Empirical Approaches to Text and Language Analysis

dedicated to Luděk Hřebíček on the occasion of his 80th birthday

edited by

Gabriel Altmann, Radek Čech,
Ján Mačutek, Ludmila Uhlířová

2014
RAM-Verlag

Studies in quantitative linguistics

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp.
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3*. 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 4*. 2014. V + 148 pp.
15. K.-H. Best, E. Kelih (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014. VI + 163 pp.
16. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Language and Text Analysis*. 2014. V + 231 pp.

ISBN: 978-3-942303-24-8

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

A letter to Luděk Hřebíček from the editors

Dear Luděk Hřebíček,

The years have quickly passed since the time when we were happy to dedicate to you a volume *„Text as a Linguistic Paradigm: Levels, Constituents, Constructs*" (Quantitative Linguistics, Vol. 60, Wissenschaftlicher Verlag Trier)" to celebrate your 65th birthday. Now, fifteen years later, we have another great opportunity to congratulate you most sincerely: Please accept the content of the present book as a kind of scientific metaphor and paraphrase of "*Happy birthday to you*"… composed by 25 scientists from 10 countries on the occasion of your unbelievable 80th birthday. This festschrift is meant to be an expression of a deep respect for you as a prominent Czech orientalist, a theoretician of language and text structure, an author of numerous books, studies, and articles published in English, German and Czech in publishing houses across Europe, and last but not least, for you as an excellent essayist who knows perfectly how to persuade your readers that to build a theory of text is an everlasting and fascinating adventure.

The editors

A Postscript

All of us, who are Hřebíček´s readers, colleagues, students and successors, know well that perhaps the most characteristic and, probably, the most impressive feature of Hřebíček´s scientific work – from the very beginning of his career – is a unique and original methodological line of thinking and reasoning. For several decades, he has been construing his quantitative text theory as an interdisciplinary field, and as a testable theory, very near to those done/performed/ implemented in mathematics and the natural sciences. He formulates models, hypotheses, and conjectures, using various mathematical procedures to test them. He hates chaos (with the only exception of *deterministic* chaos). He loves order, rules, laws, and synergy. He views and interprets linguistics as a *science of language*. He is one of the well-known representatives of such exact thinking in the European scientific context, as well as a singular voice in the context of present-day Czech linguistics. (Of course, not a singular voice in the sense of an isolated promoter of certain linguistic ideas, but in the sense of a perfectly "sharpened" thinker – a linguist and philosopher of language.)

The festschrift mentioned above contains Hřebíček´s selected bibliography of monographs, textbooks, articles, book reviews and popular articles (not to speak about his translated novels and works for the stage, mainly from Turkish). Still, comparing the rich bibliography published there with what he has written since 2000, it is clear that the period of the last fifteen years has been most productive and most fruitful. In addition to articles and papers published both in the Czech Republic (e. g., in the journals *Slovo a slovesnost* or in *Vesmír*), and abroad (in various volumes, mostly in English), three books have appeared: *Variation in Sequences* (2000, Prague: Oriental Institute); *Vyprávění o lingvistických experimentech s textem* (2001, Prague: Academia); *Text in Semantics* (2007, Prague: Oriental Institute). In addition, Hřebíček is a co-author of a voluminous collective book written together with G. Wimmer, G. Altmann, S. Ondrejovič, and S. Wimmerová *Úvod do analýzy textov* (2003, Bratislava: Veda). The books represent a further decisive step in the development of modern text linguistics. Hřebíček´s ideas are acknowledged, accepted, approved and further developed, especially by the younger generation. The present festschrift is good evidence of it. It is not "only" a birthday present, but a manifestation of the undeniable fact that Hřebíček´s ideas represent a big challenge to follow.

Contents

The study of hrebs

Gabriel Altmann

The creator of these new linguistic units, L. Hřebíček, called them originally *sentence aggregates* or *text aggregates* and defined them as a collection of referential entities. The author re-baptized them to *hrebs* for two reasons: first, in order to make his fellow student and collaborator immortal and second, because the concept could be generalized and applied to different other phenomena than sentences. In linguistics, where all units have a vague identity and can be recognized only with the aid of operational criteria (which are conventions), it is possible to set up textual aggregates on all levels (phonology, morphology, syntax, semantics, pragmatics, argumentation, denotation, etc.), define them and set up *morpheme hrebs, word hrebs, phrase hrebs, clause hrebs, sentence hrebs, speech act hrebs, denotation hrebs, predication hrebs, frequency hrebs*, etc., and in addition, it is also possible to set up *category hrebs* encompassing e.g. all words of a text expressing the same grammatical or even semantic category; and of course, a number of other kinds the idea of which does not yet even exist. As a matter of fact, a similar act has been accomplished since antiquity in the form of parts-of-speech, a phenomenon that entered every grammar, even if there are problems in every language: many English nouns can be used as verbs and all German verbs can be used as nouns, many adjectives are at the same time adverbs, etc. Hence, this classification is not unequivocal but all grammars use it. In texts, the identification is more plain, nevertheless, here new problems may arise (synonymy, polysemy, etc.) whose solution can (and sometimes must) be sought without recourse to etymology. For example, in German "er *ging* auf dem *Gang*", the verb (*walked*) and the noun (*corridor*) have in German the same etymology but do not belong to the same denotation hreb.

None the less, the discovery of hrebs by Hřebíček (1992, 1993, 1995) is something like the discovery of a new particle in physics. Unfortunately, the two introductory works on hrebs (cf. Ziegler, Altmann 2002; Ziegler 2005) were written in German and could not infiltrate into general linguistic thinking. Besides, they are restricted to denotation.

Communication is possible only because one knows what belongs to what, which of the lot of uttered entities concern the same or a cognate or a related entity, etc. But this "belonging to" can be defined in many different ways. At the beginning, the linguistic level of the hreb must be defined; then the "element of the hreb" can be defined either dichotomically (yes - no), or in a fuzzy way by scaling the membership degree of the element in the hreb. The scaling of the membership is, again, a problem of definition. For example, direct naming, synonym, pronoun, reference, anaphora, cataphora, relative clause, structural similarity, qualitative similarity, etc. may obtain different degrees of membership and what is more, units may belong to several hrebs at the same time but in different degrees. Here pre-fabricated criteria play an eminent role. Since hreb is a unit-, level- or property-dependent unit, its construction will differ in different languages, even with different researchers. Its analysis will be accompanied by the same vagueness as the definition of word or sentence (there are more than two hundred definitions of sentence!). Nevertheless, it allows a deeper insight in the inconspicuous aspects of the text than some other classical units.

It is to be noted that there is still another known super-unit, namely Köhler's motif (cf. Beliankou, Köhler, Naumann 2013; Köhler 2006, 2008a,b; Köhler, Naumann 2008, 2009, 2010; Mačutek 2009; Sanada 2010) which can be based either on numerical or qualitative entities. And since these two super-units have already been detected, it can be supposed that there are still other hierarchically ordered levels between the hreb and the text, whose discovery is a task for the future (cf. Köhler, Altmann 2009: 66). The way into the hidden interior of text is not easier than that into the interior of atoms or genes. But while in natural sciences one cannot omit facts obtained by experiments, in linguistics, and especially in semantics, one can restrict the discussion to the definition and identification of entities - a necessary but not sufficient condition for theory construction; one usually restricts the research to some kind of English though with linguists, one would expect the knowledge of more than one language; one speaks about theory omitting any kind of hypothesis testing, and the concept of law is not even known.

Types of hrebs

Whatever the level of the hreb, it can be presented in different ways (cf. Ziegler, Altmann 2002: 31):

(1) *Data hreb* containing all units belonging to it and their places in the text. If one omits the places, one obtains a *list hreb*.

(2) For units which have variants (e.g. allomorphs or word forms) one can define a *set hreb*, containing only the morpheme (no allomorphs) or only the lemma or only some other basic form. These hrebs can be presented also as ordered sets (e.g. alphabetically, according to the length, frequency, weight, centrality, semantic features, etc.)

(3) Ordered *position hreb* containing only the positions of the individual hreb elements in the text.

For illustration we show these hrebs as obtained on the level of words in the poem *Der Erlkönig* by Goethe for the unit "Kind" (*child*) (cf. Ziegler, Altmann 2002: 32)

List hreb:

Kind = [Kind, Knaben, ihn, ihn, Sohn, du, dein, Sohn, du, Kind, dir, mein, mein, mir, Kind, Knabe, du, dich, dich, mein, mein, Sohn, Sohn, dich, deine, du, mein, mein, mich, mir, Kind, Kind]

Set hreb:

Kind = [du, mein, Kind, Sohn, ich, Knabe, dein, er]

Position hreb:

Kind = [15, 19, 26, 30, 36, 39, 54, 59, 61, 72, 86, 88, 96, 104, 113, 114, 121, 136, 138, 140, 153, 155, 169, 172, 177, 184, 186, 191, 195, 211, 223].

The set hreb can be reduced - according to the definition of lemmatizing - identifying *du-dein* (and *ich-mein*) as allosemes of the same sememe.

The above sets can be characterized in different ways and at the same time, they can be used for the characterization of the given text. Needless to say, hrebs display properties or behaviour which can be linked with those of other entities, thus yielding material for a synergetic view of texts.

In the above example of list-hreb, all words designate in some way the same person. In the set-hreb only the basic lemmas are shown and in the position-hreb one obtains a set of numbers which can be further analysed. But even these kinds of hrebs can be set up in different ways by using different identifications. There are several problems associated with hreb construction on any level. Let us mention at least some of them: (a) Distinguishing specific and generic, e.g. *my hand* vs. *the human hand* containing a difference in reference*;* (b) lemma and a synonym used in quite different contexts, e.g. *Obama* vs. *the president*; (c) expression of the same grammatical category, e.g. time, person, number; (d) should one consider only the head of a compound or all parts? (e) Some words may occur in many hrebs, e.g. *are* which belongs not only to persons but also to the verb *be*; the copula *is* may concern a quite great number of nouns in text, etc. (cf. Ziegler 2005); (f) when do pronouns like *I* and *we* belong to the same class? The Indonesian *kita* (inclusive "we") denotes the speaker(s) and the hearer(s). Every text in any language will bring new problems that can be solved only by a decision (= conventional criteria).

The number of word-hrebs in the text may be smaller than that of words because some hrebs may encompass more than one word/lemma. Thus the distribution of the cardinal numbers of list hrebs may differ from the usual word spectrum or the rank-frequency distribution. The higher the level of the hreb, the fewer hrebs are in the text.

For hrebs on higher levels, e.g. phrase, clause, sentence, speech act, references, association, metaphor, it is necessary to apply either more general criteria or numerate the entities in the text and set up merely position hrebs, though using a computer everything can be done.

Hypotheses

There are many kinds of hypotheses whose testing can yield statements about text sort, text construction, the associative world of the writer, etc. Let us present at least some of them in a non formal way:

(1) The *cardinal numbers* of the complete set of hrebs in the text follow their own distribution. It can, of course, be conjectured that some of the Zipfian variants will be adequate. The kind of the hreb and the level may play the role of subsidiary or boundary conditions influencing the parameters, and perhaps also the setting up of differential equations. The basic question is always why the cardinal numbers follow the given regularity. What are the forces or requirements (cf. Köhler 2005) influencing the structuring of the text?

(2) The mean cardinal number of hrebs is linked with *text length*. The longer the text, the greater will evidently be the mean cardinal number because the addition of new words slows down and ever more and more hrebs come to have a greater size (cardinal number).

(3) *Text concentration* (thematic concentration) is linked with (a) the relative size of the greatest word hreb, (b) with the steepness of the rank-frequency distribution of hreb sizes. Thematic concentration (cf. Popescu et al. 2009) can be more exactly captured by considering hrebs (rather than words), because hrebs may be defined also for synsemantics, references, synonymous morphemes, etc. while the thematic concentration based on words takes into account only autosemantics. Of course, it depends also on the condition whether one considers some synsemantics (e.g. prepositions) as parts of the words or not.

(4) The *distances* between the elements of the same hreb in text follow a regular probability distribution or some function. The derivation of such a distribution/function must be interpreted linguistically and may lean against the “unified theory” (cf. Wimmer, Altmann 2005). In order to be able to approach this problem, one must analyze long texts in order to obtain many distances.

(5) Since the entities can simultaneously be elements of several hrebs, different hrebs are linked. If we consider the hreb as a vertex and connect it with all the other ones containing some of its elements, we obtain the *associative graph* of the text. Graphs have a number of properties whose application and evaluation may contribute to text theory (cf. Ferrer-i-Cancho 2013; Watts 2004). This holds for all types of hrebs.

(6) Graphs can be obtained in many different ways, e.g. writing all scrutinized entities (morphemes, words, sentences, references,…) as a sequence of numbers, and joining with edges those numbers whose background entities belong to the same hreb. A very simple procedure is to numerate the sentences and set up a graph representing identities, similarities, references, co-references, etc.

(7) If there are several levels of hrebs and we are able to discern them and construct a hierarchy, then *Menzerath’s law* or its adaptations will make their way in some form. But this is rather a dream of the future. Up to now merely

word and phrase hrebs have been analysed but no hierarchy has been touched. It will not be easy to find and to define a hierarchy.

(8) Having set up the graph of the text in terms of some hrebs, one can scrutinize whether it is a *small world* or does it differ from it (cf. Amaral et al. 2000; Dorogovtsev, Mendes 2002; Ferrer-i-Cancho, Solé 2001, Watts 2004, to mention only some few works). Taking a long text, how many edges are maximally necessary to come from one hreb to any other one? Do text-sorts differ from the small world point of view? How long must a texts be in order to be a small world?

(9) What is the *hubiness* of such graphs? What is the clustering coefficient? How are the graph properties of hrebs linked with one another? Etc.

(10) Do different levels of hrebs produce different *graph properties*? For example, do sentence hrebs produce graphs differing form those produced by words?

(11) What is the *chaining coefficient* of hrebs? (cf. Belza 1971; Skorochod'ko 1981). What is the *coherence* of the text if hrebs of some kind are taken into account? The topical literature does not take hrebs into account but one can adopt the results concerning coherence.

(12) How are all these properties linked? Is there a control cycle regulating at least a part of properties?

(13) Is the behaviour of hreb elements chaotic, random or ordered? This can be studied especially on position hrebs. Is the sequence of the numbers volatile or persistent?

(14) What are the fractal properties of the sequence of position hrebs?

(15) What kind of entities can be chosen as a basis of hreb definition and what kind of behaviour can be expected?

This list could be continued ad infinitum because here several sciences are cooperating (linguistics, textology, statistics, graph theory, fractals,…). For denotative word-hrebs, Ziegler and Altmann (2002) mention properties like topicality, concentration, connotative concentration, diffuseness, compactness, thema-rhema analysis, cohesion, coincidences of different kind, cliques, etc. We recommend to begin with denotation hrebs because they seem to be quite simple, though in strongly synthetic languages problems may arise. Besides, computer programs allow one to draw the respective graph and evaluate it stepwise (cf. Batagelj, Mrvar 2003). But one usually draws one's own graphs in order to display the discovered structure in a special way. Unfortunately, for a hreb-analysis of sentences one needs longer texts; but the longer the texts, the more vague can become some synonymies, polysemies, identities, similarities, references, associations, etc. In poetry, one can take even the verse as a clearly delimited unit and construct verse hrebs on the basis of identities or references etc.

For this reason we illustrate the procedure establishing hrebs in two Slovak texts by the same writer (Eva Bachletová). In the poetic text we establish *verse hrebs* associated semantically by elements belonging to the same hreb and characterize the resulting graph. In the prose text we set up sentence hrebs and compare the two analyses. This is, of course, only an illustration.

Here we shall omit synsemantics. If there are different derivations of a root (e.g. *pocit, znecitlivenie* < *cit* “feeling”) we place the given verses in the same hreb. If the same word, e.g. a pronoun, concerns two different persons, they will belong to two different hrebs. If in a verse more elements of the same hreb occur, it is sufficient for our purpose to consider only one of them because we associate whole verses or whole sentences.

Data

The data were placed at our disposal by E. Bachletová herself but they can be found also on the Internet. For the sake of illustration, we present the poem *Aby spriesvitnela.* The verses are numerated; the numbers behind words show those words which are responsible for the membership of the given verse in a verse hreb. We set up the position hrebs for the poem. As can be seen, some verses belong to several hrebs; some verses are unique members of a verse hreb and do not contribute in our sense to some poem properties. The valency of verbs and nouns is not taken into account. If a verb in the verse is a predicate of some noun (i.e. having the respective person, gender and number markers) which is placed in another verse, we consider the two verses as belonging to the same verse hreb.

Aby spriesvitnela

1. Nemám rada bielu(1)
2. dnes je(1,7) prízrakom chladu
3. znecitlivenia(2)
4. konečného verdiktu
5. nad človekom
6. nad pocitom(2)
7. nad láskou.
8. Dnes je(7) tu iná biela(1)
9. biela(1) obrazovky
10. počítača
11. tam nahadzujeme(3)
12. svoje(3) vnemy
13. čiernymi linkami
14. rýchlo a bezpečne
15. kreslíme(3) životy
16. slovami,(4)
17. ktoré(4) navždy
18. zmenili(4) bielu(1)
19. a odviedli(4) nás(3)
20. od základných farieb

21. bytia.
22. A možno stačí jedna(6)
23. nenapísaná veta(6)
24. aby „novodobá“(1)
25. biela(1) spriesvitnela.
26. Lebo čistá – biela(1) krehkosť(5)
27. prichádza(5) potichu...

We obtain the following positional verse hrebs:

1. (biela, biely) = {1,2,8,9,18,24,25,26}	(white)
2. (cit) = {3,6}	(feeling)
3. (my) = {11,12,15,19}	(we)
4. (slovo) = {16,17,18,19}	(word)
5. (krehkost’) = {26,27}	(softness)
6. (veta) = {22,23}	(sentence)
7. (byť) = {2,8}	(to be)

Since the verses of a given hreb are all connected with one another, each hreb makes up a clique. But since a verse may belong at the same time to several hrebs, we obtain a graph whose properties can be evaluated. Placing the cliques in a square we obtain Figure 1. The vertices 4,5,7,10,13,14,20,21 are isolated and not inserted in the figure. Hreb no. 7 is redundant because it occurs always with hreb no. 1; it can be taken into account or not.

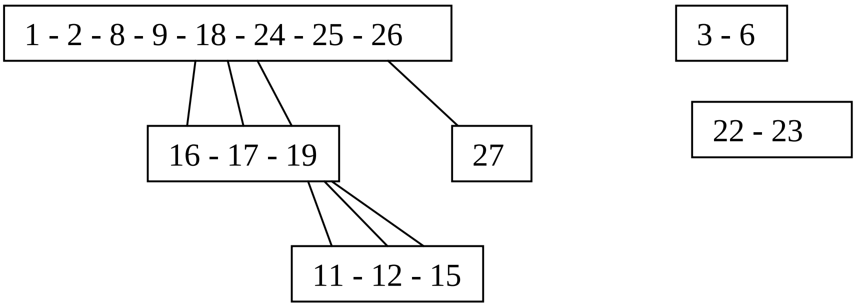


Figure 1. The connections of the verse hrebs of the poem.

In order to evaluate this state of affairs we first look at the degrees of vertices. Degree of a vertex is the number of other vertices which are adjacent to it. In the above Figure 1, we obtain the result resumed in Table 1.

Table 1
The degrees of vertices (verses) in *Aby spriesvitnela*.

Degree d_i	**Number** n_i	**Vertices (Verses) No.**
0	8	4,5,7,10,13,14,20,21
1	5	3,6,22,23,27
3	5	11,12,15,16,17
6	1	19
7	6	1,2,8,9,24,25
8	1	26
10	1	18

Here $n = \Sigma n_i = 27$, (number of verses)
$m = \Sigma d_i n_i = 86$ (sum of edges of all vertices)
$N = n(n\text{-}1)/2 = 27(26)/2 = 351$ (number of possible edges)
$m/n = 3.1851$ (mean number of edges per vertex).

Using these numbers we compute the indicator of *connotative concentration* defined as

(1) $$CC = \frac{m}{N},$$

yielding in or case $CC = m/N = 86/351 = 0.2450$.

That means, connotative concentration is the ratio of the number of realized degrees to that of possible ones. We can consider it as a proportion. Hence the variance of *CC* is

(2) $$Var(CC) = CC(1 - CC)/N,$$

in our case, $Var(CC) = 0.2450(1 - 0.2450)/351 = 0.000698$.

A comparison with other texts for *CC* is possible using the asymptotic normal test. Another possibility is that of comparing *means* using standard methods.

Taking a prose text, *Sila ľudského ducha,* in which we set up sentence hrebs, we obtain $m = 242$, $N = 496$, $CC = 0.4879$. Here, *CC* is the double of that in the poem. Nevertheless, we perform an asymptotic normal test and compute

(3) $$P = \frac{m_1 + m_2}{N_1 + N_2}$$

in order to obtain a common *CC*. The variance of *P* is

(4) $$Var(P) = P(1-P)\left(\frac{1}{N_1}+\frac{1}{N_2}\right),$$

hence

(5) $$u = \frac{|CC_1 - CC_2|}{\sqrt{Var(P)}}.$$

Comparing the above mentioned texts, we obtain *P* = (86 + 234)/(351 + 496) = 0.2932, *Var*(*P*) = 0.2932(1 - 0.2932)(1/351 + 1/496) = 0.001008, hence √0.001008 = 0.031752. Inserting these numbers in (5) we obtain

$$u = \frac{|0.2450 - 0.4718|}{0.031752} = 7.14,$$

displaying a very highly significant difference between the given poetic and prosaic works of the author. However, one cannot say whether this difference will be the same after having analysed a great number of texts, or those by other authors.

In Table 2 one finds the analysis of some poems by E. Bachletová. As can be seen, the very personal poem *Rozt'atá pritomnost'* has the highest *CC*.

Table 2
Connotative concentration in some poems by E. Bachletová.

Text	n	N	m	CC	Var(CC)
Moje určenie	52	1326	198	0.1493	0.000096
Podobnost' bytia	29	406	66	0.1626	0.000335
Aby sprievitnela	29	351	86	0.2450	0.000526
Malý ošiaľ	27	351	102	0.2905	0.000567
Iba neha	54	1431	542	0.3788	0.000154
Tak málo úsmevu	20	190	96	0.5053	0.001316
Stály smútok pre šesť písmen	48	1128	608	0.5390	0.000220
Rozt'atá prítomnost'	36	630	444	0.7048	0.000330

Runs

Selecting one of the many possibilities to characterize the text by means of hrebs we decided to show some possibilities of the analysis by means of the theory of runs. One may ask whether the elements of individual hrebs occur rather in the immediate mutual neighbourhood or are randomly dispersed in the text. If there is certain *local thematic concentration*, then we obtain long runs of the same elements and a (significantly) small number of runs in the text. The text may, of course, be thematically concentrated but the concentration need not be local. On the other hand, even if the text does not display a strong general thematic concentration, it may display local concentrations.

The theory of runs can help us to scrutinize some of these questions. Let us first set up the sequence of verse hrebs in the poem *Aby spriesvitnela*. As can be seen above, the sequence is

1, 1, 7, 2, 2, 7, 1, 1, 3, 3, 3, 4, 4, 4, 1, 3, 6, 6, 1, 5, 5

It is irrelevant whether one uses numbers or hreb names. We have k = 7 different signs and the length of the sequence is n = 21. (This is not the number of verses but of those verses in which an element of a hreb occurred) As can easily be seen, there are r = 12 runs. The individual frequencies are k_1 = 6, k_2 = 2, k_3 = 4, k_4 = 3, k_5 = 2, k_6 = 2, k_7 = 2. It would be possible to take into account also the verses in which there are only isolated words, i.e. the verse is not linked with the other ones. Such verses (or sentences) could be marked as 0.

Since the methods of analysis of runs are easily accessible (cf. e.g. Bortz, Lienert, Boehnke 1990; Gibbons 1971; Fu, Wendy Lou 2003) we shall mention here merely some hypotheses in the form of questions which could be substantiated linguistically and tested using hrebs.

(1) Is the distribution of runs a product of chance or does it display some regularity? If there is some regularity, it may testify to the rise of order in text.

(2) Does the number of runs differ from the expected one? The greater it is, the smaller is the concentration of the text. Hence text books will have fewer runs of hrebs than poems; perhaps a typology of texts could be made up starting from different hreb runs.

(3) It is a small step from runs to Markov models. One can study the order of the chain, the existence of embedded chains and other Markov properties.

(4) If one prepares a table of transitions from one hreb to another, one can study the diagonal of the contingency table. If the diagonal is significantly preferred, it is a sign of local concentrations of the text. Here a number of interpretations are possible. It can be conjectured that such a preference will occur in speech act hrebs observed in a stage play.

(5) It can be conjectured that the distances between individual elements of identical hrebs abide by a regularity which can be modelled. If the distances are random, they follow the Zörnig model (1987), otherwise they can be derived from the unified theory.

(6) In the sequence of (quantified) hrebs, phases may be detected whose number and length may be rather regular than random. The number and length of the phases is another sign of the interrelation between (quantified) hrebs.

(7) The longest run, if it is significantly long, is a sign of a hreb concentration in some part of the poem. The same holds for a sequence of sentences

belonging to the same hreb, etc. A special question is which of the hrebs displays a significant length. It is most probably the central hreb of the text.

(8) If, e.g. in sequences of sentence or verse hrebs, there is a significantly small number of runs, it may testify to the activity of the Skinner effect in the semantic domain.

(9) Do sequences of hrebs of different kinds differ from text to text, from text sort to text sort, from language to language, or do they display a similar behaviour?

(10) Are there other sequential regularities in hrebs displaying strong tendencies?

Again, the number of conjectures, questions and hypotheses could be enlarged and the domain could be further developed.

Conclusion

Hreb analysis is a method for studying denotative, connotative, referential, associative, etc. structures of texts. For the time being, we do not know if special levels are identical in all languages, if there are languages in which unique hrebs can be set up, if hrebs of some kind are indicators of personal style or text sort, or how do hrebs of the same kind differ in a strongly synthetic and strongly analytic language, etc. It is important not to consider a special language as a standard. Hreb analysis cannot replace classical disciplines of linguistics based on rules;of it is a look at a text from many different points of view. It will yield indicators, their correlations or stronger dependencies, it will yield frequency distributions, graphs, control cycles, time series, and at a very high level, one will be able to derive also laws. The future theory of texts will be possible only after a thorough scrutinizing dozens of aspects of hreb analysis. Not before having done it one will one be able to achieve the next dimension positioned somewhere between hreb and text

In any case, later on, some kind of scaling of hreb elements will be necessary. One will surely find various aspects, for example semantic centrality, inclusivity, referentiality, etc. The set hreb will not be merely a set of elements but a set of scaled elements for which possibly laws will be found. Maybe the theory of rough sets (cf. Pawlak 1991; Thomas, Nair 2011) will be useful in this domain. But this is a task for the (remote) future.

References

Amaral, N.A.L., Scala, A., Barthélémy, M., Stanley, H.E. (2000). Classes of behavior of small-world networks. *Proc. Nat. Acad. Sci. USA 97, 11149-11152. (Condmat 0001458)*

Batagelj, V., Mrvar, A. (2003). Pajek - Analysis and Visualization of Large Networks. In: Jünger, M., Mutzel, P. (eds.) *Graph Drawing Software*: 77-103. Berlin: Springer. (http://vlado.fmf.uni-lj.si/pub/networks/pajek/ 06.20.2013).

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012: 35-43.* Belgrade: Academic Mind..

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

Bortz, J., Lienert, G.A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

Dorogovtsev, S.N., Mendes, J.F.F. (2002). Evolution of networks. *Advances in Physics 51, 1079-1187 (Condmat 0106144)*

Ferrer-i-Cancho, R. (2013). Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics 25, 1-21.*

Ferrer-i-Cancho, R., Solé, R.V. (2001). The small world of human language. *Proceedings of the Royal Society Ser. B 268, 2261-2265. (SFI 01-03-016)*

Fu, J.C., Wendy Lou, W.Y. (2003). *Distribution theory of runs and patterns and its application*. Singapore: World Scientific.

Gibbons, J.D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bo--chum: Brockmeyer.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. B. (eds.), *Contributions to Quantitative Linguistics: 33-39.* Dordrecht: Kluwer.

Hřebíček, L. (1995). *Text levels. language constructs, constituents and the Menzerath-Altmann law*. Trier: WVT.

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

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

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

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

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics, Vol 2.* Lüdenscheid: RAM.

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: 637-646*. Berlin, Heidelberg: Springer.

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

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

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

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer Academic Publishing

Popescu, I.-I., Altmann, G. (2011). Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), *Issues in Quantitative Linguistics Vol. 2: 110-116*. Lüdenscheid: RAM.

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

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

Skorochod'ko, E.F. (1981). *Semantische Relationen in der Lexik und in Tex--ten.* Bochum: Brockmeyer.

Thomas, K., Nair, L. (2011). Rough intuitionistic fuzzy sets in a lattice, *International Mathematical Forum 6(27), 1327-1335*.

Watts, D.J. (2004*). Small worlds. The dynamics of networks between order and randomness.* Princeton, N.J.: Princeton University Press.

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

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 423-447*. Berlin: de Gruyter

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Zörnig, P. (1987). A theory of distance between like elements in a sequence. *Glottometrika 2, 1-22.*

Representations of Tutchev’s style: one poet or two?

Sergey Andreev

1. Introduction

The studies of style very often do not take into account the difference in the importance of the author’s works in his creative activity. This importance cannot, of course, be measured directly and depends on many factors: the degree of reading public estimation (at the time when the author wrote it and afterwards), the attitude of critics (at the time of publishing the work and later), the estimation by the author himself, etc.

It is possible to suggest that depending on the choice of the data-source one can establish different realizations (representations) of the style of an author. If we are interested in the perception of an author’s style by his contemporaries, it is reasonable to analyze all his works which were published during his life. In case the task is to find out the author’s own judgment, one should focus on collections, selected by the author himself. On the other hand, if we want to analyze the present-day vision of the author’s style, it would seem natural to limit the data source to those works which became canonical and are published in modern anthologies, textbooks, etc.

These suggestions, of course, will remain mere speculations until tested and differences in the style of the same author depending on the approach to selection of the data-source are found.

2. Hypothesis

The lyrics of Fedor Ivanovich Tutchev[1], a prominent Russian poet, reflect two styles – a “canonized” style, well-known to Russian readers, and the style which is unfamiliar now but was known in the 19th century to the readers of his works. There are certain formal differences between these two styles.

3. Data-base

For this study (a) “popular” and (b) at present “obscure” lyrics were taken. The former are included in numerous selections of his poems, in anthologies of Russian literature, and are learnt by heart and recited by schoolchildren during

[1] Another variant of the poet’s name in English is Fyodor Ivanovich Tyutchev.

literature lessons. The latter are practically unknown to the general public and can be found only in special sources or complete collections.

The division into these two categories of popular (Class 1) and unpopular (Class 2) poems in this study was made on the basis of two collections of Tutchev's poems (Pigarev 1966; Rozman 2011). Differing slightly in the general number of poems which they contain, both collections coincide in how they divide Tutchev's poems into what is here designated as popular and unpopular, placing them separately in different sections (Rozman 2011) or volumes (Pigarev 1966).

To observe homogeneity of the data, the following metrical and stanza restrictions were set. For the present study only iambic four feet lyrics with stanza organization were chosen. One more restriction was for the length of stanzas – they were to be of four lines and the number of stanzas should be more than one. It is worth noting that such verses constitute a very important part of the author's poetic works.

The poems which corresponded to the above-mentioned criteria and were included into the above-mentioned collections formed two classes. Class 1 contains 55 poems (892 verse lines), Class 2 consists of 23 poems (416 lines).

