases accompanied with the problem that some very short texts (e.g. private letters, poems etc.) have to be analyzed. Obviously the shortness of the text can cause problems in modeling them. In another context it has been shown that text and word length are interrelated in a systematic way [9].

vate letters by Achmatova written to Brodsky, Chardžiev and Maksimov (texts no. 2, 8, 13 and 20) consisting of 130, 63, 48 and 75 words, respectively, and a poem by Nekrasov "Muza" (text no. 73) with 302 words. The $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ parameter regions (cf. Figure 5, right panel) are different from those of the Slovenian texts shown in Figure 4 (right), mostly for private letters, poems and prose.

Plotting the parameters of the SP model versus each other leads to a good discrimination of the three Slovenian text groups, as can be seen in Figure 6, left. Although some Russian texts are located in clearly defined areas, there are many overlappings. However, some general tendency of journalistic texts to build a sep-

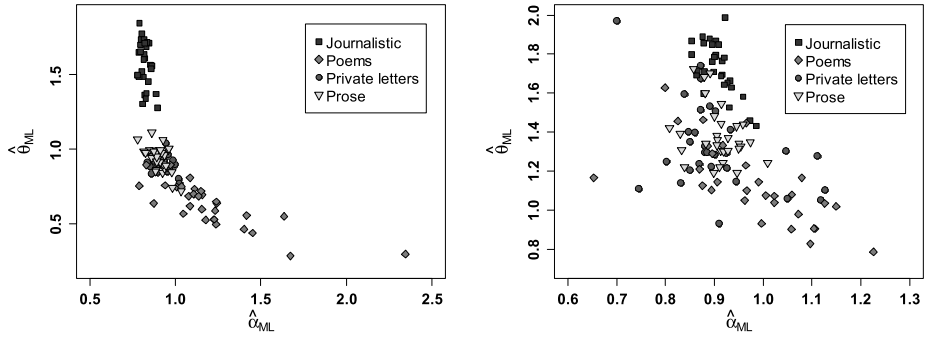


Fig. 6. Text types discrimination by (α, θ) parameter range of Singh-Poisson model of Slovenian (left) and Russian (right) language

arate category can still be observed (cf. Figure 6, right). One possible explanation can be found in the fact that the journalistic texts originate from a Russian quality newspaper, being as such free of colloquial speech, for which the use of nouns and compound words is quite typical. Also, these types of word are distinguished by longer word forms, and hence the location of the journalistic texts in the upper area of Figure 6, right, seems plausible. Although determined by the use of poetic meters, the Russian poems cannot be as clearly distinguished from letters and prose texts, as the Slovenian poems can. This quite surprising phenomenon happens possibly due to higher heterogeneity of the Russian poems compared to the Slovenian ones.

Figure 7, left panel, clearly shows that the generalized Poisson model is not appropriate for almost half (46.67%) of the Slovenian journalistic texts. However, the existence of three different parameter patterns, namely for journalistic, poems and joint group of letters and prose texts, is evident from Figure 8, left, regardless of the fact that the model fit for journalistic texts is not satisfactory.

As compared to this, the Poisson model provides more or less good fits for Russian texts (cf. Figure 7, right panel), with the exception of Achmatova's short letters no. 2, 8, 9, 13, 16, and 20 of 130, 63, 151, 48, 201, and 75 words respectively, as well as poems no. 62, 73 and 83 of length 157, 302, and 109 respectively. Yet, the discrimination of Russian texts is again not as obvious as it is in the case of the Slovenian text material (cf. Figure 8, right).

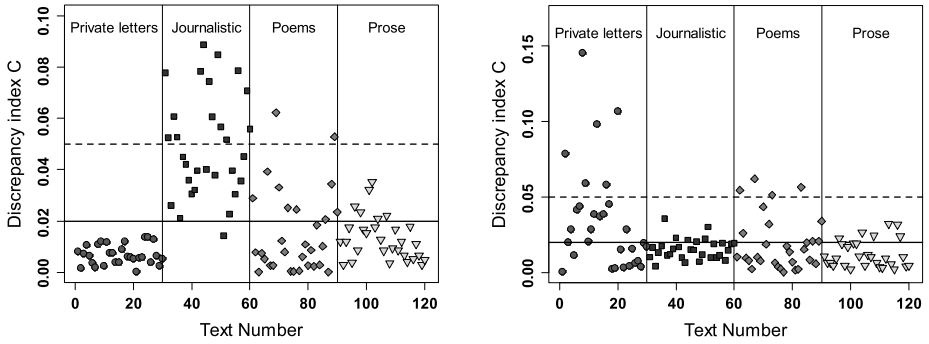


Fig. 7. Results of fitting the Generalized Poisson model to texts under study: Slovenian (left) and Russian (right)

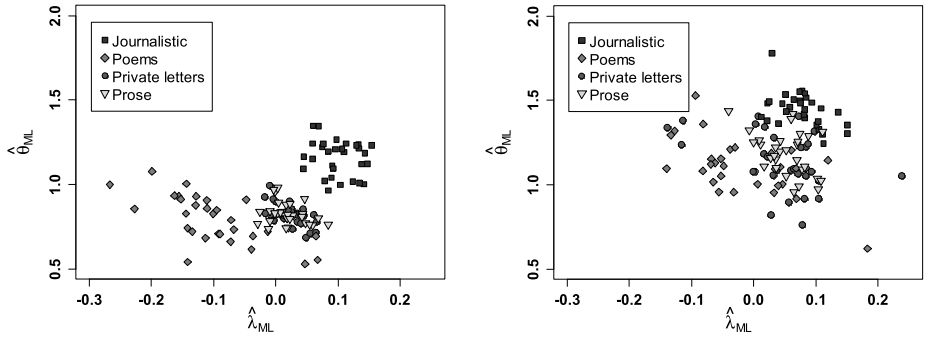


Fig. 8. Text types discrimination by (α, θ) parameter range of the Generalized Poisson model of Slovenian (left) and Russian (right) language

7 Summary

The analyzed Slavic languages, Slovenian and Russian, although belonging to the same family of Indo-European languages, are members of two different subgroups, namely the East Slavic (Russian) and the South Slavic (Slovenian) branches. Throughout history they have shown, however, a rather different development on all linguistic levels, in particular on the phonological, morphological and lexical level. Slovenian, contrary to Russian, was for instance lexically and syntactically heavily influenced by other non-Slavic languages, such as German or Italian. Despite these obvious disparities we have proven that the word length frequency distribution of our sampled texts can be described by one theoretical model. This model, known as the Singh-Poisson model, is a simple two-parametric generalization of the Poisson distribution with parameter θ. The additional parameter α tunes the type of dispersion. It allows us to model under-dispersion $(1 < \alpha \leqslant \alpha_{\max})$, equi-dispersion (Poisson case $\alpha = 1$) and over-dispersion $(0 < \alpha < 1)$. Therefore, the proposed model offers a unified approach for all dispersion cases. An additional benefit is that the maximum likelihood estimation leads, in case of the Singh-Poisson distribution, to the same estimates as the method based on the sample mean and the first

frequency class. For this reason, the calculation of maximum likelihood estimates is a very simple task. In a simulation study we have demonstrated the usefulness of the parameter estimates under three data-driven dispersion scenarios. Finally, the Singh-Poisson model was applied to 120 Slovenian and 120 Russian texts, and in all cases we obtained reasonable and stable estimates.

References

1. Antić, G., Kelih, E., Grzybek, P.: Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Language*, pp. 117-156. Springer, Dordrecht (2006)
2. Consul, P.C.: Generalized Poisson Distributions. Properties and Applications. Marcel Dekker, Inc., New York (1989)
3. Consul, P.C., Famoye, F.: *Lagrangian Probability Distributions*. Birkhäuser, Boston (2006)
4. Đuraš, G., Stadlober, E.: Modeling word length frequencies by the Singh-Poisson distribution. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures · Functions · Interrelations · Quantitative Perspectives*, pp. 37-48. Praesens, Wien (2010)
5. Đuraš, G.: *Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages*. TU Graz, Dissertation, Graz (2012)
6. Grzybek, P., Stadlober, E., Kelih, E., Antić, G.: Quantitative text typology: The impact of word length. In: Weihs, C., Gaul, W. (eds.), *Classification – The Ubiquitous Challenge*, pp. 53-64. Springer, Heidelberg (2005)
7. Kelih, E.: Slawisches Parallel-Textkorpus: Projektvorstellung von ”Kak zakaljalas’ stal’ (KZS)”. In: Kelih, E., Levickij, V. V., Altmann, G. (eds.), *Methods of Text Analysis*, pp. 106-124. ČNU, Chernivtsi (2009)
8. Kelih, E.: Zum Analytismus und Synthetismus in slawischen Sprachen: Morphologische Wortstruktur in slawischen Sprachen. In: Fischer, K., Krumbholz, G., Lazar, M. (eds.), *Beiträge der Europäischen Slavistischen Linguistik (Polyslav 14)*, pp. 99-107. Sagner, München/Berlin. (Die Welt der Slaven. Sammelbände, Sborniki, 43) (2011)
9. Kelih, E.: On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In: Grzybek, P., Naumann, S., Vulanović, R., Altmann, G. (eds.), *Festschrift für Reinhard Köhler*, pp. 67–80. Praesens, Wien (2012)
10. Puig, P., Valero, J.: Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101 (473), 332-340 (2006)
11. Stadlober, E.: *Sampling from Poisson, binomial and hypergeometric distributions: Ratio of uniforms as a simple and fast alternative* (Bericht No. 303). Mathematisch-Statistische Sektion, Forschungsgesellschaft Joanneum, Graz (1989)

A Method for Extracting Translational Equivalents from Aligned Texts

Ivan Obradović

University of Belgrade, Faculty of Mining and Geology
Đušina 7, 11000 Belgrade, Serbia
ivano@rgf.bg.ac.rs

Abstract. In this paper we present a simple method for extraction of semantically related word pairs, ideally translational equivalents, from aligned texts. The method was experimentally tested on SELFEH, a Serbian-English corpus of texts related to education, finance, health and law, aligned at the sentence level. The corpus was lemmatized and the method applied on lemmas of word forms from the corpus, by extracting candidate translational equivalents through a ranking based on lemma frequencies. A probabilistic explanation of the ranking method is given, as well as a heuristic to improve its efficiency. The results obtained are promising and future refinement of the method could bring further improvement.

Keywords: translational equivalents, aligned texts, word frequencies, Serbian-English aligned corpora.

1 Introduction

Development of multilingual language resources has become a major concern of many initiatives aimed at overcoming language barriers, such as the current European network META-NET[1]. Translational equivalents are essential for building certain types of multilingual resource, such as aligned wordnets, and hence the growing interest in methods and procedures for generating translational equivalents for a given language pair. Many of these methods are based on some sort of automatic or semi-automatic extraction of translational equivalents from aligned corpora, large collections of aligned texts, which also belong to core multilingual resources. Namely, aligned texts are obtained from parallel texts, which consist of an original text and its translation (or translations). These texts are aligned through identification of corresponding parts (most often sentences) in the original and its translation. Alignment, even at the sentence level, is far from a trivial task, as translators often split or merge sentences during translation, reorder them, or even delete

[1] http://www.meta-net.eu/

or insert parts of sentences. A large number of sentence alignment algorithms has been developed, generally based on various statistical approaches [1].

Extraction of translational equivalents, namely words that describe the same concept in different languages from parallel texts, is an even more complicated task. This topic has been the subject of intensive research for the last two decades, starting with the groundwork results of Church and Hanks [2] and Gale and Church [3]. Word alignment, which is what extraction of translational equivalents from a parallel text is often called, is typically performed after the parallel text has been aligned on the sentence level, although different approaches have also been taken [4]. It is one of the most important tasks in the majority of statistical machine translation methods, as word alignment is typically used for estimating the parameters of statistical machine translation models [5].

The method for automatic extraction of semantically strongly related or equivalent word pairs from aligned texts outlined in this paper relies on a simple ranking method based on word co-occurrence. Its probabilistic interpretation is similar in form to mutual information of two random variables, the quantity that measures the interdependence of random variables in probability theory and information theory, as well as to the dice combinatorial measure that represents the ratio between the size of the intersection of two sets and the sum of the sizes of these two sets [6]. The method outlined in this paper, however, proceeds in a somewhat specific manner, with an additional heuristic to improve its basic efficiency.

The Human Language Technology Group at the University of Belgrade has developed various language resources for Serbian, as well as language tools, among which a prominent place is taken by Leximir, a software environment for management and exploitation of language resources [7]. Among other things, Leximir enables the handling of the Serbian wordnet (SWN), developed within the scope of the Balkanet project, and aligned with the English wordnet via the inter-lingual index [8]. Thus the two aligned wordnets became one of the basic multilingual resources involving Serbian in addition to Serbian-English and Serbian-French aligned corpora [9]. Although SWN is already of a considerable size, it needs to be further expanded and refined. Translational equivalents obtained by extraction from Serbian-English aligned corpora can be easily used for defining new Serbian synsets or refining the existing ones. The use of aligned corpora for developing SWN has already been tackled in [10], and the method presented in this paper represents a further advancement in this approach.

In the following section, we describe the aligned corpus used for testing our method and the preprocessing of this resource. Section 3 gives a detailed overview of the proposed method and its probabilistic interpretation. Results obtained by applying the method to the available aligned corpus are presented in Section 4, followed by Section 5 outlining the conclusions and propositions for future work.

2 The resource and its preprocessing

Akin to other Slavic languages, Serbian is a morphologically rich language, with complex inflection rules, a number of verb tenses, seven noun cases and numerous

irregularities in gender and number agreement. Processing of Serbian texts thus presents many challenges to researchers in computational linguistics, one of them being the production of aligned bilingual corpora of even a moderate size. The Serbian-English aligned corpus SELFEH, used in our research, has been developed in the course of the participation of the University of Belgrade HLT group in the INTERA project (Integrated European language data Repository Area). One of the aims of this project was the production of new multilingual resources (parallel corpora and terminologies) for the less widely spoken languages, including Balkan languages, which generally suffer from the same problem of lack of available language resources [11].

The SELFEH corpus consists of some 100+ documents with a total of a little less than one million words per language, pertaining to education, finance, health and law. Parallel texts are composed of English originals and Serbian translations (e.g. of various international treaties and covenants), but also of Serbian originals and English translations (e.g. of Serbian legislation), and are aligned at sentence level using ACIDE, an integrated development environment for generating aligned parallel texts [12]. For each pair of texts four XML (Extensible Markup Language)[2]documents were generated and stored within the corpus. The first two XML documents contained the original text and its translation, segmented at the sentence level. Serbian texts were also tokenized, lemmatized and tagged with POS (Part-Of-Speech). The results were manually checked and a third XML document for these texts was generated. The fourth XML document was the result of the alignment process and contained aligned Serbian-English texts in TMX (Translation Memory eXchange) format.

As the English texts in SELFEH were not tagged and lemmatized during the INTERA project, some further preprocessing was needed in order to prepare it for the application of the proposed method. To that end we used the TreeTagger [13] in batch processing for lemmatization and tagging of English texts. Due to a shortage of resources the results obtained by TreeTagger were not manually checked, as opposed to the lemmatization and tagging of Serbian text, thus allowing for possible incorrect lemmas and POS tags in English texts. This possibility was, however, taken into account and its impact on the extraction method will be commented on later.

With both texts lemmatized, the alignment information from TMX files was used to create a plain-text sentence-aligned corpus in which the surface forms of words were substituted with their lemmas, and then the ranking method was applied to extract translational equivalents from this corpus.

3 Description of the method

The proposed method for extracting translational equivalents from aligned texts is based on a fairly simple strategy. We will denote the set of all words in their canonical form (lemmas) appearing in the source text as V_s, which will, for the purpose

[2]http://www.w3.org/XML/

of this paper, be referred to as the vocabulary of the source language. Similarly, we will denote the set of all words (lemmas) appearing in the target text as V_t, and refer to it as the vocabulary of the target language. Our strategy is aimed at finding $w_t \in V_t$ which is the "best match" for $w_s \in V_s$, by looking for it among the words appearing in sentences of the target language that are translational counterparts of sentences containing w_s in the source language.

Following the strategic goal adopted, the method starts by what seems to be the most straightforward idea, namely, for a given word w_s, all source language sentences containing this word are selected, and target language sentences that are aligned with these source language sentences are identified. Then the set of words occurring in those target language sentences is created. We will denote this subset of the vocabulary V_t as the relevant set $R_t \subset V_t$, and for each word x from the relevant set, $x \in R_t$, we will denote the number of its occurrences in sentences of the target language aligned with sentences that contain w_s as its conditional frequency $F(x|R_t)$.

The simplest approach would be to rank all words from the relevant set R_t in decreasing order of their conditional frequency and proceed by looking for w_t, the translational counterpart of w_s, at the top of the list obtained by such ranking. Although it might seem logical that the words from the top of the list are the most likely translational counterparts of w_s this is most often not the case. Let us illustrate this by way of an example. As we have already mentioned, the language pair in our case was Serbian and English, where both languages could be either the source or the target language. Let us use English as the source language and create the list of possible translational equivalents of the word *crime* according to the statistics collected in the aforementioned manner on the appropriate subset of SELFEH. When we look at the occurrence count, we obtain the following Serbian tokens as possible translations (number of occurrences given in brackets):

1. *jesam* (57)
2. *u* (41)
3. *i* (39)
4. *da* (35)
5. *zločin* (34)
6. ...

One could argue that this approach has some merit, as the correct translation, namely the word *zločin*, is among the first five words on the list. However, the top of the list is occupied by some of the most frequent words in Serbian, such as the extremely frequent auxiliary verb *jesam* (Eng. *am*), the conjunction *i* (Eng. *and*) and prepositions *u* (Eng. *in*) and *da* (Eng. *to*), which have no semantic relation with the English word *crime*.

One way of propelling *zločin* to the top of the list would be to remove the first four words as function words. However, we were looking for a general method for extracting translational equivalents, which would not depend on this constraint, namely on the removal of function words from the text. Thus we took another approach to circumvent the problem of frequent semantically unrelated words, by giving preferentiality to words appearing more frequently in sentences that are translations of original sentences containing the source word w_s than in the remainder of

the corpus. This could be achieved by introducing $RF(x|R_t)$, the relative frequency of the word x, obtained as the ratio of its conditional frequency $F(x|R_t)$, and the total frequency of that word in the entire corpus $F(x)$:

$$RF(x|R_t) = \frac{F(x|R_t)}{F(x)} \tag{1}$$

and ranking words in decreasing order of their relative frequency. However, ranking target language candidate words in this manner has a serious adverse effect, namely, it propels the least frequent target language words to the top. For example, the relative frequency of words that appear only once in the corpus is always 1, which is the highest possible value for $RF(x|R_t)$, and hence puts such words at the top of the ranking list. Consequently, a high rank is typically assigned to words that appear just a few times in the entire corpus. These words, on the other hand, in the majority of cases need not be also the desired translational equivalent.

In order to overcome this deficiency the two approaches were combined: ranking based on conditional frequencies and ranking based on relative frequencies. To that end the conditional frequency $F(x|R_t)$ was first normalized by dividing it with the sum of conditional frequencies for all words appearing in R_t, that is, the total occurrence count of all words appearing in these sentences, thus obtaining:

$$NF(x|R_t) = \frac{F(x|R_t)}{\sum_{y \in R_t} F(y|R_t)} \tag{2}$$

as the normalized conditional frequency of the word x. The final measure $MF(x|R_t)$ obtained for ranking purposes was defined as the multiple of the relative frequency $RF(x|R_t)$ and the normalized conditional frequency $NF(x|R_t)$, namely:

$$MF(x|R_t) = RF(x|R_t) \cdot NF(x|R_t) = \frac{F(x|R_t)}{F(x)} \cdot \frac{F(x|R_t)}{\sum_{y \in R_t} F(y|R_t)} \tag{3}$$

The selection of the measure (3) can be corroborated by a probabilistic interpretation. We will denote the probability that a randomly selected word x from the target vocabulary $V_t (x \in V_t)$ belongs to the relevant set R_t as $P(R_t|x)$. We will also denote the probability that a word x will be selected at random from the set of relevant words R_t by $P(x|R_t)$. For a relevant set R_t obtained for the word w_s from the source vocabulary V_s, $(w_s \in V_s)$, the first probability $P(R_t|x)$, that a word from the target vocabulary will belong to this particular relevant set can be estimated by

$$\frac{F(x|R_t)}{F(x)}. \tag{4}$$

On the other hand, the second probability $P(x|R_t)$, that the word x will be selected from the relevant set R_t can be estimated by

$$\frac{F(x|R_t)}{\sum_{y \in R_t} F(y|R_t)}. \tag{5}$$

According to our proposed measure used for ranking (3), the word x from the relevant set $x \in R_t$ is thus ranked by estimated values of the probabilities within the product:

$$P(x|R_t) \cdot P(R_t|x). \tag{6}$$

The ranking can further be justified in general terms related to manipulating conditional probabilities. Conditional probability of a random event A, given B is defined as

$$P(A|B) = \frac{P(AB)}{P(B)}, \tag{7}$$

where $P(B) \neq 0$ and $P(AB)$ is the *joint* probability of events A and B occurring simultaneously, that is, the probability of the intersection of A and B. Therefore,

$$P(A|B) \cdot P(B|A) = \frac{P(AB)}{P(B)} \cdot \frac{P(AB)}{P(A)} = \frac{P(AB)^2}{P(A)P(B)}. \tag{8}$$

The value of this expression is maximized when $A = B$, since then the expression evaluates to 1. Therefore, the expression (8) might be interpreted as a measure of equivalence between A and B. In our case, we are looking at the measure of equivalence between the event A defined by the statement: "x belongs to R_t" and event B defined by the statement: "a randomly picked word from V_t is x".

The relevant word x that ranks the best according to our ranking method is ideally the translational equivalent w_t we are looking for, in which case w_s and w_t belong to the same part-of-speech and bear the same semantic meaning in the two languages. However, due to synonymy, the relation between semantically corresponding words in two languages is very often not one-to-one but rather many-to-many. Thus, we expect the top ranking candidates in R_t to either be a translational equivalent w_t of the source language word w_s or at least bear a strong semantic correspondence to w_s, e.g. as a part of a compound expression, an adjective commonly used in addition to the source language noun, a common constituent of verb phrases, etc.

Finally, one should observe that the product of conditional probabilities (6) is actually the square of the geometric mean of $P(x|R_t)$ and $P(R_t|x)$. It might then be rewarding to attempt to apply other means to these two factors, such as harmonic mean, weighted harmonic mean and others. We leave this as a potential future effort to better explore this ranking method.

Once we have settled for the ranking function (3), we proceeded in the following way. For each word in the source language w_s, the top five candidates from the target language were selected, according to the chosen ranking function. Then we reversed the ranking process. Namely, for each of the top five candidates obtained by the first round of ranking, we applied the ranking function in the reverse direction, namely by changing the place of the source and target languages, and selecting the top five candidates for each of them (that we will call reverse candidates). As reverse candidates belong to the source vocabulary, it is possible that the initial source language word w_s appears among top ranking reverse candidates. This is

clearly a strong indication that the corresponding candidate word bears semantic correspondence with the word w_s. We finally filtered the candidate words by leaving out target candidates whose best reverse candidates did not contain the source word w_s, and reordered them taking into account the position of the word w_s on the list of reverse candidates.