4. Characteristics and method

Choosing characteristics for the analysis we were guided by the following criteria. Characteristics should be well distinguished formally. Besides, which we consider as very important, these characteristics must reflect the essential features of lyrical poems – be poetically and linguistically relevant. This last condition, of course, introduces certain restrictions into our research, changing its perspective from looking for any feature that possesses discriminant force to testing the discriminant value of an a priori formed feature scheme.

Due to these requirements and based on our preliminary empirical data experiments the following list of characteristics was formed.

Rhythmic characteristics. We understand by rhythm a concrete realization of a metric scheme in a verse where meter is the idealized scheme of the ordered alteration of strong (predominantly stressed) and weak (predominantly unstressed) syllabic positions. The strong positions in this paper, according to tradition, are called ictuses. Since in the actual verse the metrical scheme is sometimes violated, ictuses can be unstressed (omission of an ictus stress). Such deviations of rhythm from the metrical scheme are taken into account in this study by using the following features: the number of unstressed first, second and third ictuses. The fourth, last strong position in Russian verse is always stressed. One more characteristic reflecting the rhythm deviation from the metrical scheme is the number of stresses falling on the anacrusis (one or several syllables preceding the first strong position which metrically are not to be stressed in an iamb).

Parts-of-speech characteristics. Those which are used in this study are based on the count of nouns, verbs, adjectives, adverbs and personal pronouns (1st, 2nd, 3rd persons). In the verse line we distinguish between several different positions. Two positions are located at the beginning and the end of the line – the first and the last ictuses. Other positions in the line are the second and the third ictuses. In this study we count separately the parts of speech which are placed (a) in the first and/or the last ictus positions and (b) all the other positions (positions in the middle of the line). The distinction of the first and last positions is due to their special role in the verse structure (Gasparov 2012).

Word-length characteristics. This parameter is based on the count of words of different syllabic structure, ranging from one-syllable to five-syllable words. Longer words in the corpus are so rare that they will be ignored.

Syntactic characteristics. These include the number of inversions (full type – when the predicate or a part of it precedes the subject), and the number of syntactic pauses which take place in a verse line, breaking its unity. Two more features reflect the number of emphatically marked lines (those which end in exclamation marks, question marks or dots) and the number of emphatically marked sentences (when the sentence ends in an exclamation mark, a question mark or dots in the middle of the line).

The examples below as well as the names of the poems in English are taken from the translation of Tutchev's poems made by F. Jude (Jude 2000).

Pause

There's a familiar voice, a wondrous voice, / sometimes a lyre's note, at times a woman's sigh, but I, unwakeable sluggard, / suddenly could not reply. (*To countess E.P. Rostopchina – in reply to her letter*).

Emphatically marked lines

Always the same flow and ebb of the seas, / always that spectre of empty unease... (*The wave and the thought*).
Where are you, sons of Harmony? (*Spring – dedicated to my friends*)

Emphatically marked sentences

So why hang on? And why go yellow? (*Leaves*).
It may suit you or it may not, / why should it ask? – Move on, move on! (*From land to land, from town to town*).

Besides, a poetic syntax feature such as enjambement was used. By enjambement we understand cases when a syntactic pause does not coincide with the rhythmic pause at the end of a verse line. Three generally recognized types of enjambement were counted in this study: rejet, contre-rejet and double-rejet. Rejet is observed when the pause is moved to the middle of the second line, contre-rejet takes place when the pause is moved to the middle of the first line. In case of double-rejet pauses are observed in both the first and the second lines.

Enjambements below are given in italics.

Rejet
A glowing sphere rolls / into the ocean, which enfolds / the calm, evening red (*A Summer Evening*).

Contre-rejet
I love May's first storms: / chuckling, *sporting spring / grumbles in mock anger* (*A Spring Storm*).

Double-rejet
Carnations peak slyly, *nestling beside / more fragrant*, warmer roses (*Hide and Seek*).

The characteristics in the study are formal, but it should be mentioned that all the so-called formal features of verse (rhythm, syntax) have semantic relevance.

All the characteristics after counting were normalized over the number of lines in the poems. As a result a database was formed in which lines are poems and columns are characteristics. The data was processed with the help of discriminant analysis which has been very effective in many stylometric experiments (Baayen et al. 2002; Mikros et al. 2000; Murata 2000).

5. Results

The analysis showed that there is a clearly-marked difference between these two classes of popular and non-popular poems. A post-hoc test showed the following results (Table 1). Rows represent the classification observed in reality; columns show how the texts were grouped on the basis of the discriminant model characteristics.

Table 1
Classification Matrix (Class 1 vs. Class 2).

	Percent Correct	**Class 1**	**Class 2**
Class 1	98.2	54	1
Class 2	87	3	20
Total	94.9	57	21

The number of correctly classified cases (over 94%) proves that the discriminant model is rather effective. It includes such characteristics as omission of stress on Ictuses 1, 2 and 3, the number of words of different length (2, 3, 5 syllables), the number of nouns and verbs in the marginal positions and nouns, in non-marginal positions (in the middle of the line), the number of adverbs in the middle of the line and at the end of the line, the number of enjambements (rejet and counter-rejet types).

The poem of Class 1 which was placed into the other class is *Den' vechereet*, *noch' blizka* (Day turns to evening. Night approaches). The poems of Class 2 which were misclassified are *Knyazyu Suvorovu* (To Prince Suvorov), *Rassvet* (Daybreak) and *Nad russkoi Vil'noi starodavnoi* (Over ancient, Russian Vilnius).

The following question arises – could the difference between the classes be caused by the changes of the style of Tutchev over time?

In order to address this problem Class 1 was divided into two subclasses. Subclass 1A includes poems, written before or during the year 1840, which marks the first half of the author's creative activity (25 texts, 416 lines), all the other poems of Class 1, written later, form Subclass 1B (30 texts, 476 lines).

Class 2 vs. Subclass 1A

The results of discriminant analysis of these two groups showed obvious differences. The features which differentiate the groups have much in common with the discriminant model, given above: omission of stress on Ictuses 1, 2 and 3, the number of words of different length (2, 5 syllables), nouns in the marginal and non-marginal positions, the number of rejets. Besides there is one more characteristic which possesses discriminant force – the number of adverbs in the middle of the line. On the other hand such characteristics from the first discriminant models as verbs in the marginal positions and contre-rejet are missing.

An ad hoc test gave the results (Table 2) which display the substantial correctness of the classification (over 97%). This time only one poem was wrongly classified – the poem *Nad russkoi Vil'noi starodavnoi* (Over ancient, Russian Vilnius) from Class 2.

Table 2
Classification Matrix (Class 2 vs. Subclass 1A).

	Percent correct	**Subclass 1A**	**Class 2**
Subclass 1A	100	25	0
Class 2	95.7	1	22
Total	97.9	26	22

Since in Class 2 there are no poems which were written before 1849 and all the poems of Subclass 1A, as it was mentioned above, belong to the first period of the creative activity of Tutchev, the difference of style in these two text groups can really be explained by the time difference of writing the poems. But this is not the case with Class 2 and Subclass 1B whose works were both written during the period of the late 40s through the early 70s.

Class 2 vs. Subclass 1B

The comparison of Class 2 and subgroup B of Class 1 showed explicit differences between them. The results of a post-hoc test are shown in Table 3. Here again the automatic classification coincides closely with the "natural" grouping.

Table 3
Classification Matrix (Class 2 vs. Subclass 1B).

	Percent correct	**Subclass 1B**	**Class 2**
Subclass 1B	93.3	27	2
Class 2	95.7	1	22
Total	94.3	29	24

The discriminant model here includes the omission of stress on Ictuses 1, 2, and 3, the number of words of 2 and 5 syllables, the number of nouns and verbs in the beginning and the middle of the line, and the number of rejet and emphatic lines.

In all three models the following characteristics are present: unstressed Ictuses 1 and 2, 2-syllable words, nouns in the middle of the verse line and enjambements of the rejet type.

In all three models the following characteristics are present: unstressed Ictuses 1, 2 and 3, 2 and 5-syllable words, nouns in the middle of the verse line and enjambements of rejet type.

6. Conclusions

The results obtained in this study show that classes of popular and non-popular texts have certain formal differences which cannot be explained only by style changes over time. A culturally significant image of the style of the poet was created partially at the end of the 19th and mostly in the 20th century and has been preserved till now. "The other" Tutchev, on the contrary is not known at all, nor is he studied.

The choice of texts for modern collections was made on the basis of preferences of the reading public and critics of the 20th century. The principles and criteria of such evaluation were, of course, implicit, but as our experiment showed, they have a systematic character and deal not only with the contents of the poems, but also with their formal linguistic characteristics. As a result, style description, based only on the texts from selected collections of poems by

Tutchev, will be a priori imprecise: it will reflect only the style of the accepted vision of Tutchev's creative activity.

There is no reason to think that the situation with other poets really differs. The existence of two poetries – known, culturally accepted at some given period, and relatively unknown which may considerably differ from the author's canonical works, can occur in numerous other cases. Stylometric analysis should take into account this aspect and specify the aims of analysis to make the correct selection of data sources.

Acknowledgment

This research is supported by a grant from the Russian Humanitarian Scientific Fund (project # 14-04-00266).

References

Baayen, R.H., van Halteren, H., Neijt, A., Tweedie, F. (2002). An experiment in authorship attribution. In: *Proceedings of JADT 2002: 29–37*. Université de Rennes, St. Malo.

Gasparov, M.L. (2012). Ritmiko-sintaksicheskie klishe i formuli v epiloge "Ruslana i Ludmily" (Rythmic-syntactic clichés and formulas in the epilogue of *Ruslan and Ludmila*). In: *Lingvistica Stiha*. Moscow: Yaziki slavyanskoy kul'tury.

Jude F. (2000). *The Complete Poems of Tyutchev in an English Translation by F. Jude*. Durham, http://www.dur.ac.uk/~dem8fj/ [accessed 28.11.2009]

Mikros, G., Carayannis, G. (2000). Modern Greek corpus taxonomy. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluations, Vol. 3: 129–134*. Athens, Greece (31 May – 2 June).

Murata, M. (2000). Identify a text's genre by multivariate analysis – using selected conjunctive words and particle-phrases. *Proceedings of the Institute of Statistical Mathematics 48(2), 311–326.*

Pigarev, K.V. (ed.) (1966). *F.I. Tutchev. Lyrika*. (Lyrics), *Vol. 1–2*. Moscow: Nauka.

Rozman, N. (ed.) (2011). *Tutchev F.I. Lyublyu grozu v nachale maya … stihotvoreniya, pis'ma* (I love May's first storms … Poems, letters). Moscow: Eksmo.

Supplement

Raw data

poems	number of lines	rejet	contre-rejet	double-rejet	pause	emphatic sentence	emphatic line	inversion (full)
1	32	0	1	1	10	0	8	1
2	16	0	0	0	1	0	0	1
3	12	0	1	0	4	0	0	1
4	16	0	0	0	4	0	1	4
5	24	0	5	0	8	0	5	0
6	8	0	0	0	0	0	0	1
7	16	1	0	0	4	0	0	4
8	8	0	1	0	2	0	0	3
9	8	0	0	0	1	0	0	3
10	24	1	2	0	6	0	2	8
11	16	0	1	1	5	0	1	1
12	12	1	0	0	5	0	4	1
13	12	0	1	1	6	0	3	1
14	12	0	1	0	5	0	2	3
15	8	0	1	0	3	0	2	0
16	28	0	2	1	17	2	10	1
17	24	0	1	1	11	0	1	4
18	16	0	0	1	5	0	7	1
19	12	0	0	0	3	0	3	3
20	8	0	0	0	5	0	4	1
21	8	0	0	0	4	0	1	1
22	24	1	1	0	12	0	7	3
23	28	0	3	0	10	0	2	2
24	20	0	0	0	5	0	2	2
25	24	1	0	0	10	0	6	4
26	20	1	0	1	9	0	4	4
27	20	2	0	2	10	0	1	0
28	8	0	0	0	8	1	3	1
29	16	0	1	0	5	0	2	3
30	16	0	0	0	6	0	3	4
31	16	0	1	0	8	0	2	2
32	12	1	0	1	6	0	1	0
33	12	0	0	0	4	0	1	2
34	16	0	2	1	8	0	3	0
35	16	0	0	1	4	0	3	3

36	12	1	2	0	6	0	5	1
37	12	0	1	0	7	0	3	1
38	24	0	0	0	6	1	8	6
39	12	1	0	0	4	0	4	3
40	8	1	0	0	3	0	2	4
41	12	1	0	0	4	0	2	0
42	16	0	0	0	3	0	2	2
43	12	0	1	0	2	0	1	3
44	20	0	0	1	7	0	0	3
45	8	0	1	0	1	0	2	2
46	16	0	1	0	5	1	1	3
47	8	0	0	0	0	0	2	0
48	12	1	0	0	1	0	0	1
49	8	1	2	0	5	0	0	0
50	12	1	0	0	2	0	2	1
51	20	0	1	0	4	0	0	3
52	12	0	0	0	2	0	1	4
53	20	1	1	0	5	0	3	3
54	20	1	0	1	4	0	2	6
55	20	1	2	0	5	0	4	4
56	32	0	0	0	12	2	7	3
57	36	2	2	0	6	0	0	4
58	32	2	1	0	8	0	5	1
59	8	1	1	0	1	0	0	1
60	16	0	1	0	5	0	4	2
61	12	1	0	0	4	0	3	5
62	20	1	1	1	7	0	5	3
63	12	1	1	0	3	0	4	1
64	20	1	2	0	9	0	2	1
65	8	1	1	0	4	0	2	0
66	20	1	0	0	9	0	1	0
67	16	0	0	2	12	0	2	0
68	8	0	1	0	2	0	0	1
69	8	0	0	0	3	0	1	1
70	20	0	0	0	9	1	7	1
71	8	0	0	0	3	0	0	0
72	32	0	1	0	9	1	6	6
73	16	2	1	0	2	0	4	4
74	12	1	0	0	2	0	2	0
75	20	0	0	0	2	0	0	4
76	20	0	1	1	5	0	2	1
77	20	0	1	0	4	0	3	2
78	20	0	0	0	7	0	3	4

poems	number of lines	Parts of speech in the marginal position				Parts of speech in the middle of the line				Personal pronouns		
		nouns	verbs	adjectives	adverbs	nouns	verbs	adjectives	adverbs	person-1	person-2	person-3
1	32	26	18	9	3	14	3	10	2	2	1	2
2	16	14	9	3	2	7	1	7	3	1	0	0
3	12	8	5	4	1	7	0	4	0	0	4	0
4	16	12	12	6	0	11	4	6	1	0	1	0
5	24	16	7	10	4	15	3	5	0	7	0	0
6	8	10	2	2	0	2	3	3	1	0	0	0
7	16	12	6	9	2	9	3	4	0	0	0	0
8	8	8	2	2	3	3	3	2	0	0	0	1
9	8	7	2	2	3	3	3	3	2	0	0	0
10	24	17	13	6	4	9	5	9	1	4	0	1
11	16	17	7	4	0	4	3	10	3	0	0	0
12	12	14	3	3	0	6	4	3	0	0	0	4
13	12	9	7	3	2	5	7	4	1	2	0	3
14	12	9	5	2	3	3	6	2	0	2	0	0
15	8	5	3	5	0	4	1	4	4	0	0	0
16	28	24	12	7	3	17	5	4	3	3	5	2
17	24	18	13	7	4	15	7	6	4	4	3	0
18	16	12	6	10	2	8	3	2	2	0	0	0
19	12	11	8	3	0	5	5	1	1	1	0	2
20	8	6	6	1	0	3	2	2	3	3	1	2
21	8	7	3	1	2	5	3	1	2	2	0	3
22	24	21	15	2	1	17	6	7	0	0	2	11
23	28	19	13	10	8	10	6	3	7	0	2	1
24	20	13	12	4	4	8	2	7	0	0	1	6
25	24	15	11	10	4	8	8	13	2	1	0	7
26	20	15	10	6	3	15	10	6	3	5	0	0
27	20	19	9	7	1	19	9	7	1	0	0	2
28	8	4	9	1	1	4	9	1	1	0	0	0
29	16	9	6	1	2	9	6	1	2	3	4	3
30	16	10	5	8	2	10	5	8	2	3	4	0
31	16	11	6	5	2	11	6	5	2	0	0	1
32	12	11	5	2	0	11	5	2	0	0	2	2
33	12	9	7	3	0	9	7	3	0	0	0	0
34	16	15	5	5	0	15	5	5	0	4	0	3
35	16	17	4	5	0	17	4	5	0	1	0	4

36	12	8	4	1	0	8	4	1	0	4	0	0
37	12	13	2	4	0	13	2	4	0	0	2	0
38	24	27	5	5	3	27	5	5	3	4	1	3
39	12	12	7	3	0	12	7	3	0	0	2	0
40	8	11	3	6	4	11	3	6	4	0	0	0
41	12	7	8	3	1	7	8	3	1	2	0	2
42	16	12	7	2	3	12	7	2	3	2	0	4
43	12	9	4	4	3	9	4	4	3	1	0	0
44	20	15	10	5	11	15	10	5	11	1	6	1
45	8	5	2	5	0	5	2	5	0	0	1	0
46	16	14	11	2	1	14	11	2	1	0	1	2
47	8	6	3	6	0	6	3	6	0	1	0	3
48	12	5	3	5	0	5	3	5	0	0	0	4
49	8	5	7	2	1	5	7	2	1	0	0	1
50	12	14	4	2	0	14	4	2	0	1	0	0
51	20	18	5	9	2	18	5	9	2	4	0	1
52	12	10	5	2	3	10	5	2	3	4	1	0
53	20	17	6	7	1	17	6	7	1	3	0	2
54	20	16	9	4	0	16	9	4	0	7	1	2
55	20	16	7	3	4	6	8	4	5	4	3	0
56	32	8	0	4	1	28	18	20	2	11	0	2
57	36	29	13	12	3	13	9	11	7	4	0	6
58	32	21	9	22	2	13	7	5	4	3	2	2
59	8	3	2	4	4	2	3	1	0	2	0	2
60	16	16	4	3	5	2	5	6	1	0	1	1
61	12	10	2	3	3	2	5	5	2	0	1	0
62	20	20	4	8	1	8	5	6	0	0	5	0
63	12	12	4	1	3	2	2	5	1	0	3	0
64	20	13	5	7	2	9	3	9	0	3	0	7
65	8	8	0	4	1	4	2	4	1	0	0	2
66	20	14	10	4	0	6	10	3	2	0	1	0
67	16	10	6	3	1	5	5	3	0	1	0	6
68	8	7	2	3	0	1	2	0	0	0	2	0
69	8	4	5	3	2	6	1	1	0	1	0	1
70	20	12	14	3	2	16	5	7	2	0	1	2
71	8	7	5	1	1	3	3	1	0	2	0	0
72	32	13	17	8	2	14	8	4	4	1	7	11
73	16	13	7	5	0	4	6	3	3	0	0	3
74	12	11	3	2	1	3	3	2	0	0	5	0
75	20	19	6	7	4	10	5	6	1	0	1	0
76	20	17	8	6	3	8	6	4	0	1	1	0
77	20	13	8	5	2	12	4	7	0	0	0	5
78	20	8	11	5	2	4	9	4	2	5	3	2

poems	number of lines	1 syllable words	2 syllable words	3 syllable words	4 syllable words	5 syllable words	anacrusis	unstressed ictus 1	unstressed ictus 2	unstressed ictus 3
1	32	38	44	25	11	2	1	8	2	23
2	16	18	24	13	5	1	2	3	3	11
3	12	13	17	13	3	1	1	3	3	7
4	16	18	29	13	4	1	0	3	0	9
5	24	34	34	19	5	5	2	7	8	11
6	8	8	8	8	5	0	1	2	1	6
7	16	15	18	17	6	2	0	4	2	13
8	8	12	8	8	4	0	0	1	3	4
9	8	9	10	13	0	0	0	0	0	5
10	24	31	27	24	8	3	0	8	4	17
11	16	22	21	15	3	3	2	6	1	9
12	12	15	22	11	1	0	0	3	2	6
13	12	15	30	7	0	0	0	4	1	4
14	12	25	19	6	4	1	3	5	3	7
15	8	11	8	9	2	0	3	3	1	4
16	28	54	56	14	4	0	5	5	8	13
17	24	31	40	19	9	0	3	7	3	11
18	16	19	17	15	7	2	2	4	2	10
19	12	6	18	12	6	0	1	2	2	8
20	8	20	7	10	1	0	1	2	1	4
21	8	15	17	5	0	0	1	1	0	5
22	24	56	42	17	5	0	1	8	3	14
23	28	26	37	24	14	2	6	8	1	22
24	20	24	35	16	7	0	0	10	0	16
25	24	31	31	23	8	2	5	13	1	19
26	20	22	29	20	6	0	2	7	2	14
27	20	28	37	17	2	2	0	3	1	17
28	8	13	17	7	0	0	1	2	0	3
29	16	21	25	14	4	2	0	4	1	15
30	16	28	24	10	3	2	2	4	3	8
31	16	21	24	14	2	1	2	6	2	9
32	12	15	13	15	4	0	2	3	3	8
33	12	11	14	15	3	1	1	3	1	7
34	16	29	20	11	6	2	2	8	2	12
35	16	29	20	9	5	4	1	8	3	8
36	12	28	18	7	3	1	1	6	2	7
37	12	19	14	11	3	2	3	3	1	9
38	24	40	45	17	6	0	1	6	4	13

39	12	22	15	15	0	0	2	1	3	8
40	8	15	15	6	2	1	1	0	2	3
41	12	19	16	13	3	0	1	4	0	10
42	16	25	25	25	25	25	4	5	1	12
43	12	26	14	8	6	0	3	6	1	9
44	20	39	30	15	5	1	2	5	7	13
45	8	8	14	8	2	0	0	2	2	4
46	16	22	26	15	3	0	2	2	0	10
47	8	9	10	8	1	1	0	2	1	7
48	12	24	14	11	3	1	5	1	9	6
49	8	20	11	6	2	0	0	2	2	5
50	12	14	18	13	3	0	0	3	4	5
51	20	30	23	17	5	2	1	8	3	12
52	12	24	20	9	3	0	3	1	9	6
53	20	37	31	11	3	4	6	1	13	10
54	20	36	21	14	6	4	9	3	14	10
55	20	118	84	41	9	12	3	3	11	10
56	32	67	36	27	8	4	1	11	1	23
57	36	64	48	27	11	3	6	14	2	26
58	32	38	51	17	9	9	4	12	3	21
59	8	12	9	10	2	0	0	3	1	7
60	16	23	23	13	5	1	0	2	2	8
61	12	12	21	7	5	1	0	3	0	9
62	20	26	29	14	6	4	2	5	2	13
63	12	20	13	9	6	1	4	4	0	8
64	20	34	35	12	6	0	1	4	0	12
65	8	13	17	3	3	0	2	1	1	3
66	20	39	29	14	4	2	2	6	2	14
67	16	30	20	18	3	0	4	4	1	10
68	8	10	5	7	0	3	0	4	2	6
69	8	15	13	4	2	0	0	2	1	3
70	20	30	32	14	7	1	2	3	2	9
71	8	20	10	8	1	0	0	4	0	5
72	32	62	48	27	8	0	3	7	1	18
73	16	25	17	12	9	1	1	6	0	9
74	12	16	12	12	5	0	1	2	1	9
75	20	13	31	19	7	2	3	4	1	14
76	20	20	32	19	2	4	2	2	3	12
77	20	26	28	21	5	1	3	3	2	11
78	20	44	28	17	5	0	3	7	1	11

Poems

1. Problesk (The Gleam); 2. Letni vecher (A Summer Evening); 3. Lebed' (The Swan); 4. Vesennyaya groza (A Spring Storm); 5. Bessonnitsa (Insomnia); 6. Utro v gorakh (Morning in the Mountains); 7. Snezhnye gory (Snowy Mountains); 8. Vecher (Evening); 9. Polden' (Midday); 10. Eshchyo shumel vesyolyi den' (The happy day was loud); 11. Bezumie (Madness); 12. Strannik (The Wanderer); 13. Vesennie vody (Vernal Waters); 14. Kak nad goryacheyu zoloi (As a piece of paper); 15. Zdes', gde tak vyalo svod nebesnyi (Here the sky stares inert); 16. Iz kraya v krai, iz grada v grad (From land to land, from town to town); 17. Ya pomnyu vremya zolotoe (I remember a golden time); 18. I grob opushchen uzh v mogilu (Into the grave the coffin's lowered); 19. Vostok belel. Lad'ya katilas' (The east whitened); 20. Kakoe dikoe ushchel'e (What a wild ravine!); 21. S polyany korshun podnyalsya (The kite lifts from the field); 22. Ne to, chto mnite vy priroda (Nature is not what you think it is); 23. Vchera, v mechtakh obvorozhyonnykh (Last night in enchanted dreams); 24. Ne ver', ne ver' poetu, deva (Don't believe the poet, girl!); 25. Zhivym sochuvstviem priveta (With a lively, sympathetic greeting); 26. Glyadel ya, stoya nad Nevoi (I stood by the Neva, my gaze); 27. Poshli, Gospod', svoyu otradu (Lord, send your comfort); 28. Ne rassuzhdai, ne khlopochi! (Forget all cares, don't reason deep!); 29. Ne raz ty slyshala priznan'e (You've often heard the admission); 30. Den' vechereet, noch' blizka (Day turns to evening. Night approaches); 31. Smotri, kak na rechnom prostore (Across the river's broad expanse you see); 32. Chemu molilas' ty s lyubov'yu (What you guarded in your heart); 33. Siyaet solntse, vody bleshchut (The sun is shining, waters glisten); 34. Ya ochi znal – o eti ochi! (I knew a pair of eyes. Oh, what a sight!); 35. Bliznetsy (The Twins); 36. Leto 1854 (Summer 1854); 37. O veshchaya dusha moya! (Oh, my prophetic soul!); 38. 1856 "Stoim my slepo pred sud'boyu" (1856 "Blindly we face fate"); 39. Nad etoi tyomnoyu tolpoi (Above this ignorant crowd); 40. Est' v oseni pervonachal'noi (There is a fleeting, wondrous moment); 41. Smotri, kak roshcha zeleneet (Look at the coppice!); 42. Ona sidela na polu (She was sitting on the floor); 43. Dekabr'skoe utro (A December Morning); 44. Pri posylke Novogo Zaveta (On Sending the New Testament); 45. Inym dostalsya ot prirody (Nature has endowed some with a sense); 46. Utikhla biza.... Legche dyshit (The breeze has dropped and lighter is the breath); 47. Kak nerazgadannaya taina (Like an unresolved mystery); 48. Kak ni besilosya zlorech'e (Let foul slander rage); 49. Net dnya, chtoby dusha ne nyla (Not a day relievs the soul of pain); 50. Pevuchest' est' v morskikh volnakh (The sea is harmony); 51. Kogda dryakhleyushchie sily (When our decrepit energies turn traitor); 52. Opyat, stoyu ya nad Nevoi (Once more by the Neva I stand); 53. Yu. F. Abaze (To Yu F. Abaza); 54. Kak nas ni ugnetai razluka (No matter how we're crushed by separation); 55. K.B.; 56. Grafine E.P. Rostopchinoi "V otvet na eyo pis'mo" (To Countess E.P. Rostopchina "In Reply to her Letter"); 57. Kak letnei inogda poroyu (Just as now and then during summer); 58. Na yubilei N.M. Karamzin (On the Jubilee of N.M. Karamzin); 59. Velen'yu vyshemy pokorny (Submissive to a high command); 60.

Knyazyu P.A. Vyazemskomu “Teper' ne to, chto za polgoda” (To Prince P.A. Vyazemsky “It's not the same now as it was six months back”); 61. Ty l'dolgo budesh' za tumanom (Russian star, will you always seek); 62. Teper' tebe ne do stikhov (You're not in the mood for verses); 63. Po sluchayu priezda avstriiskogo ertsgertsoga na Pokhorony imperatora Nikolaya (On the Occasion of the Arrival of the Austrian Archduke at the Funeral of the Emperor Nicholas); 64. Est mnogo melkikh, bezymyannykh (There are many tiny; unnamed); 65. Grafine Rostopchinoi “O, v eti dni – dni rokovye” (To Countess Rostopchina “Oh, in these days, these fateful days”); 66. Chemu by zhizn' nas ni uchila (Whatever life might have taught us); 67. 12-oe aprelya 1865 (April 12th. 1865); 68. Knyazyu Suvorovu “Dva raznorodnye stremlen'ya” (To Prince Suvorov “Two disparate tendencies”); 69. I v Bozh'em mire to zh byvaet (In God's world it can happen); 70. Rassvet (Daybreak); 71. Grafine A.D. Bludovoi (To Countess A.D. Bludova); 72. Slavyanam “Oni krichat, oni grozyatsya” (To the Slavs “They shout, they threaten”); 73. Po prochtenii depesh imperatorskogo kabineta; napechatannykh v "Journal de St. Petersbourg" (On Reading the Imperial Despatches, Printed in the Journal de St. Petersbourg); 74. Vy ne rodilus' polyakom (You weren't born a Pole); 75. Nad russkoi Vil'noi starodavnoi (Over ancient, Russian Vilnius); 76. Da, vy sderzhali vashe slovo (Yes, you have kept your word); 77. Den' pravoslavnogo Vostoka (On this day of the Orthodox East); 78. Net, ne mogu ya videt vas.... (No, I can’t see you...)

The Moran-Hutchinson formula in terms of Menzerath-Altmann's law and Zipf-Mandelbrot's law

Jan Andres

Dedicated to Luděk Hřebíček
on the occasion of his 80th birthday

1. Introduction

Postulating the self-similarity property as the definition of a fractal, Luděk Hřebíček conjectured a *weak version of a fractal structure of language* in the sense that formally the same Menzerath–Altmann law holds on every linguistic level (see e.g. Hřebíček 2002, and the references therein).

Furthermore, he noticed that the reciprocal value of the shape parameter in the truncated formula of the Menzerath-Altmann law can be expressed as a self-similarity dimension by means of the Moran-Hutchinson formula (see Hřebíček 1994, 1998).

The same relationship was already detected by Benoit B. Mandelbrot for the exponent in the Zipf-Mandelbrot law (see Mandelbrot 1983 Chapter XI, 2000 Chapter 12 and the references therein).

On the other hand, in view of these observations, the interpretation of both linguistic laws in fractal terms was only possible for structures exhibiting the very restrictive exact (strict) self-similarity property. We proposed a more realistic interpretation in Andres (2009) and Andres and Rypka (2012), for linguistic structures with a more liberal cyclic self-similarity.

In the present paper, the above discoveries will be recalled in more detail at first. Then they will be critically discussed in a new light. Our goal is not just a description of a language structure under consideration, but also an attempt at a possible explanation of the meaning of the computed fractal dimensions of some given texts. To characterize an author's style or to distinguish between originals and falsa and so on, would be, however, too ambitious. We understand that after the first steps in this field, based especially on linguistic experiments (see e.g. Hřebíček 2002, 2007; Eftekhari 2006; Köhler 2008; Andres, Benešová 2011, 2012; Andres et al. 2012a), there is still a long way to go.

2. Menzerath–Altmann law: complete formula vs. truncated formula

The Menzerath–Altmann law (hereinafter MAL) expresses the relationship between the length x of a *construct* (i.e. a unit on an upper linguistic level) and the length y of its *constituents* (i.e. units on a lower linguistic level). The *complete formula* of MAL takes the form (cf. Altmann 1980; Wimmer, Altmann, Hřebíček, Ondrejovič, Wimmerová 2003)

$$y = Ax^{-b}\,\mathrm{e}^{cx},$$

or, equivalently (for $b \neq 0$),

$$\frac{1}{b} = \frac{\log x}{\log\left(\frac{A}{y}\exp(cx)\right)} = \frac{\ln x}{\ln\left(\frac{A}{y}\exp(cx)\right)},$$

where A, b, c are real parameters and e = 2.718… is Euler’s number. It can be obtained, by integration w.r.t. x and delogarithmization, as a general solution of the linear ordinary differential equation

$$\frac{\dot{y}}{y} = \frac{-b}{x} + c, \quad \text{where } \dot{y} = \frac{dy}{dx}.$$

For $c = 0$, it reduces to the *truncated formula*, i.e.

$$y = Ax^{-b}, \qquad \text{resp. (for } b \neq 0\text{),} \qquad \frac{1}{b} = \frac{\log x}{\log \frac{A}{y}}.$$

In the particular case, when $A = y(1) = y_1$, we get the formula

$$y = y_1 x^{-b}, \qquad \text{resp. (for } b \neq 0\text{),} \qquad \frac{1}{b} = \frac{\log x}{\log \frac{y_1}{y}}$$

with only one free parameter b, usually called a *shape parameter*.