Let us clarify our procedure by way of an example. We shall denote the top five candidates in the target language for the source word w_s as $t_1, t_2, \ldots, t_5$. We shall then denote their corresponding top ranking candidates in the source language, obtained by reverse ranking as $s_{11}, s_{12}, \ldots, s_{15}, s_{21}, s_{22}, \ldots, s_{25}, \ldots, s_{51}, s_{52}, \ldots, s_{55}$. Let's say that the source word w_s appeared only among the reverse candidates for t_1, t_3 and t_5, as a_{13}, s_{31} and s_{53}, respectively. Based on this result we would eliminate the candidates t_2 and t_4 and reorder the remaining three candidates as follows. Although t_1 originally ranked better that t_3, the latter actually has a higher-ranking reverse candidate equal to *w*s. Hence, we would rank t_3 as the best candidate. As t_1 and t_5 both had w_s as their third-ranked reverse candidate, and given that t_1 originally scored better, the final ranking sequence would be t_3, t_1, t_5. This heuristic procedure devised to obtain a final reordering and elimination of some target candidates improved some of the results considerably.

There were also words from the source vocabulary for which none of the best target candidates had a reverse candidate equal to them, in which case we kept the original ranking of target candidates. However, it turned out that this usually happened when none of the words from the candidate list had any significant semantic correspondence to the source word w_s.

4 Results and evaluation

The method was validated on the SELFEH corpus, choosing Serbian for the source and English for the target language. The Serbian vocabulary, extracted from a total of 854,326 lemmas appearing in Serbian texts of the SELFEH corpus, contained 18,858 different lemmas, hence with an average lemma frequency of 45.3. The most frequent lemma was the conjunction *i* (Eng. *and*) with 38,535 occurrences among corpus lemmas, followed by the preposition *u* (Eng. *in*), with 30,130 occurrences and the verb *jesam* (Eng. *to be*) with 26,491 occurrences. Hence these three most frequent lemmas accounted for 11.1% of the entire set of corpus lemmas. On the other hand, there were 6,669 lemmas that appeared only once, thus representing 35.4% of vocabulary lemmas, but only 0.8% of the corpus lemmas.

In order to get a clearer picture between the number of lemmas with a specific frequency and their participation in the set of vocabulary lemmas and the set of corpus lemmas, we divided the 18,858 vocabulary lemmas into five disjunctive subsets according to their frequency in the corpus, as summarized in Table 1.

Taking into account the participation of lemmas distributed by frequency in the corpus and the vocabulary, as well as the fact that our ranking procedure is semi-automatic and its results need to be validated manually, we proceeded in the following manner, namely, we manually checked and validated results for each of the

2,009 lemmas with frequencies of 50 or higher, which covers 87.8% of corpus lemmas. As for the remaining, less frequent lemmas, we only made an estimate of the performance of our method by validating results for random samples of 500 lemmas from each of the four remaining groups. Thus, the results for a total of 4,009 lemmas were validated.

Table 1. Distribution of lemmas by their frequency in the vocabulary and the corpus

Frequency	Vocabulary	% Vocabulary	Corpus	% Corpus
1	6,669	35.4%	6,669	0.8%
2-5	5,550	29.4%	16,309	1.9%
6-20	3,202	17.0%	34,801	4.1%
21-49	1,428	7.6%	46,639	5.5%
$\geqslant 50$	2,009	10.7%	749,908	87.8%
Total	**18,858**		**854,326**	

Manual validation was performed by assessing whether the best ranked candidate is an acceptable translational equivalent for the source word w_s. For acceptable candidates the result was marked as OK, while in the opposite case, the mark was NOK. Hence, we deemed the method successful only if the first candidate was acceptable, although in many cases when the first candidate was unsuccessful, another candidate on the list represented an acceptable translation. However, we counted those cases as failures.

Although a translational equivalent in general presumes the same part-of-speech, we did apply some relaxations to this presumption when assigning the OK score, keeping in mind the moderate size of the corpus and the fact that the results obtained by TreeTagger were not manually checked, and that it is known that TreeTagger displays bias towards some lemmas. For example, if a candidate translational equivalent for an adjective was a verb whose past participle is a proper translational equivalent of the source adjective, we assessed it as OK, as in the case of the pair (*utvrđen*, *determine*), where *determined* would be the correct translation. Also, there were a few cases when the best candidate could not be considered as the most appropriate translational equivalent in a general context, but given that SELFEH covers only documents of a specific content, we marked it as OK, assessing it as appropriate in the given context. For example (*naručilac*, *client*) was marked as OK, although in Serbian *naručilac* is generally a more specific term related to "someone who has ordered some goods or services", and its hypernym *mušterija* is semantically closer to *client*. However, this distinction is by itself not very big and in the context of the SELFEH corpus it may be deemed as irrelevant.

The final results of the validation procedure for the five subsets were as follows:

- for lemmas occurring 50 or more times: 83.1% OK,
- for words occurring 21-49 times: 61.8% OK,
- for words occurring 6-20 times: 52.4% OK,
- for words occurring 2-5 times: 31.2% OK,
- for words occurring only once: 19.6% OK.

For these five subsets Figure 1 displays the percentage of their respective participation in the vocabulary lemmas, the corpus lemmas and the success rate of the method applied for each particular subset. It should, however, be stressed once again that the success rate for the four subsets containing lemmas with a frequency less than 50 are only an estimate based on random samples. The 83.1% of successful results for lemmas with a frequency of 50 account for 72.9% of corpus lemmas, whereas the successful candidates in the remaining four subsets cover approximately 6.3% of the corpus lemmas. We can thus conclude that the method succeeded in offering an acceptable translational equivalent for a little less than 80% of the SELFEH corpus.

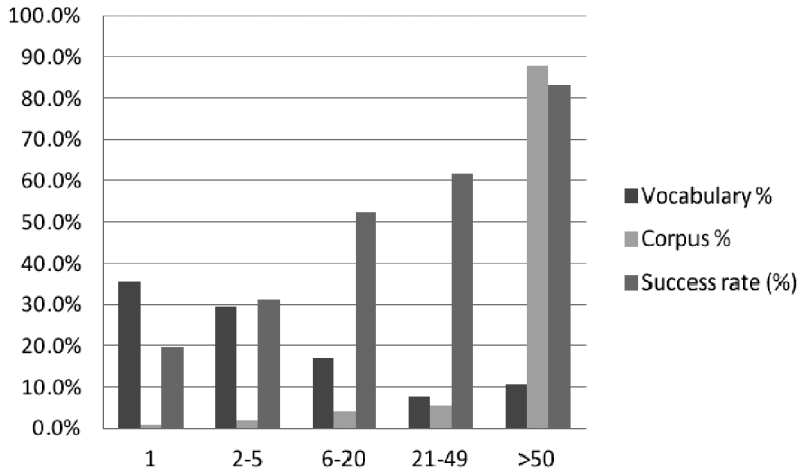


Fig. 1. Lemmas distributed by frequency

Finally, it should be noted that the chart in Figure 1 suggests a strong positive correlation between the frequency of the source word in the corpus and the success of the best candidate for translational equivalent, which was only to be expected.

5 Conclusions and future work

It is safe to say that the method yielded promising results, especially bearing in mind the moderate size of the available corpus. Also, many cases in which the results were assessed as a failure were caused by the relatively strict success criterion, which discarded results when the correct word was among top scoring candidates, but not the first one. Therefore, potential research directions include improvements of the ranking method and heuristics for further refinements.

Also, as was mentioned before, it might be beneficial to try to aggregate the two factors in the ranking with some other mean, possibly the weighted one, and see if one mean yields better results than the other.

Finally, one application would be to try to substitute some other tool for word alignment in a complete SMT system by this method and see if it improves the results, according to some established metric, for example the BLEU score [14].

Regarding language resources, clearly more effort needs to be put into assembling multilingual aligned corpora, as then it would be possible to use surface word forms as well, and fall back to lemmas only if necessary.

Acknowledgments. The author would like to thank Aljoša Obuljen for his valuable contribution to the research outlined in this work.

This research was supported by the Serbian Ministry of Education and Science under the grant #III 47003.

References

1. Yu, Q., Max, A., Yvon, F.: Revisiting sentence alignment algorithms for alignment visualization and evaluation. In: Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC2012, pp. 10–16. Istanbul, Turkey (2012)
2. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: Proceedings of the 27th annual conference of the Association of Computational Linguistics, pp. 76–82. Vancouver, Canada (1989)
3. Gale, W.A., Church. K.W.: Identifying word correspondences in parallel texts. In: Proceedings of the workshop on Speech and Natural Language, pp. 152–157. Stroudsburg, PA (1991)
4. Fung, P., Church, K.W.: K-vec: A new approach for aligning parallel texts. In: Proceedings of the Fifteenth international conference on Computational Linguistics, COLING 94, pp. 1096–1102. Kyoto, Japan (1994)
5. Brown, P.F., Della Pietra, S A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pp.263–311 (1993)
6. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the International conference on machine learning, ICML, pp. 296–304. Madison, WI (1998)
7. Stanković, R., Obradović, I.: An Integrated Environment for Management and Exploitation of Linguistic Resources. In: Proceedings of the International Multiconference on Computer Science and Information Technology, Computational Linguistics - Applications Workshop, CLA'09, pp. 287–294. Mragowo, Poland (2009)
8. D. Tufis , D. Cristea , S. Stamou: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology, 7(1-2), pp. 3–4 (2004)
9. Vitas, D., Krstev, C., Obradović, I., Popović, Lj., Pavlović-Lažetić, G.: An Overview of Resources and Basic Tools for Processing of Serbian Written Texts. In: Proceedings of the Workshop on Balkan Language Resources and Tools, pp. 97–104. Thessaloniki, Greece (2003)
10. Krstev, C., Pavlović-Lažetić, G., Vitas, D., Obradović, I.: Using Textual and Lexical Resources in Developing Serbian Wordnet. Romanian Journal of Information Science and Technology, 7(1-2), pp. 147–161 (2004)
11. Gavrilidou, M., Labropoulou, P., Desipri, E., Giouli, V., Antonopoulos V., Piperidis, S.: Building parallel corpora for eContent professionals. In: Proceedings of the Workshop

on Multilingual Linguistic Resources, 20th International Conference on Computational Linguistics, COLING, pp. 90–93. Geneva, Switzerland (2004)

12. Obradović, I., Stanković, R., Utvić, M.: An Integrated Environment for Development of Parallel Corpora (in Serbian). In: B. Tošović (ed.) Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen, pp. 563–578. Lit Verlag, Berlin (2008)
13. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference of New Methods in Language Processing, NeMLaP, pp. 44–49. Manchester, United Kingdom (1994)
14. Papineni, K., Roukos, S., Ward, T., Zhu W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02, pp. 311–318. Philadelphia, PA (2002)

Thematic concentration in Japanese prose

Haruko Sanada

Rissho University, Faculty of Economics
4-2-16, Osaki, Shinagawaku, Tokyo 141-8602, Japan
hsanada@ris.ac.jp

Abstract. We empirically scrutinize an indicator expressing the characteristics of texts called *Thematic Concentration* (*TC*). The *TC* index [18] is based on the *h*-index originally proposed by Hirsch [3]. In the present paper 10 Japanese prose texts are selected and their *TC*s are calculated. *TC* is found to have a relationship to the number of different words in a text. The thematic words are located more in pre-*h* domain, i.e. on the ranks smaller than *h*. We also offer another indicator called *Modified TC*. We investigate how Japanese prose texts are similar from the point of view of *Thematic Concentration* using the *u* test of *Modified TC*.

Keywords: *h*-index, thematic concentration, Japanese, frequency of occurrence.

1 Introduction

Under thematic concentration we understand the bringing into focus of an entity which determines the content of a text. We know intuitively what it is but we want to express it by an indicator which measures its degree. In general, we can proceed in three ways:

1. We derive the concentration from the frequency of occurrence of all *word-forms*.

2. Since word-forms may miss some commonalities, especially in strongly synthetic languages where one unit can have several morphological forms which are counted separately, we use the *lemmatized units*. This is, in any case, a better approximation to the problem.

3. We use the so called *hrebs* established by L. Hřebíček [5] (cf. also Ziegler, Altmann [22]) consisting not only of the lemmas derived by omitting affixes but also of all references to the given word or concept. For example in the German poem *Der Erlkönig* by J.W.v.Goethe the main person is a son but the *hreb* containing it consists of {child, boy, son, I, you, him, he, your, me} and some more forms in German.

The last way (3) of searching for concentration is, of course, the most appropriate one, but it presupposes a complete referential analysis of a text. In the case of long

texts very complex programs would have to be written and the error ratio would be relatively great. The second way (2) may reduce many word-forms to homonyms. thereby distorting the frequencies and omitting the synonyms of the thematic word; we thus approximate the problem by taking into account all word-forms.

2 Thematic concentration

The procedure is as follows. A word counter computes the frequencies of individual forms (form-types) and sets up the rank-frequency sequence. The ranks range from $r = 1$ to $r = V$ (V being the number of different word/forms) and for each rank there is an empirical frequency derived from the given text. Then, following the method presented in Popescu et al. ([18]: 18) we compute the h-index which bisects the rank-frequency sequence defined as

$$h = \begin{cases} r & \text{if } r = f(r) \text{ exists.} \\ \dfrac{f(r_1)r_2 - f(r_2)r_1}{r_2 - r_1 + f(r_1) - f(r_2)} & \text{if } r = f(r) \text{ does not exist.} \end{cases} \tag{1}$$

The second part of the formula interpolates between two (mostly neighbouring) ranks. The h-index is a fixed point separating in a fuzzy way synsemantics occurring mostly at lower ranks from autosemantics occurring at the higher ranks. However, some autosemantics occur also at lower ranks, i.e. with high frequencies, and these are exactly those which signal the thematic concentration of the text. The ranks of the given pre-h autosemantics will be marked with r'. The thematic concentration of the text will be computed using the frequencies $f(r')$ which will be weighted by the differences between h and the given ranks. A relativized indicator will be obtained by dividing the given sum by the sum of all possible weights and the greatest frequency $f(1)$. In this way we obtain (cf. Popescu et al. [18]: 96)

$$TC = 2 \sum_{r'=1}^{[h]} \frac{(h - r')\, f(r')}{h(h-1)\, f(1)}, \tag{2}$$

where $[h]$ is the integer part of h. The summation cannot surpass $[h]$.

For the sake of illustration let us consider the text of (1) Kobayashi[1] [8]. Table 1 shows the word frequency table with accumulations of synsemantic and autosemantic words up to h-index for (1) Kobayashi [8].

TC for (1) Kobayashi [8] is obtained as follows:

$$TC = 2\frac{5995}{29 \times 28 \times 318} = 0.0464 \tag{3}$$

with h-index $= 29$, sum of $(h - r') * f(r') = 5995$, $f(1) = 318$.

[1]The authors are marked with numbers according to the rank of their text size in order to easily distinguish them from each other.

Table 1. Frequency table up to *h*-index for (1) Kobayashi [8]

Rank	Word	Frequency	Part of speech	Type	*TC* (*h*-*r'*)**f*(*r'*)
1	の	318	postposition	Synsemantic	
2	は	254	postposition	Synsemantic	
3	て	205	postposition	Synsemantic	
4	に	175	postposition	Synsemantic	
5	を	162	postposition	Synsemantic	
6	と	159	postposition	Synsemantic	
7	た	145	auxiliary verb	Synsemantic	
8	が	116	postposition	Synsemantic	
9	で	109	auxiliary verb	Synsemantic	
10	な	92	auxiliary verb	Synsemantic	
11	ある	79	verb	Autosemantic	1422
12	し	78	verb	Autosemantic	1326
13	いう	75	verb	Autosemantic	1200
14	も	66	postposition	Synsemantic	
15	の	63	postposition	Synsemantic	
16	だ	62	auxiliary verb	Synsemantic	
17	いる	62	verb	Autosemantic	744
18	事	57	noun	Autosemantic	627
19	で	44	postposition	Synsemantic	
20	的	43	suffix	Synsemantic	
21	絵	42	noun	Autosemantic	336
22	が	41	postposition	Synsemantic	
23	ない	39	auxiliary verb	Synsemantic	
24	する	35	verb	Autosemantic	175
25	ない	34	adjective	Autosemantic	136
26	に	33	auxiliary verb	Synsemantic	
27	か	32	postposition	Synsemantic	
28	もの	29	noun	Autosemantic	29
29	私	29	prenoun	Synsemantic	
30	それ	29	prenoun	Synsemantic	

3 Analysis of Japanese texts

We analyze the *TC* in 10 Japanese texts of prose ([1] [4] [6] [7] [8] [9] [10] [12] [13] [14]). The morphological analyzers *MeCab* [2] developed by the Graduate Schools of Informatics in Kyoto University and NTT Communication Science Laboratories, and *UniDic* [11] developed by the National Institute for Japanese Language and Linguistics are used. These softwares partition the sentences into vocabulary items, though Japanese has no space as a word boundary. Using their dictionary the softwares also provide additional information for all vocabulary items, namely lexemes, parts of speech, the origin of the word (Japanese, Chinese, European), etc. Then we made tables like Table 1 from output of the software.

The 10 selected texts and their computed *TC*s are given in Table 2 and Figure 1[2].

Table 2. Thematic Concentration of some Japanese texts ordered by *TC*

Author	Text	Number of different word	*N*	*h*	*TC*
(5) Okamoto	Pari Sai	3724	17700	44.67	0.0291
(7) Oe	Shiiku	3322	22467	50.25	0.0297
(4) Nosaka	Amerika hijiki	4215	17230	45.67	0.0298
(10) Kawabata	Yujiguni	5048	41922	75.00	0.0321
(8) Hori	Utsukushii Mora	3068	25011	57.00	0.0363
(6) Kaiko	Panikku	3716	21250	49.00	0.0392
(2) Miura	Kikyo	2509	13138	43.33	0.0401
(9) Mushakoji	Yujo	3662	41278	79.60	0.0409
(3) Ariyoshi	Sumi	2673	15315	43.20	0.0455
(1) Kobayashi	Guzo suhai	1338	5613	29.00	0.0464

The counts exclude punctuation. In our last paper on the *h*-index [21] homonyms were not sufficiently discriminated; in this paper we include part-of-speech information for each word, and the number of different words and *h*-index for each text have been calculated.

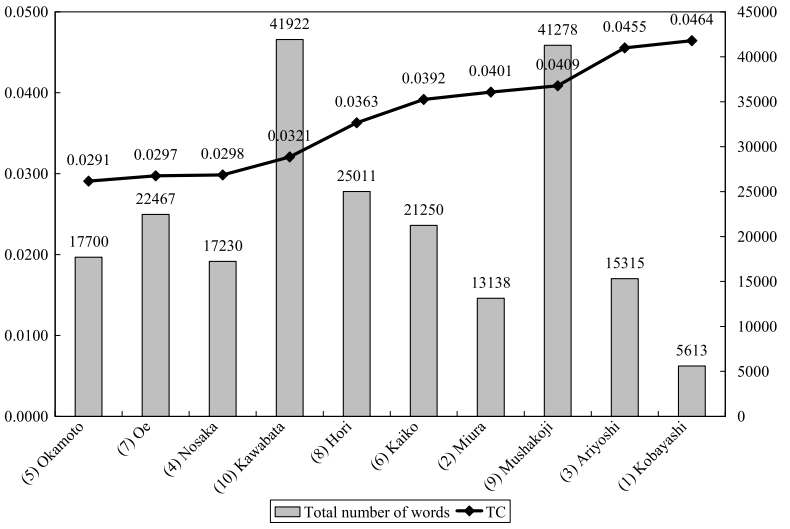


Fig. 1. Thematic Concentration in some Japanese texts

As can be seen, the *TC* does not depend on *N*, hence we obtain the sequence of writers ordered according to decreasing *TC* as follows:

(1) Kobayashi > (3) Ariyoshi > (9) Mushakoji > (2) Miura > (6) Kaiko > (8) Hori > (10) Kawabata > (4) Nosaka > (7) Oe > (5) Okamoto.

[2]Cf. Footnote 1.

We also analyze the relationships of *TC* and text size, and *TC* and the number of different words which are shown in Figures 2 and 3. We can say that *TC* is linked

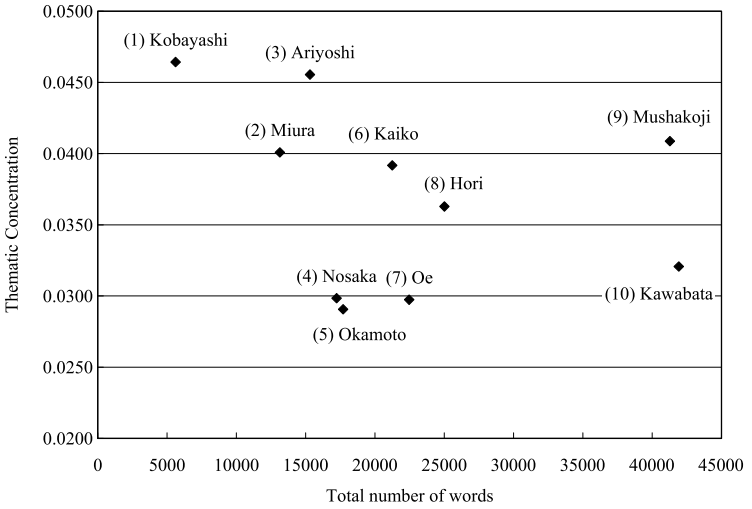


Fig. 2. The relationship of *TC* and the text size

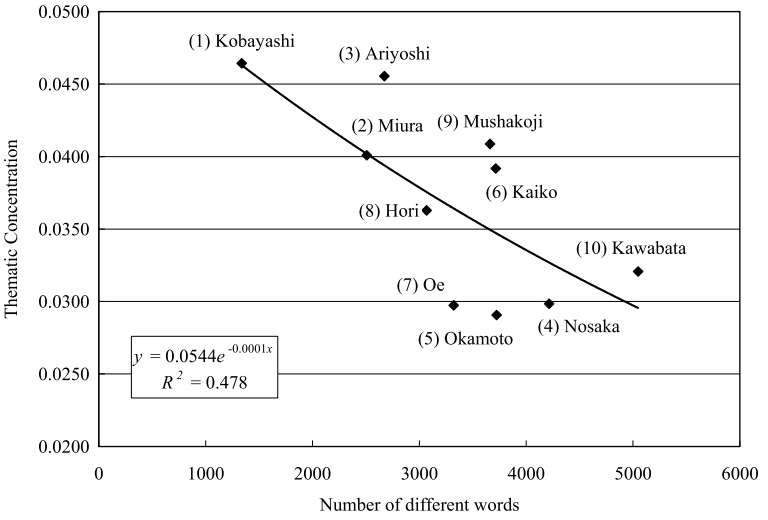


Fig. 3. The relationships of *TC* and the number of different words

to the number of different words, but it has no relationship with text size, that is, the greater the number of different words, the smaller the *TC*. If the number of different words is greater, the tail of the distribution is longer. The thematic words are expressed by synonyms or references, and there are only few thematic words in the pre-*h* domain. We have too few texts to show that this relationship is significant,

but the trend is visible. In Figure 2 the dependence of *TC* on *N* is shown. The trend has a large dispersion.

In Figure 3 the stronger dependence of *TC* on the number of different words is shown. The determination coefficient is merely $R^2 = 0.48$, and we suppose that by adding further texts the trend can be strengthened.

4 Modified Thematic Concentration

The concentration of autosemantics in the words with pre-*h* rank is expressed by the sum of their weights, that is, the distance of the rank from $h(h-r')$, and sum of their frequencies $(f(r'))$. To relativize, the sum of weights $\times$ the sum of frequencies of autosemantics is divided by the sum of all weights of the words with up to *h* rank and by the greatest possible frequency $f(1)$. However, the relativization is not perfect as it employs $f(1)$.

In this paper we offer a *Modified TC* employing the sum of frequencies of all words with up to *h* rank instead of $f(1)$. *Modified TC* is expressed by the relative weight $\times$ the relative frequency of autosemantics in the words with pre-*h* rank.

$$\text{Modified\_}TC = 2\sum_{r'=1}^{[h]} \frac{(h-r')f(r')}{h(h-1)F(h)}, \tag{4}$$

where $[h]$ is the integer part of *h* and $F(h)$ is the sum of frequencies up to *h*. *Modified TC* considers both the ratio of the sum of frequencies and the sum of weights of autosemantics in the words with pre-*h* rank.

Table 3. *Modified Thematic Concentration* of some Japanese texts ordered by *Modified TC*

Author	Text	Number of different word	*N*	*h*	*Modified TC*
(10) Kawabata	Yujiguni	5048	41922	75.00	0.002786
(4) Nosaka	Amerika hijiki	4215	17230	45.67	0.003099
(7) Oe	Shiiku	3322	22467	50.25	0.003449
(9) Mushakoji	Yujo	3662	41278	79.60	0.003460
(6) Kaiko	Panikku	3716	21250	49.00	0.003642
(2) Miura	Kikyo	2509	13138	43.33	0.003724
(8) Hori	Utsukushii Mora	3068	25011	57.00	0.004096
(5) Okamoto	Pari Sai	3724	17700	44.67	0.004241
(3) Ariyoshi	Sumi	2673	15315	43.20	0.004304
(1) Kobayashi	Guzo suhai	1338	5613	29.00	0.005455

The order of authors has slightly changed from Table 1 and Figure 1. (5) Okamoto [14] has lower rank and (9) Mushakoji [10] has higher rank with greater *Modified TC* than those with *TC*.

The relationships of *Modified TC* and text size or the number of different words are shown in Figures 5 and 6. The determination coefficient $R^2 = 0.78$ holds for the latter, and we should consider the background of this relationship.

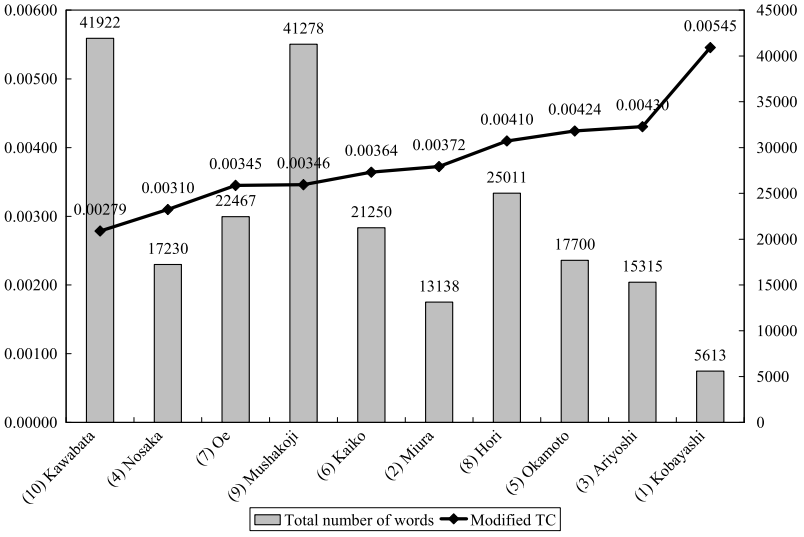


Fig. 4. *Modified Thematic Concentration* in some Japanese texts

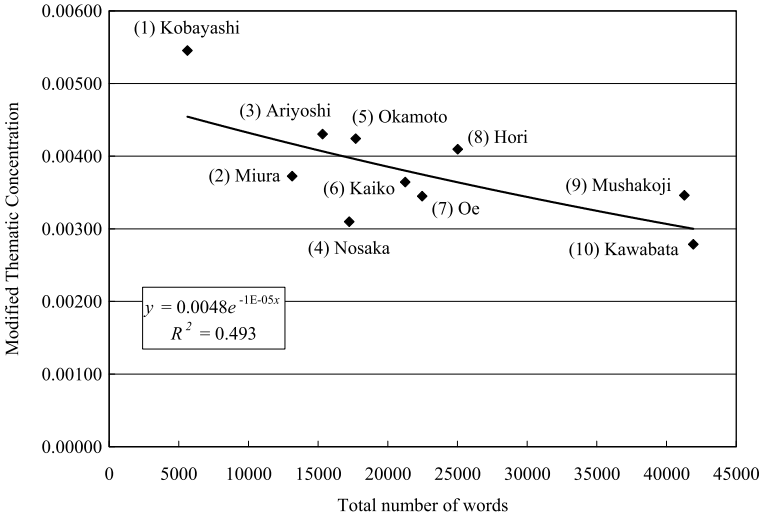


Fig. 5. The relationships of *Modified TC* and text size

As can be seen, the *Modified TC* also does not depend on the text size, hence we obtain the sequence of writer as follows:

(1) Kobayashi > (3) Ariyoshi > (5) Okamoto > (8) Hori > (2) Miura > (6) Kaiko > (9) Mushakoji > (7) Oe > (4) Nosaka> (10) Kawabata.

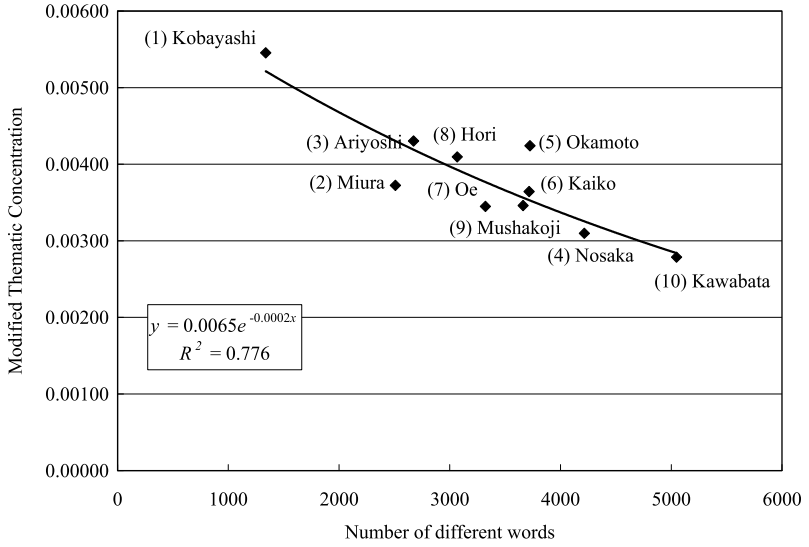


Fig. 6. The relationships of *Modified TC* and the number of different words

However, the sequence does not show whether the difference is significant. This can be established by means of a normal test.

The variance of each author's *Modified TC* is given as follows:

$$Var(Modified\_TC) = \frac{(Modified\_TC - Mean\_of\_Modified\_TC)^2}{n} \tag{5}$$

where the number of authors $n = 10$. The u statistics between two authors is given as follows:

$$u = \frac{(Modified\_TC_1 - Modified\_TC_2)}{\sqrt{Var(Modified\_TC_1) + Var(Modified\_TC_2)}}. \tag{6}$$

For example the difference between (3) Ariyoshi $(= TC_1)$ and (5) Okamoto $(= TC_2)$ yields

$$u = \frac{(0.004304 - 0.004241)}{\sqrt{(0.0000000229 + 0.0000000173}} = 0.31. \tag{7}$$

The difference of these two authors is not significant, and we can say that the degree of the thematic concentration of these two authors is similar. In order to obtain a complete view of all authors we perform the u-test for each pair using formula (5). The results are presented in Table 4.

Among 10 texts there are some values smaller than 1.96 corresponding to $\alpha > 0.05$ of the normal distribution (two-sided test), which are grouped into two classes. One of these is:

Table 4. Differences between *Modified TC* of 10 Japanese texts

(1) Kobayashi	(8)	(3)	(5)	(10)	(2)	(4)	(6)	(7)	(9)
(8) Hori	2.60								
(3) Ariyoshi	2.14	1.20							
(5) Okamoto	2.28	0.93	0.31						
(10) Kawabata	4.37	3.86	4.19	4.11					
(2) Miura	3.35	4.07	3.75	3.82	2.84				
(4) Nosaka	4.18	4.07	4.38	4.31	0.78	2.70			
(6) Kaiko	3.50	4.39	4.08	4.17	2.56	1.24	2.29		
(7) Oe	3.79	4.41	4.44	4.47	1.90	2.23	1.35	1.46	
(9) Mushakoji	3.78	4.42	4.43	4.46	1.94	2.20	1.41	1.41	0.07

{(8) Hori – (3) Ariyoshi (1.20)}, {(8) Hori – (5) Okamoto (0.93)}, {(3) Ariyoshi – (5) Okamoto (0.31)},

and the other is:

{(10) Kawabata – (7) Oe (1.90)}, {(10) Kawabata – (4) Nosaka (0.78)}, {(10) Kawabata – (9) Mushakoji (1.94)}, {(2) Miura – (6) Kaiko (1.24)}, {(4) Nosaka – (7) Oe (1.35)}, {(4) Nosaka – (9) Mushakoji (1.41)}, {(6) Kaiko – (9) Mushakoji (1.41)}, {(7) Oe – (9) Mushakoji (0.07)}.

The improved ordering of authors is now

(1) Kobayashi $>$ {(8) Hori, (3) Ariyoshi, (5) Okamoto} $>$ {(10) Kawabata, (2) Miura, (4) Nosaka, (6) Kaiko, (7) Oe, (9) Mushakoji}.

The tests in Table 4 show the mutual relationships of all writers. Since the value of *u* indicates the dissimilarity between authors, the most central is that author whose sum of dissimilarities is the smallest. Using Table 4 and adding all values of each writer we obtain the ordering as shown in Table 5:

Table 5. Sum of *u* calculated from *Modified TC*

Author	*Modified TC*	*Var* (*Modified TC*)	Sum of *u*
(1) Kobayashi	0.005455	0.0000002654	29.99
(8) Hori	0.004096	0.0000000073	29.95
(3) Ariyoshi	0.004304	0.0000000229	28.94
(5) Okamoto	0.004241	0.0000000173	28.87
(10) Kawabata	0.002786	0.0000001081	26.55
(2) Miura	0.003724	0.0000000010	26.20
(4) Nosaka	0.003099	0.0000000528	25.46
(6) Kaiko	0.003642	0.0000000034	25.10
(7) Oe	0.003449	0.0000000142	24.12
(9) Mushakoji	0.003460	0.0000000133	24.11

In general, the degree of thematic concentration depends on the text type, that is, scientific texts have greater *TC*, and art, social sciences poetry follow it ([18]: 100). Prose has relatively smaller *TC*.

5 Further work

We intend to investigate more texts to find a stable function for the relationships of *TC* and the number of different words.We also suppose that *TC* has a relationship to other properties, too, but to find them is a task for future research. We also intend to compare our results with those obtained from other languages and other types of text.

References

1. Ariyoshi, S.: Sumi (Ink Stick). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1961)
2. Graduate Schools of Informatics in Kyoto University and NTT Communication Science Laboratories: MeCab (Software, version 0.97),
   http://mecab.sourceforge.net/
3. Hirsch, J. E.: An index to quantify an individual's scientific research out put. Proceedings of the National Academy of Science of the United States of America, 102(46), 16569–16572 (2005)
4. Hori, T.: Utsukushii Mura (Beautiful Village). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1933)
5. Hřebíček, L.: Lectures on Text Theory. Oriental Institute, Prague (1997).
6. Kaiko, K.: Panikku (Panic). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1957)
7. Kawabata, Y.: Yukiguni (Snow Country). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1932)
8. Kobayashi H.: Guzo suhai (Idolatry). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1950)
9. Miura, T.: Kikyo (Going home). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1962)
10. Mushakoji, S.: Yujo (Friendship). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1920)
11. National Institute for Japanese Language and Linguistics: UniDic (Software, version 1.3.9), http://www.tokuteicorpus.jp/dist/
12. Nosaka, K.: Amerika Hijiki (American "Hijiki"). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1967)
13. Oe, K.: Shiiku (Prize Stock). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1957)

14. Okamoto, K.: Pari Sai (Quatorze Juillet). In: Shinchosha (ed.) Shincho Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shincho Bunko series). Shinchosha, Tokyo (1995). (The first edition is in 1938)
15. Popescu, I.-I.: Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.) Exact methods in the study of language and text, pp. 553–562. Mouton de Gruyter, Berlin-New York (2006)
16. Popescu, I.-I.; Altmann, G.: Some aspects of word frequencies. Glottometrics 13, 23–46 (2006).
17. Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.: Word frequency studies. Mouton de Gruyter, Berlin-New York (2009)
18. Popescu, I.-I., Mačutek, J., Altmann, G.: Aspects of Word Frequencies. RAM-Verlag, Lüdenscheid (2009)
19. Popescu, I.-I., Mačutek, J., Altmann, G.: Word forms, style and typology. Glottotheory 3(1), 89–96 (2010)
20. Sanada, H.: *h*-shihyo wo mochiita goi no keiryoteki bunseki (Quantitative Analysis of Japanese Vocabulary using *h*-index). Rissho Daigaku Keizaigaku Kiho (The quarterly report of economics of Rissho University), 6(3/4), 95–110 (2011)
21. Sanada, H.: Vocabulary richness in some Japanese texts. In: Altmann, G., Grzybek, P., Naumann, S., Vulanović, R. (eds.) Synergetic Linguistics: Text and Language as Dynamic Systems, pp.197–208. Praesens, Wien (2012)
22. Ziegler, A., Altmann, G.: Denotative Textanalyse. Praesens, Wien (2002)

PART IV

METHODOLOGICAL ISSUES

Empirical Evidence for Hilberg's Conjecture in Single-Author Texts

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
ldebowsk@ipipan.waw.pl

Abstract. Hilberg's conjecture is a statement that the mutual information between two adjacent blocks of text in natural language scales as n^β, where n is the block length. Previously, this hypothesis has been linked to Herdan's law on the levels of word frequency and of text semantics. Thus it is worth a direct empirical test. In the present paper, Hilberg's conjecture is tested for a selection of English prose using the Lempel-Ziv algorithm. An upper bound for the exponent β is found to be 0.949.

Keywords: single-author texts, universal coding

1 Introduction

Texts typically produced by humans diverge from both pure randomness and simple determinism. If we investigate predictability of such texts borrowing tools from information theory, we should observe some particular behavior of their optimal compression rate. Namely, the compression rate as a function of the text length should neither tend very quickly to zero (the case of determinism) nor tend very quickly to a constant greater than zero (the case of pure randomness). Concurring with this intuition, German telecommunications engineer Wolfgang Hilberg [7] supposed that the optimal compression rate of a text in natural language scales as $n^{-1+\beta}$, where n is the length of the text and β is close to 0.5. Hilberg's conjecture was motivated largely rationally but was partly based on an extrapolation of Shannon's seminal experimental data [10], which contained the estimates of conditional entropy for blocks of $n \leqslant 100$ characters.

As can be easily shown, Hilberg's conjecture implies that mutual information between two adjacent text blocks of length n is proportional to n^β. Using more involved mathematical modeling, the latter property can be linked with the distribution of words appearing in texts and the distribution of facts described by texts.

First, Dębowski [4] has proved a theorem by which the power-law growth of mutual information implies that the number of distinct set phrases (words) in a text of length n roughly exceeds n^{β} divided by a logarithmic term, cf. Dębowski [5]. The claim of this theorem is actually observed and known as Herdan's law [6]. Second, Dębowski [4] has proved a proposition which says that the power-law growth of mutual information is obeyed if a text of length n describes more than n^{β} independent facts in a repetitive fashion. Hence Hilberg's conjecture may be linked to power-laws on the levels of word frequency and of text semantics.

In view of these mathematical results, Hilberg's conjecture deserves experimental validation. Whereas it seems dubious that the optimal compression rate or the conditional entropy tends to zero for (con)text lengths tending to infinity, it is plausible that the mutual information between large adjacent text blocks grows according to a power law. Ever since Shannon the entropy of natural language has often been the object of scientific investigation, but the exact scaling of the compression rate is a little-investigated issue. Therefore, we decided to devote the present paper to the specific topic of verifying Hilberg's conjecture. The findings of this paper have a preliminary character.

Reviewing earlier research, we first mention Cover and King [2], who found an estimate of the asymptotic conditional entropy of English texts as 1.25 bpc (bits per character). This estimate was obtained using human subjects who were instructed to gamble on consecutive letters of the text and an estimate of entropy was computed from the accumulated capital. Modern compression algorithms compare with these estimates favorably. PPM (prediction by partial matching), being one of the best-performing compression algorithms, achieves the compression rate of 1.46 bpc for selected English texts [11]. Similar studies have been done for languages other than English, cf. e.g. Behr et al. [1], and for other compression algorithms, cf. Mahoney [9].

A graph depicting how PPM's compression rate depends on the amount of training text is also given by Teahan and Cleary [11]. We are looking for somewhat different graphs, namely, how the compression rate and the block mutual information depend on the amount of compressed text. For this reason we have decided to perform an independent compression experiment. In any such experiment there are two variables to be fixed. The first one is the compression algorithm, the second is the selection of texts.

For simplicity, we evaluate the compression rate and the mutual information using the Lempel-Ziv code [12], which is the simplest of universal codes. Universal codes are compression algorithms which asymptotically get the optimal compression rate for stationary sources. It can be shown that the estimates of mutual information given by universal codes are greater than the true mutual information. Moreover, the difference between the estimate of mutual information and the true mutual information is the smaller, the better the compression rate is. Hence we may use a universal code to upper bound the true mutual information.

Another important issue is the range of texts for which Hilberg's conjecture can reasonably be verified. One can consider either single (i.e., nonconcatenated) texts produced by single authors or concatenations of such texts (i.e., corpora). We have decided to consider first only single-author texts since the compression rate for

concatenated texts may depend on the specific choice of the text collection. Thus we consider a selection of single-author texts in English downloaded from the Project Gutenberg.

In a nutshell, the findings of this paper can be described thus. In the range of text lengths n from 10^3 to 10^7 characters, we observe a power-law relationship for both the compression rate and the mutual information computed for the Lempel-Ziv code. The fitted exponent for the compression rate is close to $\beta \approx 0.949$. This observation does not exclude Hilberg's conjecture with a very high exponent β. However, if we used a better universal code then we might obtain a tighter bound. For this reason it is advisable to repeat our experiment using better codes than the Lempel-Ziv, such as the PPM code.

The subsequent organization of the paper is as follows. In Section 2, we introduce some necessary concepts from information theory. In Section 3, we outline Hilberg's conjecture. In Section 4, we reformulate this conjecture using mutual information. In Section 5, we discuss the experiment. The paper is concluded in Section 6.

2 A bit of information theory

We first give a brief primer on information theory, cf. Cover and Thomas [3]. The fundamental concept of information theory is the entropy of a random variable. For a random variable $X_1^n = (X_1, X_2, \ldots, X_n)$, where X_i are consecutive characters of a random text, the entropy is defined as

$$H(X_1^n) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n). \tag{1}$$

If we have a uniquely decodable code C for variable X_1^n, then the expectation of its length $|C(X_1^n)|$ cannot be smaller than the entropy, i.e.,

$$\sum_{x_1^n} P(X_1^n = x_1^n)|C(x_1^n)| \geqslant H(X_1^n). \tag{2}$$

It can be shown that there exist a uniquely decodable code C with lengths $|C(x_1^n)| \leqslant -\log P(X_1^n = x_1^n) + 1$, called the Shannon-Fano code. For this code we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n)|C(x_1^n)| \leqslant H(X_1^n) + 1. \tag{3}$$

The length of the Shannon-Fano code $|C(x_1^n)|$ could be considered the information content of an individual text x_1^n.

However, we cannot evaluate the Shannon-Fano code if a proper probability distribution P is not specified or does not exist. As noticed by Kolmogorov [8], this may be well the case of natural language. In such a case, Kolmogorov proposed to define the information content of an individual text x_1^n as the length of the shortest program

for a simple universal computer (a Turing machine) that makes the computer produce x_1^n on its output. This quantity is called Kolmogorov complexity $K(x_1^n)$. For any computable code C there exists a constant c such that

$$K(x_1^n) \leqslant |C(x_1^n)| + c. \tag{4}$$

Since Kolmogorov complexity is itself a length of a computable code, we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n) K(x_1^n) \geqslant H(X_1^n) \tag{5}$$

for a random variable X_1^n on a definite probability space. In case of a computable probability distribution the Shannon-Fano code is also computable so, from (3) and (4), we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n) K(x_1^n) \leqslant H(X_1^n) + c + 1. \tag{6}$$

Hence the expectation of Kolmogorov complexity for computable distributions is close to entropy. In contrast, the difference between Kolmogorov complexity and entropy can be arbitrarily large for noncomputable distributions.

The problem with Kolmogorov complexity is, however, that it is not computable. Therefore we will rather take a middle path to measuring information content of individual texts, which is universal coding. A universal code is a uniquely decodable computable code C which for any stationary stochastic process $(X_1, X_2, \ldots)$ achieves the optimal compression rate

$$\lim_{n\to\infty} \frac{1}{n} \sum_{x_1^n} P(X_1^n = x_1^n) |C(x_1^n)| = h, \tag{7}$$

where the asymptotic entropy rate is

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1^n). \tag{8}$$

Some example of a universal code is the Lempel-Ziv code [12]. Subsequently, we will measure the information content of an individual text as the length of this code.

3 Flavors of Hilberg's conjecture

We are now in a position to introduce Hilberg's conjecture. The original form of this hypothesis deals with conditional entropy

$$H(X_n/X_1^{n-1}) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_n = x_n/X_1^{n-1} = x_1^{n-1}). \tag{9}$$

Hilberg replotted Shannon's (1951) estimates of conditional entropy for English in the double logarithmic scale and observed an approximate power-law relationship

$$H(X_n/X_1^{n-1}) \propto n^{-1+\beta}, \tag{10}$$

where $\beta \approx 0.5$ and $n \leqslant 100$. When extrapolated to arbitrary n, this relationship implies

$$H(X_1^n) = \sum_{m=1}^{n} H(X_m/X_1^{m-1}) \propto \int m^{-1+\beta} dm \propto n^{\beta}. \tag{11}$$

Hence we obtain a power law for the entropy rate

$$\frac{H(X_1^n)}{n} \propto n^{-1+\beta}. \tag{12}$$

Relationship (12) is the original Hilberg conjecture.

The original Hilberg conjecture is a bit far-fetched. Having derived (12), Hilberg conjectured that the entropy rate (8) of natural language is zero. This proposition seems unrealistic since it implies asymptotic determinism of human utterances. Thus it may be better to assume

$$\frac{H(X_1^n)}{n} \approx An^{-1+\beta} + h, \tag{13}$$

where constant h can be positive.