At first glance, one might expect that the complete formula is with no doubt more suitable for applications (because of the three free parameters) than the last, simpler, formula, where $A = y(1) = y_1$ and $c = 0$ were fixed. Nevertheless, as demonstrated both theoretically as well as practically (by means of an illustrative linguistic example in Andres et al. (2012b), the situation is rather delicate.

More precisely, although the application of the complete formula is certainly better from the point of view of the optimal approximation of the given data, it is quite the opposite from the point of view of the accuracy of the estimates of parameter *b* (which is crucial for our investigation). Thus, the suitability of the usage of the above formulas must be evaluated from two perspectives at least. In our linguistic experiment in Andres et al. (2012b), rather surprisingly, the simplest formula is optimal from the statistical point of view. We believe that this particular case holds in general, provided the structure of the given data is statistically significant.

Moreover, since the shape of the graph of the function $y = Ax^{-b}e^{cx}$ with $c > 0$ does not often reflect the verbal form of MAL, for larger values of *x* (cf. Andres et al. 2012b), it can only be used for small integers *x*. On the other hand, the "optimal" usage of the simple formula creates a serious difficulty, for the related regression model. Namely, since the value y_1 represents the realization of a random variable, it contains the measurement error which is automatically transferred into the regression model. Subsequently, all the resulting values must only be considered as those conditioned by y_1.

The second criticism about the usage of the complete formula concerns its derivation. It seems to us that the implementation of an additional parameter *c* via the exponential function e^{cx} is rather artificial (i.e. without any satisfactory linguistic explanation), just because of a simple application of a linear regression technique. For more details concerning this technique, see e.g. Montgomery et al. (2006).

Perhaps more appropriately, instead of the initial equation

$$\frac{\dot{y}}{y} = \frac{-b}{x} + c,$$

one should rather start from the one of the form

$$\frac{\dot{y}}{y} = \frac{-b}{(x+c)},$$

because in this way parameter *c* can be interpreted as something like a "mental over-ride" in our speech as we complete a construct. Thus, the related formula would take the form

$$y = A(x+c)^{-b},$$

or, equivalently (for $b \neq 0$),

$$\frac{1}{b} = \frac{\log(x+c)}{\log\frac{A}{y}}.$$

Unfortunately, a direct application of the linear regression technique fails here, because after its logarithmization, the equation becomes nonlinear. On the other hand, the resulting nonlinear equation can be still linearized, and subsequently the linear regression technique can be applied to a linearized equation.

3. Zipf-Mandelbrot law: isomorphism of form

The Zipf-Mandelbrot law (hereinafter ZML) describes the distribution of word frequencies for *regular lexicographic trees*. It takes the form (see e.g. Mandelbrot 1983 Chapter XI, 2000 Chapter 12; Manin 2009; Montemurro 2001, 2004; Wimmer, Altmann 2005, and the references therein)

$$U = P(\rho + V)^{-1/D}$$

or, equivalently,

$$D = \frac{\log(\rho + V)}{\log\frac{P}{U}},$$

where ρ denotes the order of a word (words were ordered in a decreasing way according to their frequencies) with probability U and P, V, *and* D are real constants. The regularity of trees means that each branching is related to a single word and the probability weight on the k-th level takes the form $U = U_0\, r^k$, where $0 < r < 1$, and U_0 is such that the sum of all probability weights is equal to 1.

Observe that, despite their meaning, the last two formulas, i.e. $y = A(x+c)^{-b}$ and $U = P(\rho + V)^{-1/D}$, are formally the same. Thus, ZML can be again obtained as a general solution of the linear ordinary differential equation

$$\frac{\dot{U}}{U} = \frac{-1/D}{(\rho + V)}, \text{ where } \dot{U} = \frac{dV}{d\rho}.$$

As pointed out in Altmann et al. (1989), the effect of MAL can be detected not only in various domains of linguistics, but also in nonlinear biology, sociology and psychology. The same is true for ZML, as explained in economic terms in Mandelbrot (1983 Chapter 13, 2000 Chapter 12), for a mathematically equivalent form of *Pareto's law*, i.e.

$$\rho = -V + U^{-D} P^{D}.$$

Such an isomorphism documents the well-known ***principle of least effort*** *reflecting the economization of nature* (see Wimmer, Altmann 2005, and the references therein) or a certain sort of a ***conservation law***.

There is, however, still one more remarkable isomorphism to the *Moran-Hutchinson formula* (for more details, see e.g. Barnsley 1993 Chapter V), for the computation of a *self-similarity dimension* of a fractal, namely

$$D = \frac{\log m}{\log 1/r},$$

where *m* denotes the number of parts on each scale (number at contractions of the generating iterated function system) and *r* stands for the length of each part (contraction factor), provided the *open set condition* is satisfied (i.e. if the fractal set is either totally disconnected or its parts are at least just touching). Moreover, because of the above interpretation, and subsequently a possible visualization by means of iterated function systems, (x + c) resp. (ρ + V) should be a positive integer.

This isomorphism was observed for the first time by Mandelbrot (see Mandelbrot 1983 Chapter XI, 2000 Chapter 12, and the references therein) w.r.t. ZML and, independently, by Hřebíček (see Hřebíček 1994) w.r.t. MAL.

Despite the evident correspondences

$$m \sim \rho + V, \qquad m \sim \frac{U}{P} = (\rho + V)^{-1/D_i}$$

or

$$D \sim \frac{1}{b}, \qquad m \sim x + c, \qquad r \sim \frac{y}{A} = (x+c)^{-b}$$

(or, for the complete formula of MAL, D ~ 1/b, m ~ x, $r \sim \frac{y}{A\exp(cx)} = x^{-b}$), the application of fractal geometry to ZML or MAL might be only theoretical, because the exact self-similarity would require unrealistic regular trees or the same values of parameters *A*, *b*, *c* on all linguistic levels, respectively.

4. Fractal analysis of texts: cyclic self-similarity hypothesis

Nevertheless, assume that on *n* scaling levels $i = 1, 2, \ldots, n,$ we have

$$U_i = P_i(\rho_i + V_i)^{-1/D_i}$$

or

$$y_i = A_i(x_i + c_i)^{-b_i}$$

(or, for the complete formula of MAL, $y_i = A_i x_i^{-b_i} \mathrm{e}^{c_i x_i}$) and that (cyclicity hypothesis)

$$U_i = U_{i+rn},\ P_i = P_{i+rn},\ \rho_i = \rho_{i+rn},\ V_i = V_{i+rn},\ D_i = D_{i+rn}, \qquad \text{for } r = 1, 2, \ldots,$$

or

$$y_i = y_{i+rn},\ A_i = A_{i+rn},\ x_i = x_{i+rn},\ c_i = c_{i+rn},\ b_i = b_{i+rn}, \quad \text{for } r = 1, 2, \ldots,$$

i.e. that the situation is cyclically repeated in blocks of *n*-levels (which is a purely mathematical but more realistic assumption). In this way, we arrive at the following interpretation of ZML or MAL in fractal terms (in the case of MAL, see Andres 2009; Andres, Rypka 2012):

ZML:

$$D := \frac{n}{\dfrac{1}{D_1} + \dfrac{1}{D_2} + \cdots + \dfrac{1}{D_n}},$$

$m := (\rho + V)^{kn}$, provided $\rho + V = \rho_1 + V_1 = \rho_2 + V_2 = \ldots = \rho_n + V_n$ are such that $(\rho + V)$ is a positive integer,

$$r = r_1 \ldots r_n := \prod_{i=1}^{n} \left(\frac{U_i}{P_i} \right)^k = (\rho + V)^{-k \sum_{i=1}^{n} \frac{1}{D_i}},$$

where $k \geq \max\limits_{i=1,2,\ldots n} D_i$ is a suitable positive integer,

$$D := \frac{n \log(\rho + V)}{\sum\limits_{i=1}^{n} \log\left(\dfrac{P_i}{U_i} \right)},$$

resp.

$$\rho = -V + \exp\left(\frac{D}{n} \sum_{i=1}^{n} \ln\left(\frac{P_i}{U_i} \right) \right),$$

which is a generalized *Mandelbrot's* resp. *Pareto's formula*, because for $P = P_1 = P_2 = \ldots = P_n$ and $U = U_1 = U_2 = \ldots = U_n$, we get

$$D = \frac{\log(\rho + V)}{\log \frac{P}{U}},$$

resp.

$$\rho = -V + U^{-D} P^{D}.$$

Analogously,

$$D := \frac{n}{b_1 + \cdots + b_n} = \frac{n \log(x+c)}{\sum_{i=1}^{n} \log\left(\frac{A_i}{y_i}\right)},$$

$m := (x+c)^{kn}$, provided $x + c = x_1 + c_1 = x_2 + c_2 = \ldots = x_n + c_n$ are such that $(x+c)$ is a positive integer,

$$r = r_1 \ldots r_n := \prod_{i=1}^{n} \left(\frac{y_i}{A_i}\right)^{k} = (x+c)^{-k\sum_{i=1}^{n} b_i},$$

where $k \geq \max_{i=1,2,\ldots n} \frac{1}{b_i}$ is again a suitable positive integer.

MAL:

$$D := \frac{n}{b_1 + \cdots + b_n} = \frac{n \log x}{\sum_{i=1}^{n} \log\left(\frac{A_i}{y_i} \exp(c_i x)\right)},$$

$m := x^{kn}$, provided $x = x_1 = x_2 = \ldots = x_n$,

$$r = r_1 \ldots r_n := \prod_{i=1}^{n} \left(\frac{y_i}{A_i \exp(c_i x)}\right)^{k} = x^{-k\sum_{i=1}^{n} b_i},$$

where $k \geq \max\limits_{i=1,2,\dots n} \frac{1}{b_i}$ is a suitable (sufficiently large) positive integer in order for the open set condition to be satisfied.

Roughly speaking, if $1/D_i > 0$ resp. $b_i > 0$, for every $i = 1,2,\dots,n$ then we can speak about the given language structures as *language fractals*, because their models represent approximations of the particular associated mathematical fractals whose self-similarity dimension D is described above. For more details, see Andres (2009) and Andres, Rypka (2012).

If, reversely, $1/D_i \leq 0$ resp. $b_i \leq 0$ holds, for some i, then either ZML resp. MAL fails on the i-th level or there is an intermediate linguistic level missing which should have been taken for the appropriate constituents, instead of the former ones (cf. Andres 2010). Thus, the correct application of ZML resp. MAL can be tested (even for discovering possible new linguistic levels) by means of the sign of $1/D_i$ resp. b_i, $i = 1,2,\dots,n$. By new linguistic levels we mean, for instance, suprasentence levels which might possibly be higher than semantic constructs detected by Hřebíček (see Hřebíček 2002, 2007; Wimmer et al. 2003, and the references therein).

5. Computation of fractal dimension of structures: universality vs. specificity

For the simplest formula of MAL, when $A_i = y_i(1) = y_{1i}$, $i = 1,2,\dots,n$, and $c = c_1 = c_2 = \dots = c_n = 0$, the self-similarity dimension D of the associated mathematical fractals can be estimated simply in terms of the length of the given constructs x_{ji} and the constituents y_{ji}, $i = 1,2,\dots,n$; $j = 1,2,\dots,p_i$, as follows (cf. Andres et al. 2012b):

$$D := \frac{n}{\sum\limits_{i=1}^{n} b_i},$$

where

$$b_i \approx \frac{\ln y_{1i} \sum\limits_{j=1}^{p_i} \ln x_{ji} - \sum\limits_{j=1}^{p_i} \ln x_{ji} \ln y_{ji}}{\sum\limits_{j=1}^{p_i} (\ln x_{ji})^2}, \qquad i = 1,2,\dots,n.$$

Incorporating the weights

$$w_{ji} := \frac{z_{ji}}{\sum_{j=1}^{p_i} z_{ji}}, \ i = 1,2,\ldots,n,$$

corresponding to the *j*-th relative frequency $z_{ji} \Big/ \sum_{j=1}^{p_i} z_{ji}$, $j = 1,2,\ldots, p_i$, into the approximate formula, we obtain

$$b_i \approx \frac{\ln y_{1i} \sum_{j=1}^{p_i} w_{ji} \ln x_{ji} - \sum_{j=1}^{p_i} w_{ji} \ln x_{ji} \ln y_{ji}}{\sum_{j=1}^{p_i} w_{ji} (\ln x_{ji})^2}, \qquad i = 1,2,\ldots,n.$$

The same is formally true for ZML in terms of trees, provided $P_i = U_i(1) = U_{1i}$, i =1,2,…,n, and $V = V_1 = V_2 = \ldots = V_n = 0$. For the complete formula of MAL, we can obtain more complicated, but still explicit formulas estimating *D* (see Andres et al. 2012b). On the other hand, for ZML in general, only numerical calculations can be made, because the linear regression technique does not apply there.

In linguistic experiments, the calculated values of *D* for novels and newspaper texts, and so on, are higher than those for poems. Introductory chapters possess higher values of *D* than subsequent chapters, etc. On the basis of such empirical arguments which we believe to be so mainly due to semantics, we decided to call *D* the *measure* or *degree of semanticity* of a given text, provided the lowest linguistic level is “reasonable” from the semantic point of view, such as the length of words calculated in the average number of syllables.

Mandelbrot (2000, Chapter 12) discusses the meaning of two possibilities: $D \geq 1$, when the text (in his case, the number of words) is finite, and $D < 1$, when it is infinite (which is in practice not realistic). In order to simplify the calculations, we put in this section $V = V_1 = V_2 = \ldots = V_n = 0$. According to Mandelbrot, this means that the hierarchy trees are symmetric. Otherwise, i.e. in the asymmetric cases, he explains the role of V_i, i =1,2,…,n, in terms of fractal lacunarity.

Eftekhari’s fractal analysis of Shakespeare’s works (see Eftekhari 2006) consists of computing the fractal dimension D_F (always less than 1) and the Zipf dimension D_Z (always less than 2). In our analysis, the values of *D* are typically one or even two orders higher. The reason of this discrepancy is simply the fact that Eftekhari computes different fractal dimensions, where the open set condition need not be satisfied, and so takes into account different relations. In a similar way, for a *box-counting dimension*, i.e. when refining the grid scales, for instance, the fractality of architectural objects can be calculated in order to clarify their aesthetic preference. The images with a higher dimension were considered complex, while those with a lower dimension uninteresting. The preferable box-counting dimension of objects, whose supporting space is a two-dimensional plane, was found around 1.35 in den Heijer and Eiben (2010), 1.52 ± 0.23 in Draves et al. (2008) which is in agreement with measurements in quoted papers there (1.51 ± 0.43, around 1.54, etc.).

Because of the universality of MAL resp. ZML, one can speak with no problems about the *language of architecture*, when specifying appropriately the terms of constructs/constituents resp. trees. Of course, provided the related numbers are statistically relevant. Then it would be also interesting to detect preferable self-similarity dimensions of given objects. On the other hand, it is a question whether to consider the lengths of constructs in the mean but the absolute number of constituents (as we always have done) or to "measure" the objects on concrete linguistic levels by means of suitable physical units, as proposed by Köhler (1997) (cf. also Leopold 2001; Andres 2010).

If, for instance, the length of semantic constructs would be measured with kilosomethings, the length of clauses with somethings, the length of words with centisomethings and the length of syllables in millisomethings, then after rescaling to the same unit (e.g. millisomethings), the dimensions would dramatically change. Perhaps the most appropriate physical units in quantitative linguistics might be with this respect suitable (normalized) time units. Nevertheless, observe that in all the above formulas the average number of constituents was already normalized by means of A_i (= y_{1i}) resp. P_i (= U_{1i}).

6. Experience from experiments

As an illustrative example, we can recall the data related to the original of Poe's Raven, presented in Andres, Benešová (2012):

Table 1
Length of semantic constructs vs. length of clauses. x_{j1} the length of semantic constructs (in clauses), z_{j1} their frequency, y_{j1} the length of clauses (the average length in words).

x_{j1}	z_{j1}	y_{j1}
1	251	9.7750
2	70	9.8357
3	28	9.2738
4	11	9.7045
5	9	8.6000
6	5	9.7000
7	7	8.2245
8	5	8.1750

10	2	11.0000
11	2	10.9091
14	2	9.7500
15	1	10.9333
17	1	8.9412
18	1	8.9444
23	1	10.3043
38	1	8.8947
41	1	6.4634
56	1	8.9821

Table 2
Length of clauses vs. length of words. x_{j2} the length of clauses (in words), z_{j2} their frequency, y_{j2} the length of words (the average length in syllables).

x_{j2}	z_{j2}	y_{j2}
1	9	2.5556
2	18	1.2222
3	18	1.4444
4	11	1.5682
5	23	1.4870
6	17	1.3922
7	4	1.6071
8	8	1.6250
9	10	1.5556
10	10	1.7300
11	6	1.5606
12	1	2.0000
13	3	1.6667
14	3	1.6905
15	2	1.6667
16	2	1.8750
17	2	1.8529
18	2	1.7222
26	1	1.7308

Table 3
Length of words vs. length of syllables. x_{j3} the length of words (in syllables), z_{j3} their frequency, y_{j3} the length of syllables (the average length in phonemes).

x_{j3}	z_{j3}	y_{j3}
1	559	2.6494
2	257	2.6284
3	113	2.3717
4	30	2.3917

Applying the formula for the shape parameters b_1, b_2, b_3 in the foregoing section, we can easily obtain, for $n = 3$ and $p_1 = 18$, $p_2 = 19$, $p_3 = 4$, that

$$b_1 \approx 0.023578, \qquad b_2 \approx 0.177895, \qquad b_3 \approx 0.074535.$$

Taking into account the weights w_{j1}, w_{j2}, w_{j3}, the values of b_1, b_2, b_3 take the form (again, $n = 3$, $p_1 = 18$, $p_2 = 19$, $p_3 = 4$)

$$b_1 \approx 0.029206, \qquad b_2 \approx 0.257059, \qquad b_3 \approx 0.061153.$$

Thus, the original of Poe’s Raven is a language fractal such that¨

$$D = \frac{3}{b_1 + b_2 + b_3} \approx 10.86923$$

in the first case, resp. $D \approx 8.63513$, in the latter case.

We know from our experiments (see Andres, Benešová 2011, 2012) that there is unfortunately a too-sensitive dependence of the values of D on the way of segmenting a given text. The same is true for various translations of the same text.

Hence, distinguishing by prime the values of shape parameters b_i, $i = 1, 2, \ldots, n$, of another text under consideration (like its translation) or the same text but segmented in a different way, it is useful not only to compare the related degrees of semanticity D but also to measure the *distance* d *between two language fractals* characterized by vectors $(b_1, b_2, \ldots, b_n)$ and $(b'_1, b'_2, \ldots, b'_n)$ as follows:

$$d\left(\left(\frac{1}{b_1},\frac{1}{b_2},\ldots,\frac{1}{b_n}\right),\left(\frac{1}{b_1^{'}},\frac{1}{b_2^{'}},\ldots,\frac{1}{b_n^{'}}\right)\right)=\sqrt{\sum_{i=1}^{n}\left(\frac{1}{b_i}-\frac{1}{b'_i}\right)^2}\ .$$

In particular, for $n=3$, we have

$$d\left(\left(\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}\right),\left(\frac{1}{b_1^{'}},\frac{1}{b_2^{'}},\frac{1}{b_3^{'}}\right)\right)=\sqrt{\left(\frac{1}{b_1}-\frac{1}{b'_1}\right)^2+\left(\frac{1}{b_2}-\frac{1}{b'_2}\right)^2+\left(\frac{1}{b_3}-\frac{1}{b'_3}\right)^2}\ .$$

Thus, in our case, the distance between two variants of the original of Poe’s Raven segmented in different ways (see Andres, Benešová 2011, 2012), characterized by vectors

$$(b_1,b_2,b_3)\approx(0.023578,\ 0.177895,\ 0.074535)$$

and

$$(b'_1,b'_2,b'_3)\approx(0.010735,\ 0.221931,\ 0.063182)$$

of shape parameters, equals:

$$d\left(\left(\frac{1}{0.023578},\frac{1}{0.177895},\frac{1}{0.074535}\right),\left(\frac{1}{0.010735},\frac{1}{0.221931},\frac{1}{0.063182}\right)\right)\approx$$
$$\sqrt{(42.4118-93.1507)^2+(5.62127-4.50591)^2+(13.4164-15.8272)^2}\approx 50.80844,$$

resp., when incorporating the weights into calculations,

$$d\left(\left(\frac{1}{0.029206},\frac{1}{0.257059},\frac{1}{0.061153}\right),\left(\frac{1}{0.015771},\frac{1}{0.288223},\frac{1}{0.048660}\right)\right)\approx 29.47166\,.$$

At the same time, for the difference ΔD of the related degrees of semanticity, we get $\Delta D\approx 10.86923-10.14034=0.72889$, resp., when incorporating the weights into calculations, $\Delta D\approx 8.63513-8.50692=0.12821$.

7. Concluding remarks

Despite the above criticism, we understand that in quantitative linguistics something like the complete formula of MAL is necessary, because its reduced form does not often fit the given data on lower linguistic levels (from the words downwards).

According to the personal experience of Prof. Reinhard Köhler with linguistic data on all levels, the power law form is more appropriate to higher linguistic levels (i.e. to those associated with more semantic units) whereas the pure exponential part of the formula corresponds to the "material" or "energetic" levels (e.g. when a sound duration is involved). The application of a complete formula is so, in view of his arguments, appropriate especially to intermediate levels.

On the other hand, the complete formula of MAL might be sometimes replaced by the following one:

$$y = A(x+c)^{-b}\,.$$

Roughly speaking, one should make a suitable choice between easy calculations, when a linear regression technique directly applies in a theory, and perhaps linguistically a more justified law, when all parameters must be calculated in a more complicated way. A compromise could be possible, provided the value of parameter c in the above formula can be detected, at least for some sorts of texts, empirically. This assumption seems to be, however, in practice unrealistic.

After all, the unified formula, involving four free parameters A, b, c_1, c_2, which is a general solution of the differential equation

$$\frac{\dot{y}}{y} = \frac{-b}{(x+c_1)} + c_2\,,$$

takes the form

$$y = A(x+c_1)^{-b} e^{c_2 x}\,.$$

Summing up briefly the above comments, for fractal analysis of texts, there is still a challenge to make a suitable choice among the possibilities concerning:

- the "optimal" formula of MAL (resp. ZML),
- whether or not to incorporate the weights into calculations,
- the appropriate segmentation of a given text.

In our note, among other things, we tried to present plausible arguments for at least some possible answers.

Acknowledgement

The author was supported by the IGA UPOL grant IGA_PrF_2014028.

References

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

Altmann, G., Schwibbe, M.H., Kaumanns, W. (eds.) (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Andres, J. (2009). On de Saussure's principle of linearity and visualization of language structures. *Glottotheory 2(2), 1–14.*

Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics 17*(2)*, 101–122*.

Andres, J., Benešová, M. (2011). Fractal analysis of Poe's Raven. *Glottometrics 21, 73–98.*

Andres, J., Benešová, M. (2012). Fractal analysis of Poe's Raven, II. *Journal of Quantitative Linguistics 19(4), 301–324.*

Andres, J., Rypka, M. (2012). Self-similar fractals with a given dimension and the application to quantitative linguistics. *Nonlinear Analysis – B: Real World Applications 13(1), 42–53*.

Andres, J., Benešová, M., Kubáček, L., Vrbková, J. (2012a). Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics 19(1), 1–31.*

Andres, J., Kubáček, L., Machalová, J., Tučková, M. (2012b). Optimization of parameters in the Menzerath–Altmann law. *Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica 51(1), 5–27*.

Barnsley, M. F. (1993). *Fractals Everywhere* (Second Edition). San Diego: Morgan & Kaufmann (an imprint of Academic Press).

Draves, S., Abraham, R., Viotti, P, Abraham, F.D., Sprott, J. C. (2008). The aesthetics and fractal dimension of electric sheep. *International Journal of Bifurcation and Chaos 18(4), 1243–1248*.

Eftekhari, A. (2006). Fractal geometry of texts: First attempt to Shakespeare's. *Journal of Quantitative Linguistics 13(2–3), 177–193*.

den Heijer, E., Eiben, A. E. (2010). Comparing aesthetic measures for evolutionary art. In: Di Chio, C. et al. (eds.), *Applications of Evolutionary Computation. Proceedings of the Conference on Applications of Evolutionary Computation EvoApplications 2010: EvoCOMNET, EvoENVIRONMNET, EvoFIN, EvoMUSART, and EvoTRANSLOG (Istanbul, Turkey, April 2010), Part II, LNCS 6025: 311–320.* Berlin: Springer.

Hřebíček, L. (1994). Fractals in language. *Journal of Quantitative Linguistics 1(1), 82–86.*

Hřebíček, L. (1998). Language fractals and measurement in texts. *Archiv orientální 66, 233–242*.

Hřebíček, L. (2002). *Vyprávění o lingvistických experimentech s textem*. [Stories about Linguistics Experiments with the Text] Praha: Academia.

Hřebíček, L. (2007). *Text in Semantics – The Priciples of Compositeness*. Prague: The Academy of Sciences of the Czech Republic (Oriental Institute).

Köhler, R. (1997). Are there fractal structures in language? Units of measurement and dimensions in linguistics. *Journal of Quantitative Linguistics 4(1–3), 122–125.*

Köhler, R. (2008). The fractal dimension in script: an experiment. In: Altmann, G., Fengxiang, F. (eds.): *Analyses of Script, Properties of Characters and Writing Systems: 115–119.* Berlin: de Gruyter.

Leopold, E. (2001). Fractal structure in language. The question of the imbedding space. In: Köhler, R., Uhlířová, I., & Wimmer, G. (eds.): *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček: 163–176.* Trier: Wissenschaftlicher Verlag.

Mandelbrot, B.B. (1983). *The Fractal Geometry of Nature*. New York: Freeman and Comp.

Mandelbrot, B.B. (2000). *Les objects fractals. Forme, hasard et dimension*. Paris: Flammarion.

Manin, D. Yu. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language? *Journal of Quantitative Linguistics 16(3), 274–285.*

Montemurro, M.A. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A 300, 567–578.*

Montemurro, M.A. (2004). A generalization of the Zipf–Mandelbrot law in linguistics. In: Gell-Mann, M., Tsalis, C. (eds.): *Nonextensive Entropy: Interdisciplinary Applications: 347–356.* Oxford: Oxford University Press.

Montgomery, D.C., Peck, E.A., Vining, G.G. (2006). *Introduction to Linear Regression Analysis* (4th Edition). Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons.

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

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. [Introduction to the Analysis of Texts] Bratislava: Veda.

Adverbials in Czech: Models for their frequency distribution

Radek Čech, Ludmila Uhlířová

„Text is a virtual transformation of a set of words from lexical units to semantic elements“
(Hřebíček 2007: 74)

1. Introduction

Luděk Hřebíček, whose 80th birthday we celebrate this year, is an internationally respected author of general text theory, which he formulated and elaborated in the course of his scientific path through life. His text theory (cf. Hřebíček 1992, 1995, 1997, 2000, 2007), in which he brilliantly presented his original ideas and also significantly developed some discoveries made by Altmann in the eighties (partly as Altmann´s co-author and co-editor), is based on the idea that fundamental principles of text structures can be explained by the means of Menzerath-Altmann´s law (Altmann 1980) as its core, and with the numerous interrelations between two kinds of fundamental linguistic units, *constructs* and *constituents*. The present contribution, inspired by Hřebíček´s philosophy of language, deals with one concrete class of constituents and a class of their constructs.

Let us say that *words/lexical units* are constituents of constructs on the immediate higher level, i.e. of *phrase* or *clause*. If we omit a level and jump to sentence, the relation may become more complex. There are complex interrelations between them which can be defined and described in various terms. In the present contribution we deal with just one. We ask the following question: Do syntactic categories of a certain type, as will be specified below, show similar distributional features as other types of constituents and do they abide by the same statistical law(s)?

2. Development of syntactic analysis – from a description to an explication

As Köhler (2012) pointed out recently, the field of syntax remained – for a long time – less affected by quantitative methods than the lower levels of language. It was concentrated mainly on descriptive and applied aspects rather than on a theoretical (explanatory) analysis. Only today syntax became one of the fields which come more and more in the centre of interest of quantitative studies. The interests of quantitative researchers are expanding rapidly now; effective and sophisticated mathematical methods and procedures of study are being developed.

There are many reasons why traditional syntactic statistics dealt mainly with the counting of particular syntactic phenomena, and not with general hypotheses and laws, which would “explain why languages are as they are” (Köhler 2012:7). The crucial ones should be sought in the very nature of syntax, namely in the manifold and complicated interrelations both between its constituents and constructs and between interrelations to other levels of language. Let us compare: In the Czech language, according to Ludvíková (1987), there are 36 speech sounds: 11 vowels and 26 consonants, and from the sum of all theoretically possible bigrams only about 60 per cent are attested, with different functional loads in the phonological system and with different frequencies in texts. The limited inventory of phonemes made it possible to perform phonological statistics already in the first half of the 20th century, such as those done by linguists of the pre-war Prague school generation, e.g. Trnka (1935), Mathesius (1947), Vachek (1940), and later on Ludvíková (1968) and others. On the other hand, there does not exist any “inventory”, or any “list” of Czech sentences. What we have is a finite set of abstract sentence (or clause) patterns (schemes, types) and rules for their application. They are stored in our minds and described in grammars. However, due to our linguistic competence we are able to create (or “generate”), theoretically, an infinite number of utterances. This fact is considered to be one of the most important (and the most astonishing) aspects of human language. But it poses both theoretical and methodological problems. First, a huge variability of utterances leads to relatively big differences among particular syntactic descriptions, even within the same or similar methodological framework. Further, for an adequate statistical syntactic analysis, appropriate data must be prepared which is not a trivial task. Fortunately, because of the rapid development of computational linguistics in the last few decades, syntactically annotated corpora of Czech are available now (e.g., Králík, Uhlířová 2007; Vidová Hladká et al. 2008; Hajič et al. 2006).

To sum up, after a period of syntactic research that has struggled with many difficult problems (in comparison to phonology or morphology), it is time to take steps to a deeper understanding of syntactic functioning. So, we are trying to follow this research direction. Our analysis may be taken as an attempt to show how a descriptive analysis (Uhlířová 1975) can be reinterpreted in the light of current quantitative linguistic knowledge.