Striving for even more realism, we notice that there is no good probability distribution for texts in natural language. Hence, it seems more correct to speak of Kolmogorov complexity$K(x_1^n)$ of an individual text x_1^n rather than the entropy $H(X_1^n)$ of a random text X_1^n. Thus another plausible modification of Hilberg's conjecture reads

$$\frac{K(x_1^n)}{n} \approx An^{-1+\beta} + h. \tag{14}$$

This proposition may be called a relaxed Hilberg conjecture for individual texts. In the following, we will try to check whether (14) applies to texts in natural language. Prior to this, we will however discuss some bounds for mutual information that arise for universal codes.

4 Bounds for mutual information

It is insightful to rephrase Hilberg's conjecture using mutual information. There are three kinds of mutual information that are important for our considerations. First, the Shannon mutual information between random blocks is defined as

$$I_H(X_1^n; X_{n+1}^{2n}) = H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}). \tag{15}$$

Second, the algorithmic mutual information between individual texts is defined as

$$I_K(x_1^n; x_{n+1}^{2n}) = K(x_1^n) + K(x_{n+1}^{2n}) - K(x_1^{2n}). \tag{16}$$

Third, the mutual information based on a universal code C is

$$I_C(x_1^n; x_{n+1}^{2n}) = |C(x_1^n)| + |C(x_{n+1}^{2n})| - |C(x_1^{2n})|. \tag{17}$$

The nice feature of mutual information is that when we rephrase the modified Hilberg conjecture using this concept then the linear terms will cancel. Assuming that $H(X_1^n) \approx H(X_{n+1}^{2n})$, for the conjecture (13) we have

$$I_H(X_1^n; X_{n+1}^{2n}) = 2An^\beta + 2hn - A(2n)^\beta - 2hn \propto n^\beta. \tag{18}$$

Similarly, supposing that $K(x_1^n) \approx K(x_{n+1}^{2n})$, for the conjecture (14) we obtain

$$I_K(x_1^n; x_{n+1}^{2n}) = 2An^\beta + 2hn - A(2n)^\beta - 2hn \propto n^\beta. \tag{19}$$

In our application, we are going to estimate the algorithmic mutual information $I_K(x_1^n; x_{n+1}^{2n})$ using the code-based mutual information $I_C(x_1^n; x_{n+1}^{2n})$. It is important to know what an error is that we make by such an approximation. In determining this error the following proposition is helpful:

Lemma 1 [4]. *Let a function G satisfy $\lim_{k\to\infty} G(k)/k = 0$ and $G(n) \geqslant 0$ for all n. Then $2G(n) - G(2n) \geqslant 0$ for infinitely many n.*

A nice feature of universal codes, which follows from the above lemma, is that they yield an upper bound for the Shannon mutual information. Consider a stationary process $(X_1, X_2, \ldots)$. By Lemma 1, from (2) and (7), we obtain

$$\sum_{x_1^n} P(X_1^{2n} = x_1^{2n}) I_C(x_1^n; x_{n+1}^{2n}) \geqslant I_H(X_1^n; X_{n+1}^{2n}) \tag{20}$$

for infinitely many n. For the algorithmic mutual information, we can obtain a similar statement. Consider an infinite individual text $(x_1, x_2, \ldots)$. Suppose plausibly that $|C(x_1^n)| \approx |C(x_{n+1}^{2n})|$, $K(x_1^n) \approx K(x_{n+1}^{2n})$, and

$$\lim_{n\to\infty} \frac{1}{n}(C(x_1^n)) = \lim_{n\to\infty} \frac{1}{n} K(x_1^n). \tag{21}$$

Then by Lemma 1, from (4) we obtain

$$I_C(x_1^n; x_{n+1}^{2n}) \geqslant I_K(x_1^n; x_{n+1}^{2n}) \tag{22}$$

for ininitely many n. Hence when $I_C(x_1^n; x_{n+1}^{2n})$ obeys a power law with a given exponent then $I_K(x_1^n; x_{n+1}^{2n})$ may only obey a power law with a smaller exponent.

The bound given in (22) is the tighter, the better the code compresses the data. Suppose that we have a code D that satisfies $|D(x_1^n)| \leqslant |C(x_1^n)|$ and the analogue of (21). Then by Lemma 1, we have

$$I_D(x_1^n; x_{n+1}^{2n}) \leqslant I_C(x_1^n; x_{n+1}^{2n}) \tag{23}$$

for infinitely many n. Hence if we look for a good estimate of algorithmic mutual information, we should use the shortest code available.

5 Empirical findings

For the sake of testing Hilberg's conjecture, we have compressed 10 texts written in English by single authors. The texts were downloaded from Project Gutenberg [10] and are listed in Table 1. We have deleted the preambles of the text files and reduced the alphabet to 27 symbols (26 capital letters and a space), as it has been usually done in previous publication concerning the entropy of English. Subsequently, we have measured the length of the Lempel-Ziv code for exponentially growing initial text blocks.

The dependence of the compression rate on the block length is given in Figure 1, whereas the dependence of the mutual information on the double block length is given in Figure 2. Using the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm, we have fitted the following simple model for the compression rate:

$$\frac{|C(x_1^n)|}{n} \approx 6.22 n^{-1+0.949} [bpc]. \tag{24}$$

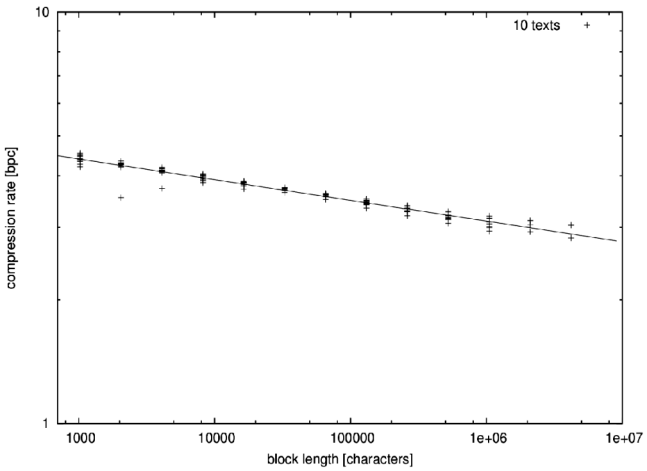


Fig. 1. Compression rate vs. block length. The solid line corresponds to the curve (24)

From formula (24), we derive mutual information

$$I_C(x_1^n; x_{n+1}^{2n}) \approx 0.432 n^{0.949} [bits]. \tag{25}$$

In Figures 1 and 2, we can observe that both models fit the data very well.

It may be somewhat surprising that model (24) fits so well although it contains no constant term $h > 0$ supposed in conjecture (14). We know, however, from independent studies that the asymptotic entropy rate h for English is less than 1.25 bpc [2].

In contrast, the lowest compression rate that we observe in Figures 1 is about 3.0 bpc. Thus a constant term of the order of 1.25 bpc cannot be reliably identified in the considered data.

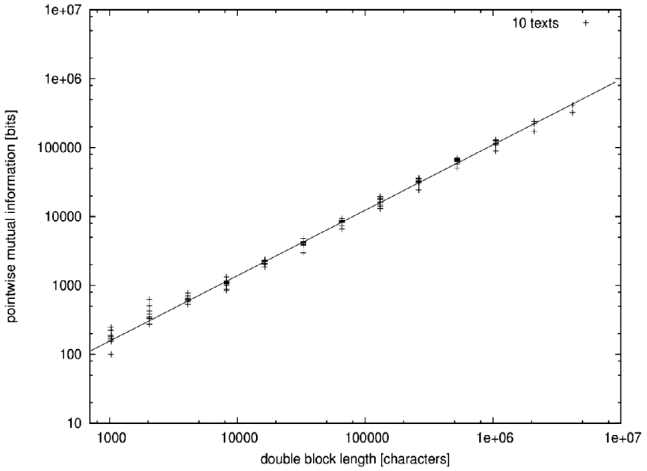


Fig. 2. Mutual information vs. double block length. The solid line corresponds to the curve (25)

Relationships (22) and (25) suggest that this bound holds for the algorithmic mutual information of texts in English:

$$I_K(x_1^n; x_{n+1}^{2n}) \leqslant 0.432n^{0.949} + c[bits]. \tag{26}$$

The above relationship does not exclude Hilberg's conjecture with a very high exponent β.

Table 1. The selection of compressed texts

Title	Author
First Folio/35 Plays	W. Shakespeare
Critical & Historical Essays	T. B. Macaulay
The Complete Memoirs	J. Casanova
Memoirs of Comtesse du Barry	E. Lamothe-Langon
The Descent of Man	C. Darwin
Gulliver's Travels	J. Swift
The Mysterious Island	J. Verne
Mark Twain, a Biography	A. B. Paine
The Journal to Stella	J. Swift
Life of William Carey	G. Smith

6 Conclusion

In this paper, we have first presented an approach how to understand Hilberg's conjecture using Kolmogorov complexity and algorithmic mutual information. Putting Hilberg's conjecture in this setting escapes the problem of deciding what is an appropriate probability distribution for human language production. In the second turn, we have tried to verify Hilberg's conjecture using the Lempel- Ziv code. Our findings do not exclude Hilberg's conjecture with an exponent β close to 1. However, if we used a better universal code than the Lempel-Ziv code, such as the PPM code, then we might obtain a tighter bound for the exponent. We leave this problem for the future research.

References

1. Behr, F., Fossum, V., Mitzenmacher, M., Xiao, D.: Estimating and comparing entropy across written natural languages using PPM compression. Tech. Rep. TR 12-02, Harvard University (2002)
2. Cover, T.M., King, R.C.: A convergent gambling estimate of the entropy of English. IEEE Transactions on Information Theory 24, 413–421 (1978)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd ed. New York: John Wiley (2006)
4. Dębowski, Ł.: On the vocabulary of grammar-based codes and the logical consistency of texts. IEEE Transactions on Information Theory 57, 4589–4599 (2011)
5. Dębowski, Ł.: Maximal lengths of repeat in English prose. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.) Synergetic Linguistics. Text and Language as Dynamic System, pp. 23–30. Wien: Praesens Verlag (2012)
6. Herdan, G.: Quantitative Linguistics. London: Butterworths (1964)
7. Hilberg, W.: Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente? Frequenz 44, 243–248 (1990)
8. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission 1(1), 1–7 (1965)
9. Mahoney, M.V.: Text compression as a test for artificial intelligence. In: Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference, AAAI'99/IAAI'99, p. 970 (1999)
10. Project Gutenberg, http://www.gutenberg.org/
11. Shannon, C.: Prediction and entropy of printed English. Bell System Technical Journal 30, 50–64 (1951)
12. Teahan, W.J., Cleary, J.G.: The entropy of English using PPM-based models. In: Proceedings of the Conference on Data Compression, DCC'96. pp. 53–62. Washington, DC: IEEE Computer Society (1996)
13. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Transactions on Information Theory 23, 337–343 (1977)

An Assessment of String Similarity Methods for Cognate Identification

Antonella Delmestri[1], Liviu P. Dinu[2]

[1]C.T.S.U., University of Oxford
Richard Doll Building, Old Road Campus, Roosevelt Drive, Oxford, OX3 7LF, U.K.
Antonella.Delmestri@ctsu.ox.ac.uk

[2]Faculty of Mathematics and Computer Science, Center for Computational Linguistics
University of Bucharest, Str. Academiei 14, 010014, Bucharest, Romania
LDinu@fmi.unibuc.ro

Abstract. We investigate and compare the performance of several manually-designed procedures and data-driven models in the task of cognate identification. The static schemes include *Manhattan Distance*, *Jaro and Jaro/Winkler Distances* and *ALINE*. The learning methods consist of *Pair Hidden Markov Models*, *Dynamic Bayesian Networks* and a measuring string similarity system, which generates substitution matrices using several techniques. Utilising *Point Accepted Mutation-like* matrices and orthographic data, this learning system shows superior performance and higher consistency across different Indo-European language pairs, when assessed against all comparable phonetic and orthographic methods reported in the literature. This result proves to be statistically significant and impressively stable when increasing the training dataset dimension by a factor of approximately 100. This suggests that learning algorithms outperform static procedures in cognate identification. Moreover, this outcome reinforces the hypothesis that orthographic learning methods may accurately detect traces of sound changes left in the orthography and outperform static phonetic systems.

Keywords: Cognate identification, scoring matrices, string similarity measures.

1 Introduction

Languages that originate from a common ancestor are genetically related and the study of language relatedness has been historically based on the detection of strict or genetic *cognates* [1], which are words that derive from the same predecessor through a *vertical* transmission [2] and share an identical etymological origin. In other contexts, such as many disciplines of natural language processing, the term *cognates* has a wider meaning and also includes *borrowings* [1], which are words loaned from other languages through a *horizontal* transmission [2].

Dialectology [3-6], proto-language reconstruction [7-10] and phylogenetic inference [11-20] are tasks of computational historical linguistics where cognate identification has been successfully applied. Areas of natural language processing that have benefited from cognate identification include semantic word clustering [21], bilingual lexicography [22-23], machine translation [24-25], lexicon induction [26-29], parallel corpora sentence alignment [30-33], parallel corpora word alignment [34-35], cross-lingual information retrieval [36] and confusable drug name detection [37]. Many different approaches to the cognate identification problem have been proposed: static procedures and learning systems have been used as well as orthographic and phonetic models.

2 Manually-Designed Procedures

We have applied and compared, in the task of cognate identification, four different static methods: the *Manhattan Distance* [38], the *Jaro Distance* [39], the *Jaro/Winkler Distance* [40] and the phonetic aligner *ALINE* [41].

2.1 Manhattan Distance

The *Manhattan Distance* (*MD*) [38], known with several different names including *taxicab metric, city-block distance* and *rectilinear distance*, is a metric on $\Re^n$ that calculates the distance between two points in an *n*-dimensional space as the sum of the absolute differences of their coordinates. Formally, if $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are *n*-dimensional real vectors, *MD* is defined as:

$$MD(p,q) = \sum_{i=1}^{n} |p_i - q_i|. \tag{1}$$

In order to apply the *Manhattan distance* to the calculation of word distance, we have employed the Roman alphabet and we have prepared a vectorial representation of each word in $\aleph_0^{26}$ through the computation of the occurrences of each letter, where 0 means no occurrence.

2.2 The Jaro Distance and the Jaro/Winkler Distance

The *Jaro Distance* (*JD*) [39] and its variant, the *Jaro/Winkler Distance* (*JWD*) [40], have been designed to calculate the similarity between short strings, such as in name-matching tasks, and they have been adopted in probabilistic record linkage. These distances, based on the number and order of common characters between two strings, are normalised to reach rates in the range [0, 1] so that, the higher the distance, the lower the similarity. Given two strings $S_1 = (a_1, a_2, \ldots, a_m)$ and $S_2 = (b_1, b_2, \ldots, b_n)$, the number of common characters in S_1 and S_2 is the number of characters in S_1 that satisfy the following:

$$a_i \text{ is common in } S_2 \leftrightarrow \exists j : a_i = b_j \text{ and } |i - j| \leqslant \frac{\max(m,n)}{2} - 1. \tag{2}$$

The number of character transpositions in S_1 and S_2 is the number of common characters, but in different positions, divided by 2. Formally, if c is the number of common characters and t the number of character transpositions, JD is defined by:

$$JS(S_1,S_2) = \frac{1}{3}\left(\frac{c}{m} + \frac{c}{n} + \frac{c-t}{c}\right). \quad (3)$$

The variant *JWD* [40] considers also the length L of the longest common prefix of S_1 and S_2 up to a maximum of 4 characters and a scaling factor $P = 0.1$, so that:

$$JWD(S_1,S_2) = JD(S_1,S_2) + L * P * (1 - JD(S_1,S_2)). \quad (4)$$

2.3 ALINE

ALINE [41] is a manually-designed algorithm developed by *Kondrak* for phonetic sequence alignment. The aligner represents phonetic segments as vectors of feature values and calculates their similarity through a procedure of local alignment, performed by dynamic programming. Twelve phonetic features are considered and weighted according to their salience, which was established manually by trial and error. For example, *Place* and *Manner* are the most significant features and they are assigned higher weights than less important features, like *High* and *Long*. The numerical values of each feature are based on data reported in the literature with the aim of reflecting the distances between vocal organs during verbal emission.

3 Data-Driven Models

We have considered three learning methods in the task of cognate identification: *Pair Hidden Markov Models* [42-43], *Dynamic Bayesian Networks* [44] and a measuring string similarity system [45], which uses several training techniques, including *Absolute Frequency Ratio* [46], *Pointwise Mutual Information* [6] and *Point Accepted Mutation*-like (*PAM-like*) [47].

3.1 Pair Hidden Markov Models

Mackay [42] developed a cognate identification orthographic learning system using *Pair Hidden Markov Models* (*PHMMs*). His system was inspired by a model for biological sequence analysis originally introduced by *Durbin et al.* [48] and consisted of a suite of *PHMMs* utilising several alignment algorithms in order to calculate the word pair similarity. The training dataset of about 120,000 word pairs was extracted from the *Comparative Indo-European corpus* by *Dyen et al.* [11], described in Section 4. All the 95 languages present in the digital file were employed together with the reverse of each word pair, with the aim of avoiding possible bias due to the word ordering. To reduce the large dimension of the training dataset, *Mackay* finally considered only the word pairs where both words where at least 4 characters long. In order to determine several parameters for the model, the author built a development

dataset using two examples of language pairs showing distant and close relatedness, respectively: Italian and Serbo-Croatian; Polish and Russian. The test dataset consisted of 10 language pairs extracted from the 200-word *Swadesh lists* prepared by *Kessler* [49] for Albanian, English, French, German and Latin, described in Section 4.

Mackay and Kondrak [43] compared four of the *PHMMs* proposed by *Mackay* [42] with other systems in the task of cognate identification by employing the same test dataset that *Mackay* [42] used. The authors tested the *PHMMs* against the *Levenshtein distance with Learned Weights* (*LLW*) [26], *ALINE* [41], introduced in Section 2.3, and *Longest Common Subsequence Ratio* (*LCSR*) [33]. The authors showed that all the four *PHMMs* outperformed *LCSR*, *LLW* and *ALINE* in the task of cognate identification. The one that performed better in identifying cognate words utilised the *log-odds* version of the *Viterbi algorithm* [48], with uniform gap and transition probabilities, and showed a significant improvement compared with the others. This model is called hereinafter *PHMM* only.

3.2 Dynamic Bayesian Networks

Kondrak and Sherif [44], working on orthographic data in the task of cognate identification, developed four learning *Dynamic Bayesian Networks* based on a method previously proposed by *Filali and Bilmes* [50]. They prepared a training dataset of about 180,000 word pairs extracted from the *Comparative Indo-European corpus* by *Dyen et al.* [11]. In order to set several parameters for their model, the authors built up a development dataset composed of three language pairs representing distant, medium and close relatedness, respectively. The language pairs were: Italian-Croatian, Spanish-Romanian and Polish-Russian. The test dataset, extracted from the *Kessler lists* [49], was the same used by *Mackay* [42] and *Mackay and Kondrak* [43]. *Kondrak and Sherif* tested their *DBNs* in the task of cognate identification against other orthographic and phonetic systems, including *ALINE* [41] and *PHMM* [42-43]. Only one of the four *DBNs*, called hereinafter *DBN*, reached very good results in recognising cognates and outperformed the other models including *PHMM*, but not significantly.

3.3 A String Similarity Measuring System

Delmestri [45] developed an orthographic learning system for measuring string similarity, inspired by biological sequence analysis. The author extracted a training dataset of about 650 word pairs from the *Comparative Indo-European corpus* by *Dyen et al.* [11] considering only six languages (Italian, Portuguese, Spanish, Dutch, Danish and Swedish) and their word pairs classified as certain cognates. *Delmestri* then applied pairwise global alignment [51] to these cognate pairs, with the aid of a novel linguistic-inspired substitution matrix [45], [47] in order to produce a sensibly aligned training dataset. On it, the author used several learning techniques to infer increasingly complex scoring matrices, including *Absolute Frequency Ratio* [46], *Pointwise Mutual Information* [6] and *PAM-like* matrices [47]. The produced substitution matrices were then utilised to measure word similarity,

employing global and local alignment algorithms [51-52] and a novel family of parameterised string similarity measures [47]. These similarity measures were the result of different normalisations of a generic scoring algorithm considering the similarity of each string with itself, with the aim of reducing the bias due to different string length. Given a rating algorithm, there was no significant difference in the performance of these similarity measures, but all of them outperformed the basic algorithm on which they were based. The test dataset was the same one utilised by *PHMM* [42,43] and *DBN* [44].

Absolute Frequency Ratio Matrices

A simple statistical method for estimating a substitution matrix from aligned data is to calculate the absolute frequencies [46] of the characters present in the employed alphabet. Formally, given a set of aligned word pairs from an alphabet $\mathcal{A}$ with $|\mathcal{A}| \geqslant 2$, the *Absolute Frequency Ratio* (*AFR*) of the character pair $(\mathcal{A}_i, \mathcal{A}_j)$ is the observed absolute frequency of character $\mathcal{A}_i$ being transformed into character $\mathcal{A}_j$, divided by the absolute frequency of character $\mathcal{A}_i$ and character $\mathcal{A}_j$:

$$AFR(i,j) = AFT(A_i, A_j) = \frac{\#(A_i, A_j)}{\#A_i * \#A_j}. \tag{5}$$

Pointwise Mutual Information Matrices

Pointwise Mutual Information (*PMI*) is another statistical method for estimating a scoring matrix from aligned data. *PMI* is a measure of association between two events described by discrete probability distributions and derives from the *Mutual Information*, originally introduced by *Fano* [53] in the field of information theory. *PMI* is defined as the *log-odds ratio* of the joint probability of observing two events together, to the marginal probabilities of observing them independently. Formally, given a dataset of aligned words from an alphabet $\mathcal{A}$, with $|\mathcal{A}| \geqslant 2$, each entry $PMI(i,j)$ of the substitution matrix is obtained by *the log-odds ratio* of the joint relative frequencies of the two characters $\mathcal{A}_i$ and $\mathcal{A}_j$, over the product of their disjoint relative frequencies:

$$PMI(i,j) = \log_2 \frac{f(i,j)}{f(i) * f(j)}, \tag{6}$$

$$f(i,j) = \frac{\#(A_i, A_j)}{\sum_{h,k} \#(A_h, A_k)}, \tag{7}$$

$$f(i) = \frac{\#A_i}{\sum_h A_h}. \tag{8}$$

PAM-like Matrices

The *PAM* method [54-56] has been successfully designed for amino acid sequence analysis by *Margaret Dayhoff* and co-workers. This approach assumes a constant rate of evolution and, by inferring substitution parameters from global alignments between closely related sequences, extrapolates from them longer evolutionary divergences. The *PAM-like* model [47] derives from it, though including some adaptations to the linguistic environment due to the differences existing between cognate

words and biological sequences (e.g. the amount and structure of available data, the string length, etc.). Formally, the construction of a family of *PAM-like* matrices is based on a non-symmetric matrix M of mutation probabilities $M(i,j)$ containing the probability that character $\mathcal{A}_j$ mutates to character $\mathcal{A}_i$ in 1 *PAM* unit:

$$M(i,j) = \frac{\mu * m(j) * A(i,j)}{\sum_i A(i,j)} \quad \forall i \neq j, \tag{9}$$

$$M(i,i) = 1 - \mu * m(i) \quad \forall i. \tag{10}$$

where μ is a proportionality constant set to 1; $m(j)$ is the relative mutability of each character $\mathcal{A}_j$, calculated as the ratio of observed changes to the frequency of occurrence; A is a matrix of accepted point mutation, where $A(i,j)$ and $A(j,i)$ were incremented every time character $\mathcal{A}_i$ was replaced by $\mathcal{A}_j$ or vice-versa. This matrix M is multiplied by itself n times to calculate the probability that any particular character mutates to another one in *n PAM units*. The following *log-odds ratio* allows the *PAMn* matrix to be obtained, where $f(i)$ and $f(j)$ are the observed frequencies of character $\mathcal{A}_i$ and $\mathcal{A}_j$, normalized by the number of all mutations:

$$PAMn(i,j) = 10 * \log_{10} \frac{f(j) * M^n(i,j)}{f(i) * f(j)} = 10 * \log_{10} \frac{M^n(i,j)}{f(i)}. \tag{11}$$

In order to assess the robustness of the *PAM-like* approach, *Delmestri and Cristianini* [57] evaluated the influence of the training dataset dimension on the performance of the cognate identification system. They prepared a second training dataset extracted from the *Dyen et al. corpus* [11] introduced in Section 4, considering all the word pairs reported as certain cognates for all the 76 languages included in the monograph that did not overlap with the test dataset, described in Section 4. This second training dataset included a total of about 62,000 cognate pairs. The *PAM-like* matrices based on it showed an extremely similar performance to that achieved by the *PAM-like* matrices trained by the first training dataset of about 650 cognate pairs from 6 languages [57]. *Delmestri and Cristianini* also investigated and assessed the statistical significance of the *PAM-like* results for both the training datasets.

4 Experimental Design

In order to make an assessment of different systems in the task of cognate identification, we have chosen only comparable models using the same training database source, the same test dataset and the same evaluation methodology. The training and test datasets have no intersection between the languages of which they are composed.

The *training dataset* for all the systems considered has been extracted from the *Comparative Indo-European corpus* by *Dyen et al.* [11], which consists of *Swadesh lists* [58] of 200 universal, non cultural and stable meanings from contemporary Indo-European languages. The digital version of the dataset covers 95 speech varieties, but only 84 were considered accurate enough to be included in the monograph. The lexical data are represented in orthographic format using the 26 letters

of the Roman alphabet without diacritics. The words are clustered by meaning and cognateness, which are coded as certain or doubtful.

The *test dataset* consists of the 200-word *Swadesh lists* of English, German, French, Latin and Albanian prepared by *Kessler* [49], augmented by his cognateness information. It is provided in orthographic format with phonetic transcriptions, which allows the assessment of orthographic as well as phonetic systems.

The evaluation methodology used in this assessment addresses the problem of measuring the accuracy of cognate identification as a classification task, with the aim of comparing the system classifications with the correct cognateness judgements. We have employed the *11-point interpolated average precision* [59], which has been specifically designed to evaluate rankings in the field of *Information Retrieval* and has been frequently used in the evaluation of cognate recognition systems.

5 Experimental Results

We have compared the results achieved by the manually-designed procedures, introduced in Section 2, and by the data-driven models, presented in Section 3. The table below shows an assessment in the task of cognate identification of these comparable orthographic and phonetic methods in terms of *11-point interpolated average precision* [59] over ten language pairs of an Indo-European test dataset [49], introduced in Section 4. The table also displays the averaged *11-point interpolated average precision*, its *standard deviation*, *variance* and *median* [46].

Table 1. *11-point interpolated average precision* for several methods

Languages		Cognate proportion	MD	NEDIT	JD	JWD	ALINE	AFR	PHMM	DBN	PMI	PAM like
English	German	0.590	0.883	0.907	0.912	0.912	0.912	0.909	0.930	0.927	0.925	0.934
French	Latin	0.560	0.866	0.921	0.908	0.912	0.862	0.924	0.934	0.923	0.925	0.924
English	Latin	0.290	0.605	0.703	0.676	0.676	0.732	0.776	0.803	0.822	0.795	0.826
German	Latin	0.290	0.564	0.591	0.568	0.564	0.705	0.706	0.730	0.772	0.745	0.772
English	French	0.275	0.676	0.659	0.693	0.693	0.623	0.768	0.812	0.802	0.790	0.830
French	German	0.245	0.545	0.498	0.567	0.551	0.534	0.700	0.734	0.645	0.757	0.788
Albanian	Latin	0.195	0.440	0.561	0.566	0.566	0.630	0.584	0.680	0.676	0.676	0.721
Albanian	French	0.165	0.369	0.499	0.526	0.538	0.610	0.557	0.653	0.658	0.621	0.625
Albanian	German	0.125	0.244	0.207	0.233	0.242	0.369	0.486	0.379	0.420	0.470	0.552
Albanian	English	0.100	0.229	0.289	0.272	0.272	0.302	0.280	0.382	0.446	0.404	0.518
AVERAGE		0.284	0.542	0.584	0.592	0.593	0.628	0.669	0.704	0.709	0.711	0.749
Std. deviation		0.168	0.229	0.231	0.225	0.224	0.193	0.197	0.194	0.176	0.173	0.144
Variance		0.028	0.052	0.054	0.051	0.050	0.037	0.039	0.038	0.031	0.030	0.021
Median		0.260	0.555	0.576	0.568	0.565	0.627	0.703	0.732	0.724	0.751	0.780

The edit distance with unitary costs [60] normalised by the length of the longer string (*NEDIT*) has been used as a baseline. The *11-point interpolated average precision* achieved by *ALINE* [41], *PHMM* [43], *DBN* [44] and *PAM-like* is reported as in the literature. *PAM-like* [47] shows the best results achieved by the *PAM-like*

method with the first training dataset of about 650 cognate pairs, where the learnt parameters include the gap penalties and the *Smith-Waterman* algorithm [52] for local alignment is used to rate the word pairs.

The *Manhattan Distance* [38] produces a negative outcome showing a performance lower than *NEDIT* [60]. The *Jaro Distance* [39] and the *Jaro/Winkler Distance* [40] generate results only slightly higher than *NEDIT* and the *Absolute Frequency Ratio* [46] performs a little better than *ALINE* [41]. The *Pairwise Mutual Information* [6] reaches good results, which are comparable to those achieved by *PHMM* [43] and *DBN* [44]. The *PAM-like* method [47], [57] achieves the higher accuracy in cognate identification, with an averaged *11-point interpolated average precision* [59] approximately 5% higher than *PHMM* [43], *DBN* [44] and *PMI* [6], 18% higher than *ALINE* [41] and 28% higher than *NEDIT*. Not only the average of the *11-point interpolated average precision* and its median are higher, but also the standard deviation and variance [46] are much lower. This would suggest that the performance of the *PAM-like* system across the various language pairs of the test dataset is more stable than that achieved by the compared methods.

6 Conclusions

In the task of cognate identification, a phonetic approach is supposed to be more accurate than an orthographic one because of its insight and understanding of phonetic changes. However, several recent studies and comparative evaluations of different systems [43], [44], [47], [57] have shown that orthographic learning models may outperform phonetic static procedures. Our investigation has confirmed this tendency, estimating more static and learning systems on orthographic data. In particular, the *PAM-like* method has achieved a statistically significant higher accuracy in cognate identification regardless of the training dataset dimension, overcoming one of the limits of learning systems, which is the need for a large training dataset. This would suggest that phonetic changes can leave enough traces in the word orthography to be successfully utilised by orthographic learning systems. This idea is very constructive taking into account that accurate phonetic transcriptions are difficult to generate and are frequently produced manually, with the consequent loss of time and the potential lack of uniformity and accuracy. Moreover, this assessment shows that all the learning systems considered perform better than the static methods, even if the latter may have been specifically designed for the task of phonetic alignment.

Our future plans include the investigation of several other learning techniques developed for biological sequence analyses and their application to the tasks of cognate recognition. We are particularly interested in training *BLOSUM-like* matrices on the same training datasets employed by the *PAM-like* method and evaluating their performance on the test dataset described in Section 4.

Acknowledgments. We thank Grzegorz Kondrak for providing his version of the test dataset.

References

1. Kondrak, G.: Identification of Cognates and Recurrent Sound Correspondences in Word Lists. Traitement Automatique des Langues et Langues Anciennes, 50(2), 201-235 (2009)
2. Wang, W.S-Y., Minett, J.W.: Vertical and Horizontal Transmission in Language Evolution. Transactions of the Philological Society, 103(2), 121-146 (2005)
3. Kessler, B.: Computational Dialectology in Irish Gaelic. In Proceedings of EACL'95, pp. 60-66. Morgan Kaufmann Publishers Inc., San Francisco, California, USA (1995)
4. Nerbonne, J., Heeringa, W.: Measuring Dialect Distance Phonetically. In Proceedings of ACL SIGPHON-97, pp. 11-18 (1997)
5. Wieling, M., Leinonen, T., Nerbonne, J.: Inducing Sound Segment Differences using Pair Hidden Markov Models. In Proceeding of SIGMORPHON'07, pp. 48-56. ACL, Stroudsburg, PA, USA (2007)
6. Wieling, M., Prokić, J., Nerbonne, J.: Evaluating the Pairwise String Alignment of Pronunciations. In Proceedings of EACL LaTeCH-SHELT&R'09, pp. 26-34. ACL, Stroudsburg, PA, USA (2009)
7. Covington, M.A.: An Algorithm to Align Words for Historical Comparison. Computational Linguistics, 22(4), 481-496 (1996)
8. Covington, M.A.: Alignment of Multiple Languages for Historical Comparison. In Proceedings of COLING'98, pp. 275-280. ACL, Stroudsburg, PA, USA (1998)
9. Oakes, M.P.: Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages. Journal of Quantitative Linguistics, 7(3), 233-243 (2000)
10. Kondrak, G.: Algorithms for Language Reconstruction, University of Toronto, Canada, PhD Thesis (2002)
11. Dyen, I., Kruskal, J.B., Black, P.: An Indoeuropean Classification: A Lexicostatistical Experiment. Transactions of the American Philosophical Society, vol. 82, part 5 (1992)
12. Gray, R.D., Jordan, F.M.: Language Trees Support the Express-Train Sequence of Austronesian Expansion. Nature, 405, 1052-1055 (2000)
13. Ringe, D.A., Warnow, T., Taylor, A.: Indo-European and Computational Cladistics. Transactions of the Philological Society, 100(1), 59-129 (2002)
14. Gray, R.D., Atkinson, Q.D.: Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin. Nature, 426, 435-439 (2003)
15. Rexová, K., Frynta, D., Zrzavý, J.: Cladistic Analysis of Languages: Indo-European Classification Based on Lexicostatistical Data. Cladistics, 19(2), 120-127 (2003)
16. Nicholls, G.K., Gray, R.D.: Dated Ancestral Trees from Binary Trait Data and their Application to the Diversification of Languages. Journal of Royal Statistical Society, Series B: Statistical Methodology, 70(3), 545-566 (2008)
17. Serva, M., Petroni, F.: Indo-European Languages Tree by Levenshtein Distance. Europhysics Letters, 81(68005), 1-5 (2008)
18. Brown, C.H., Holman, E.W., Wichmann, S., Vilupillai, V.: Automated Classification of the World's Languages: a Description of the Method and Preliminary Results. STUF - Language Typology and Universals, 61(4), 285-308 (2008)
19. Downey, S.S., Hallmark, B., Cox, M.P., Norquest, P., Lansing, S.J.: Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. Journal of Quantitative Linguistics, 15(4), 340–369 (2008)
20. Delmestri, A., Cristianini, N.: Linguistic Phylogenetic Inference by *PAM-like* Matrices. Journal of Quantitative Linguistics, 19(2), 95-120 (2012)

21. Adamson, G.W., Boreham, J.: The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. Information Storage and Retrieval, 10(7-8), 253-260 (1974)
22. Brew, C., McKelvie, D.: Word-Pair Extraction for Lexicography. In Proceedings of NEMLAP, pp. 45-55 (1996)
23. Inkpen, D., Frunza, O., Kondrak, G.: Automatic Identification of Cognates and False Friends in French and English. In Proceedings of RANLP'05, pp. 251-257 (2005)
24. Guy, J.B.M.: An Algorithm for Identifying Cognates in Bilingual Word-Lists and its Applicability to Machine Translation. Journal of Quantitative Linguistics, 1(1), 35-42 (1994)
25. Kondrak, G., Marcu, D., Knight, K.: Cognates Can Improve Statistical Translation Models. In Proceeding of HLT-NAACL 2003, pp. 46-48. ACL, Stroudsburg, PA, USA (2003)
26. Mann, G.S., Yarowsky, D.: Multipath Translation Lexicon Induction via Bridge Languages. In Proceedings of NAACL'01, pp. 151-158. ACL, Stroudsburg, PA, USA (2001)
27. Koehn, P., Knight, K.: Knowledge Sources for Word-Level Translation Models. In Proceedings of EMNLP 2001, pp. 27-35 (2001)
28. Schulz, S., Markó, K., Sbrissia, E., Nohama, P., Hahn, U.: Cognate Mapping - A Heuristic Strategy for the Semi-supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In Proceedings of COLING'04, vol. 2, pp. 813-819. ACL, Stroudsburg, PA, USA (2004)
29. Mulloni, A., Pekar, V.: Automatic Detection of Orthographic Cues for Cognate Recognition. In Proceedings of LREC 2006, pp. 2387-2390 (2006)
30. Simard, M., Foster, G.F., Isabelle, P.: Using Cognates to Align Sentences in Bilingual Corpora. In Proceedings of TMI-92, pp. 67-81 (1992)
31. Church, K.W.: Char_align: a Program for Aligning Parallel Texts at the Character Level. In Proceedings of ACL-1993, pp. 1-8 (1993)
32. McEnery, A., Oakes, M.P.: Sentence and Word Alignment in the CRATER Project. In Using Corpora for Language Research, Thomas, J. and Short, M., Eds., pp. 211-231. Longman (1996)
33. Melamed, D.: Bitext Maps and Alignment via Pattern Recognition. Computational Linguistics, vol. 25, no. 1, pp. 107-130. MIT Press Cambridge, MA, USA (1999)
34. Tiedemann, J.: Automatic Construction of Weighted String Similarity Measures. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 213-219 (1999)
35. Kondrak, G.: Cognates and Word Alignment in Bitexts. In Proceedings of MT Summit X, pp. 305-312 (2005)
36. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., Järvelin, K.: Fuzzy Translation of Cross-Lingual Spelling Variants. In Proceedings of ACM SIGIR'03, pp. 345–352. ACM, New York, USA (2003)
37. Kondrak, G., Dorr, B.J.: Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In Proceedings of COLING 2004, pp. 952-958. ACL, Stroudsburg, PA, USA (2004)
38. Deza, E., Deza, M.M.: Dictionary of Distances. Elsevier (2006)
39. Jaro, M.A.: Advances in Record Linkage Methodology as Applied to the 1985 Census of Tampa Florida. Journal of the American Statistical Society, 84(406), 414-420 (1989)
40. Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In Proceedings of the Section on Survey Research, pp. 354-359 (1990)

41. Kondrak, G.: A New Algorithm for the Alignment of Phonetic Sequences. In Proceedings of ANLP-NAACL 2000, vol. 4, pp. 288-295. ACL, Stroudsburg, PA, USA (2000)
42. Mackay, W.: Word Similarity Using Pair Hidden Markov Models, University of Alberta, Master's thesis (2004)
43. Mackay, W., Kondrak, G.: Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models. In Proceedings of CONLL 2005, pp. 40-47 (2005)
44. Kondrak, G., Sherif, T.: Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In Proceedings of COLING-ACL 2006 Workshop on Linguistic Distances, pp. 43-50 (2006)
45. Delmestri, A.: Data Driven Models for Language Evolution. Lambert Academic Publishing (2011)
46. Ott, L.R., Longnecker, M.T.: An Introduction to Statistical Methods and Data Analysis, 5th ed. Duxbury Press, Pacific Grove, California, USA (2001)
47. Delmestri, A., Cristianini, N.: String Similarity Measures and PAM-like Matrices for Cognate Identification. Bucharest Working Papers in Linguistics, XII(2), 71-82 (2010)
48. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis. Cambridge University Press, Cambridge, UK (1998)
49. Kessler, B.: The Significance of Word Lists. CSLI Publications, Stanford, California, USA (2001)
50. Filali, K., Bilmes, J.: A Dynamic Bayesian Framework to Model Context and Memory in Edit Distance Learning: An Application to Pronunciation Classification. In Proceedings of ACL'05, pp. 338-345. ACL, Stroudsburg, PA, USA (2005)
51. Needleman, S.B., Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, 48(3), 443-453 (1970)
52. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147(1), 195-197 (1981)
53. Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, Massachusetts, USA (1961)
54. Dayhoff, M.O., Eck, R.V.: A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence and Structure 1967-1968, 3, 33-41 (1968)
55. Dayhoff, M.O., Eck, R.V., Park, C.M.: A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence and Structure, 5, 89-99 (1972)
56. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence and Structure, 5(3), 345-352 (1978)
57. Delmestri, A., Cristianini, N.: Robustness and Statistical Significance of PAM-like Matrices for Cognate Identification. Journal of Communication and Computer, 7(12), 21-31 (2010)
58. Swadesh, M.: Lexico-Statistic Dating of Prehistoric Ethnic Contacts. In Proceedings of the American Philosophical Society, vol. 96, no. 4, pp. 452-463 (1952)
59. Manning, C.D.: Schütze, H., Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts, USA (1999)
60. Gusfield, D.: Algorithms on Strings, Trees and Sequences. Cambridge University Press (1997)

Rank-frequency Relation & Type-token Relation: Two Sides of the Same Coin[1]

Jiří Milička

Institute of Comparative Linguistics, Charles University, Prague
Jana Palacha 2, Praha 1, 116 38
Milicka@centrum.cz

Abstract. This paper shows that the type-token relation, the hapax-token relation and, generally, the relation between types of certain frequency and tokens can be computed from the rank-frequency relation or from any type of frequency distribution and that the type-token relation can be computed from the hapax-token relation. This paper shows that there is no need for any approximation or assumptions and that the formulae can be derived purely algebraically. The second part of the paper observes that, for a very large corpora, the ratio between the number of hapax legomena and types converges to a constant Z; $Z > 0$. Under this assumption an approximation is built that enables us to predict the type-token relation and other aforementioned relations from the single parameter Z. This approximation is only valid for very large corpora. As the last section shows, this assumption implies that for an infinitely increasing number of tokens, the number of types increases beyond any limit.

Keywords: Type-token relation, hapax-token relation, rank-frequency relation, words frequency distribution.

1 Introduction

The type-token relation (TTR) and the rank-frequency relation (RFR) are two of the most popular ways to quantify a text. They are used in empirical linguistics, NLP, literary theory, etc. Many approximative models of these relations have been introduced since the dawn of quantitative linguistics. The most famous ones are Herdan's Law for TTR (mostly referred to as Heaps' Law) and Zipf's or the Zipf-Mandelbrot Law for RFR.

In the first two chapters of this paper we won't see those relations through the prism of any approximation; on the contrary, the first section shows that mere means of algebra are sufficient to transform any measured distribution of types (and thus RFR) into a TTR curve, or into a hapax legomena – token relation (T_1TR), or a dis legomena – token relation (T_2TR), or any other relation between the number of types of certain frequency and the number of tokens (T_gTR). In the second section

[1]This research was supported by GA UK 595212.

we use these formulae to derive a formula that exactly transforms a T_1TR curve into a TTR curve, and introduce some other formulae related to T_1TR and T_gTR.

The next section is based on these formulae and discusses consequences of an empirical observation that the ratio between the number of hapax legomena and the number of types asymptotically tends to a constant larger than zero. This works for even very large corpora.

2 Computation of TTR from a frequency distribution of types (or RFR)

The following formulae, which enable us to compute TTR, T_1TR and generally T_gTR of a text from a mere frequency distribution of types in a text, were derived 5 years ago [4]. In those days, the idea that it must be possible to compute TTR from RFR was quite common among researchers (e.g. [3]) but none of them approached this task without any assumptions. Dieter Müller even states that TTR cannot be derived from the general distribution without an approximation. [2]

But it can be. The formula transforming RFR (or any absolute frequency distribution of types, Zipfian or otherwise) into TTR is the following one[3]:

$$V(N) = \sum_{i=1}^{M} \left(1 - \frac{(d-N)!(d-f)i)!}{d!(d-N-f_i)!}\right) \tag{1}$$

We are able to compute T1TR from RFR using the following formula[4]:

$$V_1(N) = \sum_{i=1}^{M} \frac{\frac{N(d-N)!}{(f_i-1)!(d-N-f_i+1)!}}{\frac{d!}{f_i!(d-f_i)!}} \tag{2}$$

And in the most general case, we can compute the relation between types of a certain frequency and tokens (T_gTR) from a types distribution (RFR) according to this formula[5]:

$$V_g(N) = \sum_{i=1}^{M} \frac{\frac{N!}{g!(N-g)!}\frac{(d-N)!}{(f_i-g)!(d-N-f_i+g)!}}{\frac{d!}{f_i!(d-f_i)!}} \tag{3}$$

[2] "The general case of a vocabulary V with arbitrary type probabilities w_j requires an approximation" [5, page 204].

[3] $V(N)$ represents the number of types after measuring N tokens, i is a control variable that represents the order of a type in the lexicon, f_i is the number of occurrences (absolute frequency) of the type in the text, M is the total number of types in the text, d is the total number of tokens in the text.

[4] $V_1(N)$ represents the number of hapax legomena.

[5] $V_g(N)$ is the number of types represented by g tokens.

The main idea on which the formulae are based is that (technically), although we cannot make all permutations of a text and measure TTR etc. for these permutations and average them, we can simulate this process by means of algebra.

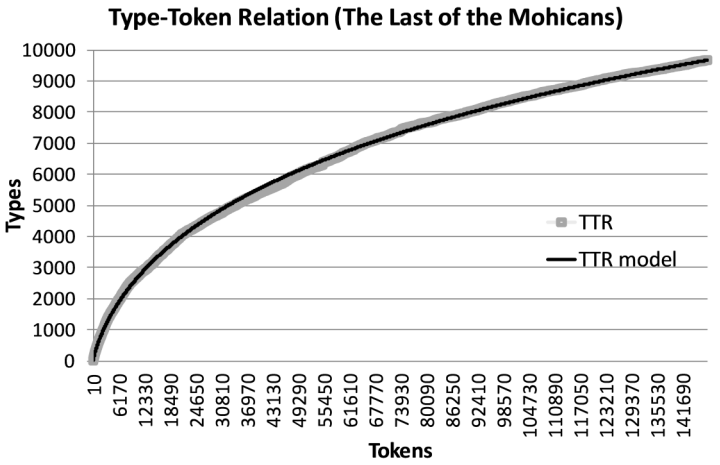


Fig. 1. Illustration of the usage of the model. Type-token relation measured and computed for The Last of the Mohicans by J. F. Cooper [9]

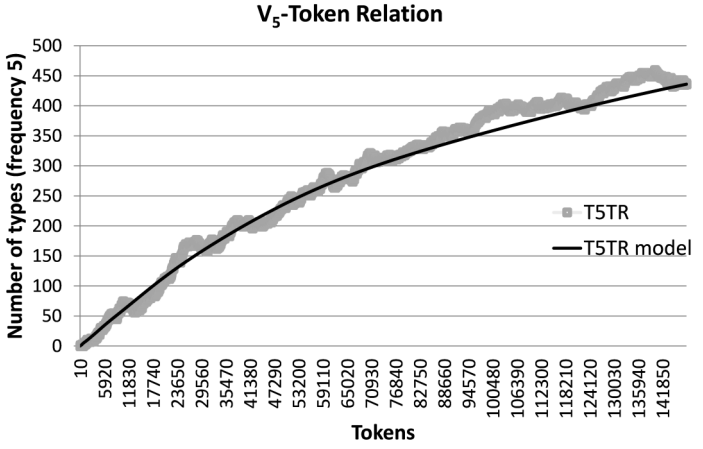


Fig. 2. Illustration of the usage of the model. The relation between types with frequency 5 and tokens, measured and computed for The Last of the Mohicans by J. F. Cooper [9]

The fact that the formula corresponds to the quantities measured on natural texts tells us that the text is a homogenous one, i.e. the writer chose words similarly as if he chose them randomly from a multiset of words.