3. Specification of the task

In the following, we deal with one kind of sentence constituent – the adverbial. A plausible assumption to be tested, namely that the frequency distribution of adverbials abides by the same/similar law(s) to those that are already known for syntactic and/or other constituents, can be based on the hitherto achieved experience: 1. There exist already a number of empirical frequency distributions as well as their probability models in the field of syntax; see Köhler (2012) for the present state of the art. 2. There exist typical frequency distributions of constituents of other levels of language which have been theoretically derived and attested on hundreds of languages; see Altmann (1980, 1988, 1993, 2001), Best (2005), Altmann, Köhler (1996), Grzybek (2006), Köhler (2005), Köhler, Altmann (2000), Köhler, Naumann (2009), Wimmer (2005), Wimmer, Altmann (1999, 2006) and many other authors. 3. No uniform frequency distributions of any language phenomenon/unit (not only in the field of syntax) have been found so far; it seems that nothing is uniformly distributed in language. 4. All frequency distributions commonly known are shifted in some way or other, but none display symmetry. 5. Generally: Human language is a phenomenon which abides by laws of a probabilistic nature – the idea which was once proclaimed by Zipf (1935, 1949), Mathesius (1947), Halliday (1993), and other great personalities in the course of the whole 20th century and which is being successfully formulated in a strict mathematical way now by Hřebíček, Altmann, Köhler and those mentioned already above. 6. The existence of underlying regularities in frequency distributions can be interpreted as a result of a diversification process (Altmann 2005). This process – alongside the opposite process of unification (cf. Zipf 1949) – has a decisive impact on language form and it follows from general principles which control human language behavior, such as the least effort principle (Zipf 1949) or a self-regulation in the synergetic model of language (Köhler 1986, 2005).

4. Syntactic framework and language material

We start from the "classical" dependency grammar formalism, according to which the finite verb is the centre of the sentence; the arguments headed by the finite verb are subject(s), object(s), or adverbial(s). We accept the definition and classification of adverbials given in the well-known grammar by Šmilauer (1966), inspired, basically, by the well-known views of Tesnière. Šmilauer´s syntactic framework was applied to the oldest Czech treebank, the Czech Academic Treebank corpus, compiled (and tagged) as early in the seventies by a team of linguists at the Czech Language Institute; in the seventies, it was the best structural syntactic description available for the given purpose (cf. Králík, Uhlířová 2007). Due to the consistency of tagging, the Czech Academic Treebank could be technically modernized later (see Czech Academic Corpus 2.0 Guide 2008; Vidová Hladká et al. 2008), and it is still used and still respected, even though today we work with a number of other syntactic frameworks as well as with huge corpora and treebanks. The data used in the analysis are taken from four non-fiction text samples from the Czech Academic Treebank: history of architecture, psychology, sociology and communication engineering, 250 adverbials from the beginning of each text. We took only adverbials expressed by a noun, adverb or adverbial clause, not by pronouns. (For more details, see Uhlířová 1975). Each adverbial was labeled with one of the thirteen possible labels: place, time, manner, degree, means, aspect, cause, purpose, condition, concession, origin, originator, and result.

5. Statistical procedures and interpretation of the data

Frequencies of the adverbial classes are given in Table 1. The data in the columns reads as follows (from left to right): adverbial class r = rank, f = absolute frequency, f_r = relative frequency in per cent.

Table 1
Frequencies of adverbials.

Adverbial	***r***	***f***	f_r
Place	1	273	27.3
Time	2	204	20.4
Manner	3	172	17.2
Means	4	68	6.8
Aspect	5	61	6.1
Condition	6	59	5.9
Measure	7	52	5.2
Cause	8	30	3.0
Result	9	18	1.8
Origin	10	18	1.8
Purpose	11	17	1.7
Concession	12	16	1.6
Originator	13	12	1.2
∑		1 000	100

The data can be interpreted in two steps.

First step: On the primary level of interpretation, we can simply count the occurrences of adverbials attested in the samples, as is done in the second and third columns. According to their frequencies, we may divide all adverbials into three groups. (a) Adverbials with frequencies higher than 10 per cent (in the four samples taken together). Here go the adverbials of place, adverbials of time and adverbials of manner, which, in the total, make roughly two thirds of all adverbials. (b) Adverbials with frequencies within the interval from 5 up to 10 per cent; they express means, aspect, condition and measure. (c) The lowest frequencies are attested with the adverbials of cause, origin, result, purpose, concession and originator. It can hardly be denied that even such an elementary result of counting is of a certain descriptive value. The absolute as well as the relative frequencies (given in per cent in the third column) show different quantitative weights of adverbials in non-fiction texts: With the increasing rank the frequency of the adverbial class decreases, and so does, implicitly, its relevance in the semantic structure of the text (in the respective non-fiction field). Absolute and/or relative frequency may serve as a quantitative indicator of the content of the text.

Moreover, in each of the four samples, a certain diversification in frequencies may be seen, due to different thematic and stylistic factors. For example, in the text on the history of architecture (dealing with the development of building styles) a higher frequency of the adverbials of time is found, whereas in the text on communication engineering, in which processes in electronic circuits are described and explained, there is a higher frequency of the adverbials of condition. Unfortunately, the data from our four texts are too small to allow considering “boundary” conditions in more detail.

However, such a “traditional” interpretation could be considered as sufficient and useful, let us say, only forty years ago.

Second step: On the advanced level we make a step from the empirical frequencies of adverbials to their probabilities. This step is theoretically decisive and reflects the present demands on quantitative linguistics as “empirical science” in the sense of Altmann´s, Hřebíček´s and Köhler´s theoretical approach to linguistics. The questions sound: What are the numerical representations of word class frequencies and do they abide by a probabilistic law? What model should be used? What are the (dis)advantages and limits of particular models? etc. In the following sections three approaches to a modeling of the distribution of adverbials will be discussed.

5.1. Models for the distribution of adverbials

It has been shown by Hammerl (1990), Liu (2009), and Köhler (2012) that the Zipf-Alexeev approach seems to be a good model for a representation of different word classes (parts of speech, dependency type, motifs). Here we start from the assumption that the frequency in class *r* takes on values proportional to the preceding class, *r* – 1. This is based simply on the a priori diversification (which decreases) and the a posteriori ranking which captures this process. Hence the relative rate of change of frequency *y’/y* is proportional to the relative rate of change of the rank (*r*) in the following way:

$$\frac{dy}{y} = \frac{a + b\ln r}{r}dr\,.$$

The solution of this differential equation yields

$$y = Kr^{a + b\ln r},$$

where K is simply a constant, or, if taken as a probability distribution (discrete or continuous), it may be considered a normalization constant.

Therefore, we expected that the distribution of adverbial classes (see Table 1) should follow this model. We have found out that the Zipf–Alekseev distribution fits to our data with a very good result: We used it simply as a continuous function. The goodness-of-fit, tested with the help of the determination coefficient, yields R^2 = 0.97 (with parameters K = 273.1124, a = -0.0383, b = -0.49745), see Fig. 1.

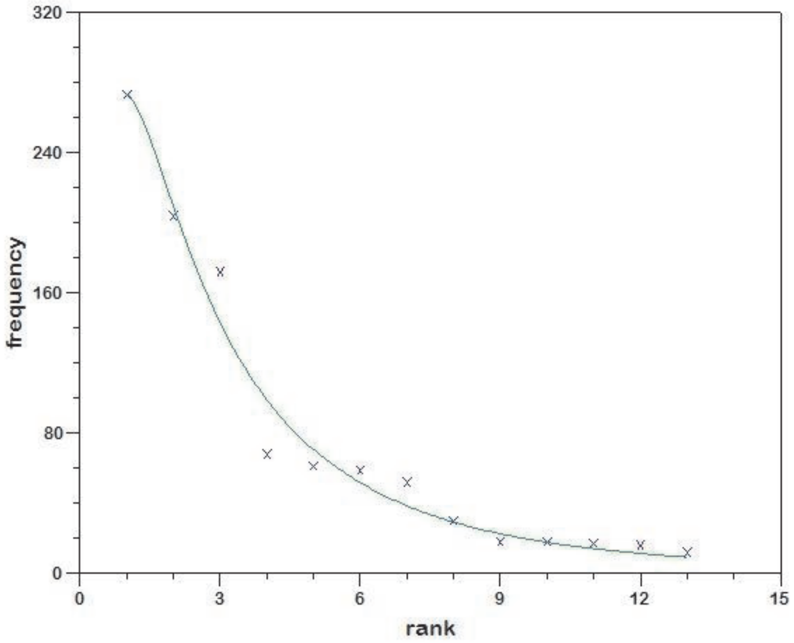


Figure 1. The distribution of all adverbials and the result of the fitting of the Zipf-Alekseev function to the data.

Now let us have a look at our thirteen classes of adverbials with regard to their part-of-speech appurtenances, and, once more, let us ask about their frequency within the parts-of-speech classes. An adverbial expressing place, may belong to a noun, adverb or be a complete clause, etc. Let us test the numbers of adverbial frequencies separately within each of their parts of speech. Though some of the differences in absolute frequencies seem to be quite large, the Zipf–Alekseev function can be fitted to the data with very good results again: the goodness-of-fit is tested with the help of the determination coefficient R^2. The values of R^2 are the following: R^2 = 0.98 for nouns (with parameters K = 260.4290, a = -1.1856, b = -0.0078), R^2 = 1 for adverbs (with parameters K = 104.1095, a = 0.5208, *b* =-1.4542), R^2 = 0.96 for dependent clauses (with parameters K = 28.3545, a = 0.0356, b = -0.5672), see Figures 2 to 4.

At first sight, we may repeat our conclusion already achieved in the previous point: The assumption that it is possible to find a model of the frequency distribution of adverbials is valid also with regard to their parts of speech; the model is fully compatible with the data. However, the course of the function, as is presented in Figure 3 and 4, does not seem to be plausible for the modeling of observed distributions – there is first an increase and then a monotonically decreasing part. It means that the "mere" good fitting does not mean automatically the best choice of a model. Consequently, we have looked for other models.

Table 2
Part-of-speech frequencies of adverbial classes.
R^2 expresses results of the goodness-of-fit of the Zipf–Alekseev function.

Adverbial	**Noun**	**Adverb**	**Clause**
Place	263	9	1
Time	96	104	4
Manner	79	75	18
Means	68	-	-
Aspect	46	13	2
Condition	30	-	29
Measure	21	30	1
Cause	11	-	19
Result	18	-	-
Origin	18	-	-
Purpose	10	-	7
Concession	4	-	12
Originator	12	-	-
∑	676	231	93
R^2	0.98	1	0.96

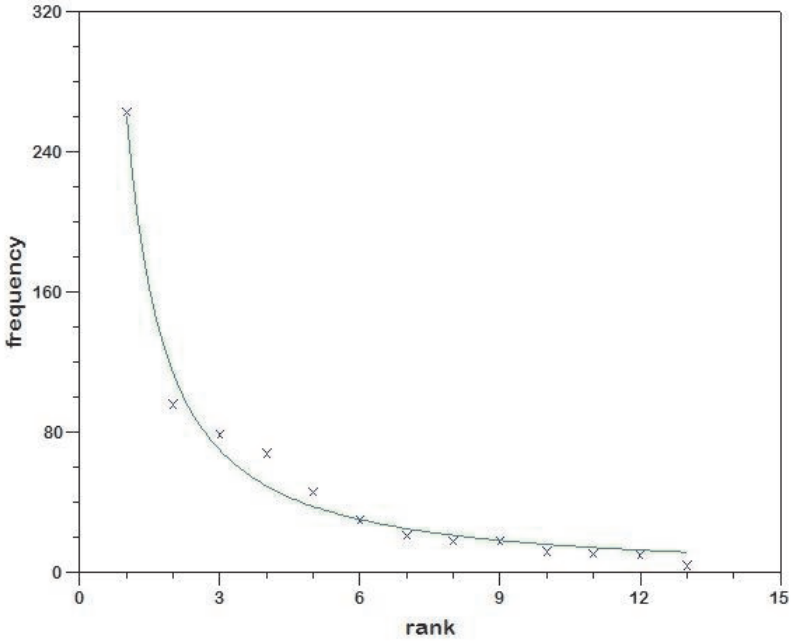


Figure 2. The distribution of adverbials expressed by nouns and the result of the fitting of the Zipf-Alekseev function to the data.

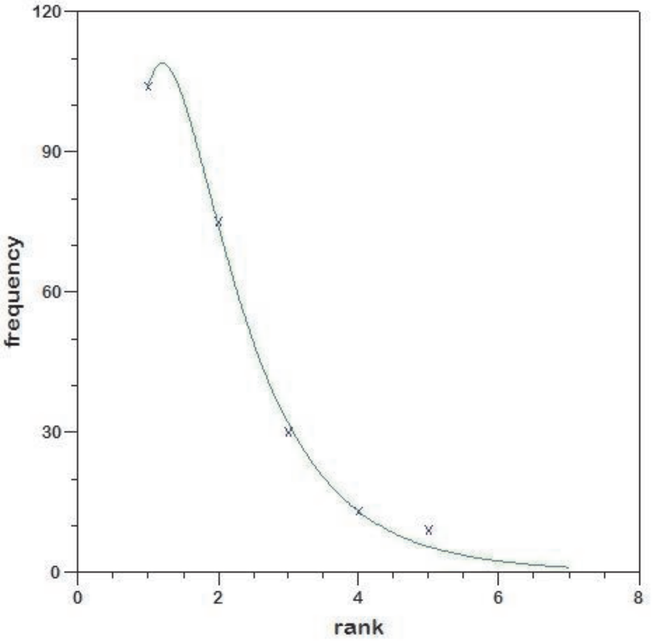


Figure 3. The distribution of adverbials expressed by adverbs
and the result of the fitting of the Zipf-Alekseev function to the data.

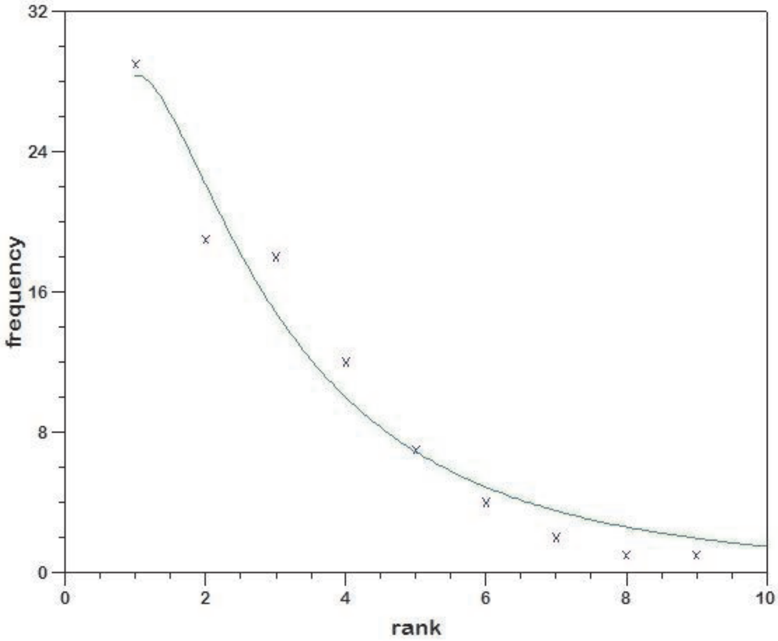


Figure 4. The distribution of adverbials expressed by clauses
and the result of the fitting of the Zipf-Alekseev function to the data.

Since Zipf, it has been well-known that a simple power-law function can be used as an acceptable model for different kinds of distributions; it is based on the assumption that the relative rate change of frequency *y'/y* is proportional to the relative rate of change of the rank (*r*) in the following way:

$$\frac{dy}{y} = \frac{b}{r} dr ,$$

where *b* is the constant. The solution of the equation yields the power-law function

$$y = Kr^b ,$$

where *K* is a well interpreted constant – it usually corresponds approximately to the highest frequency.

The results of fitting the function to the data are presented in Table 3. Except for adverbs, the fits bring worse results than the Zipf-Alekseev function. This result seems to corroborate Köhler's statement according to which this model is more appropriate for data with a bigger inventory size (e.g., word forms or lemmas) (cf. Köhler 2012: 75). Even though the results can be considered acceptable, we are striving for better results.

Table 3
The results of the fitting of the power-law function to the data; the fitting of all adverbials is visualized in Figure 5.

	K	b	R^2
All	298.7741	-0.9023	0.90
Adverb	109.5577	-1.0807	0.98
Noun	260.6419	-1.1985	0.88
Clause	31.0882	-0.8874	0.87

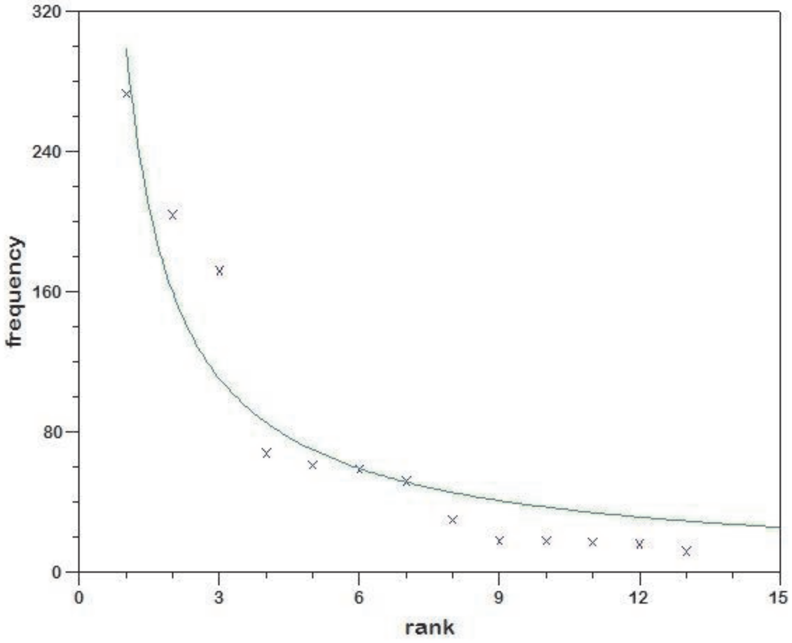


Figure 5. The distribution of all adverbials and the result of the fitting of the power-law function to the data.

Based on a unified theory (Wimmer, Altmann 2005), we assume that the relative change of frequency y could be proportional to the change of rank in a degree which is represented by a function expressing mutual relations between speaker's and hearer's impacts on a process of communication. Specifically,

$$\frac{dy}{y} = \frac{a + br}{cr} dr ,$$

where a is a constant (it differs with regard to a specific class of language units, e.g., parts of speech, clauses, morphs), br is the impact of the speaker (he changes *r* constantly according to *b*) and *cr* is the impact of language community (it restricts speaker’s tendency to perform too much change in his speech). The solution of the equation is a function

$$y = Ce^{\frac{\frac{br}{c}+a\ln r}{c}},$$

after a simplification

$$y = Ce^{ar}r^{b},$$

where *C* is a constant of integration. Despite the fact that this function is identical with a “long” version of the well-known Menzerath-Altmann law, the identity is purely formal because there is no identity in theoretical derivation (cf. Köhler 2012: 75). Fitting the function to the data yields excellent results, see Table 4. Consequently, we consider this model to be the best one for an analysis of the frequency distribution of adverbials.

Table 4
The results of the fitting of the function derived from the unified theory to the data; the fitting of all adverbials is displayed graphically in Figure 6.

	C	*a*	*b*	R^2
All	385.5941	-0.3337	0.0189	0.97
Adverb	440.5010	-1.4413	1.5788	0.99
Noun	267.4394	-0.0304	-1.1157	0.98
Clause	45.1149	-0.4662	0.3009	0.97

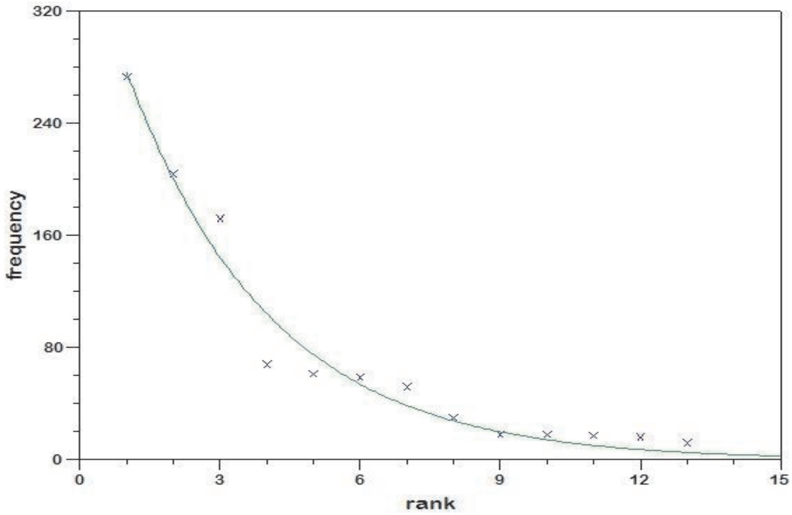


Figure 6. The distribution of all adverbials and the result of the fitting of function derived from the unified theory to the data.

6. Discussion

No new model has been “discovered” (by the way, it was not our aim). Anyway, we did not find anything contrary to what was intuitively expected. We may conclude that we have - most modestly - contributed to the general quantitative understanding of the syntactic constituents and their constructs: Adverbials manifest the same distributional universals as other types of constituents, and they abide by the same statistical law(s). As for models, we tried to demonstrate that the choice of the model is not a trivial task. Obviously, we are still at the beginning of this kind of syntactic research and only further research will reveal which models are more useful than others.

Our results can be interesting also for “traditional” (i.e. mostly descriptive) grammarians: The meaningfulness and logic of the classical dependency framework, especially of the classification of adverbials into the thirteen semantic classes, which we chose for our empirical counting, was fully supported by the results.

But at the same time, we have shown the beginnings of a particular research strategy which is well known from physics: In the first step, one describes a class of entities and shows that they follow some regularity. In the next step, one analyzes an individual class and shows that it is not unique but forms again a hierarchically lower stratum; e.g. one has the set of place adverbials whose elements can be classified as nouns, adverbs or clauses, and their distribution is again an expression of some regularity. In the third step one looks at the nouns, classifies them and states that the same conjecture (in best cases a law) holds again, but the parameters are different. This step will be repeated, so to say, ad infinitum, just as done in physics where from time to time a new, smaller particle will be discovered. The problem in linguistics is that this way is not possible without mathematics, but if it is done by means of mathematics, the time will come in which we shall be able to forecast the result at the next lower level.

Acknowledgments

We are deeply indebted to Gabriel Altmann for his substantial help and advice in mathematical matters, generously given to us during the writing of this text. We would also like to thank to Reinhard Köhler for inspiring ideas and comments.

References

Altmann, G. (1980). Prolegomena to Menzerath’s Law. In: Grotjahn, R. (ed.), *Glottometrika 2: 1–10.* Bochum: Brockmeyer.

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

Altmann, G. (1993). Science and linguistics. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to quantitative linguistics: 3–10.* Dordrecht: Kluwer.

Altmann, G. (2001). Theory Building in Text Science. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček: 10–20*. Trier: WVT.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 646–659*. Berlin, New York: de Gruyter.

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

Best, K.-H. (2005). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quan-titative Linguistics. An International Handbook: 298–304.* Berlin, New York: de Gruyter.

Czech Academic Corpus 2.0 Guide (2008). [Available at: http://ufal.mff.cuni.cz/rest/CAC/doc-cac20/cac-guide/eng/html/index.html]

Grzybek, P. (2006). On the science of language in light of the language of science, In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 1–14.* Dordrecht: Springer Verlag.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M. (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. In: Hoey, M. (ed.), *Data, Description, Discourse: 1–25.* London: Harper Collins Publishers.

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, L. (ed.), *Glottometrika 11: 142–156.* Bochum: Brockmeyer.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. B. (eds.), *Contributions to Quantitative Linguistics: 33–39.* Dordrecht: Kluwer.

Hřebíček, L. (1995). *Text levels. language constructs, constituents and the Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in Sequences. Contributions to general text theory.* Prague: Oriental Institute.

Hřebíček, L. (2007). *Text in semantics. The principle of compositeness.* Prague: Oriental Institute.

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

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

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin, Boston: Mouton de Gruyer.

Köhler, R., Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7, 189–200.*

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

Králík, J., Uhlířová, L. (2007). The Czech Academic Corpus (CAC), its History and Presence. *Journal of Quantitative Linguistics 14, 265–285.*

Liu, H. (2009). Probability distribution of dependencies based on Chinese dependency treebank. *Journal of Quantitative Linguistics 16, 256–273.*

Ludvíková, M. (1968). Kombinatorika českých fonémů z kvantitativního hlediska. *Slovo a slovesnost 29, 56–65.*

Ludvíková, M. (1987). Čísla o hláskách. In: Těšitelová et al., *O češtině v číslech: 91–108.* Praha: Academia.

Mathesius, V. (1947). Úvod do fonologického rozboru české zásoby slovní. In: *Čeština a obecný jazykozpyt:* 59–86. Praha: Melantrich.

Šmilauer, V. (1966). *Novočeská skladba*, Praha: Státní pedagogické nakla--datelství.

Trnka, B. (1935). *A phonological analysis of present-day Standard English.* Praha: Univerzita Karlova.

Uhlířová, L. (1975). O frekvenci příslovečného určení v souvislém textu. *Naše řeč 58, 133–142.*

Vachek, J. (1940), Poznámky k fonologii českého lexika. *Listy filologické 67,* 395–402.

Vidová Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., Raab, J. (2008). *Czech Academic Corpus 2.0*. Philadelphia: Linguistic Data Consortium.

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

Wimmer, G., Altmann, G (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 760–775*. Berlin, New York: de Gruyter.

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 329–337*. Dordrecht: Springer Verlag.

Zipf, G. K. (1935). *The psycho-biology of language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin. Cambridge: M.I.T. Press (2nd edition, 1968).

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison–Wesley.

The use of the *POR* in macro-lexical analyses

Fan Fengxiang, Zhou Pianpian, Su Hong

1. Introduction

In the quantitative approach to the study of language, linguists either rely on the traditional (traditional in the sense of quantitative linguistics) measurements, such as word frequency, position and length of sentential elements etc, or create new units of analysis, e.g., the *hreb* (Hřebíček 1992, 1997, 2000), the motif (Köhler 2006; Köhler, Naumann 2008), the arc length (Popescu et al. 2008), the *h*-point (Popescu et al. 2007), and other points related to it and so on. The *hreb* is a semantic entity, referring to sentences within a text containing a synonym, a reference or some other identifying semantic connection among sentences. The motif consists of non-decreasing sequences of some measured entities of a text. The arc length can be regarded as the sum of Euclidean distances between points corresponding to consecutive frequencies f_i and f_{i+1}. The *h*-point is the place in a word frequency spectrum where the word frequency rank *r* equals the frequency of the word. These old and new measurements serve as empirical yardsticks and prove most effective in developing and applying mathematical models such as the Zipf-Alekseev law (Alekseev 1978), the Menzerath-Altmann law (Altmann, Schwibbe 1989), etc, and for revealing and describing textual and lexical characteristics. In this contribution, we introduce a new measurement for macro-lexical analyses, the *POR* (Point Of Return), and consider its possible applications in macro-lexical analyses. This point refers to the place on a hapax/vocabulary ratio (*HVR*) curve of a text or collection of texts from where the curve displays a general upward trend all the way to the end (Fan 2010).

2. Causation of the *POR*

The existence of such a point seems contrary to common sense. Intuitively, the *HVR* curve of a text or collection of texts should be monotonically decreasing as the text length increases just like the *TTR* (type/token ratio). Suppose there is a text *T* with a length *N*, a vocabulary size *V* consisting of *H* hapaxes. If the text has only one word, i.e., $N = 1$, then $V = H = N = TTR = HVR = 1$. As *N* gradually increases, this relationship may still maintain for a short while, until *N* reaches a certain value *K*, e.g., 20, from which this relationship would start to change; both the *TTR* and *HVR* would be smaller than 1. As *N* keeps increasing, the *TTR* and *HVR* would keep decreasing. As *N* approaches ∞, all the words in the language would have occurred more than once, and the number of hapaxes would be zero and so would the *HVR*, and *V* would no longer increase and the *TTR* would gradually approach its horizontal asymptote 0.

A simulation of this hypothetical changing relationship between *N*, *V*, *TTR* and *HVR* as *N* keeps increasing seems to corroborate this conjecture. In the simulation, Lewis Carroll's *Through the Looking-glass and What Alice Found There* was used as a miniature language which contains its entire set of vocabulary. The novel has 30,566 word tokens and 2,754 word types, i.e., the entire set of vocabulary of the miniature language. An initial sample of 100 words was randomly drawn with replacement from this miniature language, with its vocabulary size, *TTR* and *HRV* computed. Then this initial sample was continuously enlarged by a sample of 100 words randomly drawn with replacement from the miniature language until the vocabulary size of the enlarged sample equals that of the miniature language 2,754, exhausting the inventory of the miniature language's vocabulary. As *N* of the sample increases, *V* increases but its *TTR* and *HVR* keeps decreasing until the *TTR* approaches its horizontal asymptote 0 and the *HVR* becomes 0. Figure 1 shows the result of this simulation.

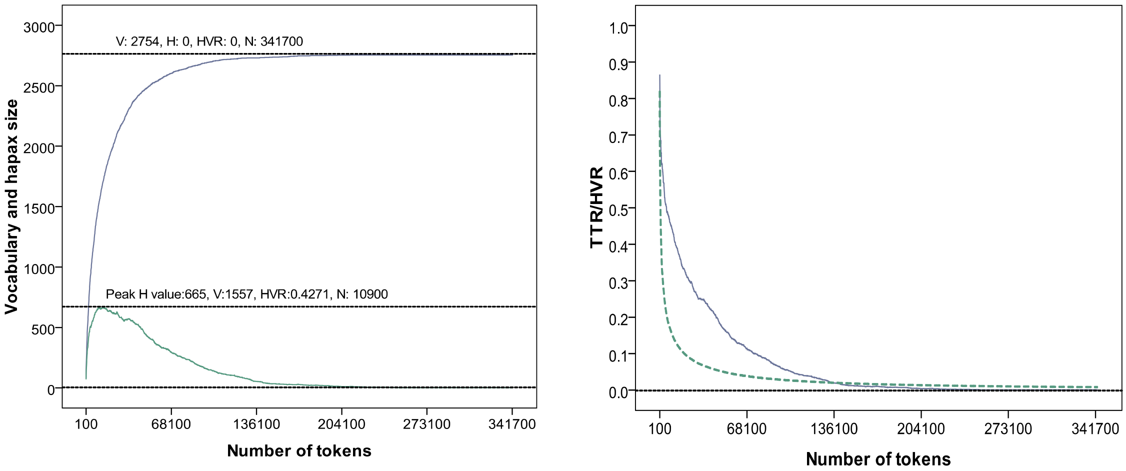


Figure 1. Simulation of *V*, *H*, *TTR*, *HVR* as *N* keeps increasing until the entire vocabulary of the miniature language is exhausted. The left panel: the upper curve and the lower curve are respectively the simulated vocabulary growth and hapax growth. The right panel: the solid curve and the dotted curve are respectively the *HVR* curve and the *TTR* curve.

In the simulation, the number of hapaxes reaches its peak at *N* = 10, 900, at which point *H* = 665 and *HVR* = 0.4271. Then the number of hapax and *HVR* starts to decrease monotonically until both become zero; the *TTR* is 0.00806.