How we arrived at these formulae can be found in [4]. They were introduced here because the next section is based on them.

3 Computation of TTR from T_1TR

The following formula is based on the previous ones and thus is also distribution-independent. It transforms T_1TR into TTR and vice versa exactly (without any approximation). The underlying idea is very simple:

Consider a pack consisting of the cards with these suits:

♥ ♠ ♣ ♥ ♦ ♣ ♥ ♥

And imagine that you take away one card randomly. The probability that the number of suits in the pack decreases is equal to the number of suits that are only once in the pack divided by the number of cards[6].

$$V(N) - V(N-1) = \frac{V_1(N)}{N} \tag{4}$$

A similar formula has been published by Baayen (p. 115, [1])

We can also derive the formula directly from the formulae described in the previous chapter and thus complete them into one consistent framework.[7]

$$\sum_{i=1}^{M}\left(1-\frac{(d-N)!(d-f_i)!}{d!(d-N-f_i)!}\right)-\sum_{i=1}^{M}\left(1-\frac{(d-N+1)!(d-f_i)!}{d!(d-N+1-f_i)!}\right)=$$

$$=\frac{1}{N}\sum_{i=1}^{M}\frac{\frac{N(d-N)!}{(f_i-1)!(d-N-f_i+1)!}}{\frac{d!}{f_i!(d-f_i)!}}$$

$$\sum_{i=1}^{M}\frac{(d-N)!\,f_i(d-f_i)!}{(d-N-f_i+1)!}=\sum_{i=1}^{M}\frac{(d-N)!\,f_i(d-f_i)!}{d!(d-N-f_i+1)!} \tag{5}$$

Now, a measured T_1TR curve can be transformed into TTR easily:

Another form of the formula is[8]:

$$\frac{V(N)}{V(N-1)}=\frac{N}{N-Z(N)}. \tag{6}$$

An inverse transformation (TTR into T_1TR) is not easy when using real data, because the real world difference between $V(N)$ and $V(N-1)$ is heavily influenced by random deviations. However, this does not mean that the formula is not valid and that we cannot use it to derive other formulae.

---

[6] Given a multiset A and a multiset B; and a multiset C; $C = B \ominus supp.(A \ominus B)$. The probability that the cardinality of a multiset B decreases when decreasing cardinality of A by one is equal to cardinality of C divided by cardinality of A.

[7] The whole proof can be found at www.milicka.cz/kestazeni/beograd/dukaz1.pdf.

[8] Where $Z(N) = V_1(N)/V(N)$.

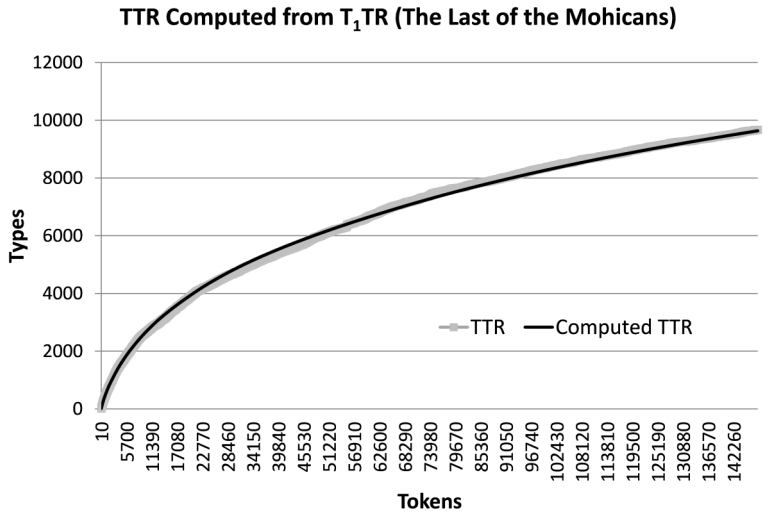


Fig. 3. Type-token relation measured on a text (The Last of the Mohicans by J. F. Cooper [9]) and the same curve computed from hapax-token relation measured on the same text

The following formula is based on the same idea as the previous one and we can also prove it using the combinatorial model[9]. It expresses the exact relation between the number of types represented by g tokens and the number of types represented by tokens.

$$V_g(N) - V_g(N-1) = \frac{gV_g(N) - (g+1)V_{g+1}(N)}{N} \tag{7}$$

The same formula in another form:

$$V_g(N) = \frac{(g-1)V_{g-1}(N) - N\big(V_{g+1}(N) - V_{g-1}(N-1)\big)}{g} \tag{8}$$

And the most general one:

$$V_g(N) = N\frac{V(N) - V(N-1) - \sum_{i=1}^{g-1}\left(V_{i(N)} - V_{i(N-1)}\right)}{g} \tag{9}$$

Because of the random deviations, these formulae are not very useful directly for the real life data, but we will need them in the next section.

[9]Proof available at www.milicka.cz/kestazeni/beograd/dukaz2.pdf

4 The approximation

In this section we expand outside pure algebra and take into account an assumption that (for large monolingual corpora) the ratio between the number of hapax legomena and the number of all types converges to a constant larger than zero.

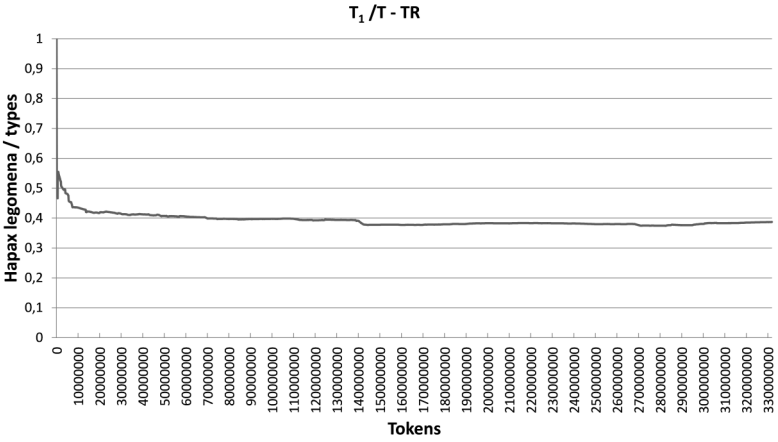


Fig. 4. The ratio between the number of hapax legomena and the number of all types converges (for Arabic corpora CLARA [10] and CLAUDia [11]) to cca 0.38

This assumption was widely discussed in [2], where Cvrček even claims that after an initial drop the ratio slightly increases (measured for wordforms and lemmas in large corpora of European languages). This assumption allows us to modify formula 6, where $Z(N)$ is a function, resulting in the following formula:

$$\frac{V(N)}{V(N-1)} = \frac{N}{N-Z} \tag{10}$$

where Z is a constant[10], at least for very large corpora consisting of hundreds of millions of tokens. Here is the non-recurrent form of the formula (using Pochhammer's symbol)

$$V(N) = \frac{V(M)(1)_{(N-M)}}{(1-Z)_{(N-M)}}. \tag{11}$$

The parameter M expresses the number of tokens by which we assume that the hapax-type ratio reached its constant value. $V(M)$ is the number of types after measuring M tokens. It is our initial position, starting point.

We can use the formula in real life to predict TTR for very large amounts of data, simply by measuring TTR and T_1TR until $Z(N)$ is satisfactorily stable. Then we

[10] $Z = \lim_{N\to\infty} V_1(N)/V(N)$.

consider the token reached as an initial position M, the number of types so far measured as initial $V(M)$ and from this initial position further we can model the growth of the number of types.

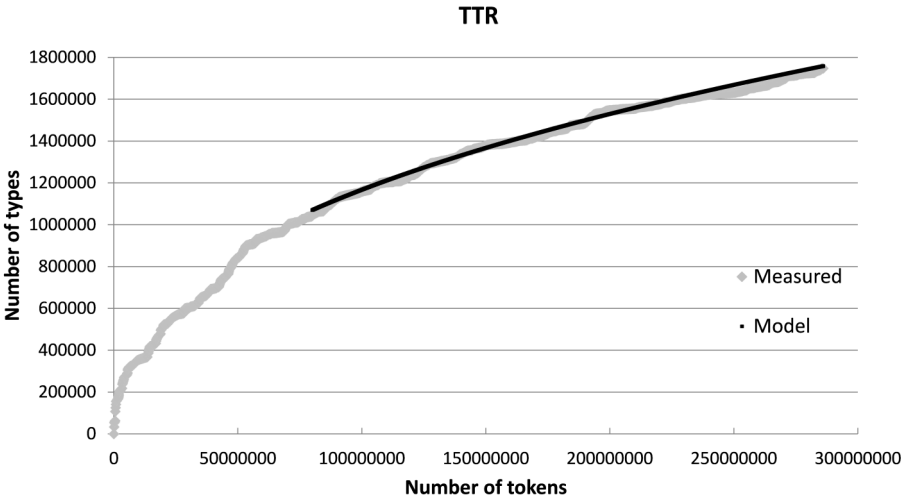


Fig. 5. Illustration of usage of the model. Type-token relation measured and computed for the CLAUDia corpus [10]

It can be interesting for some linguists that this formula tells us that if the constant Z is larger than zero then the number of types grows beyond any limit.

$$\lim_{N\to\infty} \frac{V(M)(1)_{(N-M)}}{(1-Z)_{(N-M)}} = \infty. \tag{12}$$

The similar formula is also valid for T_1TR[11]:

$$\frac{V_1(N)}{V_1(N-1)} = \frac{N}{N-Z}. \tag{13}$$

And by mathematical induction[12]we can prove that the following similar formula is also valid for the growth of the number of any types of any other frequency

$$\frac{V_g(N)}{V_g(N-1)} = \frac{N}{N-Z}. \tag{14}$$

From these formulae we can arrive[13] at the final formula that enables us to calculate the number of types of a certain frequency (V_g) using the constant Z as the only

[11]For the proof see www.milicka.cz/kestazeni/beograd/dukaz3.pdf
[12]For the proof see www.milicka.cz/kestazeni/beograd/dukaz4.pdf
[13]For the proof see www.milicka.cz/kestazeni/beograd/dukaz5.pdf

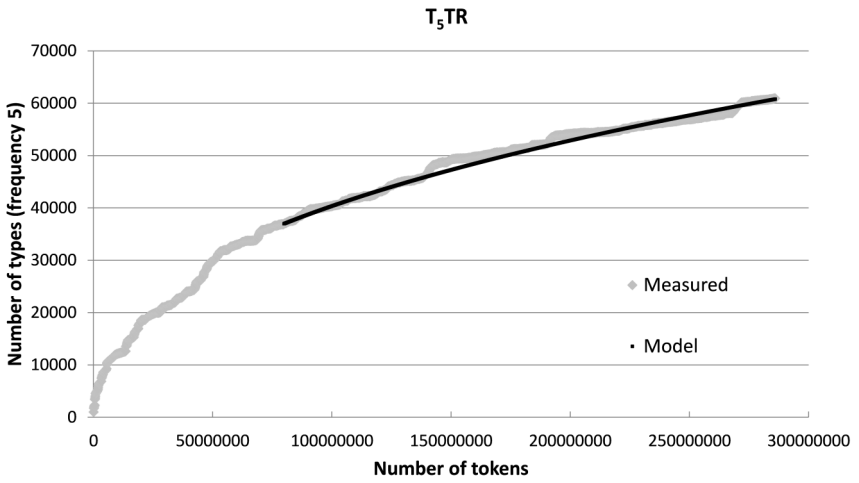


Fig. 6. Illustration of the usage of the model. The relation between types with frequency 5 and tokens, measured and computed for the CLAUDia corpus [10]

parameter

$$\frac{V_g}{V_{g-1}} = \frac{g - Z - 1}{g}. \tag{15}$$

This formula enables us to transform the number of types into a frequency density function, which it is possible to transform into distribution of frequencies of types or RFR.

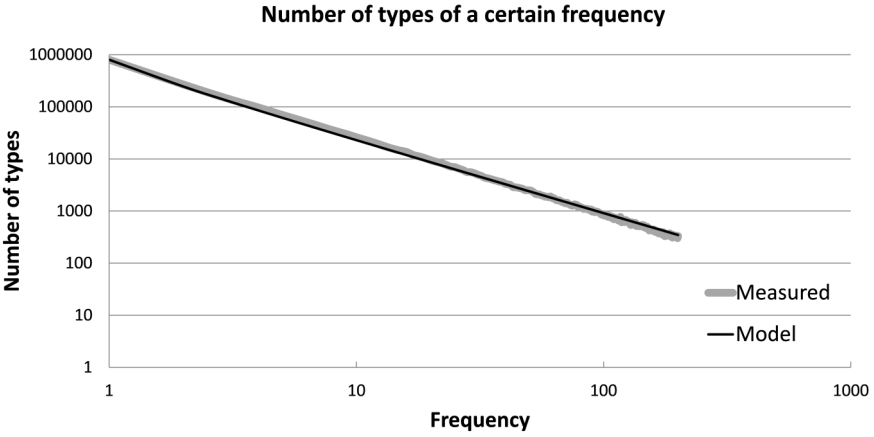


Fig. 7. Dependency of the number of types with certain frequency on the frequency (log-log). Measured on CLARA [10] and CLAUDia [11]

5 Conclusions

1. RFR, TTR, T_1TR, T_gTR and also the average frequency of types etc. are one phenomenon. If two texts substantially differ in one of these quantities, they would also be different in other ones.
2. We can exactly calculate TTR, T_1TR and T_gTR from RFR or a distribution of types. We can also exactly calculate TTR from T_1TR. For inverse relations we still need an approximation.
3. A good approximation of V_1/V would enable us to calculate TTR, T_gTR, RFR, etc.

   Even if we consider V_1/V to be a constant, we can successfully model TTR, T_gTR, RFR, etc. for very large corpora (because V_1/V seems to converge to a constant). This constant is the only parameter for all of these approximations.
4. V_1/V converging to a constant larger than zero implies that the number of types (and the number of types of a certain frequency) does not converge to any constant.

References

1. Baayen, R. H.: Quantitative aspects of morphological productivity. In: G. E. Booij and J. van Marle (eds), Yearbook of Morphology 1991, pp. 109–149. Dordrecht (1992).
2. Cvrček, V.: How Large is the Core of Language? In: Proceedings from the sixth international corpus linguistics conference 2011, Birmingham. [url: http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2011/Paper-145.pdf] (2011)
3. Leijenhorst,D. C. van – Weide, Th. P. van der: A formal derivation of Heaps' Law. In: Information Sciences 170, pp. 263–272. New York (2005).
4. Milička, J.: Type-token & Hapax-token Relation: A Combinatorial Model. In: Glottotheory 2/1, pp. 99–110 (2009).
5. Müller, D.: Computing the Type Token Relation From the A Priori Distribution of Types. In: Journal of Quantitative Linguistics, Vol. 9, No 3, pp. 193–214 (2002).
6. Syropoulos, A.: Mathematics of Multisets. In: Multiset Processing, pp. 347–358.: Springer-Verlag, London (2001).
7. Wimmer G.: The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R. (eds.) Quantitative Linguistics: An International Handbook, pp. 361-368. Walter de Gruyter, (2005).
8. Wyllys, R. E.: Empirical and Theoretical Bases of Zipf's Law. In: Library Trends 30(1) no. 53–64, pp. 53–64, (1981).

Corpora

9. *CLARA* (cca 40M tokens) Synchronic Arabic Corpora (MSA)
10. *CLAUDia* (cca 300M tokens) Diachronic Arabic Corpora
11. Cooper, J. F.: *The Last of the Mohicans*

Measurement of nonlinear distance in data derived from linguistic corpora

Hermann Moisl

School of English Literature, Language, and Linguistics
Newcastle University
Newcastle upon Tyne NE1 7RU
United Kingdom
Hermann.moisl@ncl.ac.uk

Abstract. Most science and engineering disciplines recognize that application of linear analytical methods to data containing nonlinearities can distort results, and in response have developed mathematically and statistically based methods for dealing with nonlinearity. In linguistics, however, there has thus far been little recognition of the possibility that there might be nonlinearity in data abstracted from speech and text corpora or, where found, what the implications for analysis are. The present paper addresses this issue in three main parts. The first part outlines the nature of data nonlinearity, the second reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects, and, using these methods, the third identifies and quantifies the degree of nonlinearity present in data abstracted from the *Diachronic Electronic Corpus of Tyneside English*, a dialect speech corpus.

Keywords: Corpus linguistics, corpus data analysis, nonlinearity, geodesic distance, graph distance

Introduction

Most science and engineering disciplines recognize that application of linear analytical methods to data containing nonlinearities can distort results, and in response have developed methods for dealing with nonlinearity [1]. In linguistics, however, there has thus far been little recognition of the possibility that there might be nonlinearity in data abstracted from speech and text corpora or, where found, what the implications for analysis are [2]. The present paper addresses this issue in three main parts. The first part outlines the nature of data nonlinearity, the second reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects, and, using these methods, the third identifies and quantifies the degree of nonlinearity present in data abstracted from the *Diachronic Electronic Corpus of Tyneside English*, a dialect speech corpus [3].

1 Nonlinearity

1.1 Nonlinearity in Natural Processes

In natural processes there is a fundamental distinction between linear and nonlinear behaviour. Linear processes have a constant proportionality between cause and effect. If a ball is kicked x hard and it goes y distance, then a $2x$ kick will appear to make it go $2y$, a $3x$ kick $3y$, and so on. Nonlinearity is the breakdown of such proportionality. In the case of our ball, the linear relationship increasingly breaks down as it is kicked harder and harder. Air and rolling resistance become significant factors, so that for, say, $5x$ it only goes $4.9y$, for $6x$ $5.7y$, and again so on until eventually it bursts and goes hardly any distance at all. Such nonlinear effects pervade the natural world and gives rise to a wide variety of complex and often unexpected – including chaotic – behaviours [4].

1.2 Nonlinearity in Data

Data is a description of objects involved in a natural process of interest in terms of a set of variables. Given m objects described by n variables, a standard representation of data for computational analysis is a matrix M in which each of the m rows represents a different object, each of the n columns represents a different variable, and the value at $M_{i,j}$ describes object i in terms of variable j, for $i = 1,\ldots,m$, $j = 1,\ldots,n$. The matrix thereby makes the link between the researcher's conceptualization of the process in terms of the semantics of the variables s/he has chosen and the state of the world, and allows the resulting data to be taken as a representation of the process based on empirical observation. Assuming that the representation is a faithful one, any nonlinearity in the process will be reflected in the data.

M is linear when the functional relationships between all its variables, that is, the values in its columns, conform to the mathematical definition of linearity. In mathematics, a linear function f is one that satisfies the following properties, where x and y are variables and a is a constant [5]:

- Additivity: $f(x+y) = f(x) + f(y)$ – adding the results of f applied to x and y separately is equivalent to adding x and y and then applying f to the sum.
- Homogeneity: $f(ax) = af(x)$ – multiplying the result of applying f to x by a constant is equivalent to multiplying x by the constant and then applying f to the result.

A function which does not satisfy these two properties is nonlinear, and so is a data matrix in which the relationship between two or more of its columns is nonlinear.

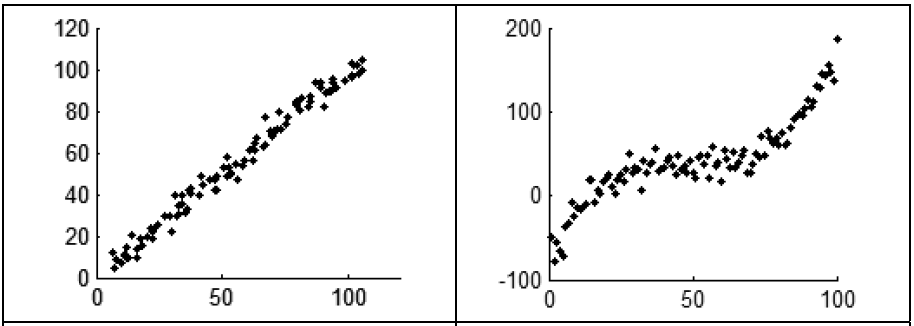
2 Nonlinearity Detection

It is not in general obvious whether a given data matrix contains nonlinearity, and the only way to find out if it does is to test for it. In practice, data abstracted from observation is likely to contain at least some noise, and it is consequently unlikely

that strictly linear relationships between variables will be found. Instead, one is looking for degrees of deviation from linearity. Three ways of doing this are presented. Two of them are well-established, and the third is a proposal based on graph distance measurement.

2.1 Graphical Identification of Nonlinearity

The graphical method is based on pairwise scatter-plotting of variables and subsequent visual identification of deviation from linearity. In figure 1a, for example, the essentially linear relationship of variables *v*1 and *v*2 is visually clear despite the scatter, and the nonlinear relationship in figure 2b equally so.



a: Linear relationship of *v1* and *v2* | b: Nonlinear relationship of *v1* and *v2*

Fig. 1. Scatter plots of linear and nonlinear bivariate data

Looking for nonlinearity in this way involves plotting of all possible distinct pairings of data variables and visual identification of any nonlinearity. This can be a fairly onerous but generally not insuperable undertaking where the number of variables is large. A more serious problem is that visual interpretation of scatter plots is subjective, and where the shape of the relationship between variables is not as un-

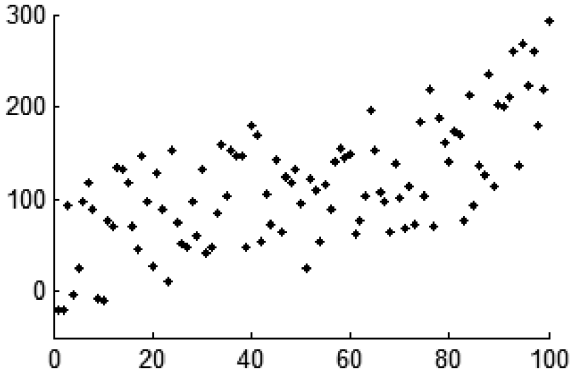


Fig. 2. Possibly noisy linear, possibly nonlinear bivariate data

ambiguous as those in figure 1 different observers are likely to draw different conclusions. For example, is the relationship in figure 2 linear with substantial noise, or nonlinear?

2.2 Identification of Nonlinearity Based on Regression

To be fully useful, graphically-based identification of nonlinearity needs to be supplemented by quantitative measures of the degree of nonlinearity. Regression analysis provides this [6, 7]. Regression attempts to model the relationship between one or more independent variables and a dependent variable whose values are hypothesized to be causally determined by the independent one(s), by finding a mathematical function which best fits the data distribution. Because the aim is simply to decide whether given data is linear or nonlinear rather than to find the optimal mathematical fit for it, the discussion confines itself to parametric regression.