However, this would not happen in reality. In a study on the *HVR* of the 100-million-word BNC (the British National Corpus), Fan (2010) observed a *U*-shaped *HVR* curve; that is, as *N* reached a certain value, the *HVR* curve reached its *POR* and reversed it downward trend and started to go upwards all the way until the end of the curve. The *POR* of the BNC is 3,820,340, at which point the vocabulary size is 63,449, the number of hapaxes is 24,740 and the *HVR* is 0.3899. Figure 2 shows BNC's vocabulary and hapax growth curves and the *U*-shaped *HVR* curve

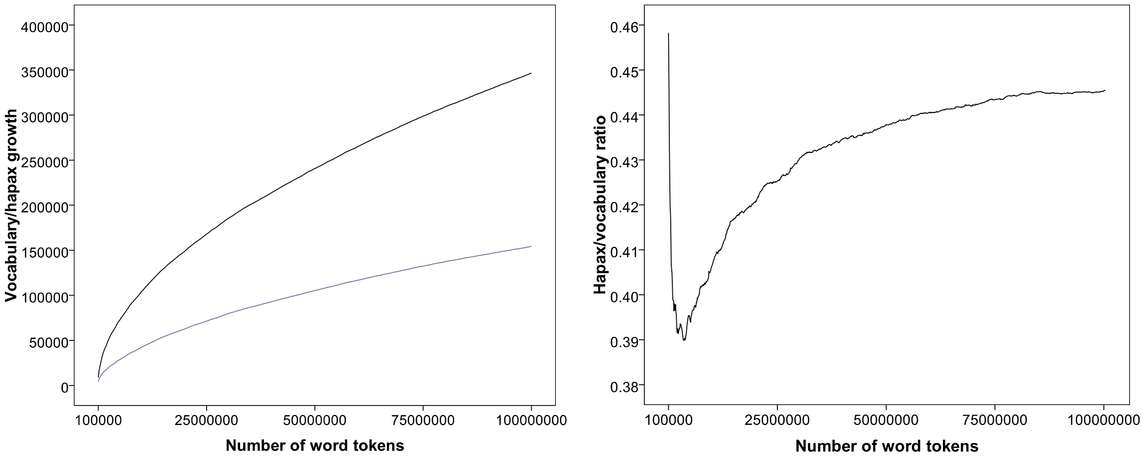


Figure 2. Vocabulary and hapax growth curves (the left panel) and the *U*-shaped *HVR* curve (the right panel) of the BNC.

What causes the existence of the *POR*? In a natural language its lexicon can not possibly be exhausted, however large a sample drawn from the language population may be. The vocabulary of a language can be divided into two parts, the core vocabulary and peripheral vocabulary, the number of the former would be more limited and the latter is far more numerous (Miguel 2007). Here we would rather use the term main vocabulary since the core vocabulary generally refers to the vocabulary that is commonly shared in all the domains of a language. We now give a tentative working definition of the main vocabulary of a corpus or a domain: it is the set of words that accounts for 95% or more of the words of the language carrier in question. We adopt this lexical threshold because failure to reach this percentage would result in inadequate comprehension of the target texts (Nation, Waring 1997; Schmitt et al. 2011). It generally consists of within-dictionary words and some out-of-dictionary words used in a particular field. Apart from the main vocabulary, there is also the peripheral vocabulary which contains rare within-dictionary words and random out-of-dictionary alpha-numeric strings of extremely low probability whose number is practically infinite. The existence of the *POR* is generally due to the accumulation of extremely rare alpha-numeric strings in a language as *N* increases. As *N* reaches a certain value, most of the main vocabulary of the corpus or domain under study would have been used more than once and the accumulation of the rare alpha-numeric strings would become large enough to reverse the decreasing of the *HVR* curve and make it go upwards. This would not happen in the simulation in which the entire lexicon has appeared more than once in the sampled texts.

3. Use of the *POR* in macro-lexical analyses

As previously mentioned, the *POR* of the BNC is 3,820,340, at which point the vocabulary size is 63,449, the number of hapaxes is 24,740 and the *HVR* is 0.3899. This suggests that the size of the main vocabulary of the population from which the BNC was drawn is around the *POR*, i.e., 63,449. Further computation reveals that these 63,449 words account for 94.7% of the word tokens of the BNC. So the *POR* can be used as a macro-lexical indicator of a corpus or a domain.

We first examined the *POR* of technical English, represented by the JDEST Corpus containing texts mainly in the physical sciences, engineering and technology (Kennedy 1998:44) and the Marine Engineering Corpus compiled by researchers at Dalian Maritime University consisting of texts on ship engines, on-board machinery, the propelling system, fuel etc. The two corpora, each containing about one million words, were combined together, totalling 2,000,000 words and is referred to here as the Combined Technical Corpus. The vocabulary size of the Combined Technical Corpus is 32,932, the number of hapaxes is 14,392, and the overall *HVR* is 0.437. Figure 3 is the vocabulary and hapax growth curves, the *HVR* curve and the *POR*.

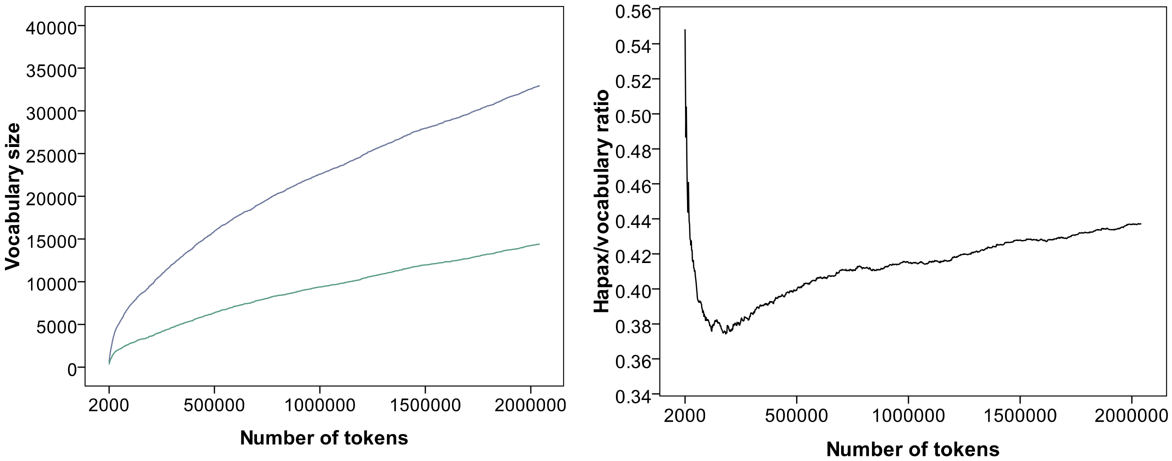


Figure 3. The vocabulary and hapax growth curves (left panel) and the *HVR* curve (right panel) of the Combined Technical Corpus.

What is unusual of the Combined Technical Corpora is that it has a very small *POR*, only 196,505, from which the *HVR* curve turns upwards. The vocabulary size at the *POR* is 9,165 and the *HVR* is 0.3741. Although these 9,165 words are less than one third of the corpus's vocabulary, they account for 97.39% of the words of the Corpus.

To see whether other corpora of the same domain display similar *POR* and the *HVR* curve, we examined the technical English sub-corpus of the ANC (the American National Corpus), which has about four million words. To be compatible with the Combined Technical Corpus, we randomly drew a 2000000-word sample from it and computed its vocabulary size, vocabulary growth, hapax growth, the *HVR* and the *POR*. The vocabulary size of the sample is 44,286, the number of hapaxes is 19,060 and the overall *HVR* is 0.4304. Figure 4 is the vocabulary and hapax growth curves, the *HVR* curve and the *POR*.

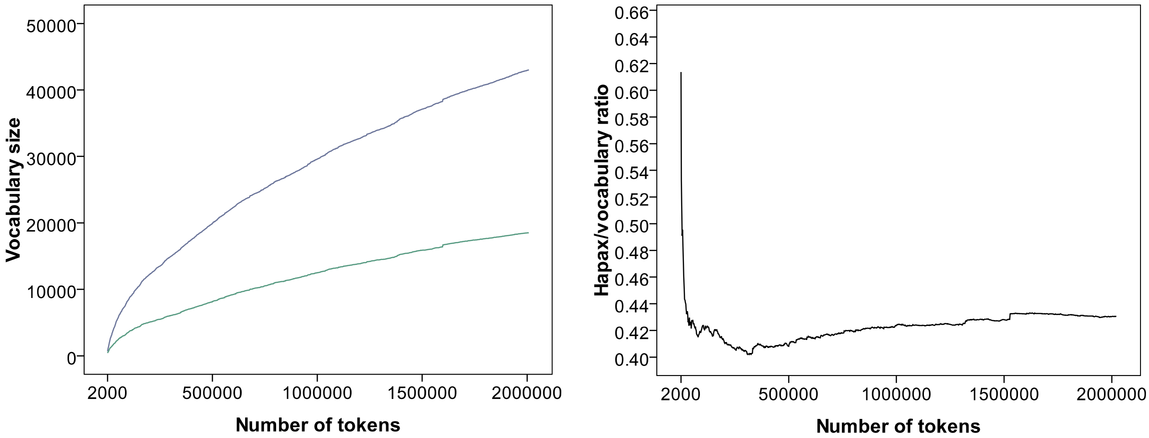


Figure 4. The vocabulary and hapax growth curves of the ANC technical English sample (left panel) and the *HVR* curve (right panel).

As shown in Figure 4, the *HVR* curve displays a pattern similar to that of the Combined Technical Corpus. The curve drops until *N* = 325,720, from which point it starts going upwards. Although the *POR* is larger than that of the Combined Technical Corpus, it is still much smaller than that of the BNC. The vocabulary at *POR* is 15,501, accounting for only 35% of the vocabulary but 95.37% of all the words of the sample. The *HVR* at the *POR* is 0.4019, larger than that of the Combined Technical Corpus.

To see whether other domains have similar *HVR* curve patterns, we drew samples from the nine domains of BNC, i.e., Imaginative, Natural Sciences, Applied Sciences, Social Sciences, Belief, Art, World Affairs, Commerce and Leisure. We drew a 2000000-word sample from each of the domains. In addition, we also drew a 2000000-word sample from ANC's spoken English section, here referred to as ANC Spoken. The following were computed: the vocabulary size, number of hapaxes and the overall *HVR*.

As shown in Table 1, one thing that is noteworthy is the vocabulary size of ANC Spoken, which is only 17,738, only about half of that of BNC Imaginative, whose vocabulary size is 33,519, the second smallest of the ten domains.

Table 1
Lexical statistics of the samples from the ten domains of the BNC
and the ANC Spoken.

Domain	Vocabulary size	Hapax	*HVR*
ANC Spoken	17738	5830	0.3287
BNC Imaginative	33519	11226	0.3349
BNC Belief	35911	12605	0.3510
BNC Commerce	35149	13190	0.3753
BNC Social Sciences	35635	13453	0.3775
BNC Natural Sciences	43891	17178	0.3914
BNC Applied Sciences	43082	17047	0.3957
BNC Leisure	50373	20134	0.3997
BNC Art	53264	21365	0.4011
BNC World Affairs	51629	21481	0.4161

Secondly the *HVR* generally has a positive correlation with vocabulary size as shown in Figure 5. The ANC Spoken has both the smallest vocabulary and *HVR*; while Art and World Affairs have the largest vocabulary and *HVR*. However, all the *HVR* curves display a downward trend to the end and none of them have a distinctive *POR*, as shown in Figure 5. This means that for these domains at *N* = 2,000,000 the rate of reduction of the hapaxes is still faster than that of the increase of the hapaxes as *N* increases, indicating that many of the hapaxes are not rare enough to stay as *N* keeps increasing. In other words, possibly many of the hapaxes in these domains at this sampling time belong to the main vocabulary, whose frequency would increase if *N* keeps increasing, and the *HVRs* would keep dropping as well, until the *POR*s are reached.

In the case of the Combined Technical Corpora and the ANC Technical, the positive correlation between vocabulary size and *HVR* (as seen in the ten domains) does not hold. The vocabulary sizes of the two are respectively 32,932 and 44,286, which would rank second and eighth (assuming the smallest value has a rank of 1), but their *HVR*s are respectively 0.437 and 0.4304, larger than any of the *HVRs* of the ten domains. In addition, their *HVR*s at their respective *POR*s are relatively high, 0.3741 for the Combined Technical Corpus and 0.4019 for the ANC Technical. A logical explanation of this phenomenon is that the technical domain has a relatively small main vocabulary and relatively large set of peripheral words. As *N* reaches the *POR*, the major part of the main vocabulary has exhausted itself and words belonging to this part have all occurred more than once; at the same time the frequency of a majority of the rare within-dictionary words and low probability alpha-numeric strings still remains 1, hence causing the reversal of the *HVR* curve.

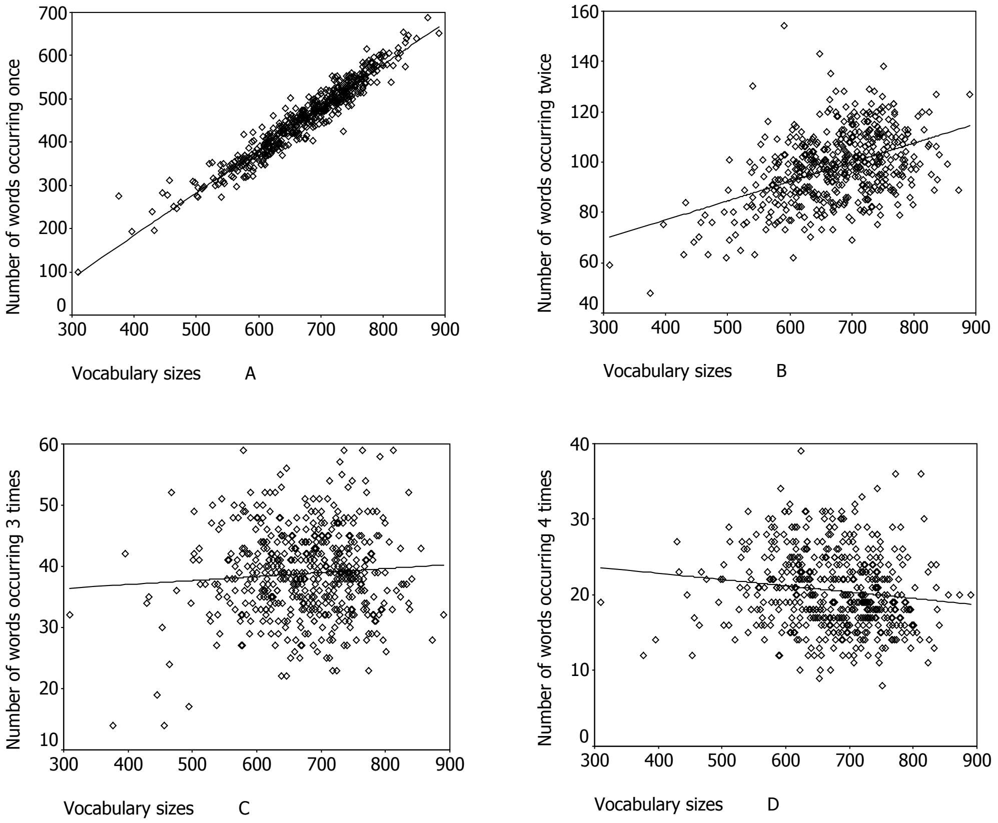


Figure 5. The relationship between vocabulary size and the number of words occurring 1-4 times in 500 2000-word texts randomly drawn from the BNC. The little squares in panels A-D are respectively the number of words occurring 1-4 times. Panel A reveals a clear positive linear relationship between the number of hapaxes and vocabulary size.

4. Concluding remarks

The *POR* can be used as a macro-lexical indicator providing a quick view of a macro-lexical composition of a corpus or a domain. An early, or a small *POR* indicates a relative small main vocabulary and a relatively large peripheral vocabulary, which is characteristic of technical English, as revealed in this study. If in samples of texts from a domain there is no distinctive *POR*, it would imply that the major part of the main vocabulary has not yet been exhausted. This suggests that in corpus compilation, for it to be minimally lexically adequate, the corpus in question should display a distinctive *POR*. Further work could be done on the *POR*s of other domains as well on the possible relationships between the *POR* and other quantitative measurements, such as word length, word frequency spectrum, the arc length, and the *h*-point and its related points.

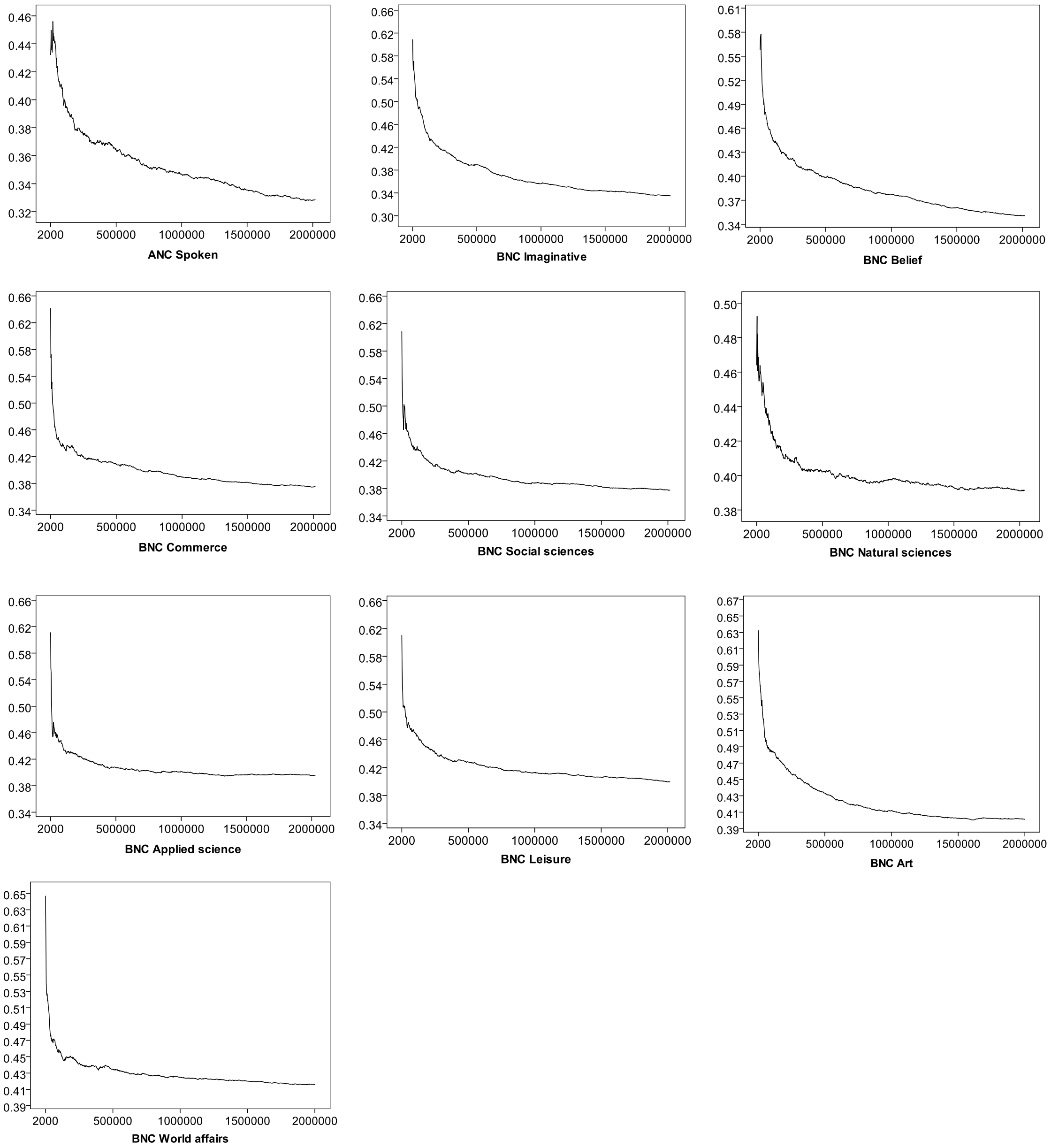


Figure 6. The *HVR* curves of the nine domains of the BNC and the ANC Spoken.

References

Alekseev, M. (1978). O nelinejnych formulirovkach zakona Cipfa. In: Piotrovskij, R. G. (ed.), *Statistika reči i avtomatičeskij analiz teksta: 53–65*. Moskva/Leningrad: Naučnyj sovet po kompleksnoj probleme "Kibernetika" AN SSSR.

Altmann, G., Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildescheim: Olms.

Fan, F. (2010). An Asymptotic Model for the English Hapax/Vocabulary Ratio. *Computational Linguistics 36*(*4*)*, 631–637*.

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1997). Persistence and other aspects of sentence-length series. *Journal of Quantitative Linguistics 4(1-3)*, *103–109*.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Addison Wesley.

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

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. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637–646*. Berlin, Heidelberg: Springer.

Miguel, F. M. (2007). Renewal of Core English Vocabulary: A Study Based on the BNC. *English Studies*, *88 (6), 699–723*.

Nation, P., Waring, R. (1997). Vocabulary size, text coverage and word lists. In: Schmitt, N., McCarthy, M. (eds.) *Vocabulary: description, acquisition and pedagogy: 6–19*. Cambridge: Cambridge University Press.

Popescu, I.-I., Best, K.-H., Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics 14, 58–71.*

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18–44.*

Schmitt, N., Jiang, X., Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal 11, 26–43.*

The phoneme-grapheme relation in the scheme of the Chinese Phonetic Alphabet

Wei Huang, Haitao Liu

1. Introduction

It is widely known that Chinese has a script system consisting of signs. Thus, it is difficult to compare Chinese with other languages which use letter scripts. However, the Scheme of the Chinese Phonetic Alphabet (Pinyin), which uses Latin letters, makes this kind of comparison possible.

The Scheme of the Chinese Phonetic Alphabet was established in 1958, and became an international standard in 1982 named Romanization of Chinese (ISO 7098: 1982). Pinyin is not the script system of Chinese in a strict sense, but a set of letters and rules for the Romanization of Chinese Mandarin. It is an artificial Romanization scheme, and plays an important role in documenting, and transliterating for both first and second language learning. From the functional viewpoint, it is equivalent to a letter script.

From the technological perspective, one can study the phoneme-grapheme relation in Pinyin by using the methods of studying other letter scripts since these relations do exist in this planned script. The phoneme-grapheme relation is part of the technological nature of a script system, which together with the universal and legal natures are the three basic characteristics of a script system (Zhou 1983). The last two are related to the community and society while the first one is more independent as well as prerequisite. A script system can be examined and evaluated from the technological point of view, and so can Pinyin.

It is the thinking in quantitative linguistics and synergetic linguistics that answers the question of how to examine the technological characteristics of a script system. In order to build a theory of language or script under the scheme of synergetics, a number of properties of letter script systems shall be quantitatively examined. It is the exactly defined concepts of these properties and the quantification of the measurement of these properties that makes it possible to compare the script systems among different languages and to evaluate a script system (even an artificial one) with other script systems in the world.

Many languages using Latin letters have been examined quantitatively in this way (Altmann, Fan 2008). However, Chinese (whose phonetic alphabet uses Latin letters, too) has not been examined yet. We examined some quantitative properties on the phoneme-grapheme level of Pinyin and made comparisons with several other script systems using Latin letters. An early conclusion is that Pinyin has high optimized effectiveness as an artificial equivalent of script system.

2. Letters, phonemes and graphemes of Pinyin

Pinyin uses 26 Latin symbols (and their upper case) as its letters. The letter ‘v’, which is reserved for spelling foreign words, is not used in any syllables of Chinese Mandarin. Besides, there are 2 additional letters with diacritic marks in Pinyin, namely ‘ü’ and ‘ê’. The former represents the phoneme [y][1] , while the latter represents [e_o] when it forms a syllable independently, which is a rare phenomenon in Chinese Mandarin.

There are 22 consonant phonemes, 10 vowel phonemes and 4 tone phonemes in Pinyin. The consonant phonemes, consisting of 21 initial consonants and the only final consonant [ng], and the corresponding graphemes are listed in Table 1. The consonants and the graphemes map in a one-to-one correspondence, i.e. a phoneme is represented by only one grapheme while a grapheme represents only one phoneme. Most of the initial consonant phonemes are represented by the one-letter graphemes. Besides, there are 4 combined graphemes with length 2 in Pinyin, namely ‘zh’, ‘ch’, ‘sh’, and ‘ng’.

Table 1
Consonant phonemes in Pinyin.

phoneme	grapheme	example	phoneme	grapheme	example
[p]	b	ba(爸/father)	[ts/]	j	ji(鸡/chicken)
[p_h]	p	pa(爬/climb)	[ts/_h]	q	qi(七/seven)
[m]	m	ma(妈/mother)	[s/]	x	xi(西/west)
[f]	f	fa(法/law)	[ts]	z	zi(子/seed)
[t]	d	da(大/big)	[ts_h]	c	ci(词/word)
[t_h]	t	ta(他/he)	[s]	s	si(思/think)
[n]	n	na(哪/where)	[ts`]	zh	zhi(纸/paper)
[l]	l	la(拉/pull)	[ts`_h]	ch	chi(吃/eat)
[k]	g	ge(哥/brother)	[s`]	sh	shi(是/is)
[k_h]	k	ke(渴/thirsty)	[z`]	r	ri(日/sun)
[x]	h	he(和/and)	[ng]	ng	bang(棒/excellent)

The vowel phonemes and their corresponding graphemes in Pinyin are more complex than the consonants. Lin and Wang (1992: 203) summarized that there are only 6 vowel phonemes in Pinyin. They consider the phonemes [Ii] and [iI] as variants of [i]. However, other linguists (Huang, Liao 2007: 98) hold

[1] The phonemes are coded in Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) within square brackets in this paper.

different opinions on [Ii] and [iI], which occur only behind the [ts, ts_h, s] and [ts`, ts`_h, s`, z`] respectively and are both represented by the grapheme ‘i’ in Pinyin, which also represents the phoneme [i]. Hence, the letter ‘i’ plays a role of multifunctional grapheme which represents 3 different phonemes in Pinyin. The retroflex vowel [@r] is usually considered a phoneme in Pinyin by other Chinese linguists (Huang, Liao 2007: 97). A syllable in Chinese Mandarin always corresponds to one Chinese character. However, the retroflex suffixation is an exception, as the last entry indicates in Table 2. Besides, there is a transliteration rule in Pinyin that the phoneme [e_o] should be represented by the diacritical letter /ê/ when it is used independently, i.e. it makes up a syllable solely, although there are few syllables (combination of phoneme and tone) of this type and few words with these syllables in Chinese Mandarin. To maximize the complexity of the phoneme-grapheme relation of Pinyin, we consider all of these vowels to be 10 phonemes in the following examination. The representation of the vowel phonemes in Pinyin is also complicated, as shown in Table 2. The sign ‘∈’ means there is an abbreviatory grapheme provided by the transliteration rules of Pinyin, which is that ‘iou’, ‘uei’ and ‘uen’ should be abbreviated as ‘iu’, ‘ui’ and ‘un’ when they follow an initial consonant in a syllable. Therefore the grapheme ‘u’ represents two phonemes, just as in English, the ‘x’ represents [k] and [s] in ‘affix’.

Table 2
Vowel phonemes in Pinyin.

phoneme	grapheme	example	phoneme	grapheme	example
/A_A/	a	ba(爸/father)	[u]	u	lu(路/road)
[o]	o	po(破/break)		wu	wu(无/none)
	∈u	liu(六/six)		w	wa(瓦/tile)
[e]	e	de(的/of)		o	bao(包/bag)
	∈u	gui(贵/expensive)		∈u	liu(六/six)
[e_o]	ê	ê(欸/eh)	[y]	ü	nü(女/female)
[i]	i	li(梨/pear)		u	qu(去/go)
	yi	yi(一/one)		yu	yu(鱼/fish)
	y	ya(牙/tooth)		io	qiong(穷/poor)
[Ii]	i	zi(子/seed)		y	yong(用/use)
[iI]	i	zhi(纸/paper)	[@r]	er	er(儿/son)
-	-	-		r	huar(花儿/flo wer)

Chinese is a tonal language, in which the tone can distinguish meanings. The phoneticians (Lin, Wang 1992: 128; Huang, Liao 2007: 65) claimed that there are only 4 tones in modern Chinese Mandarin, and that the neutral tone is not the fifth one, but a type of tonal sandhi. According to the scheme of the Chinese Phonetic Alphabet, the original tone of a syllable should be marked and there is no diacritic symbol for the neutral tone. The 4 tone phonemes and the corresponding graphemes (symbols) of Pinyin are shown in Table 3. The symbol '#' means the formation of no symbol. For example, 'mā' (ma1) has no tone symbol in 'Ni3hao3 ma?' (你好吗？/How are you?). Moreover, the functional domain of the tones in Chinese is not the main vowel but the whole syllable. Hence, the main vowels with tone symbol, for example, 'i3' or 'a3', should not be considered as new diacritic graphemes that differ from 'i' or 'a'.

Table 3
Tone phonemes of Pinyin.

phoneme	[_1]	[_2]	[_3]	[_4]
grapheme	1/#	2/#	3/#	4/#
example	ma1 (妈/mother)	ma2 (麻/hemp)	ma3 (马/horse)	ma4 (骂/abuse)

3. Some quantitative properties of Pinyin

We will examine quantitatively the following properties of the Scheme of the Chinese Phonetic Alphabet: the economy of the script system, orthographic uncertainty, the distribution of graphemic representation, grapheme size, the grapheme load of letters, and letter usefulness. Comparisons with Italian (Bernhard, Altmann 2008), Slovene (Kelih 2008), Slovak (Nemcová, Altmann 2008), German and Swedish (Best, Altmann 2005) are performed.

3.1. Economy of the script system

The economy of the script system can be defined as the ratio of the number of phonemes to the number of graphemes (Best, Altmann 2005: 34). Script economy (SE) is a global index which does not express the phoneme-grapheme relationship in individual cases (Kelih 2008). We computed this index for Pinyin, SE = 0.86 (tones excluded) or SE = 0.88 (tones included). Compared with that of German (0.57), Swedish (0.63), Italian (0.60), Slovene (1.03), and Slovak (0.52), Pinyin has the second highest economy in these script systems (Figure 1).

3.2. Orthographic uncertainty of phonemes

It is reasonable to assume that the more “necessary” different graphemes are to represent a phoneme, the smaller the effectiveness of the writing system (Kelih 2008). According to Best and Altmann (2005) the orthographic uncertainty of phonemes can be expressed as

$$U_{/x/} = \log_2 n_x ,$$

where $U_{/x/}$ is the uncertainty of the phoneme $/x/$, $\log_2$ is the logarithm base 2, and n_x is the number of graphemes that can represent the phoneme $/x/$. The orthographic uncertainty of phonemes in Pinyin is showed in Table 4, x denoting the number of representing graphemes, U_x the degree of uncertainty, and f_x the number of phonemes with uncertainty U_x.

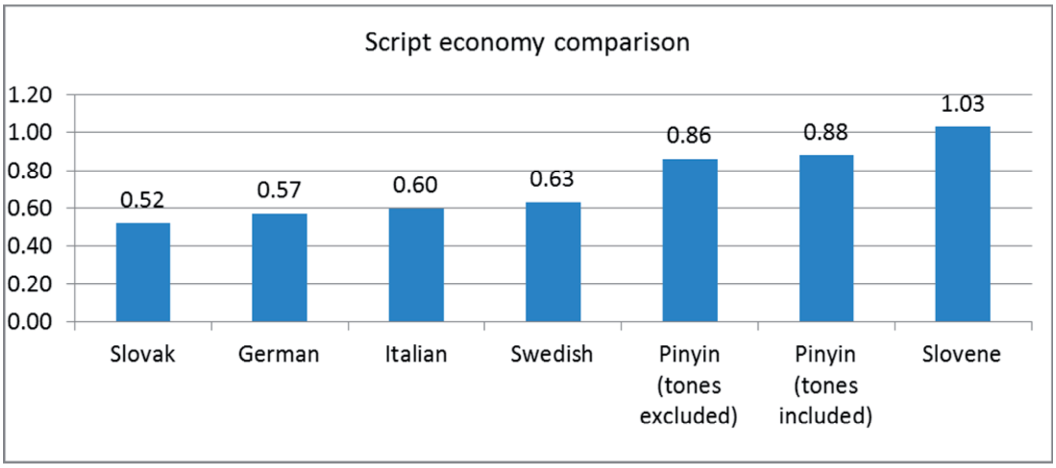


Figure 1. Script economy comparison within several languages.

Table 4
Orthographic uncertainty of Pinyin (tones included).

Phonemes	n_x	U_x	f_x
[p], [p_h], [m], [f], [t], [t_h], [n], [l], [k], [k_h], [x], [ts/], [ts/_h], [s/], [ts], [ts_h], [s], [ts`], [ts`_h], [s`], [z`], [ng], [A_A], [e_o], [Ii], [iI]	1	0	26
[@r], [o], [e], [_1], [_2], [_3], [_4]	2	1	7
[i]	3	1.58	1
[u], [y]	5	2.32	2

The majority of the phonemes are represented by one grapheme, especially the consonant phonemes. The mean orthographic uncertainty of phonemes of a given script system can be defined as (cf. Best, Altmann 2005)

$$\bar{U} = \frac{1}{N}\sum_x f_x U_x \,,$$

where N is the number of all representations. In Pinyin the $\bar{U}$ = 0.5101 (tones included) or $\bar{U}$ = 0.5739 (tones excluded).

We compared these numbers with the result from Italian ($\bar{U}$ = 0.5641), Slovene ($\bar{U}$ = 0.7841), Slovak ($\bar{U}$ = 0.7586), German ($\bar{U}$ = 0.9650) and Swedish ($\bar{U}$ = 0.7970). The $\bar{U}$ of Pinyin is almost the smallest one. In order to get a more objective image of these differences we executed an asymptotic test according to Bernhard and Altmann (2008) as

$$z = \frac{\bar{U}_1 - \bar{U}_2}{\sqrt{V\left(\bar{U}_1\right) + V\left(\bar{U}_2\right)}} \,,$$

where $V\left(\bar{U}\right)$ is the variance of $\bar{U}$ and should be

$$V\left(\bar{U}\right) = V\left(\frac{1}{N}\sum_{x=1}^{N}\log_2 n_x\right).$$

This formula can be transformed to (cf. Bernhard, Altmann 2008)

$$V\left(\bar{U}\right) = \frac{s^2}{0.48N\bar{x}^2} \,,$$

where

$$\bar{x} = \frac{1}{N}\sum_x x f_x$$

and

$$s^2 = \frac{1}{N}\sum_x \left(x - \bar{x}\right)^2 f_x \,.$$

As showed in Table 5, the uncertainty of Pinyin, whether the tones are included or not, is significantly different from that of German at the 95% confidence level.

There is no significant difference of uncertainty of phonemes in the statistical sense between Pinyin and Swedish, Pinyin and Italian, Pinyin and Slovene, and Pinyin and Slovak.

Table 5
The asymptotic test of differences of orthographic uncertainty.

	German	Swedish	Italian	Slovene	Slovak
Pinyin (tones included)	2.32	1.30	0.28	1.37	1.28
Pinyin (tones excluded)	1.80	0.93	0.05	0.95	0.85

3.3. Distribution of graphemic representation

There are three methods to build graphemes if the number of phonemes in a language is larger than the inventory size of Latin letters (Bernhard, Altmann 2008). Pinyin uses two of them. One is combining letters, eg. <zh, ch, sh>, and the other is introducing marks over the letters, eg. <ü, ê>. In these ways the phonemes in Pinyin acquire multiple representations and the distribution of representation size of phonemes can be captured formally. This type of representation size decreases geometrically in the simple hypothesis

$$P_x = pq^{x-1}, \quad x = 1, 2, 3, \ldots$$

and this 1-displaced geometric distribution fits well for Italian and Swedish script, but does not fit for German (Bernhard, Altmann 2008). Moreover, they used an expansion of the above distribution by Gram-Charlier, namely Gram-Charlier-geometric or Shenton-Skees-geometric distribution

$$P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right], \quad x = 1, 2, 3, \ldots,$$

where $q = 1 - p$, $0 < p \le 1$, and $0 \le a \le \frac{1}{q} - 1$. It fits very well for Italian, Swedish, German, Slovene, and Slovak.

Nevertheless, the empirical data of Pinyin listed in Table *4*4 can be captured by the 1-displaced geometric distribution as illustrated in Table 6 and Figure 2, but they can’t be captured by the Shenton-Skees-geometric distribution. We suspect that this is caused by the artificiality of Pinyin, but up to now there is no method to verify it.

Table 6
Fitting the 1-displaced geometric distribution
to graphemic representations in Pinyin.

x	f_x	NP_x
1	26	25.8111
2	7	7.3052
3	1	2.0676
4	0	0.5852
5	2	0.2310
p = 0.7170, χ^2 = 0.0188 , P = 0.8909, DF = 1		

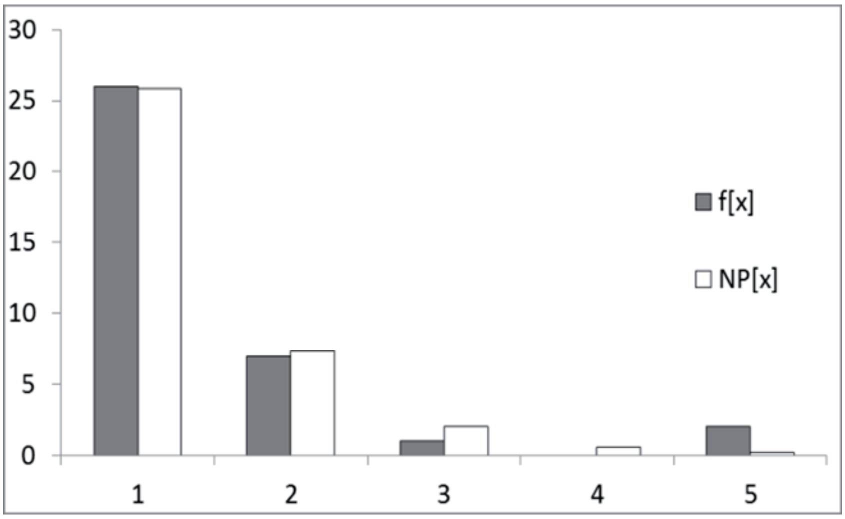


Figure 2. Fitting the 1-displaced Geometric distribution to graphemic representations in Pinyin.

3.4. Grapheme size

The grapheme size can be measured depending on whether the diacritic marks are considered independently. For Pinyin we obtain the results as shown in Table 7 (method 1, diacritic marks ignored) and Table 8 (method 2, diacritic marks considered). It is the <ê> and <ü> that makes the difference.

To compare with other languages we computed the average grapheme size of Pinyin. The result is 1.27 for method 1, and 1.32 for method 2. If the four tones with the grapheme size of 1 are included, these values will be 1.24 (method 1) and 1.29 (method 2). These values are respectively smaller than that in Italian (1.65), German (1.68), and Swedish (1.61) by method 1, and smaller than that in Italian (1.70), German (1.78), and Swedish (1.67) by method 2, while larger than that in Slovene (1.11) and Slovak (1.16) by method 1 and that in Slovene (1.25) by method 2.

The low grapheme size of Slovene can be explained by the “write as you speak” principle (Kelih 2008). Pinyin has also a low grapheme size compared with the analyzed script systems. The simple construction of graphemes is the result of the commitment to keeping things simple in the design of Pinyin.

Table 7
Grapheme size of Pinyin: method 1, tones excluded.

Size	Grapheme	Number
1	a, b, c, d, e, ê, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, ü, w, x, y, z	27
2	ch, er, io, ng, sh, wu, yi, yo, yu, zh	10

Table 8
Grapheme size of Pinyin: method 2, tones excluded.

Size	Grapheme	Number
1	a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, w, x, y, z	25
2	ch, er, io, ng, sh, wu, yi, yo, yu, zh, ê, ü	12

3.5. Grapheme load of letters

The exploitation of letters for building graphemes can be designated as grapheme load. By ignoring the diacritics, we computed the mean grapheme load of Pinyin based on the numbers in Table 9 as (27 · 1 + 10 · 2) / 25 = 1.88.

Table 9
Grapheme load of Pinyin.

Component in x graphemes	Letter	Number of letters
1	a, b, d, f, j, k, l, m, p, q, t, x	12
2	c, g, n, r, s, w, z	7
3	e, i, o	3
4	h, u, y	3
Total		25

If the diacritics are considered, the grapheme load of Pinyin will be (25 · 1 + 12 · 2) / 25 = 1.96. This means that a Latin letter is used not more than twice on the average in constructing graphemes in Pinyin. This value is smaller than that of Italian (3.92), German (3.96), Swedish (3.36) and Slovak (2.23), but larger than that of Slovene (1.32).

3.6. Letter usefulness

There is an assumption in measuring the participation of letters in building graphemes. The more peripheral the role of letter is, the later it appears in the grapheme. The definition of positional participation of a letter by Bernhard and Altmann (2008) is

$$PP_{\langle x \rangle} = \sum_{g_i \in G} p_x g_i \, ,$$

where p_x is the position of the letter $\langle x \rangle$ and g_i is the number of graphemes. All positional participations of letters in Pinyin are shown in Table 10, based on which we computed the average positional participation of Pinyin according to the formula by Bernhard and Altmann (2008)

$$\overline{PW}(Language) = \frac{1}{L} \sum_x f_x PP_{\langle x \rangle} \, .$$

The mean positional weight of letters in Pinyin is 2.28, and is smaller than that of Italian (6.48), German (6.12), Swedish (5.4), and Slovak (2.50), larger than that of Slovene (1.45), i.e. Pinyin has weak letter usefulness (small positional weight).

Table 10
Positional participation of letters in graphemes.

PP_x	Letters	f_x
1	a, b, d, f, j, k, l, m, p, q, t, x	12
2	c, n, s, w, z	5
3	e, g, r	3
4	i, y	2
5	o	1
6	u	1
7	h	1

We also examined the distance from the ideal state in which the letters are used only in mono-graphemes according to Nemcová and Altmann (2008) by

$$D = \sqrt{\sum_{x=1}^{K} \left(M_x - W_x \right)^2} \,,$$

where *K* is the number of the participative weights, 7 in our case, *W* is the vector of weight frequencies, [12, 5, 3, 2, 1, 1, 1] in our case, and *M* the ideal vector of weight frequencies, namely [25, 0, 0, 0, 0, 0, 0] in Pinyin. The D_{Pinyin} yields 14.49, which is smaller than D_{Slovak} = 17.78 and D_{Italian} = 21.6, but larger than D_{Slovene} = 6.16. Since the *D* indicates the necessity to perform a reform of orthography based on purely phonological reasons and disregarding cultural, historical and grammatical backgrounds (Nemcová, Altmann 2008), it may be true that the design of the Scheme of the Chinese Phonetic Alphabet has a relatively ideal graphemic optimization indeed.

4. Conclusions

In this article, we examined several quantitative characteristics of the Scheme of the Chinese Phonetic Alphabet. Similar to the script system of Slovene, the main property of Pinyin is a relatively small inventory of phonemes (32 consonants and vowels plus 4 tones), which is also represented by a small number of graphemes. All of the economy of the script system, the grapheme size, the grapheme load of letters, and the letter usefulness (including the D-indicator) reveals that Pinyin is more optimal than the script system of German, Italian, Swedish and Slovak respectively, and that it is less optimal than the Slovene script system. Pinyin has the lowest orthographic uncertainty, and has no significant difference with Italian, Swedish, Slovak and Slovene, but has a significant difference with German. The graphemic representation of Pinyin follows the 1-displaced geometric distribution.

The high effectiveness and optimization revealed in our examination enhances the learnability of Pinyin as a useful tool for both mother language and second language acquisition. Chinese is widely known as a non-letter script system. However, Pinyin which is a planned letter script system, is more effective and optimal than some natural ones. We assume, similar to the positive effect of human intervention in the development of Slovene, the advantages of Pinyin spring from its design. The designers proposed this Romanization scheme of Chinese based on the advantages of 14 previous projects from the 1580s to 1950s, including the Wade-Giles system, the Mandarin Romanization and the work by Matteo Ricci and others (Ma 2008). They constructed Pinyin by adopting a reasonable and effective means of expression as much as possible. Nevertheless, this assumption should be further proven by the quantitative examination and comparison of all these Romanization scheme of Chinese.

Moreover, the examination of the first conclusion that Pinyin is a highly effective and optimized script system needs further work. More languages should be investigated and compared in this way. There is another important factor. Chinese is a tonal language and Pinyin has a way without using Latin letters to represent the tone of syllables. However, not all of the quantitative properties in our examination are related to the tones. Thus these properties and measures do not cover the whole phoneme-grapheme relation of Pinyin, so the conclusion is also premature in this sense. Though the studies in this article support the conclusion, more properties should be examined in the future.

Acknowledgments
We are indebted to professor Shuiyuan Yu for his advice on SAMPA and to Huiyuan Jin for her literal improvement of the draft. This work was supported by the Municipal Social Science Foundation of Beijing (12WYC032), the National Social Science Foundation of China (11&ZD188 and 13CYY022) and the Science Foundation of BLCU (supported by the Fundamental Research Funds for the Central Universities) (13YBB20).

References

Altmann, G., Fan, F. (eds.) (2008). *Analyses of Script: Properties of Characters and Writing Systems*. Berlin, New York: Mouton de Gruyter.

Bernhard, G., Altmann, G. (2008). The phoneme-grapheme relationship in Italian. In: Altmann, G., Fan, F. (eds.) (2008). *Analyses of Script: Properties of Characters and Writing Systems: 11–22*. Berlin, New York: Mouton de Gruyter.

Best, K.-H., Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics 9, 29–39*.

Huang, B., Liao, X. (eds.) (2007). *Xiandai Hanyu* [Modern Chinese]. Beijing: High Education Press. (4th edition)

Lin, T., Wang, L. (eds.) (1992). *Yuyinxue Jiaocheng* [Course of phonetics]. Beijing: Peking University Press.

Kelih, E. (2008). The phoneme-grapheme relationship in Slovene. In: Altmann, G., Fan, F. (eds.) (2008). *Analyses of Script: Properties of Characters and Writing Systems: 23–58*. Berlin, New York: Mouton de Gruyter.

Ma, Q. (2008). Hanyu pinyin fang’an de laiyuan he jinyibu wanshan [The origin and improvement of Pinyin]. *Yuyan wenzi yingyong* [Applied Linguistics] *3, 17–18*.

Nemcová, E., Altmann, G. (2008). The phoneme-grapheme relation in Slovak. In: Altmann, G., Fan, F. (eds.) (2008). *Analyses of Script: Properties of Characters and Writing Systems: 77–85*. Berlin, New York: Mouton de Gruyter.

Zhou, Y. (1983). Pinyin he wenzi [Pinyin and script]. *Wenzi gaige* [Script Revolution] *4*.

Analysing h-point in lemmatised and non-lemmatised texts

Emmerich Kelih, Andrij Rovenchak, Solomija Buk

1. Introduction

In quantitative text analysis, the h-point plays an important role and is an important requirement for the calculation of further statistical stylistic and typological parameters. The h-point is a fixed point of a rank frequency distribution and is mainly used in lexical and word frequency studies. From a linguistic point of view the h-point is mainly understood as a fuzzy border between autosemantic and synsemantic word forms (cf. Popescu et al. 2009: 23). This paper tackles the methodologically and theoretically far-reaching consequences of the calculation of the h-point in lemmatised and non-lemmatised texts. After a short theoretical discussion and deductive derivation of the behaviour of the h-point in lemmatised and non-lemmatised texts, some empirical results from Russian, Slovene and Ukrainian will be discussed.

2. Lemmatisation – tokenisation: Quantitative and qualitative consequence

Word frequencies are studied in different branches of linguistics, in particular in quantitative stylistics, quantitative text analysis, psycholinguistics, authorship attribution and corpus linguistics. In word frequency studies, text pre-processing is obligatory and usually either word-form tokens or word-form lemmas are analysed. Whereas a study of word-form tokens is often based on a plain text approach, lemmatisation requires additional linguistic treatment of plain texts and the determination of standardised word forms, i.e. lemmas (e.g. infinitive or some basic form for verbs, nominative singular for nouns, etc.). Below, these two principally diverging approaches will be exemplified based on this Russian text:

> Kto iz vas, podlecov, kurit? Vse četvero ticho otvetili: - My ne kurim, batjuška. Lico popa pobagrovelo. - Ne kurite, merzavcy, a, machorku kto v testo nasypal? Ne kurite? A vot my sejčas posmotrim! Vyvernite karmany! Nu, živo! Čto ja vam govorju? Vyvoračivajte!
> [original Cyrillic text: Кто из вас, подлецов, курит? Все четверо тихо ответили: - Мы не курим, батюшка. Лицо попа побагровело. - Не курите, мерзавцы, а, махорку кто в тесто насыпал? Не курите? А вот мы сейчас посмотрим! Выверните карманы! Ну, живо! Что я вам говорю? Выворачивайте!]

Indeed text processing heavily depends on the problem of the word definition in general – a discussion which cannot be explored in detail here (for some recent discussions cf. Dixon, Aikhenvald 2002).

Quite often in quantitative and computational linguistics word definitions based on orthographic criteria are used. In this case usually the blank and punctuation marks are understood as a word delimiting sign, thus alphanumeric signs between two blanks in a written text are defined as one word form. Using this orthographic word definition for the Russian text presented above, the count yields the following results: 35 word-form types (counting word forms without taking their frequency into account) and 41 word-form tokens are obtained. In this case exactly five word forms appear more than once (*ne*, *a*, *kto*, *kurite*, *my*) and thus the number of word-form tokens is slightly higher than the number of word-form types. For details on the counted word form frequencies see Table 1.

Table 1
Frequency of word-form types.

No.	Types	Frequency	No.	Types	Frequency
1	*ne*	3	19	*kurit*	1
2	*a*	2	20	*lico*	1
3	*kto*	2	21	*machorku*	1
4	*kurite*	2	22	*merzavcy*	1
5	*my*	2	23	*nasypal*	1
6	*batjuška*	1	24	*nu*	1
7	*v*	1	25	*otvetil*	1
8	*vam*	1	26	*pobagrovelo*	1
9	*vas*	1	27	*podlecov*	1
10	*vot*	1	28	*popa*	1
11	*vse*	1	29	*posmotrim*	1
12	*vyvernite*	1	30	*sejčas*	1
13	*vyvoračivajte*	1	31	*testo*	1
14	*govorju*	1	32	*ticho*	1
15	*živo*	1	33	*četvero*	1
16	*iz*	1	34	*čto*	1
17	*karmany*	1	35	*ja*	1
18	*kurim*	1			

The lemmatisation (i.e. the reduction of word-form types to standardised word forms) leads again to a slightly lower number of counted lexical entities in the text. According to broadly accepted definitions (cf. Bußmann 2008: 296) the lemmatisation is the grouping together of different inflected forms to a standardised “basic form”. Morphologically similar word forms are therefore analysed as one single item. During the lemmatisation, a disambiguation of word forms is usually performed, too, and so homographs are identified and are treated in a linguistically more sophisticated way.

The quantitative consequence of a (manually performed) lemmatisation of the above Russian text is a shift of the frequency distribution of the counted lexical elements: the pronouns *vam*, *vas* are grouped together with *vy*, the inflected verb forms of *kurim* (1.P.Pl.), *kurit* (3.P.Sg.), *kurite* (2.P.Pl.) are reduced to the infinitive form *kurit'*. Finally, for the analysed text one obtains 32 lemmas, whereas in the case of tokenisation 35 word form types and 41 word tokens were obtained. For details on the counted lemma frequencies see Table 2.

Table 2
Frequency of lemmas.

No.	**Types**	**Frequency**	**No.**	**Types**	**Frequency**
1	*ne*	3	17	*lico*	1
2	*a*	2	18	*machorka*	1
3	*kto*	2	19	*merzavec*	1
4	*kurit'*	4	20	*nasypat'*	1
5	*my*	2	21	*nu*	1
6	*batjuška*	1	22	*otvetit*	1
7	*v*	1	23	*pobagrovet'*	1
8	*vy*	2	24	*podlec*	1
9	*vot*	1	25	*pop*	1
10	*vse*	1	26	*posmotret'*	1
11	*vyvernut'*	1	27	*sejčas*	1
12	*vyvoračivat'*	1	28	*testo*	1
13	*govorit'*	1	29	*ticho*	1
14	*živ*	1	30	*četvero*	1
15	*iz*	1	31	*čto*	1
16	*karman*	1	32	*ja*	1

In total, the following quantitative differences between lemmatisation and a simple tokenisation are obtained for the analysed Russian text: 41 word form tokens, 35 word form types and 32 lemmas (taking into account the frequency of lemmas equals the number of word-form types). From a purely quantitative point of view, simple tokenisation and lemmatisation are different approaches leading to a quantitative reduction of the lexical items in the texts. Beyond these quantitative consequences from a linguistic point of view, lemmatisation and tokenisation are rather different approaches, with a strong influence for any further qualitative interpretation of word form and lemma frequencies. Particularly one has to take into account that:

- Lemmatisation is predominantly relevant for inflecting languages. In this respect, the obtained quantitative differences between lemmatised and non-lemmatised texts provide in-depth information about the inflectional activity of analysed languages. As a tendency, mainly nouns, verbs, adjectives, pronouns and adverbs are characterised by inflectional forms and thus these morphosyntactic word-form classes are mainly affected by lemmatisation.

- Lemmatisation provides a significant possibility to obtain the lexical richness of texts. The problem of lexical richness – a problem mainly discussed in quantitative stylistics, authorship attribution and quantitative text analysis – cannot be analysed in a satisfactory way without the lemmatisation of texts, especially in the case of languages with inflection.
- Disambiguation, which is usually performed during the process of lemmatisation, is a necessary precondition of a sophisticated analysis of the lexico-semantic structure of texts. Further, the disambiguation of texts gives detailed information about the number of homographs in a text – without lemmatisation[1] this information is not available to the linguist.

Generally, it is important to note that the question of tokenisation and lemmatisation of texts is directly dependent on the linguistic hypothesis being explored, although for stylistic purposes and related problems, one has to favour a lemmatisation of texts. In the next section the impact of lemmatisation on the determination of the h-point will be discussed in detail.

3. h-point: lemmatisation and tokenisation

The h-point, originally introduced into scientometrics and bibliometrics by Hirsch (2005), has recently been discussed intensively in quantitative linguistics and in word frequency studies (cf. Popescu 2007; Popescu, Altmann 2006; Popescu, Altmann 2007; Mačutek et al. 2007; Popescu, Altmann 2008). The h-point is a fixed point on a rank-frequency distribution, where the rank r and the frequency $f(r)$ of a countable linguistic entity coincide. For special cases, where one cannot obtain the point where $r = f_r$, one can determine the h-point by the point where the product of rank and frequency reaches its maximum (cf. Martináková et al. 2008: 93). In both the above-mentioned cases, the h-point can be obtained rather easily and mechanically. For the exact calculation of the h-point one can use the formula proposed by Popescu and Altmann (2008: 95):

$$h = \begin{cases} r & \textit{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \textit{if there is no } r = f_r \end{cases}$$

[1] Nevertheless, in some languages general problems of lemmatisation arise that must be solved by arbitrary decisions, e.g. in Hungarian, the verb lemma is given as the third person singular; or in Indonesian languages where there is no inflection, there are “basic words” from which everything else is derived.

where f_r is the frequency of the element at rank r.

According to Popescu et al. (2009), the h-point separates the vocabulary (V) of a text into two parts, namely into a class of magnitude h of frequent synsemantic auxiliaries (prepositions, conjunctions, pronouns, articles, particles, etc.) and a much greater class ($V - h$) of autosemantics, which are not so frequent but which build the very vocabulary of the text. Thus the h-point separates the “rapid” branch of synsemantics before the h-point from the “slow” branch of autosemantics after the h-point. Without a doubt the separation of autosemantic and synsemantic word forms is not clear-cut; sometimes there are autosemantics in the rapid branch and in other cases there are synsemantics in the slow branch. Thus the h-point is not an exact border between autosemantic and synsemantic word forms, but rather a fuzzy separating point of the lexico-semantic material on a rank frequency distribution.

The main field of application of the h-point is in quantitative text analysis and language typology. In cross-linguistic analysis and language typology the h-point can be interpreted as a sign of analytism, i.e. in analytic languages the number of word forms is smaller, and the synthetic elements are replaced by synsemantics. Furthermore, the h-point is considered to be a characteristic of individual texts within a given language and a sign of analytism/synthetism in cross-linguistic comparison. The h-point furthermore can be used for the measurement of the lexical richness of texts. This seems to be valid only when one accepts the area below the h-point (as the relevant one for the lexical richness of a text. As shown above, this area is characterised by a large number of autosemantics and thus Popescu et al. (2009: 95ff.) utilise this behaviour for their concept of the thematic concentration of texts.

Despite the broad applicability of the h-point, one has to take into account that the h-point systematically depends on text lengths. Therefore, one has to analyse texts which are approximately of similar length or one uses indices which are based on the h-point but normalised by text length (cf. Popescu et al. 2009: 19).

In any case the h-point plays a crucial role for word frequency studies and related problems of quantitative text analysis and cross-linguistic studies. All in all, the h-point seems to be an appropriate and multi-sided tool for word frequency studies, in particular text analysis and language typology. Whereas without any doubt, the h-point has a high conceptual value for the linguistic analysis of rank frequency distribution, to the best of our knowledge the influence of the lemmatisation of texts for the determination of the h-point has not been systematically analysed yet.

Deductively one can state that in the case of lemmatisation in highly inflectional languages the h-point will be higher than in cases of simple tokenisation, e.g. without lemmatisation. This is explainable by the fact – as already stated – that in the area before the h-point there are mostly synsemantics, which are as generally characterised by a low degree of inflection. In particular this holds true for synsemantics such as prepositions, conjunctions, particles, etc., which usually do not have inflectional forms. Furthermore, in the area before the h-point, the occasional autosemantic word forms (as types) are characterised by an extraordinarily high frequency (as tokens). A lemmatisation thus increases the frequency of these particular word forms because word forms of verbs, nouns, pronouns, etc. are reduced to a single item, e.g. one lemma. This behaviour has already been demonstrated in Section 1, where the frequency of particular word-form tokens like *kurit’* or *vy* increased. In the case of lemmatisation, the increase of the frequency of words forms causes a systematic shift of the h-point in the rank–frequency distribution. Or in other words: due to the lemmatisation and this systematic frequency shift of word forms to lemmas, the h-point “glides” to the right side on the rank frequency curve. In this case the lemmatisation of text thus causes an increase of the h-point, which should be rather systematic and in a systematic relationship to the length of the analysed texts. The empirical verification of these deductively found assumptions regarding the behaviour of the h-point in lemmatised and non-lemmatised texts will be performed on the basis of Russian, Slovene and Ukrainian texts in the next section. The focus is on a potential systematic shift of the h-point in lemmatised and non-lemmatised texts, using texts from different languages and of varying length.

3.1. Empirical evidence from Slovene

For Slovene the ranks and frequencies of word forms, i.e. word form tokens and lemmas, were determined in the short novel *Hlapec Jernej*, written by Ivan Cankar. The novel consists of ten chapters. All texts were first automatically lemmatised[2] by special software and in the second step all entries were checked manually. To get a more in-depth insight into the behaviour of the h-point in texts of different text lengths, the ten chapters were analysed in cumulative form (chapter 1, chapter 1 + 2, …, chapter 1 + 2 + … + 10).

Fig. 1a represents the behaviour of the h-point in cumulative Slovene non-lemmatised texts and in Fig. 1b the h-point in lemmatised Slovene texts can be seen.

First of all, the systematic behaviour of the h-point in relation to text length (= number of word forms and lemmas) can be obtained. This characteristic of the h-point is well known and explored in detail in Popescu et al. (2009: 19). Furthermore, it is obvious that the h-point in the lemmatised text is slightly higher than in the non-lemmatised texts (cf. Table 3 for details), but obviously due to the relative shortness of the analysed texts in some cases (chapters 1–9) no changes of the h-point can be seen, whereas in chapters 1–4 the h-point in the lemmatised texts is even lower than in the non-lemmatised texts.

[2] The software is available for free at http://nl.ijs.si/analyse/. The error rate for the analysed text is 5%.

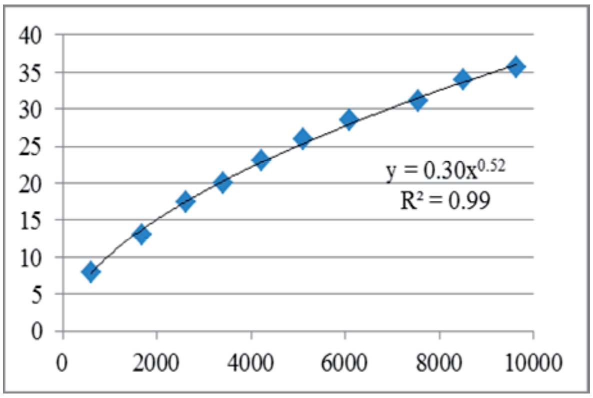


Figure 1a. The h-point in Slovene non-lemmatised texts.

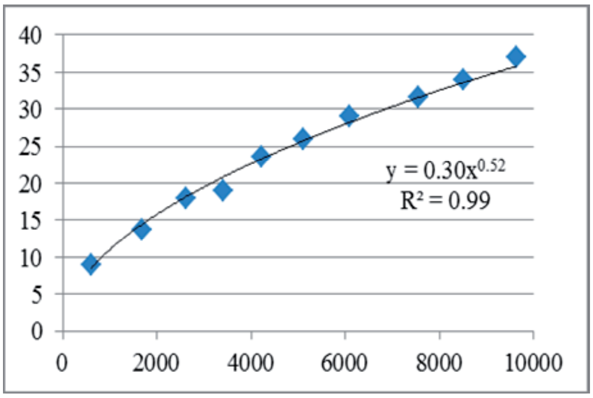


Figure 1b. The h-point in Slovene lemmatised texts.

Table 3
The h-point in lemmatised and non-lemmatised texts (Slovene).

Chapter	**Tokens**	**Types**	**Lemmas**	**h-point**	
				Tokens	**Lemmas**
1	602	343	283	8	9
1–2	1679	676	519	13	13.7
1–3	2617	938	687	17.5	18
1–4	3413	1149	814	20	19
1–5	4222	1302	910	23	23.5
1–6	5112	1477	1013	26	26
1–7	6085	1688	1133	28.5	29
1–8	7558	1969	1297	31	31.5
1–9	8497	2173	1404	34	34
1–10	9631	2361	1500	35.7	37

Thus we can state the following intermediate result for the Slovene text: as a rule of thumb, the h-point in lemmatised and non-lemmatised texts is in approximately the same position, or in other words, in our Slovene database no substantial differences can be obtained. This observation, which is in contradiction to our suggested behaviour of the h-point, confirms the already known systematic interrelation of the h-point with text length but not the systematic shift of the h-point. Nevertheless this finding seems to be explainable by the relative shortness of the texts used, which does not exceed 10,000 word forms and 1500 lemmas per chapter. Regarding the relevance of lemmatisation in quantitative text analysis and language typology, this can be understood in the following way: for a suitable interpretation at least some minimal text length is required and in the case of relatively short texts the lemmatisation does not provide any substantial additional linguistic information in comparison to a simple tokenisation.

3.2. Empirical evidence from Russian

Russian is usually considered to be a synthetic and strongly inflected language. The analysed Russian texts are slightly longer (cf. Table 4 for details) than the Slovene texts, so the impact of text length can be analysed in more detail.

The texts are taken from the Russian novel *Kak zakaljalas' stal'* (*How the Steel Was Tempered*). The lemmatisation of the ten analysed chapters was performed by TreeTagger, a lemmatiser for Russian available for free[3]. The novel *Kak zakaljalas' stal'* represents a specific form of the socialistic realism literary language of the 1930s and thus the lemmatiser we used does not recognise many of the word-form tokens in a proper way. The success rate was 90 per cent and therefore over 4,000 types were lemmatised manually.

The results of determining the h-point in lemmatised and non-lemmatised texts (cf. Table 4) show that the h-point in lemmatised texts is – as already predicted – in all cases higher than in non-lemmatised texts. Again, the predicted systematic relationship between the h-point and text length is found.

Table 4
The h-point: Non-lemmatised and lemmatised texts (Russian).