The first step in parametric regression is to select a mathematical model that relates the values of the dependent variable y to those of the independent variable x. A linear model proposes a linear relationship of the general form

$$y = ax + b \tag{1}$$

where a and b are scalar constants representing the slope of the line and the intercept of the line with the y-axis respectively; a and b are unknown and are to be determined. This is done by finding values for a and b such that the sum of squared residuals, that is, distances from the line of best fit to the dependent-variable values on the y-axis, is minimized. The line determined by the values for a and b is the best linear fit for the hypothesized relationship between x and y. A nonlinear model proposes a nonlinear relationship between x and y. Numerous nonlinear models are available. Frequently used ones in regression are polynomials with the general form

$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \tag{2}$$

where the $a_n, \ldots, a_0$ are constants and n is the order of the polynomial; where $n = 1$ the polynomial is first-order, where $n = 2$ it is second-order and so on, though traditionally orders 1, 2, and 3 are called 'linear', 'quadratic', and 'cubic' respectively. As with linear regression, nonlinear regression finds the line of best fit by calculating the coefficients $a_n, \ldots, a_0$ which minimize the sum of squared residuals between the line and the y values.

Using regression to identify nonlinearity in data would appear simply to be a matter of comparing the goodness of fit of the linear model with that of whatever nonlinear model has been chosen: the data is linear if a straight line provides as good a fit as any other mathematical function [11], and nonlinear if the nonlinear model is a significantly better fit than the linear one [7]. In figure 3, for example, the cubic model looks like it fits the data best, the quadratic less well, and the linear least well; based on visual inspection, one would say that this data is nonlinear.

Such direct visual interpretation can be corroborated by residual analysis and various goodness-of-fit statistics like the runs test, summed square of errors (SSE), root mean squared error (RMSE), and R^2 [8–12]. These statistics all look reasonable but

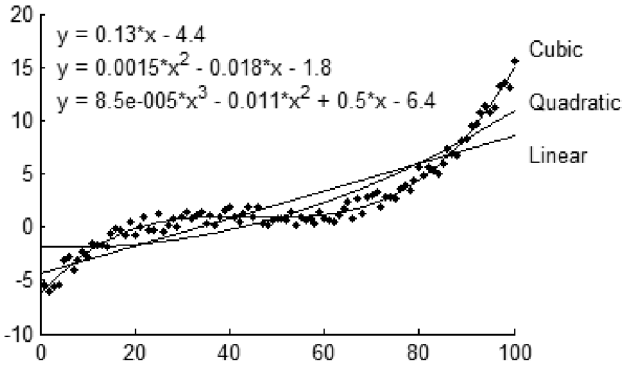


Fig. 3. Linear, quadratic and cubic polynomials with curves of best fit

have an underlying problem. For a given family of models such as polynomials, the model with more parameters typically fits the data better than one with fewer; the more parameters the more convoluted the line of best fit can be and thus the closer it can get to the data values, thereby reducing SSE and affecting RMSE and R^2. Use of the foregoing statistics for identification of nonlinearity therefore implies that the best model is always the one which comes closest to the data points. Where the relationship between variables is perfectly linear this is not a problem because increasing the number of parameters will not affect the statistics: the linear model is optimal. But, as already noted, empirical data typically contains noise, and that is where the problem lies. Given data that is not perfectly linear and a model for it with $n > 2$ parameters, there are two possible interpretations. On the one hand, it may be that the model is fitting noise and thereby obscuring a relationship between the variables which is better captured by a model with fewer than n parameters. On the other, it may be that the nonlinearity is not noise but a genuine reflection of the nonlinear relationship between those aspects of the domain which the data describes, and that the model with n parameters is the preferred one. Which interpretation is correct? Knowledge of the likelihood and scale of noise in the domain can help in deciding, but this is supplemented by an extensive range of model selection methods [13, ch.5]. Two of the more frequently used methods are the extra sum-of-squares F-test and Akaike's information criterion [14; 7, ch.22].

2.3 Identification of Nonlinearity Based on Graph Distance

An alternative to regression proposed here is to make the ratio of mean nonlinear to mean linear distances among points on the data manifold the basis for nonlinearity identification. This is motivated by the observation that the shape of a manifold represents the real-world interrelationship of objects described by variables, and curvature in the manifold represents the nonlinear aspect of that interrelationship. Linear metrics ignore the nonlinearity and will therefore always be smaller than nonlinear ones; a disparity between nonlinear and linear measures consequently indicates nonlinearity, and their ratio indicates the degree of disparity.

Given a set X, a metric [15; 16] is a function $d : X * X > R$ if, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$, the following properties hold:

- $d(\mathbf{x}, \mathbf{y}) \geqslant 0$, that is, the distance between any two vectors is non-negative.
- $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$, that is, the distance from a vector to itself is 0, and for vectors which are not identical is greater than 0.
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$, that is, distances are symmetrical.
- $d(\mathbf{x}, \mathbf{z}) \leqslant d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$, that is, the distance between any two vectors is always less than or equal to the distance between them and a third vector.

A metric space $M(V, d)$ is a vector space V on which a metric d is defined, which returns the distance between any two points in the space.

Numerous distance metrics exist [16]. For present purposes these are divided into two types: linear metrics, where the distance between two points on a manifold is the length of the straight line joining the points without reference to the shape of the manifold, and nonlinear metrics where the distance is the length of the shortest line joining them along the surface of the manifold, which need not be flat. Where the manifold is flat, linear and nonlinear measures are identical. Where it is curved, however, linear and nonlinear measurements can differ to varying degrees depending on the nature of the curvature, as shown in figure 4.

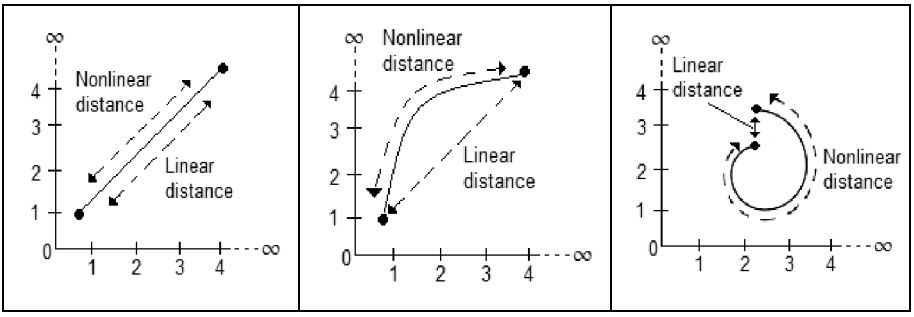


Fig. 4. Linear and nonlinear distance on flat and curved manifolds

Euclidean distance is here used for linear measurement and geodesic distance for nonlinear. Euclidean distance is well known and commonly used [16]; geodesic distance requires a little explanation. Etymologically, the word 'geodesy' comes from Greek *geodaisia*, 'division of the earth'; geodesic distance is the shortest distance between any two points on the Earth measured along its curved surface. Mathematically, geodesic distance is a generalization of linear to nonlinear distance measurement in a space: the geodesic distance $g(\mathbf{x}, \mathbf{y})$ is the shortest distance between two points x and y on a manifold measured along its possibly-curved surface [16, ch.6]. What follows develops a method for approximating geodesic distance on manifolds using graph distance. Figure 5 shows a small nonlinear matrix M and the associated scatterplot.

For a data matrix M with m rows and n columns, a Euclidean distance matrix D is an $m \times m$ matrix each of whose values $D_{i,j}$ (for $i, j = 1, \ldots, m$) is the Euclidean distance from row i to j of M in n-dimensional space. Figure 6 shows D for M in figure 5.

	V1	V2
A	0.27	0.20
B	0.32	0.50
C	0.40	0.70
D	0.50	0.80
E	0.60	0.70
F	0.68	0.50
G	0.73	0.20

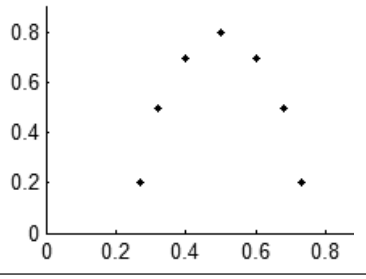


Fig. 5. Nonlinear data matrix and corresponding scatterplot

M is interpretable as a connected graph *G* each of whose arcs from *i* to *j* is labelled with the Euclidean distance between G_i and G_j, as shown in figure 6; the distance between node *A* and node *B*, for example, is given in the table as 0.30, between *A* and *G* as 0.46, and so on; only two arcs are labelled to avoid clutter.

	A	B	C	D	E	F	G
A	0	.30	.52	.64	.60	.51	.46
B	.30	0	.22	.35	.34	.36	.51
C	.52	.22	0	.14	.20	.34	.60
D	.64	.35	.14	0	.14	.35	.64
E	.60	.34	.20	.14	0	.22	.52
F	.51	.36	.34	.35	.22	0	.30
G	.46	.51	.60	.64	.52	.30	0

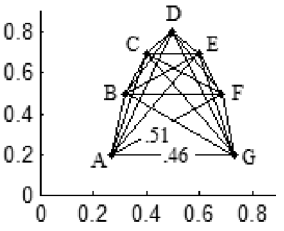


Fig. 6. Euclidean distance matrix for the data in figure 5 and interpretation of the manifold as a connected graph with Euclidean distances as arc labels

A spanning tree for *G* is an acyclic subgraph of *G* which contains all the nodes in *G* and some subset of the arcs of *G* [17]. A *minimum* spanning tree of *G*, as its name indicates, is a spanning tree which contains the minimum number of arcs

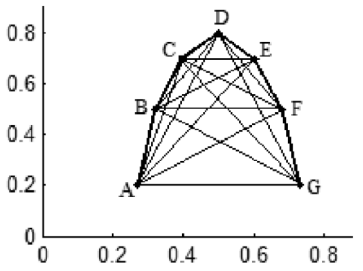


Fig. 7. Minimum spanning tree for the graph in figure 6

required to connect all the nodes in *G*, or, if the arcs have weights, the smallest sum of weights. The minimum spanning tree for *G* in figure 6 is shown in figure 7, with the arcs comprising the tree $A > B > C > D > E > F > G$ emboldened.

A minimum spanning tree can be used to approximate the geodesic distances using the Euclidean distances because the distance between any two nodes is guaranteed to be minimal; in figure 7, from A to B the geodesic and Euclidean distances are identical, but from A to C the geodesic is $AB + BC$ rather than the Euclidean AC, and so on. Figure 8 shows a table constructed in this way together with the corresponding Euclidean one.

	A	B	C	D	E	F	G
A	0	.30	.52	.64	.60	.51	.46
B	.30	0	.22	.35	.34	.36	.51
C	.52	.22	0	.14	.20	.34	.60
D	.64	.35	.14	0	.14	.35	.64
E	.60	.34	.20	.14	0	.22	.52
F	.51	.36	.34	.35	.22	0	.30
G	.46	.51	.60	.64	.52	.30	0

a. Euclidean distance matrix for M.
- Sum of distances: 16.52
- Mean distance: 0.34
- Distance A-G: 0.46

	A	B	C	D	E	F	G
A	0	.30	.52	.66	.80	1.0	1.3
B	.30	0	.22	.36	.50	.72	1.0
C	.52	.22	0	.14	.28	.50	.80
D	.66	.36	.14	0	.14	.36	.66
E	.80	.50	.28	.14	0	.22	.52
F	1.0	.72	.50	.36	.22	0	.30
G	1.3	1.0	.80	.66	.52	.30	0

b. Graph distance matrix for M.
- Sum of distances: 22.60
- Mean distance: 0.46
- Distance A-G: 1.3

Fig. 8. Euclidean and geodesic distance matrices for the data in figure 14a

The sum of distances and mean distance for the Euclidean matrix are both substantially less than for the graph, and the graph distance between A and G is almost three times larger than the Euclidean, which figure 7 confirms visually.

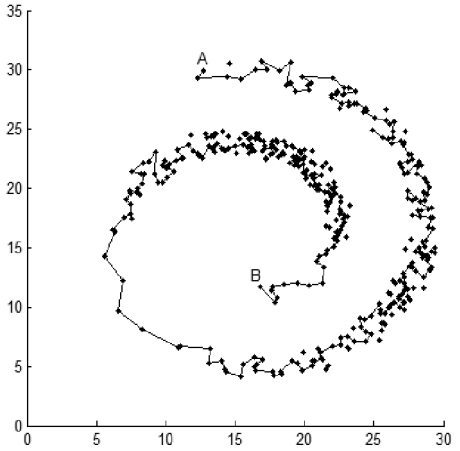


Fig. 9. Euclidean and graph distances on a nonlinear data manifold

The ratio of mean graph to mean Euclidean distance between all pairs of nodes in a graph gives a measure of the amount of nonlinearity in a data manifold. If the manifold is linear then the two means are identical and the ratio is 1; any nonlinearity makes the mean of graph distances greater than the Euclidean mean, and the ratio

is greater than 1 in proportion of the degree of nonlinearity. Figure 9 is an example based on the Swiss roll data extensively used in discussions of nonlinearity, and shows the path of the shortest graph distance from *A* to *B*.

The ratio of mean graph to mean Euclidean distance in figure 9 is 3.7, and the ratio of graph to Euclidean distance from *A* to *B* is 6.6, that is, almost seven times as far.

3 Case Study

This final section presents a case study to show that substantial nonlinearity does in fact occur in a particular speech corpus.

3.1 Corpus Data

The *Diachronic Electronic Corpus of Tyneside English* (DECTE) [3] includes phonetic transcriptions of 63 audio recordings, and the data for what follows is abstracted from these. Each speaker was represented by a 156-element vector each element of which represents a different phonetic segment in the DECTE transcription scheme, and the value at any given element is the frequency with which the speaker uses the associated segment in his or her interview. The set of speaker vectors was assembled into a matrix M in which the rows i (for $i = 1 \ldots 63$) represent the speakers, the columns j (for $j = 1 \ldots 156$) represent the phonetic segments, and the value at $M_{i,j}$ is the number of times speaker i uses segment j. M was normalized using mean document length [18] to remove the effect of variation in interview length.

3.2 Identification of Nonlinearity in M

Using graphical and regression-based methods, no strictly or even approximately linear relationships between pairs of variables were found in M. In a few cases the relationships looked random, but most showed a discernible pattern; the segment pair ɔ: and *a* : is representative and is used as the basis for discussion in what follows.

3.2.1 Graphical Identification of Nonlinearity

A scatter plot of ɔ: on the horizontal axis and *a:* on the vertical in figure 10 shows a visually clear nonlinear relationship.

3.2.2 Regression-based Identification of Nonlinearity

Using ɔ: as the independent variable and *a:* as the dependent, a selection of polynomials was used to model the relationship. These are shown in figure 11.

Visually, the linear model appears to fit least well and the 5^{th}-degree polynomial best, as expected, and this is confirmed by runs tests, residual plots, and the goodness of fit statistics in table 1.

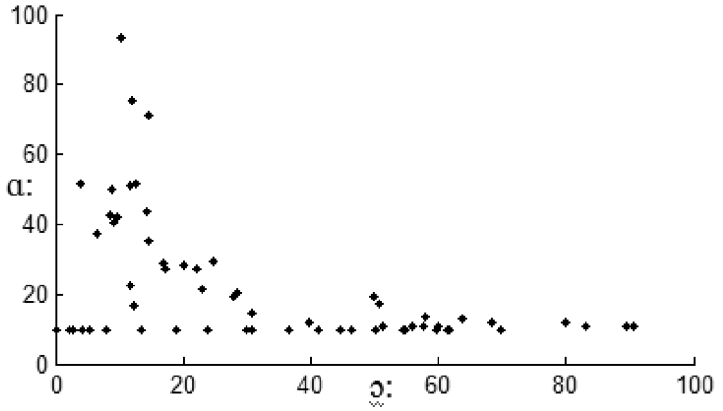


Fig. 10. Scatter plot of column values in *M* representing the phonetic segments ɔ : and *a:*

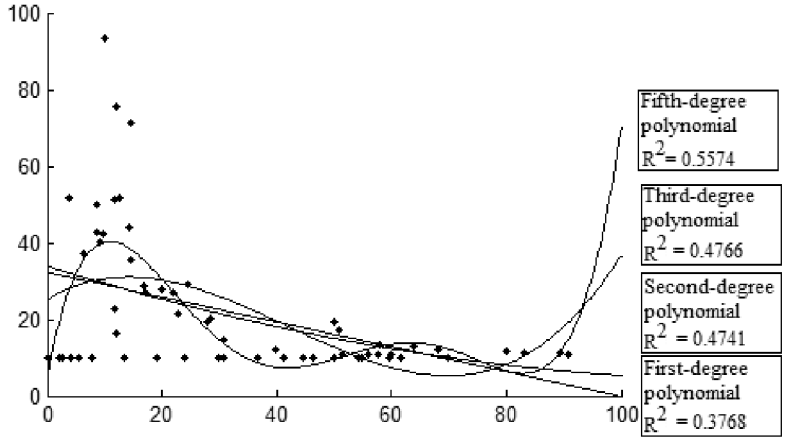


Fig. 11. Polynomial regression models of the ɔ : /*a* : relationship

Table 1. Goodness of fit statistics for figure 11

	SSE	**RMSE**	**R^2**
Degree 1	12420	15.03	0.3768
Degree 2	10480	13.93	0.4741
Degree 3	10390	14.00	0.4786
Degree 5	8821	13.15	0.5574

The extra sum-of-squares F-test and AIC test further support the indications so far: that the first-order model is worst, that second-order is better than third, but that the fifth-order model is preferred.

3.2.3 Graph Distance-based Identification of Nonlinearity

The Euclidean 63×63 distance matrix E was calculated for M, the minimum spanning tree for E was found, and the graph distance matrix G was derived by tree traversal, all as described in the foregoing discussion. The rows of both E and G were then linearized into vectors of length $63 \times 63 = 3969$, sorted, and co-plotted to get a representation of the relationship between linear and graph distances in the two matrices. This is shown in figure 12.

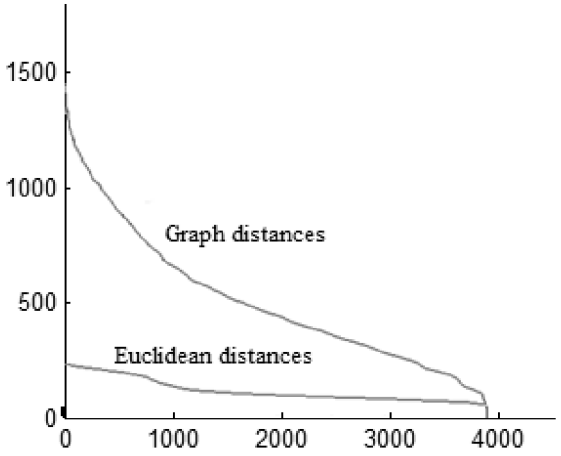


Fig. 12. Comparison of Euclidean and graph distances for *M*

The graph distances between and among the speakers in *M* are consistently larger than the Euclidean ones over the entire range. This is summarized in the ratio *mean(G) / mean(E)* of mean distances, which is 3.89. On these indicators, M can be said to contain a substantial amount of nonlinearity.

Conclusion

This discussion set out to show how nonlinearity can be detected in data derived from linguistic corpora using established graphical and regression-based methods and proposing a method based on approximation of geodesic distance measurement with graph distance. These methods were applied to frequency data abstracted from phonetic transcriptions of speech from DECTE, a dialect corpus, and all the methods agreed that substantial nonlinearity was present. DECTE is typical of the many digital electronic language corpora that have appeared in recent years [19,20], and it is reasonable to suspect that nonlinearity will be present in data abstracted from these as well. Where, therefore, measurement of distance among data objects is a factor in analysis, as it is, for example, in cluster analysis, the data should first be screened for nonlinearity and the selection of analytical method should be guided by the result.

References

1. Strogatz, S.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Perseus Books, New York (2000)
2. Moisl, H.: Data nonlinearity in exploratory multivariate analysis of language corpora. In: Nerbonne, J., Ellison, M., Kondrak, G. (eds.) Computing and Historical Phonology. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pp. 93–100. Association for Computational Linguistics (2007)
3. Corrigan, K., Moisl, H., Buchstaller, I.: The Diachronic Electronic Corpus of Tyneside English. http://research.ncl.ac.uk/decte/index.htm (2012)
4. Bertuglia, C., Vaio, F.: Nonlinearity, Chaos, and Complexity: The Dynamics of Natural and Social Systems. Oxford University Press, Oxford (2005)
5. Lay, D.: Linear Algebra and its Applications, 4th ed. Pearson, New York (2010)
6. Seber, G., Wild, C.: Nonlinear Regression. Wiley-Interscience, Hoboken, NJ (2003)
7. Motulsky, H., Christopoulos, A.: Fitting Models to Data using Linear and Nonlinear Regression. Oxford, Oxford University Press (2004)
8. Mark, H., Workman, J.: Linearity in calibration: the importance of nonlinearity. Spectroscopy 20(1), (2005)
9. Mark, H., Workman, J.: Linearity in calibration: the Durbin-Watson statistic. Spectroscopy 20(3) (2005)
10. Mark, H., Workman, J.: Linearity in calibration: other tests for nonlinearity. Spectroscopy 20(4) (2005)
11. Mark, H., Workman, J. Linearity in calibration: how to test for non-linearity. Spectroscopy 20(9) (2005)
12. Mark, H., Workman, J.: Linearity in calibration: quantifying nonlinearity. Spectroscopy 20(12) (2005)
13. Izenman, A.: Modern Multivariate Statistical Techniques. Springer, Berlin (2008)
14. Burnham, K., Anderson, D.: Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer, Berlin (2002)
15. O Searcoid, M.: Metric Spaces. Springer, Berlin (2006)
16. Deza, M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009)
17. Gross, J., Yellen, J.: Graph Theory and its Applications, 2nd ed. Chapman & Hall, London (2006)
18. Moisl, H.: Variable scaling in cluster analysis of linguistic data. Corpus Linguistics and Linguistic Theory 6, 75–103 (2010)
19. Beal J., Corrigan K., Moisl H.: Creating and Digitizing Language Corpora, Volume 1: Synchronic Databases. Palgrave Macmillan, Basingstoke (2007)
20. Beal J., Corrigan K., Moisl H.: Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases. Palgrave Macmillan, Basingstoke (2007)

A Multidimensional Generalization of the Piotrowski-Altmann Law

Relja Vulanović

Department of Mathematical Sciences,
Kent State University at Stark,
6000 Frank Ave. NW, North Canton, Ohio, USA
rvulanov@kent.edu

Abstract. The Piotrowski-Altmann Law is extended to several dimensions by generalizing the logistic differential equation to a system of partial differential equations. It is shown how this system can be solved. The solution is a multidimensional generalization of the sigmoid. The multidimensional sigmoid has been used earlier to model certain phenomena related to parts-of-speech systems, but without any reference to differential equations. These models are now discussed as examples of the multidimensional Piotrowski-Altmann Law.

Keywords: Piotrowski-Altmann Law, sigmoid, system of partial differential equations, multidimensional surface, parts-of-speech systems, grammar efficiency.

1 Introduction

The well-known *logistic differential equation*,

$$\frac{dy}{dt} = ay(m-y), \quad t \geqslant 0, \tag{1}$$

models the growth of the population size y, as a function of time t, when the population size cannot exceed a certain value m, $m > 0$. The equation in (1) is constructed under the assumption that the rate of growth of population is directly proportional to both the population size and the quantity indicating how close the population gets to m. The constant of proportionality is a, $a > 0$.

The equation (1) has two constant solutions, $y \equiv 0$ and $y \equiv m$, and two general non-constant solutions,

$$y_{\pm} = \frac{m}{1 \pm e^{b-amt}}, \tag{2}$$

which are easy to obtain after separating the variables and integrating using partial fractions; b is the integration constant. Only the *logistic curve* y_+ is an appropriate model. Unlike y_-, it satisfies $0 < y < m$ and has the desired behavior: there are two almost-flat levels as the curve approaches the horizontal asymptotes $y = 0$ (observable if t is extended onto the interval $(-\infty, \infty)$) and $y = m$, and there is a smooth increasing part connecting the two levels.[1] Because of this S-shape, the logistic curve is also known as the *sigmoid*, particularly in linguistics. In linguistics, y typically denotes the proportion or percent of a linguistic form which is in the process of replacing an old form. Many linguistic changes can be modeled this way; [1], [2], and [3] survey some of the results. Most of the changes are phonetic and lexical, but see [4] and [5] for a syntactic change. All changes governed by the differential equation (1) are said to obey the Piotrowski or Piotrowski-Altmann Law. It is interesting that the original Piotrowski's work [6] on the topic considers an S-shaped curve different from y_+. The curve is given in terms of the arctan function and is not associated with the logistic differential equation (1). It is used to fit the data concerning the genitive forms without ending of some Russian units of measurement. The data show that the proportion of these forms has increased in time. The derivation of y_+ from (1) is given in [7], where the sigmoid is successfully fitted to the same data as those used in [6].

In order to model some other types of linguistic change, modifications of the basic differential equation (1) are proposed in [2], [8], [9], [4], and [5]. In these papers, the equation (1) is generalized by making the coefficient a a function of t. Piecewise constant, linear, and quadratic functions have been chosen for a. The present paper is concerned with a different generalization of the Piotrowski-Altmann Law: y is assumed to depend on several independent variables, $y = y(x_1, x_2, \ldots, x_n)$, and a system of partial differential equations is considered instead of (1). The motivation for this kind of generalization comes from [10], [11], [12], and [13], where a multidimensional generalization of the sigmoid y_+ is used in some quantitative linguistic investigations.[2] However, in these four papers, the equation of the generalized sigmoid is formed without referring to any differential equation. The same generalized-sigmoid equation is derived here as a solution of the system of partial differential equations analogous to (1).

The system of partial differential equations and its solution are presented in Section 2. Then, in Section 3, the applications from [10], [11], [12], and [13] are revisited and discussed from the point of view of the new finding. Finally, Section 4 contains a brief conclusion.

[1] There is no loss of generality in the assumption that 0 is the minimum value which y reaches asymptotically. If some other minimum value m_0 is assumed and the maximum value is m^*, then the equation looks like $dy/dt = a(y - m_0)(m^* - y)$. However, the substitution $z = y - m_0$ reduces this equation to (1) with $m = m^* - m_0$. It is also possible that $a < 0$, but then the model represents decay from m to 0, not the growth from 0 to m.

[2] In three dimensions, the generalization looks like an S-shaped surface.

2 The Generalization

Suppose y is a function of n independent variables $x_1, x_2, \ldots, x_n$, with respect to which it has continuous partial derivatives of first order. Let each partial derivative $\partial y/\partial x_i$, as the instantaneous rate of change of y relative to x_i, be directly proportional to y itself, as well as the quantity $m-y$, where m is a positive constant representing a maximum value that y reaches asymptotically. Let the constant of proportionality be a_i. In this way, we arrive at the following system of first-order partial differential equations:

$$\frac{\partial y}{\partial x_i} = a_i y(m-y), \quad i = 1, 2, \ldots, n. \tag{3}$$

Since, like in the one-dimensional case, we are interested in the solution satisfying $0 < y < m$, we consider

$$y = \frac{m}{1+\exp(b_1 - m a_1 x_1)} \tag{4}$$

as the solution of the first equation in the system (3). This is analogous to (1), but now the integration constant b_1 is a constant only with respect to x_1, i.e., it still depends on the remaining independent variables, $b_1 = b_1(x_2, x_3, \ldots, x_n)$. By substituting (4) in the second equation of the system (3), we get an equation which simplifies to

$$\frac{\partial b_1}{\partial x_2} = -m a_2. \tag{5}$$

From here,

$$b_1(x_2, x_3, \ldots, x_n) = -m a_2 x_2 + b_2(x_3, x_4, \ldots, x_n). \tag{6}$$

In equation (6), b_2 is a new integration constant, which is now independent of both x_1 and x_2. This equation is like a recursion formula between consecutive integration constants and, therefore, in the same way we get

$$\begin{gathered} b_2(x_3, x_4, \ldots, x_n) = -m a_3 x_3 + b_3(x_4, x_5, \ldots, x_n), \\ \vdots \\ b_{n-2}(x_{n-1}, x_n) = -m a_{n-1} x_{n-1} + b_{n-1}(x_n), \\ b_{n-1}(x_n) = -m a_n x_n + b, \end{gathered} \tag{7}$$

where b is the final integration constant independent of any variables. When all these recursive equations are combined, it follows that

$$b_1(x_2, x_3, \ldots, x_n) = b - m(a_2 x_2 + a_3 x_3 + \cdots + a_n x_n). \tag{8}$$

This, substituted in (4), gives

$$y = \frac{m}{1+\exp\left(b - m\sum_{i=1}^{n} a_i x_i\right)} \tag{9}$$

for the solution of the system (3). Clearly, the sigmoid y_+ in (2) is just a one-dimensional case of (9).

By differentiating (9), we get that

$$\operatorname{sign}\frac{\partial y}{\partial x_i} = \operatorname{sign} a_i. \tag{10}$$

This means that y changes monotonically in the direction of any independent variable x_i, increasing if $a_i > 0$ and decreasing if $a_i < 0$. Also, if $a_i > 0$ and if $x_i \to \infty$, while other variables are kept fixed, then y approaches m from below, and, if $x_i \to \infty$, then y approaches 0 from above. In the case when $a_i < 0$, this behavior is reversed. Thus, the multidimensional solution (9) preserves all important properties of the sigmoid.

3 Applications

The multidimensional sigmoid (9) is used with $n = 2$ in [10], [11], and [12], whereas in [13], $n = 3$ and $n = 4$. The constant $m = 1$ in these four papers, which all deal with parts-of-speech (PoS) systems in the sense of Hengeveld; see [14] and [15]. In Hengeveld's approach to parts of speech and their systems, four propositional functions are considered: $P =$ the head of the predicate phrase, $R =$ the head of the referential (noun) phrase, $r =$ the modifier of the referential phrase, and $p =$ the modifier of the predicate phrase. Verbs, nouns, adjectives, and manner adverbs [3] are the word classes specialized for the corresponding propositional functions. However, as the 50-language sample in [15] shows, many languages possess other word classes, which are flexible, i.e., they can have more than one propositional function. Three such word classes are identified in [15]: contentives, which can have all four propositional functions; non-verbs, which can be used to fulfill any propositional function other than P; and modifiers, which function as both r and p. There are also languages which do not have all four propositional functions.

The PoS system of a language is defined by the propositional functions present and by the word classes that are used to fulfill those functions. Seven main PoS systems and six intermediate ones are identified in [15]. The main types are shown in Table 1. Each intermediate PoS system type falls in between two consecutive main types. Intermediate types either have word classes with overlapping functions, or one small, closed word class.

PoS system types 4-7 are *rigid* in the sense of only having specialized word classes. On the other hand, types 1-3 are *flexible* because they have one flexible word class. The flexibility of this word class is greater the smaller the type number. It is to be expected that the more flexible a language is, the more it needs to use grammatical markA Multidimensional Generalizationof the Piotrowski-Altmann Lawers or fixed word order, or both, in order to distinguish between the propositional functions. Hengeveld et al. (2004) discuss this by investigating the PoS systems in their

[3] Adverbs other than manner adverbs are not considered since they often modify larger units within the sentence and not just the head of the predicate phrase.

Table 1. Main PoS system types

PoS system type	P	R	r	p
1	contentive			
2	verb	non-verb		
3	verb	noun	modifier	
4	verb	noun	adjective	manner adverb
5	verb	noun	adjective	–
6	verb	noun	–	–
7	verb	–	–	–

linguistic sample. For each language in the sample, they report the presence of markers and word-order rules. This information was used in [10]: two independent variables were considered, x_1 – the number of propositional functions, and x_2 – the number of word classes in the PoS system. The main and intermediate types which share the same counts (x_1, x_2) were lumped together, thus forming six classes of PoS systems: (4, 1), (4, 2), (4, 3), (4, 4), (3, 3), and (2, 2).[4] A seventh class, (3, 1), was added to this list, cf. [16].[5] The proportion y of languages that use markers and/or fixed word order was then calculated within each (x_1, x_2)-class. The dependence of y on x_1 and x_2 was modeled by the equation (9) with $m = 1$ and $n = 2$ and the equation was successfully fitted to the data. The reason for choosing this equation was that the proportion y of languages with disambiguation devices transitions between two plateaus, $y = 0$ and $y = 1$. Based on the results of Section 2, we now know that the values of y are governed by the system

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= a_1 y(1-y), \\ \frac{\partial y}{\partial x_2} &= a_2 y(1-y). \end{aligned} \tag{11}$$

Furthermore, if the number of word classes is fixed and the number of propositional functions increases, then the need to disambiguate between propositional functions increases as well, meaning that y should become greater, i.e., $\partial y/\partial x_1 > 0$ should hold true. On the other hand, if the number of propositional functions is fixed and the number of word classes increases, then the need for disambiguation is smaller and y should decrease, thus $\partial y/\partial x_1 < 0$ should be satisfied. Then we immediately know from (10) that $a_1 > 0$ and $a_2 < 0$ should hold true in (11). This turns out to be the case when the sigmoidal surface

$$y = \frac{1}{1 + \exp(b - a_1 x_1 - a_2 x_2)}, \tag{12}$$

which is a solution to (11), is fitted to the data, see Table 2.

[4] Class (1, 1) is only a theoretical possibility; it is not present in the language sample of [15].

[5] This class is represented by Tagalog, which is in the sample, but which is treated in [15] as a type 1 language, together with the whole (4, 1)-class.

Table 2. Fitting equation (12) to the data for 7 classes of PoS systems (R^2 = coefficient of multiple determination, Ra^2 = adjusted R^2)

x_1	x_2	Proportion y of languages wit markers and/or fixed word order	
		observed	calculated from (12)
4	1	1	0.999
4	2	1	0.993
4	3	0.933	0.938
4	4	0.515	0.603
3	1	1	0.990
3	3	0.682	0.512
2	2	0.333	0.420
$a_1 = 2.663, a_2 = -2.293, b = 1.061, R^2 = 0.901, Ra^2 = 0.851$			

Four different cases of the fit are considered in [10]. Three cases concern different contexts of disambiguation (the disambiguation between the heads P and R, between R and r, and between P and p) and the fourth case takes into account all three disambiguation contexts. The results for the latter are the only ones presented here as an illustration. Table 2 shows that the fit is very good. The graph of the sigmoidal surface is given in Figure 1. Some data points can be observed; others are hidden by the surface. It should be noted that the domain $[0,1] \times [0,1]$ shown in Figure 1 is wider than the actual domain, which is $1 \leqslant x_2 \leqslant x_1$ for $x_1 = 2,3,4$.

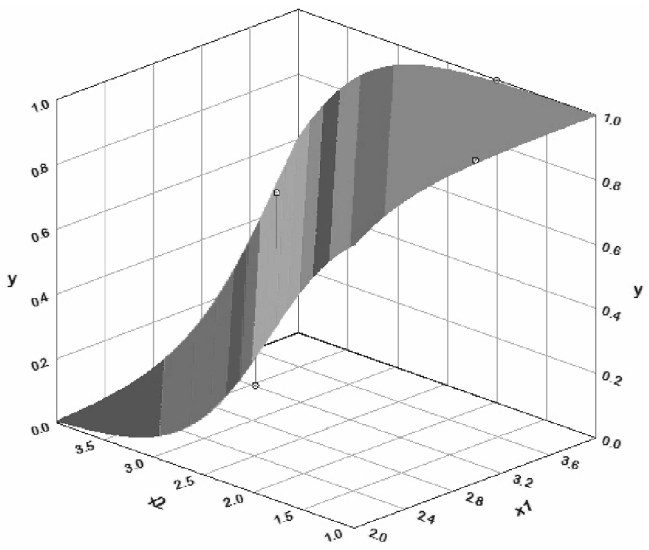


Fig. 1. The sigmoidal surface (12) fitted to the data for 7 classes of PoS systems

In paper [11], which is a continuation of [10], the counts x_1 and x_2 were modified so that intermediate PoS system types could be treated separately from the main

types. Thirteen types, belonging to the [15] sample, were distinguished. The same kind of analysis was carried out as in [10] and the sigmoidal surface (12) was fitted to the data. The results were similar to those in [10] and to what is reported here.

In papers [12] and [13], the multidimensional equation (9) was fitted to some other data related to PoS system types. The types were considered from a theoretical point of view, without using any language sample and without including grammatical markers in the analysis. All theoretically possible main PoS system types were taken into account, regardless of whether they are attested or not, and intermediate types were ignored. Using the same counts x_1 and x_2 as above, the types in [12] were represented as (x_1, x_2), $x_2 = 1, 2, \ldots, x_1$, $x_1 = 1, 2, 3, 4$. This gave 10 main PoS system types – 3 more ((3, 1), (3, 2), and (2, 1)) than in Table 1. The structure of each type was represented formally by constructing a grammar of simple intransitive sentences and the efficiency of each grammar was calculated using the formula from [17] and [18], cf. [16] as well.[6] Grammar-efficiency values depend on the word order used. All possible orders of propositional functions were considered keeping both predicate and referential phrases continuous. By this, each PoS system type was assigned an interval of efficiency values. Only the greatest values are shown here in Table 3. They cannot exceed 1, which is by scaling assigned to each grammar with the maximum possible efficiency within the class of all grammars having the same counts x_1 and x_2. The greatest grammar efficiency reaches the plateau of optimal values equal to 1 along parts of the border of the domain. Another plateau is theoretically possible, the one with efficiency values equal to 0, which happens if all sentences are ambiguous. It is because of the two plateaus that the equation (12) was used in [12] as a surface to fit the data. The fit is excellent, as the results in Table 3 show.

Table 3. Fitting equation (12) to the grammar-efficiency data for 10 PoS system types

x_1	x_2	Grammar efficiency of PoS system types calculated from	
		efficiency formula	(12)
4	1	0.250	0.259
4	2	0.786	0.767
4	3	0.914	0.969
4	4	1	0.997
3	1	1	0.987
3	2	0.875	0.999
3	3	1	1
2	1	1	1
2	2	1	1
1	1	1	1
$a_1 = -5.340$, $a_2 = 2.238$, $b = -18.074$, $R^2 = 0.962$, $Ra^2 = 0.951$			

The way grammar efficiency is defined, it certainly depends on x_1 and x_2, but some factors in the formula depend on the two variables implicitly. The equation (12)

[6] The formula would require a lengthy explanation and this is why it is not provided in this paper.

gives an explicit model of this dependence. We now know that this can be interpreted by referring to the system (11): the rate of change of grammar efficiency y, relative to either variable, is directly proportional to the efficiency itself and to the measure of the closeness of the efficiency to its theoretical maximum equal to 1. Furthermore, the following conclusions about grammar efficiency should be intuitively acceptable[7] even in the absence of the grammar-efficiency formula: (a) if the number of word classes is kept fixed and the number of propositional functions increases, it is to be expected of grammar efficiency to decrease, which, in view of (10) means that $a_1 < 0$; (b) conversely, if the number of propositional functions is fixed and the number of word classes increases, grammar efficiency should also increase, i.e. $a_2 > 0$ should hold true. The coefficients a_1 and a_2 reported in Table 3 indeed have these signs.

Some of the 10 PoS system types considered in [12] have subtypes. This means that structurally different PoS systems are classified together if they share the same counts x_1 and x_2. In order to differentiate between such subtypes, some new classification schemes were proposed in [13]. They were based on different ways of partitioning word classes into 2 or 3 disjoint groups. The number of propositional functions remained expressed by the variable x_1, but there were 2 or 3 variables which, when added together, gave the total number of word classes. One possibility considered was to count separately all rigid word classes in the PoS system (this count was x_2) and to divide the flexible word classes into two groups. The number of flexible word classes that are only used within either predicate or referential phrase was x_3 and x_4 was the number of flexible word classes that are used across the boundary between the two phrases. This classification scheme was labeled L2 in [13]. Three other classification schemes were considered as well. They all distinguished between 13 PoS system types and gave similar results when the equation (9), with $m = 1$ and either $n = 3$ or $n = 4$, was fitted to the data for the greatest grammar-efficiency value. The scheme L2, for instance, gave $R^2 = 0.926$ and $Ra^2 = 0.888$, with the expected signs of the coefficients: $a_1 = -3.806 < 0$, $a_2 = 1.950 > 0$, $a_3 = 3.196 > 0$, and $a_4 = 4.218 > 0$ (and with $b = -9.980$).

4 Conclusion

The Piotrowski-Altmann Law is represented by the logistic ordinary differential equation of first order. The solution of this equation is the sigmoid, an S-shaped curve in 2 dimensions, which is used in linguistics to model the rise of a new form/structure in the process of replacing an old form/structure. The present paper generalizes the logistic differential equation to a system of n first-order partial differential equations. This system is solved and its solution is shown to be an $(n+1)$-dimensional generalization of the sigmoid. In 3 dimensions, i.e. when $n = 2$, the graph of the generalized sigmoid is a sigmoidal surface. The system of logistic partial differential equations is suitable for modeling the change in any quantity y which satisfies that the relative rate of change in y with respect to any

[7] The values in Table 3 also confirm this to a great extent.

of its independent variables x_i, $i = 1, 2, \ldots, n$, is directly proportional to y itself and a factor which shows how close y is to a value it cannot exceed, but it can approach asymptotically. Four linguistic examples of this are discussed. They are taken from earlier papers which make use of the multidimensional generalization of the sigmoid without referring to any differential equation. In two of the examples, y represents the proportion of sample languages that have certain features related to the parts-of-speech system found in each language. The other two examples take y to be the greatest possible grammar efficiency for different parts-of-speech system types. Three examples deal with the 3-dimensional generalization of the sigmoid, and in the fourth one, the generalization is extended to 4 and 5 dimensions. None of the variables $x_1, x_2, \ldots, x_n$ in these examples is time. This is in contrast to the original Piotrowski-Altmann Law. However, the linguistic quantities represented by $x_1, x_2, \ldots, x_n$ may depend on time [19], although this is not explicitly modeled here. Thus, the dynamic nature of the Piotrowski-Altmann Law can be considered preserved.

In conclusion, with its derivation, explanation, and application, the *Multidimensional Piotrowski-Altmann Law* is fully established.

References

1. Best, K.-H., Kohlhase, J. (eds.): Exakte Sprachwandelforschung. Herodot, Göttingen (1983)
2. Altmann, G.: Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.) pp. 54-90 (1983)
3. Leopold, E.: Das Piotrowski-Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R. (eds.) Quantitative Linguistics: An International Handbook, pp. 627-633. Walter de Gruyter, Berlin/New York (2005)
4. Vulanović, R.: Fitting Periphrastic *Do* in Affirmative Declaratives. J. Quantitative Linguistics 14, 111-126 (2007)
5. Vulanović, R., Baayen, H.: Fitting the Development of Periphrastic *Do* in All Sentence Types. In: Grzybek, P., Köhler, R. (eds.) Exact Methods in the Study of Language and Text, pp. 679-688. Mouton de Gruyter, Berlin/New York (2007)
6. Piotrovskaja, A.A., Piotrovskij, R.G.: Matematičeskie modeli diachronii i tekstoobrazovanija (in Russian) In: Statistika reči avtomatičeskij analiz teksta, pp. 361-400. Nauka, Leningrad (1974)
7. Altmann, G., von Buttlar, H., Rott, W., Strauss, U.: A Law of Change in Language. In: Brainerd, B. (ed.) Historical Linguistics, pp. 104-115. Bochum: Brockmeyer, Bochum (1983)
8. Imsiepen, U.: Die e-Epithese bei starken Verben im Deutschen. In: Best, K.-H., Kohlhase, J. (eds.) pp. 119-141 (1983)
9. Best, K.-H., Beőthy, E., Altmann, G.: Ein methodischer Beitrag zum Piotrowski-Gesetz. Glottometrika 12, 115-124 (1990)
10. Vulanović, R., Köhler, R.: Word Order, Marking, and Parts-of-Speech Systems. J. Quantitative Linguistics 16, 289-306 (2009)
11. Vulanović, R.: Word Order, Marking, and a Two-Dimensional Classification of Parts-of-Speech System Types. J. Quantitative Linguistics 17, 229-252 (2010)

12. Vulanović, R., Miller, B.: Grammar Efficiency of Parts-of-Speech Systems. Glottotheory 3/2, 65-80 (2010)
13. Vulanović, R.: Classifying Parts-of-Speech Systems by Their Quantitative Properties. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.) Issues in Quantitative Linguistics Vol. 2 (Studies in Quantitative Linguistics 11), pp. 170-188. Ram-Verlag, Lüdenscheid (2011)
14. Hengeveld, K.: Non-verbal Predication. Theory, Typology, Diachrony (Functional Grammar Series 15). Mouton de Gruyter, Berlin/New York (1992)
15. Hengeveld, K., Rijkhoff, J., Siewierska, A.: Parts-of-Speech Systems and Word Order. J. Linguistics 40, 527-570 (2004)
16. Vulanović, R.: A Mathematical Analysis of Parts-of-Speech Systems. Glottometrics 17, 51-65 (2008)
17. Vulanović, R.: Grammar Efficiency and Complexity. Grammars 6, 127-144 (2003)
18. Vulanović, R.: On Measuring Language Complexity as Relative to the Conveyed Linguistic Information. SKY Journal of Linguistics 20, 399-427 (2007)
19. Smit, N.: Dynamic Perspectives on Lexical Categorisation. ACLC Working Papers 1, 49-75 (2007)

Publishers

UNIVERSITY OF BELGRADE
www.bg.ac.rs

ACADEMIC MIND
www.akademska-misao.rs

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

81'322(082)

INTERNATIONAL Conference on Quantitative Linguistics (8 ; 2012 ; Belgrade)
Methods and Application of Quantitative Linguistics : selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012 / editors Ivan Obradović, Emmerich Kelih, Reinhard Köhler ; editorial assistance Sheila Embleton, Hermann Moisl. - Belgrade : University : Academic mind, 2013 (Belgrade : Academic mind). - 193 str. : graf. prikazi ; 24 cm

Tiraž 80. - Str. 5-6: Preface / editors. - Napomene i bibliografske reference uz tekst. - Bibliografija uz svaki rad.

ISBN 978-86-7466-465-0

a) Рачунарска лингвистика - Зборници
COBISS.SR-ID 199657996