Chapter	**Tokens**	**Types**	**Lemmas**	**h-point**	
				Tokens	**Lemmas**
1	4107	1907	1376	20	21.7
1–2	8243	3538	2407	25.7	31
1–3	14567	5591	3605	35	41
1–4	18300	7003	4435	38.4	45
1–5	22070	7956	4931	42	51
1–6	28709	9822	5878	50	57
1–7	33940	11388	6693	55	63
1–8	40979	12864	7435	59	68
1–9	44275	13635	7814	62	71
1–10	49678	15053	8495	63	76.5

A graphical representation of the systematic behaviour of the h-point in lemmatised and non-lemmatised texts can be found in Fig. 2.

[3] I would like to thank Ruprecht von Waldenfels (University of Bern) for his help with the automatic lemmatisation of the Russian texts.

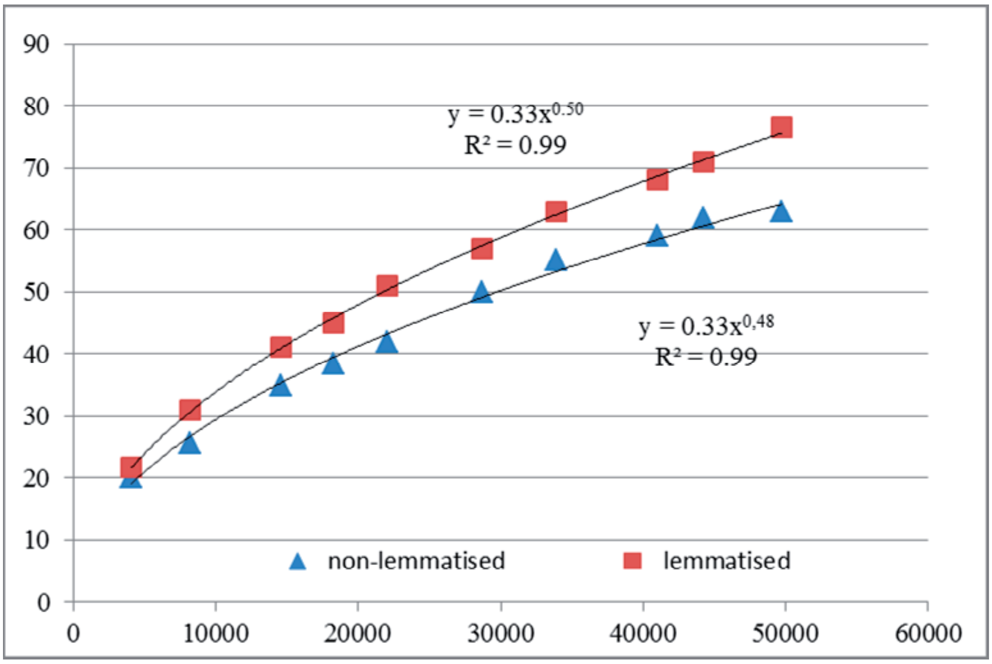


Figure 2. Text length vs. h-point.

As can be seen from Fig. 2, the distance between the h-point in lemmatised and non-lemmatised texts increases quite systematically with text length: The longer the texts, the larger the distance between the h-point in lemmatised and non-lemmatised texts. Generally, the analysis of the Russian texts confirms the deductively found claims about the h-point in lemmatised and non-lemmatised texts and furthermore the need for an analysis of texts with a suitable length is confirmed.

3.3. Empirical evidence from Ukrainian

Finally, the results for the Ukrainian texts can be presented. Compared to the Russian and Slovenian data, no cumulative chapters but eight individual novels written by the Ukrainian writer Ivan Franko were analysed. The corpus consists of the following titles (Buk 2007; 2013): *Boa constrictor* (1st edition: 1878–84; 2nd edition: 1905–07), *Boryslav smijetsja* (Boryslav Laughs) (1880–81); *Zakhar Berkut* (1883); *Ne spytavšy brodu* (Without Asking a Wade) (1885–86); *Dlja domašnjoho ohnyšča* (For the Hearth) (1892); *Osnovy suspil'nosty* (Pillars of Society) (1894–95); *Perekhresni stežky* (The Cross-paths) (1900); *Velykyj šum* (The Great Noise) (1907); *Petriji j Dovbuščuky* (2nd edition: 1909–12). In this work, the titles are referred to by the first letters of the Ukrainian transliteration, i.e.: BC, BC2, BS, ZB, NSB, DDO, OS, PS, VS and PD2. The texts of the corpus are tagged with part-of-speech and lemma information provided for each token. The tagging was performed semi-automatically. First, manual disambiguation of homographs was made. In the second step, the dictionary of word-form types was created and for each type the part-of-speech and lemma were given. This dictionary was used to tag every subsequent text and expanded as necessary.

The word forms were lemmatised, i.e., reduced to an initial (vocabulary) form: verbs to the infinitive, nouns and pronouns to nominative singular, adjectives to masculine nominative singular, etc. Suppletive forms of adjectives and adverbs (superlative and comparative) were reduced to separate (comparative) forms. Another point to be noted is the euphonic changes, which are also taken into consideration (cf. Buk, Rovenchak 2007). This affects mostly *і/й* and *в/у* alternation (word-initially and as separate words) and the verbal reflexive particle -*ся/сь*. As a rule of thumb, the vowel variant (*і* [i], *у* [u]) is used between consonants, and the consonant variant (*й* [j], *в* [v]) is used between vowels. In a mixed phonetic environment the choice is less strict as it is conditioned by phonetic harmony. The respective lemmas were joined into one type in frequency lists of all the texts apart from BC2 and PD2. The latter two novels were not included in the comparison but the preliminary analysis suggests that no substantial difference is observed: a slight shift of the h-point is caused by the fact that *і/й* and *в/у* lemmas have rather high frequencies with ranks $r < 10$. All data for the Ukrainian texts can be found in Table 5.

Table 5
The h-point in Ukrainian texts.

Texts	Tokens	Types	Lemmas	h-Point	
				Tokens	Lemmas
BC	25427	8351	5007	48	56
BS	77456	16069	8572	98.5	109.5
DDO	44841	11518	6472	71.5	78
NSB	49170	12808	7140	73.8	80
OS	67173	15437	8345	89.3	101.5
PS	93884	19425	9961	105.5	113.5
VS	37005	11058	6468	62	67
ZB	50206	12494	6520	75.8	87

It can be seen that the h-point in the Ukrainian texts (cf. Fig. 3) shows the same behaviour as already obtained in the Russian texts: (a) the h-point in lemmatised texts is slightly higher than in non-lemmatised texts and (b) in both cases the h-point systematically interrelates with the text length.

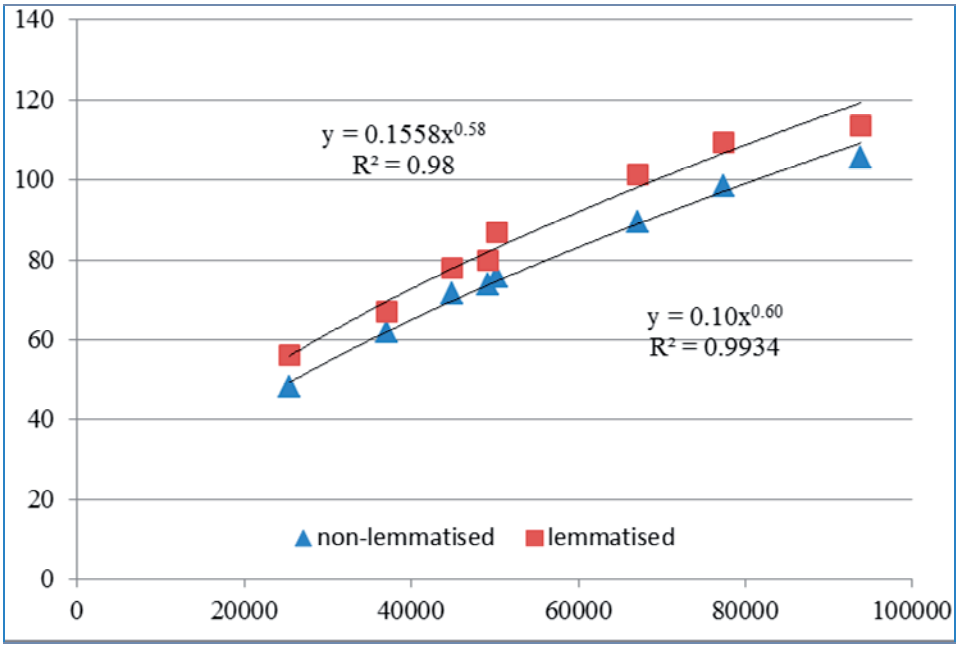


Figure 3. The h-point in Ukrainian novels

4. Summary

The main results of the systematic comparison of the h-point in lemmatised and non-lemmatised texts are as follows:

(1) In two analysed languages (Russian, Ukrainian) the h-point shows the deductively predicted behaviour. The h-point systematically interrelates with the text length, regardless of the counted lexical units (word-form tokens or lemmas).

(2) In the Russian and Ukrainian lemmatised texts without any exception, the h-point is higher than in the non-lemmatised texts, so one can claim a rather systematic effect of lemmatisation on the h-point.

(3) For Slovene no clear-cut results can be obtained, i.e. the h-point is equal in lemmatised and non-lemmatised texts. Considering the results of the Russian and Ukrainian texts, this quite specific behaviour can be explained by the relatively short length of the analysed texts. Hence in addition to the kind of text processing (lemmatisation or simple tokenisation) in h-point studies, one has to consider the sample size of analysed texts.

Altogether, the results obtained from Russian, Ukrainian and Slovene texts support the general relevance of the h-point for word frequency studies. In addition to its central function of being a fuzzy border between autosemantic and synsemantic word forms of rank–frequency distribution, which are based on word-form tokens, the h-point can indeed serve as a useful indicator for the lexical richness of texts. The latter holds true especially when sufficiently long texts are used (such as the described Russian and Ukrainian texts), whereas in relatively short texts (e.g. the Slovene database) no substantial differences of the h-point in lemmatised and non-lemmatised texts could be obtained. Thus in short texts – at least the results from the Slovene texts support this view – the h-point can be understood as a general tool for cross-linguistic comparison, stylistic analysis and particularly lexical richness as the lemmatisation of texts does not affect the position of the h-point substantially.

References

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft. Vierte, duchgesehene und bibliographisch ergänzte Auflage.* Stuttgart: Kröner.

Buk, S. (2007). Korpus tekstiv Ivana Franka: sproba vyznačennja osnovnykh parametriv [Ivan Franko text corpus: an attempt to define main parameters]. In: Šyrokov, V. A. (ed.), *Prykladna linhvistyka ta linhvistyčni tekhnolohiji: MegaLing‑2006: 72–82*. Kyiv: Dovira.

Buk, S. (2013). Kvantytatyvna parametryzacija tekstiv Ivana Franka: proekt ta joho realizacija [Quantitative parametrisation of texts written by Ivan Franko: the project and its realization]. *Visnyk Lvivs'koho universytetu. Serija filolohična 58, 290–307*; see also preprint arXiv:1005.5466v1 [cs.CL].

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel Perekhresni stežky (The Cross-Paths). In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 39–48.* Berlin, New York: de Gruyter.

Dixon, R.M.W., Aikhenvald, A.Y. (2002). Word: a typological framework. In: Dixon, R.M.W., Aikhenvald, A.Y. (eds.), *Word. A cross linguistic typology: 1–41.* Cambridge: Cambridge University Press.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA 102(46), 16569-16572.*

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics 15, 45–52.*

Martináková, Z., Mačutek, J., Popescu, I.-I., Altmann, G. (2008). Some problems of musical texts. *Glottometrics 16, 80–110.*

Popescu, I.-I. et al. (2009). *Word Frequency Studies.* Berlin, New York: Mouton de Gruyter.

Popescu, I.-I. (2007). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 555–565.* Berlin, New York: de Gruyter (Quantitative Linguistics, 62).

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23–46*.

Popescu, I.-I., Altmann, G. (2007). Writer's view of text generation. *Glottometrics 15, 71–81.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 94–108.*

The fractal structure of linguistic motifs

Reinhard Köhler

1. Introduction

It was Luděk Hřebíček who proposed to investigate linguistic structures with respect to their fractal dimensions (Hřebíček 1992, 1994, 1997, 1998), following his approach to a recursive application of the Menzerath-Altmann Law (Altmann 1980). Andres (2010) clarified the questions which were discussed in the sequel (Köhler 1995, 1997) mathematically, and he illustrated the usefulness of the approach in several publications together with Benešová (Andres, Benešová 2011, 2012).

The present paper attempts to apply fractal analyses to linguistic motifs, i.e. sequences of numerical values which represent properties of linguistic units (Köhler 2006, 2008a, b). We assume that the dimensions of motifs differ from that of the basic property and that they depend on the order of the given motifs. The *linguistic motif* (which was first called segment) is defined as

the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit.

Specifically,
An L-motif is a continuous series of equal or increasing length values (e.g. of morphs, words or sentences).
An F-motif is a continuous series of equal or increasing frequency values (e.g. of morphs, words or syntactic construction types).
A P-motif is a continuous series of equal or increasing polysemy values (e.g. of morphs or words).
A T-motif is a continuous series of equal or increasing polytextuality values (e.g. of morphs, words or syntactic construction types).

In principle, motifs can be formed on the basis of any linguistic unit and any property. An example of an L-motif segmentation is the following. The sentence

A well-known Czech linguist was the first to investigate fractal structures in language

can be mapped onto an L-motif sequence based on words as units and their lengths in terms of syllables as the property under consideration in the following way:

1 2 | 1 2 | 1 1 1 1 4 | 2 2 | 1 2,

where the vertical lines separate the individual motifs. Now, each sequence of motifs can again be used to form motifs, e.g. L- or F-motifs. By determining the lengths of the above motifs, a new sequence of L-motifs, which are called second order L-motifs, in short LL-motifs, or L^2-motifs, is formed. Our example yields the LL-motifs

2 2 5 | 2 2,

which would produce a single LLL-motif

3 2.

2. Method and experiments

In this way, starting with length values (measured in the number of syllables as in our example), three Italian[1] and three German[2] texts were converted into sequences of L-, L^2-, L^3- etc. up to L^7-motifs resp. to L^5-motifs in the case of the German texts, which were considerably shorter than the Italian ones. The fractal dimensions of these sequences were determined by means of the (accelerated) capacity method (Hunt, Sullivan 1986).

The capacity dimension is defined as

$$d_i = \frac{\ln \frac{N_{i-1}}{N}}{\ln \partial},$$

where N is the number of frames containing at least one part of the object, ∂ the contraction factor, and i the running number of iterations.

Let us demonstrate the principle of the procedure with the help of a concrete example. We choose arbitrarily a starting set of three intervals for the first step. Let us assume that we find parts of objects in all the three intervals in this first iteration cycle such that N = 3. In the next step, the interval size will be reduced by the factor ∂, for which we choose, say 0.5. Let the result of counting be N = 6 non-empty intervals in this step. If, in the next iteration cycle, after applying the same contraction factor 0.5, yielding 12 intervals, 8 non-empty intervals are found, N takes the value 8.

[1] I express my thanks to Arjuna Tuzzi, Padua University, for the access to these data.

[2] These texts were taken from the Gutenberg Project (http://www.gutenberg.org/).

With the resulting values of N and the contraction parameter $\partial = 0.5$ the first two iteration cycles of the approximation procedure with the above formula are as follows:

$$d_1 = \frac{\ln \frac{3}{6}}{\ln 0.5} = \frac{-0.69314718}{-0.69314718} = 1.0$$

$$d_2 = \frac{\ln \frac{6}{8}}{\ln 0.5} = \frac{-0.28768207}{-0.69314718} = 0.415037\,.$$

More iteration cycles approximate stepwise the fractal dimension of the object. Experience with the capacity dimension algorithm shows that the approximation process is not a smooth convergence but proceeds in abrupt jumps. Therefore, it is advisable to conduct a non-linear regression accompanying the procedure. The procedure was conducted using our program AAFD (Automated Analysis of Fractal Data).

Length values are discrete numbers, and a set of discrete (and hence unconnected) numbers should have a dimension below 1. Therefore, the algorithm was started with the embedding dimension 1. The results of the calculations can be seen in Table 1.

Table 1
The fractal dimensions of L-motifs of orders 1, 2,..., 7 of Italian novels
and of orders 1, 2, ..., 5 of the German short stories.
Length was measured in terms of the number of syllables.

Text 24: Italian novel Capacity dimension accelerated		**Text 40: Italian novel Capacity dimension accelerated**	
ORDER	DIM	ORDER	DIM
1	0.77	1	0.74
2	0.82	2	0.68
3	0.68	3	0.77
4	0.64	4	0.55
5	0.60	5	0.55
6	0.60	6	0.60
7	0.50	7	0.44
Text 11: Italian novel Capacity dimension accelerated		**Text 54: Italian novel Capacity dimension accelerated**	
ORDER	DIM	ORDER	DIM
1	0.64	1	0.71
2	0.77	2	0.77

3	0.60	3	0.64
4	0.64	4	0.55
5	0.64	5	0.55
6	0.55	6	0.60
7	0.77	7	0.44
Text PG1: German prose text (Goethe) Capacity dimension accelerated		**Text PS4: German prose text (Schnitzler) Capacity dimension accelerated**	
ORDER	DIM	ORDER	DIM
1	0.77	1	0.78
2	0.77	2	0.55
3	0.55	3	0.60
4	0.44	4	0.44
5	0.38	5	0.44
Text PS5: German prose text (Schnitzler) Capacity dimension accelerated			
ORDER	DIM		
1	0.79		
2	0.77		
3	0.55		
4	0.44		
5	0.44		

The experiment was repeated with L-motifs on the basis of character counts instead of syllables. The result of the calculations of the fractal dimensions for the three Italian novels are shown in Table 2. As can be seen, the values resemble those in Table 1.

Table 2
The fractal dimensions of L-motifs of orders 1, 2,..., 7 of Italian novels.
Length was measured in terms of the number of characters.

Text 24: Italian novel Capacity dimension accelerated		**Text 40: Italian novel Capacity dimension accelerated**	
ORDER	DIM	ORDER	DIM
1	0.55	1	0.64
2	0.85	2	0.74
3	0.74	3	0.64
4	0.64	4	0.55
5	0.60	5	0.55
6	0.64	6	0.77
7	0.77	7	0.50

Text 11: Italian novel Capacity dimension accelerated	
ORDER	DIM
1	0.60
2	0.80
3	0.77
4	0.77
5	0.60
6	0.44
7	0.44

3. Dependency of the dimensions on the order of the motifs

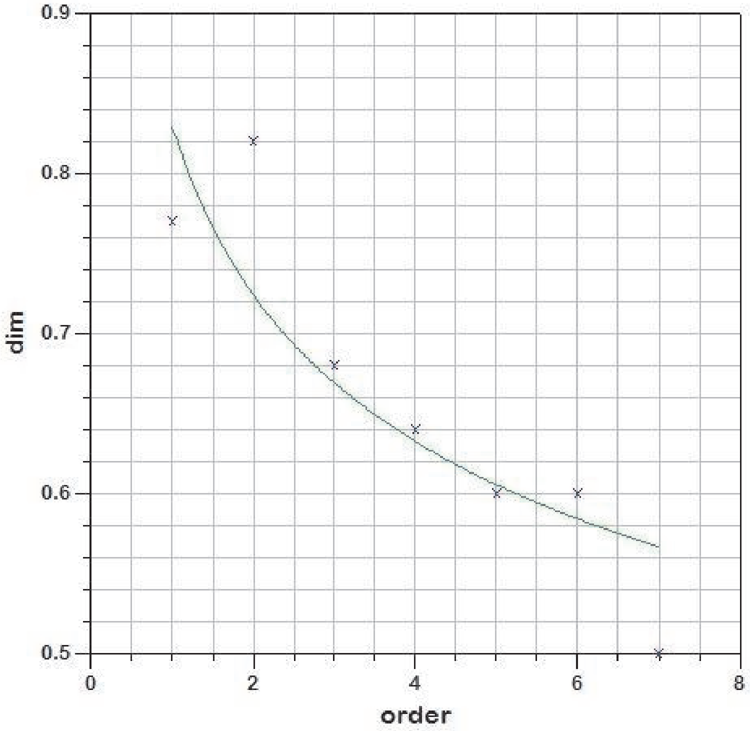
All the results presented here display a tendency of decreasing dimension values with increasing order of the L-motifs. It seems that the volatility of the series decreases with the order, which could be a result of a decreasing tendency to form varying repetition patters of length values. We will therefore study the shape of this tendency and ask the question whether it abides by a function representing a linguistic law. By analogy, we assume that the dimension values follow a power law function because this kind of interrelation is common in linguistic data. The results of fitting the power law function

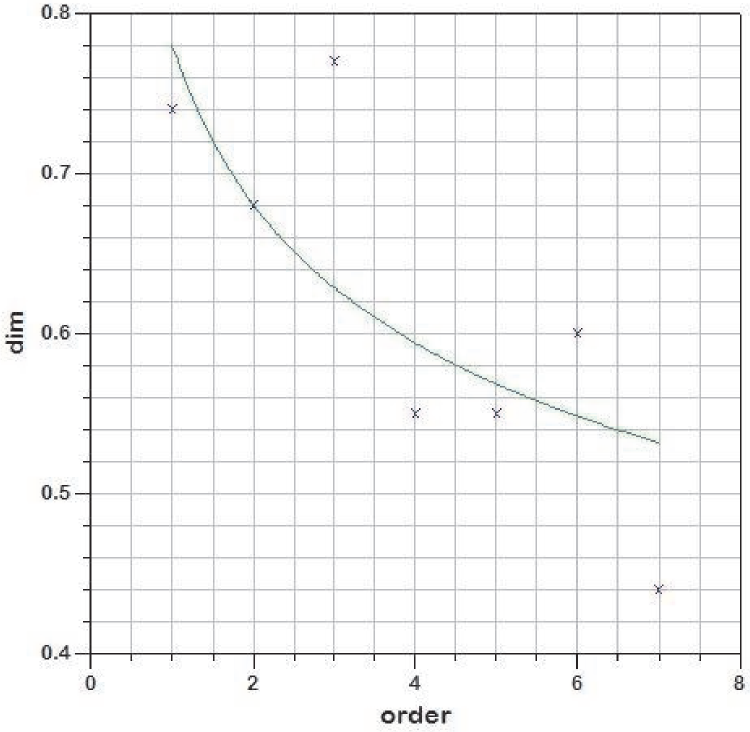
$$\dim = a \cdot order^{b}$$

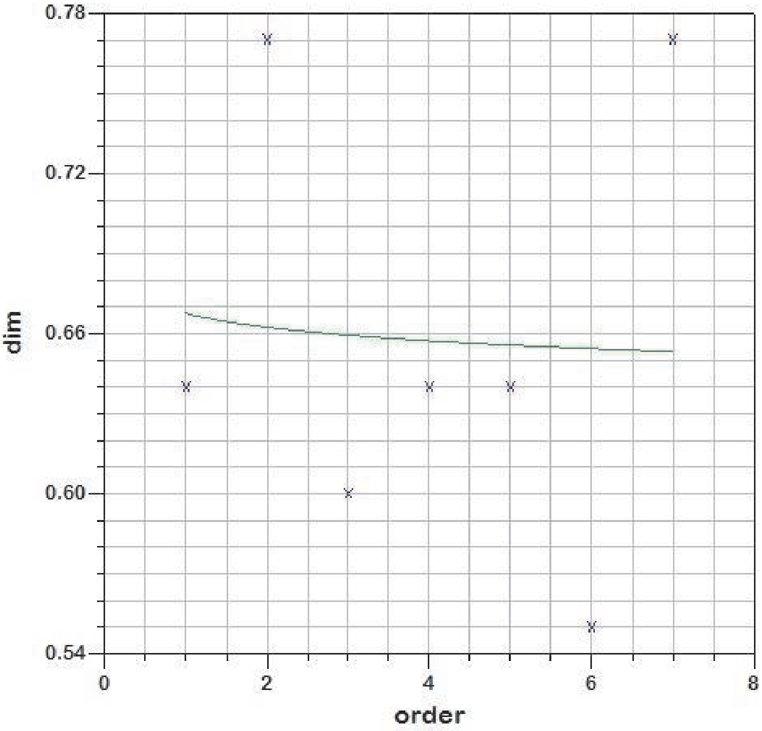
to the dimension data are given in Table 3 and in Figures 1.1–1.9.

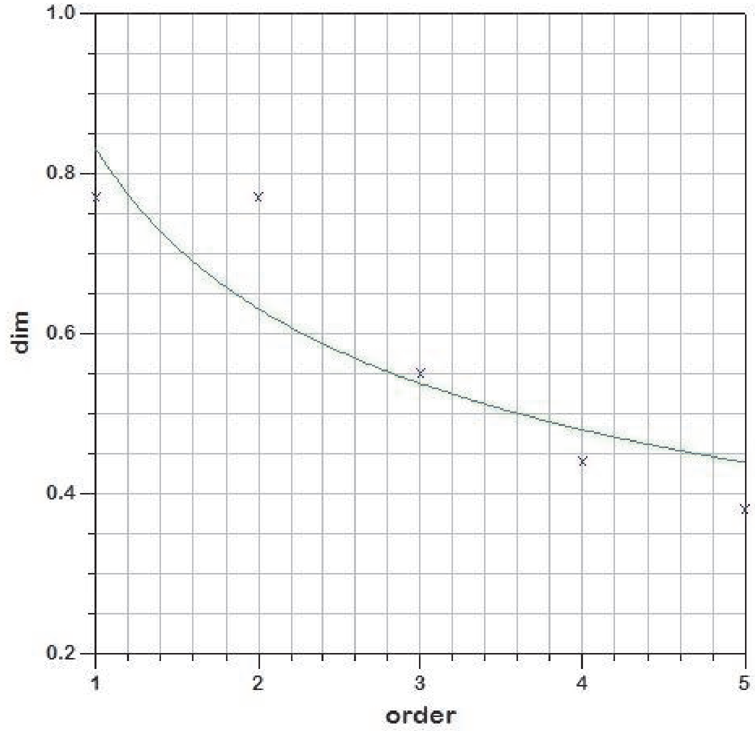
Table 3
The results of fitting the power law function.

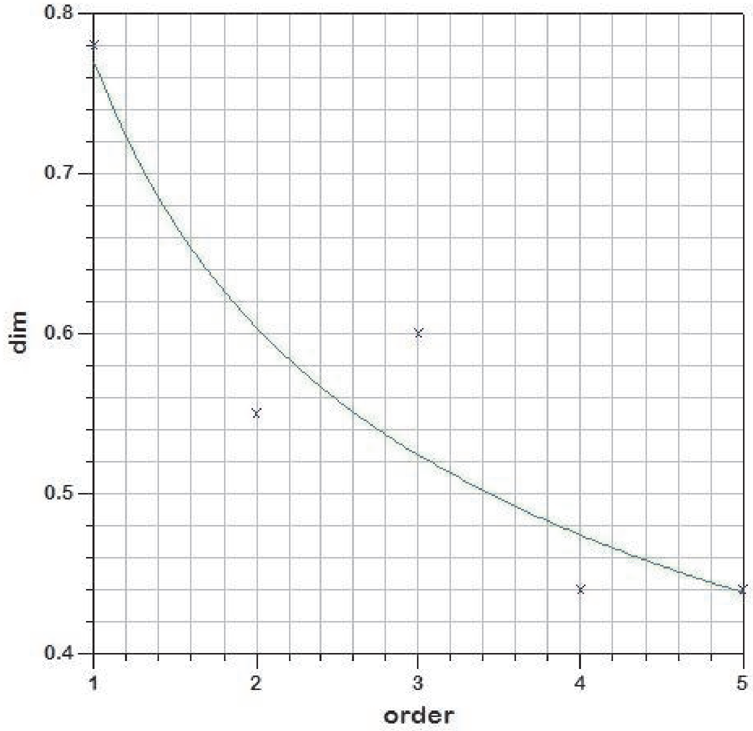
Text	a	b	R^2	Figure
24 (syllables)	0.8295	-0.1956	0.7528	1.2
40 (syllables)	0.7802	-0.1968	0.5784	1.2
11 (syllables)	0.6675	-0.0111	0.0036	1.3
G1 (syllables)	0.8296	-0.3952	0.7878	1.4
S4 (syllables)	0.7694	-0.3495	0.8741	1.5
S5 (syllables)	0.8386	-0.3765	0.8198	1.6
24 (characters)	0.6572	0.0330	0.0222	1.7
40 (characters)	0.6829	-0.0707	0.0954	1.8
11 (characters)	0.7484	-0.1423	0.1990	1.9

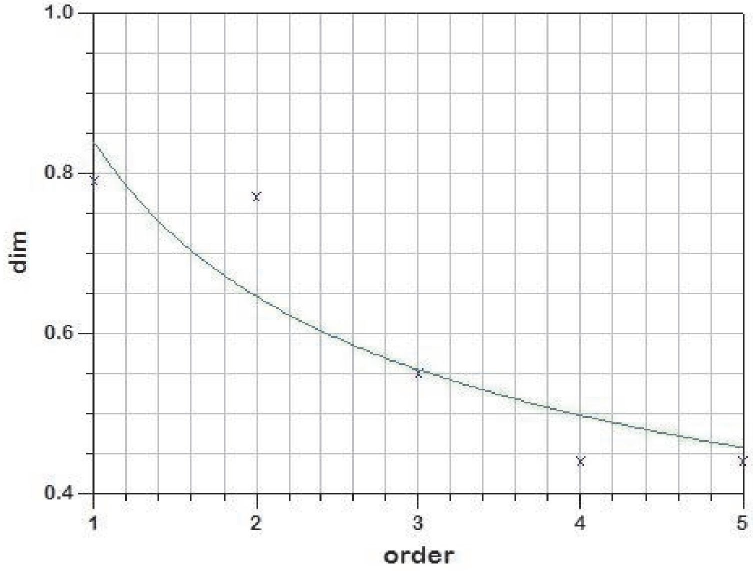

0.9
0.8
0.7
0.6
0.5
dim
0
2
4
6
8
order

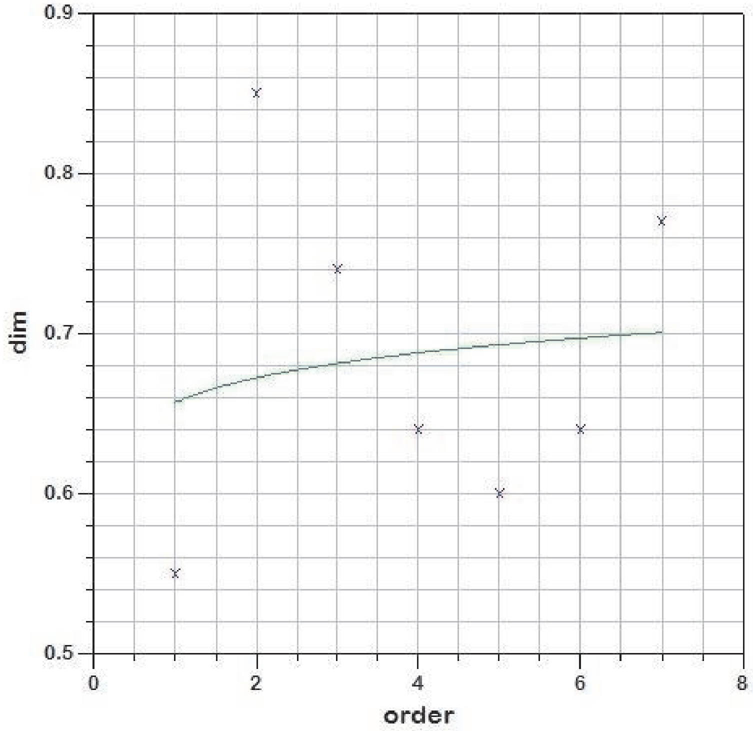

0.8
0.7
0.6
0.5
0.4
dim
0
2
4
6
8
order

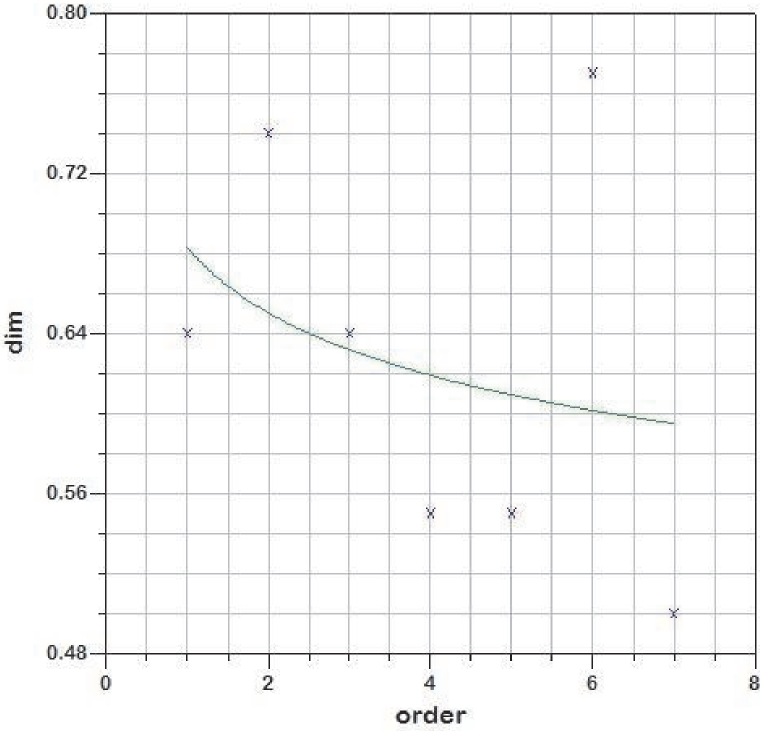

0.78
0.72
0.66
0.60
0.54
dim
0
2
4
6
8
order

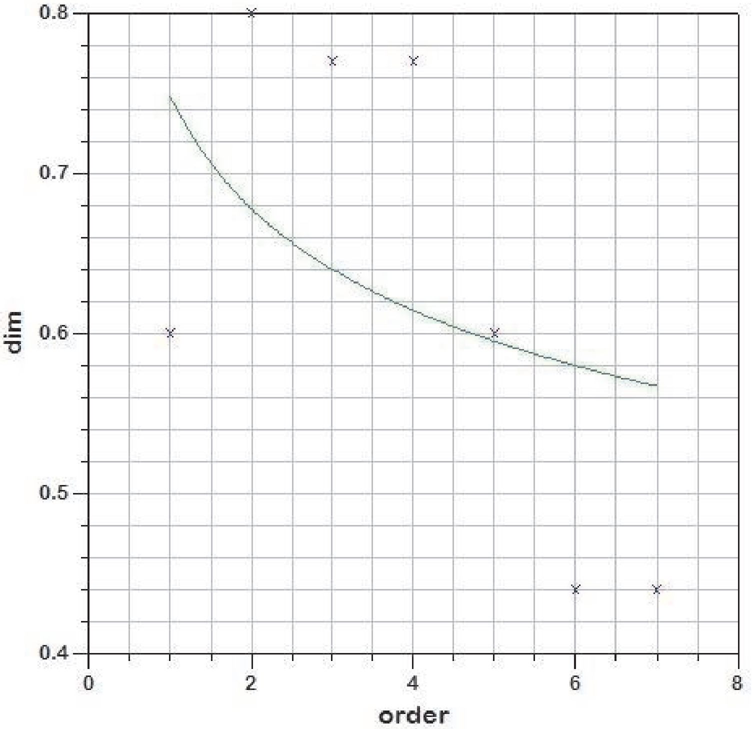

1.0
0.8
0.6
0.4
0.2
dim
1
2
3
4
5
order


0.8
0.7
0.6
0.5
0.4
dim
1
2
3
4
5
order


1.0
0.8
0.6
0.4
dim
1
2
3
4
5
order


0.9
0.8
0.7
0.6
0.5
dim
0
2
4
6
8
order


0.80
0.72
0.64
0.56
0.48
dim
0
2
4
6
8
order



Figures 1.1–1.9. The results of fitting the power law function.

4. Conclusion

Some of the results are acceptable but others are not and fail to show any regularity. Those data sets which have the assumed tendency follow in fact the power law function. At this moment, there is no clue as to the reasons for the diverging behaviour of the texts. Many more texts will have to be analysed to clarify this question. It should be marked, however, that the approximation algorithm which was used to calculate the fractal dimension of the motifs may be less reliable than expected. Some observations point in this direction, among others, the fact that the results of the calculations depend drastically on the number of iteration cycles. In a follow-up study, we plan to implement the compass dimension (Hunt, Sullivan 1986), which seems to be a more adequate model of the fractal structure of time-series data than the capacity dimension.

References

Altmann, G. (1980). Prolegomena to Menzerath’s law. *Glottometrika 2, 1–10.*

Andres, J. (2010). On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics 17, 101–122.*

Andres, J., Benešová, M. (2011). Fractal analysis of Poe's Raven. *Glottometrics 21, 73–100.*

Andres, J., Benešová, M. (2012). Fractal analysis of Poe's Raven II. *Journal of Quantitative Linguistics 19, 301–324.*

Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures*. Bochum: Brockmeyer.

Hřebíček, L. (1994). Fractals in language. *Journal of Quantitative Linguistics 1, 82–86.*

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute of Academy of Sciences of the Czech Republic.

Hřebíček, L. (1998). Language fractals and measurement in texts. *Archív orientální 66, 233–242.*

Hunt, F., Sullivan, F. (1986). Efficient Algorithms for Computing Fractal Dimensions. In: Mayer-Kress, C. (ed.), *Dimensions and Entropies in Chaotic Systems: 74–81*. Berlin: Springer.

Köhler, R. (1995). Maßeinheiten, Dimensionen und fraktale Strukturen in der Linguistik. *Zeitschrift für Empirische Textforschung 2, 5–6.*

Köhler, R. (1997). Are there fractal structures in language? Units of measurement and dimensions in linguistics. *Journal of Quantitative Linguistics 4, 122–125.*

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

Köhler, R. (2008a). The fractal dimension in script: an experiment. In: Altmann, G., Fengxiang, F. (eds.), *Analyses of Script, Properties of Characters and Writing Systems: 115–120*. Berlin: de Gruyter.

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

Moving window type-token ratio and text length

Miroslav Kubát

1. Introduction

Vocabulary richness measurement is one of the oldest and the most traditional fields of quantitative linguistics. Although researchers have proposed many methods of vocabulary richness calculation (Guiraud 1954, 1959; Tešitelová 1972; Tuldava 1977, 1995; Yule 1944; Köhler, Galle 1993; Covington, McFall 2010; Popescu et al. 2009, 2011 – to mention only some of them), this issue is still widespread. The main reason lies in the fact that researchers are still struggling with the influence of text length.

Although many formulas failed (most notable in recent years R_1 and Lambda structures proposed by Popescu et al. 2009, 2011), Moving average type-token ratio (MATTR) introduced by Covington and McFall (2010) seems to be fully independent of text length. This research aims to verify MATTR text size independence.

2. Methodology

Given that the main purpose of this study is to verify whether MATTR is influenced by text length, the term “text” must be first defined. In this research, whole collections of poetry, short stories, fairy tales, etc. are not considered as texts. It is highly unlikely that authors are able to write tens of thousands words in one consistent style. Moreover, some authors write their novels over months or even years. These long gaps between writing individual chapters must be somehow reflected in the style. In this work, only individual short stories, poems or individual chapters of a novel are considered as texts.

Moving average type-token ratio (MATTR) was proposed by Covington and McFall (2010). This method is based on the standardized type-token ratio (STTR) implemented in the *WordSmith Tools* (Scott 2013). STTR differs from MATTR in the fact that STTR is based on non-overlapping windows, whereas MATTR uses a smoothly moving window. The formula of MATTR is as follows:

$$MATTR\left(L\right)=\frac{\sum_{i=1}^{N-L}V_i}{L\left(M-L+1\right)},$$

L…arbitrarily chosen window size in tokens, $L < N$
N…number of tokens
V_i…number of types in an individual window

The corpus for testing MATTR text size independence consists of 660 texts written by the Czech writer Karel Čapek. Although all texts were published by one author, there are texts across genres: novels, short stories, fairy tales, travel books, essays and correspondence. It is important to emphasize that all texts comply with the aforementioned definition of the term "text". Only texts from one language were chosen, consciously, in order to avoid misleading results. For example, the authors in Popescu et al. (2011) demonstrate that the frequency structure lambda indicator is not influenced by text length in Figure 1.

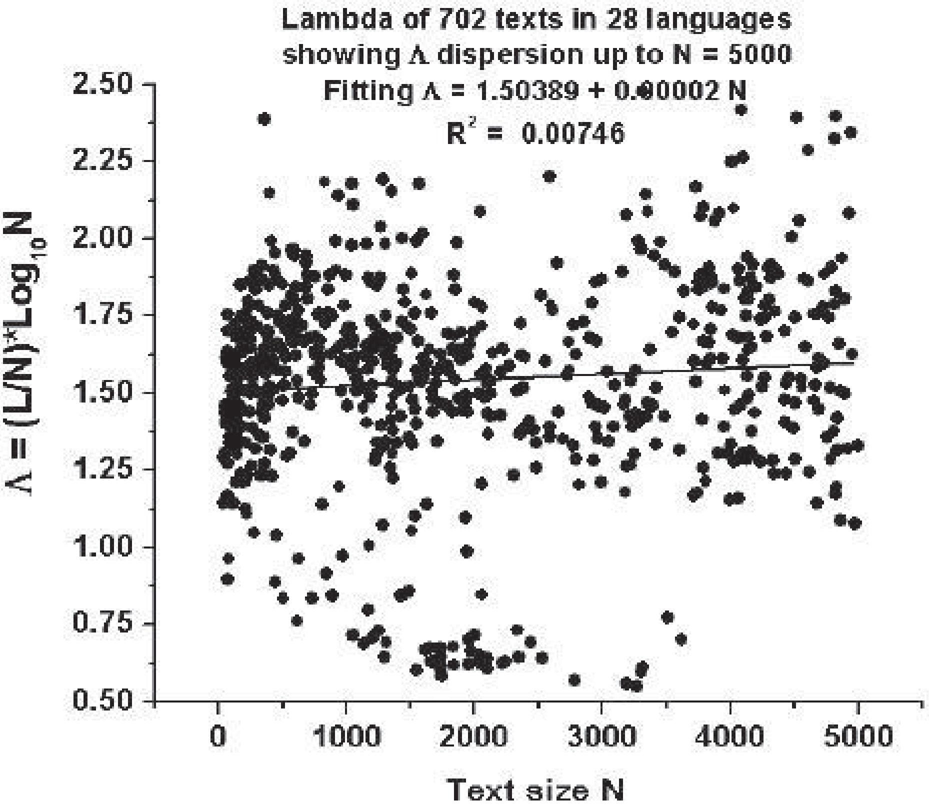


Figure 1. Lambda in 702 texts up to $N = 5000$ (Popescu et al. 2011: 11).

In Figure 1, a mixture of 702 texts from 28 languages can be seen. Although lambda seems to be quite independent of text size, this independence is because of the different structural features (typology) of the languages. This fact was pointed out by Čech (2015) who stated that "… if particular languages are analysed separately, a dependence of lambda on the text length emerged …"

In this study, the impact of text size on MATTR is verified by two methods. The first one shows the results of all 660 texts in a graph. The second method focuses on the relation between MATTR and cumulative text size in 19 relatively long texts randomly chosen from the corpus. All computations were performed by the MaWaTaTaRaD freeware (Milička 2013) and window size (*L*) was 100 tokens. In order to demonstrate how much MATTR reduces the impact of text size, TTR (type-token ratio) was also computed. From the methodological viewpoint, it must be also mentioned that a word-form is considered as a basic unit, so no text was lemmatized in this research.

3. Results

3.1. Results of whole text

The results of whole texts are presented in Figures 2 and 3. The difference between Figure 2 and Figure 3 is enormous at first sight: the simple TTR is heavily influenced by text size, whereas there is obviously no impact on MATTR. The results of TTR spread across almost the entire scale, the interval is approximately between 0.32 and 0.95. The obtained values of MATTR oscillate between 0.7 and 0.85 in the entire distribution.

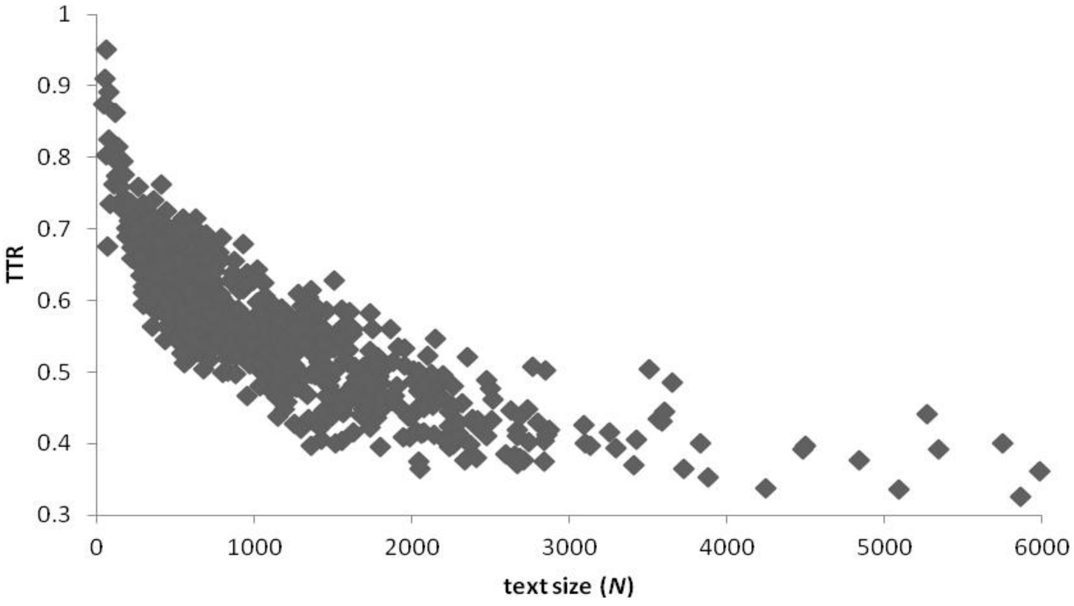


Figure 2. Results of TTR in 660 texts.

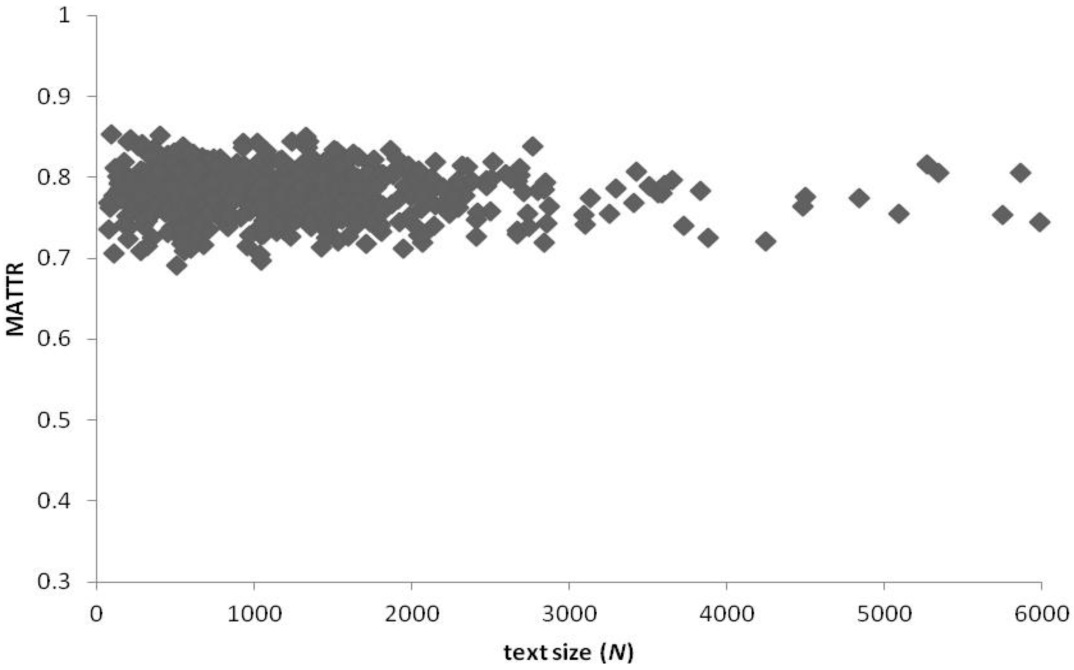


Figure 3. Results of MATTR in 660 texts.

3.2. Results of cumulative development

The results of cumulative measurement in Figures 4 and 5 confirm that MATTR is not influenced by text size.

Although MATTR seems to be a perfect tool for calculating vocabulary richness, one can ask how much the resulting values oscillate in individual windows. In other words, how much does MATTR vary in text development? If there were a big interval within one text, it would be logical to ask whether vocabulary richness measurement of a text makes sense at all. For this purpose, the 19 texts used in the cumulative MATTR were examined. This method is known as a moving window type-token ratio (MWTTR) (Köhler, Gale 1993; Covington, McFall 2010; Kubát, Milička 2013). The results of MWTTR can be seen in Figure 6 (for lack of space only 5 texts were randomly selected as an illustration).

In Figure 6, it can be seen that the values of MATTR oscillate quite a lot between individual parts of the texts. It is therefore questionable whether one average value (MATTR) is a good way of measuring vocabulary richness and at the same time whether vocabulary richness is an appropriate feature of stylometry. Kubát, Milička (2013) proposed a possible solution: the moving type-token ratio distribution (MWTTRD). In fact, MWTTRD is a distribution of all MWTTR values in a text. The results of MWTTRD in the 5 texts used above can be seen in Figure 7.

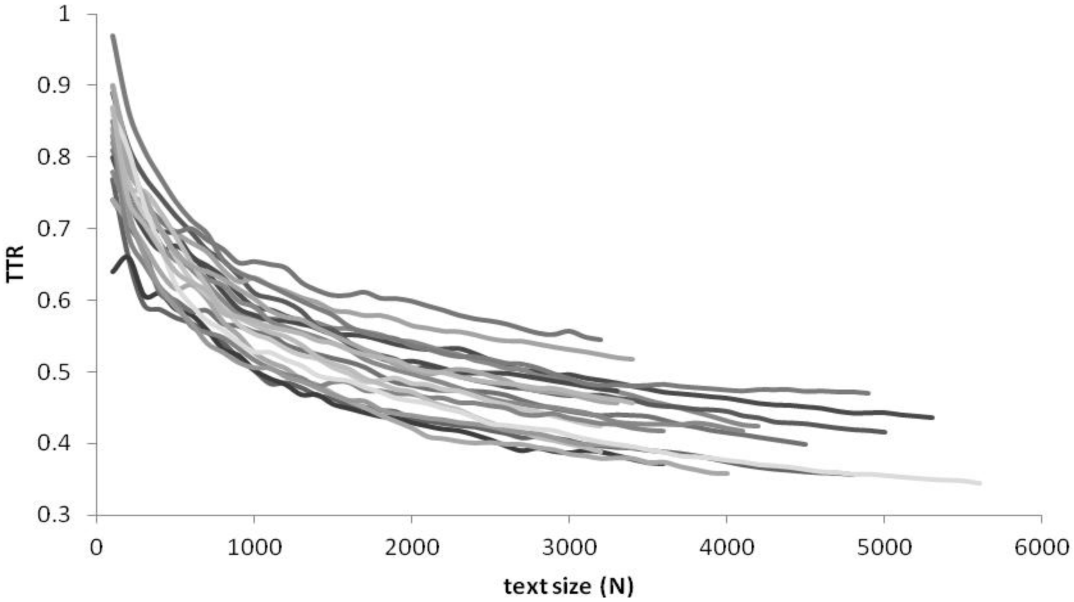


Figure 4. Results of cumulative TTR in 19 texts.

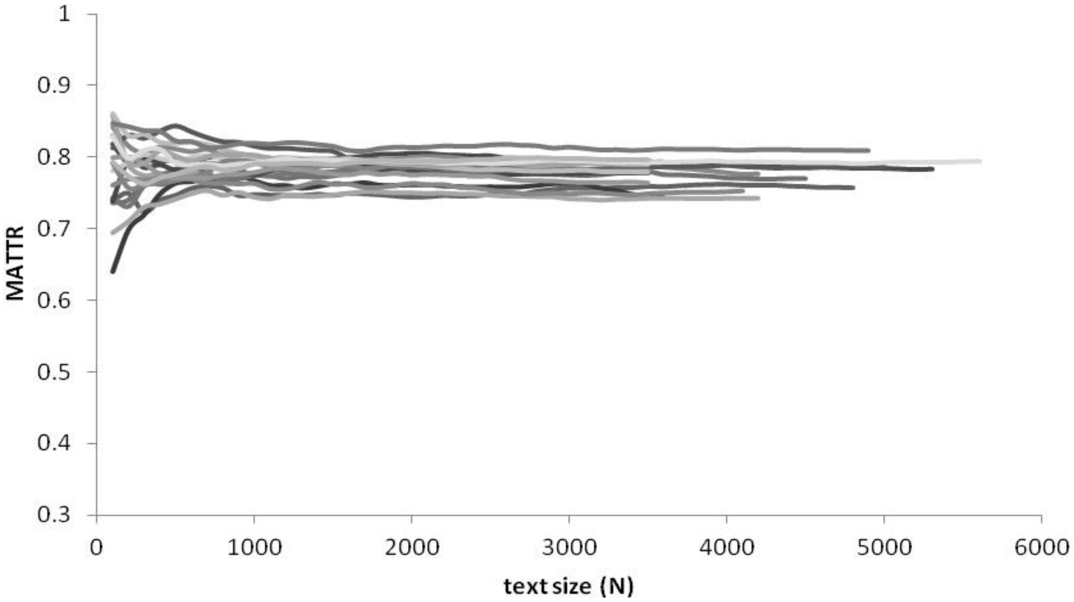
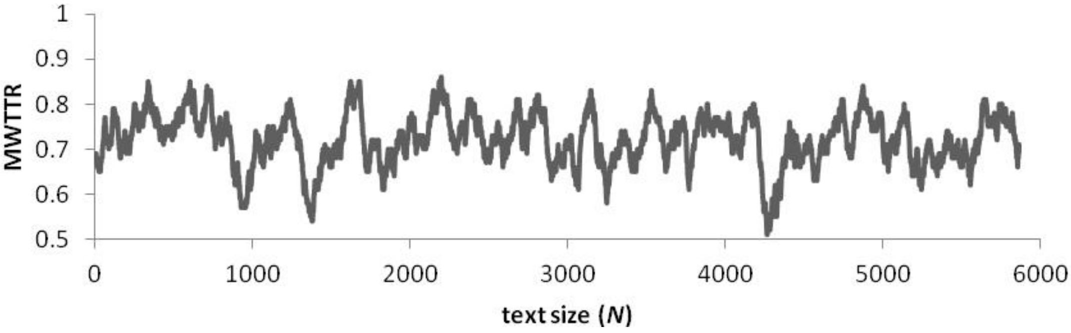
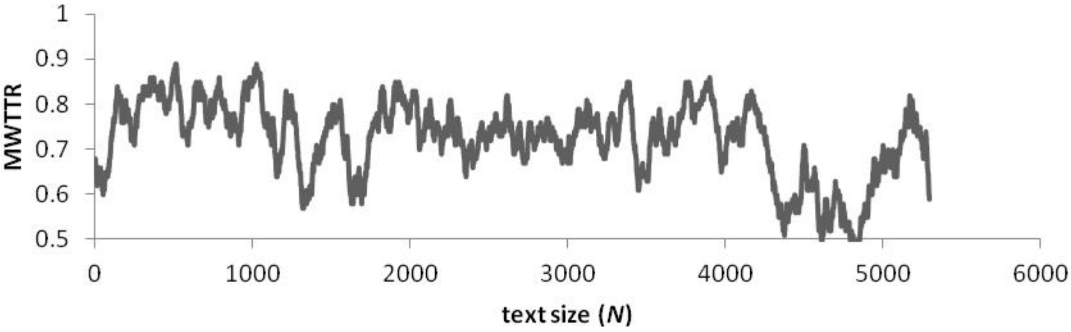
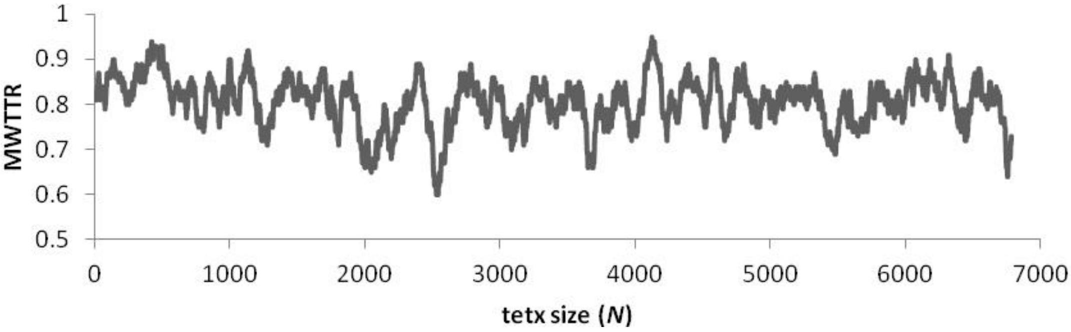
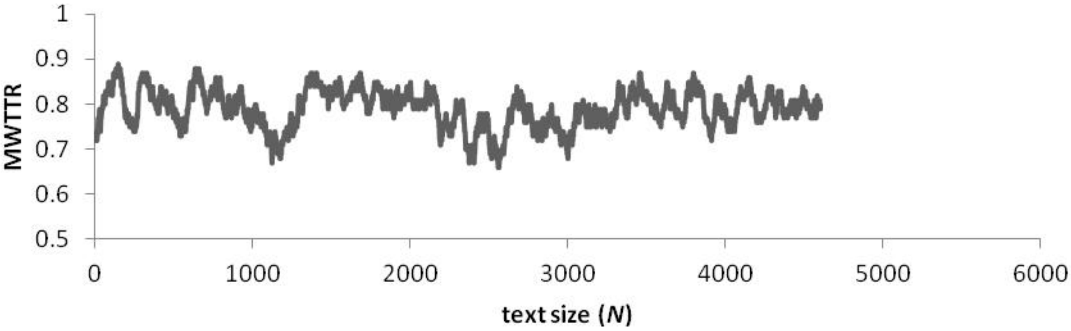


Figure 5. Results of cumulative MATTR in 19 texts.


MWTTR
1
0.9
0.8
0.7
0.6
0.5
0
1000
2000
3000
4000
5000
6000
text size (N)


MWTTR
1
0.9
0.8
0.7
0.6
0.5
0
1000
2000
3000
4000
5000
6000
text size (N)


MWTTR
1
0.9
0.8
0.7
0.6
0.5
0
1000
2000
3000
4000
5000
6000
7000
tetx size (N)


MWTTR
1
0.9
0.8
0.7
0.6
0.5
0
1000
2000
3000
4000
5000
6000
text size (N)

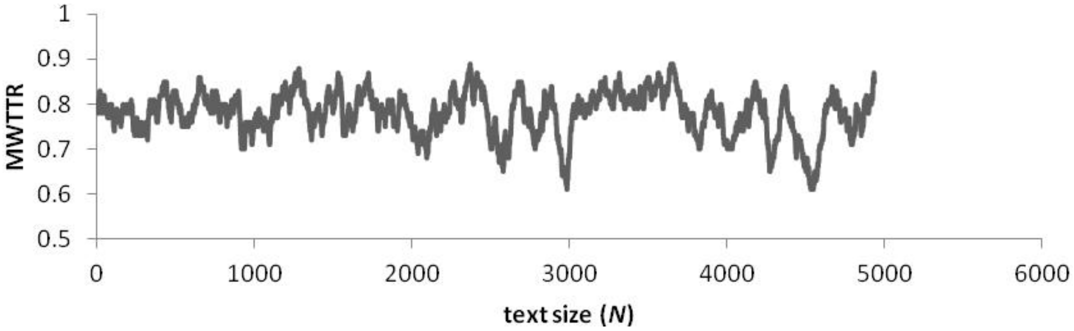


Figure 6. MWTTR in 5 texts.

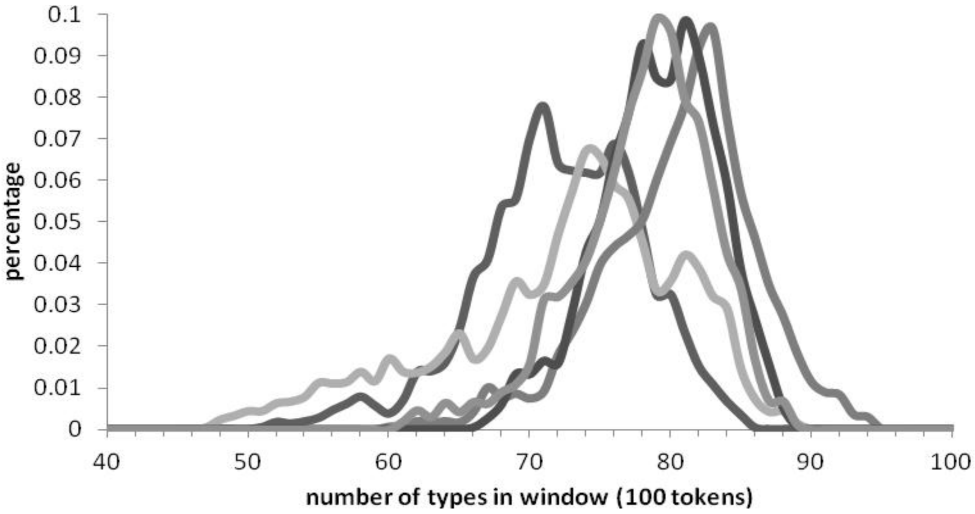


Figure 7. MWTTRD in 5 texts.

In Figure 7, it can be seen that the results of the texts are not distributed in the same way. There are considerable differences indicating specific writing styles. It is obvious that only one average value from the whole distribution can be quite misleading and inaccurate to characterize whole texts. MWTTRD seems to be a much more precise method than MATTR.

4. Conclusion and discussion

The results based on more than six hundred texts show that MATTR is not influenced by text size. Considering how many researchers have tried over several decades to discover an indicator of vocabulary richness that is independent of text length, moving window type-token ratio seems to be an important landmark in the history of quantitative linguistics. Moreover, the method of a moving window could be applied not only to TTR measurement, but also to various indicators such as lambda, repeat rate, thematic concentration, etc.

A disadvantage of the method lies in the fact that the length of the window (L) is arbitrarily chosen, there is no fixed value of L. Each researcher has to choose the size of the window. Although probably the choice depends mainly on the length of the text and on the purpose of the research, it is still arbitrary. Since TTR is influenced by text length, only results with the same window size can be compared.

Given that TTR varies quite a lot in individual windows, the results of MATTR can be misleading. Only one average value seems to be an imprecise feature of a text as a whole. Possible changes between individual parts of a text can be used for the analysis of text development (using MWTTR). MWTTRD is a much more accurate method. On the other hand, MATTR is easier for statistical comparison than MWTTR.

Finally, it must be said that this work is just one attempt to discover whether the vocabulary richness measurement based on a moving window is fully independent of text size and whether it is a suitable feature to characterize the style of a text. Thus, more texts must be analysed to support or reject these preliminary claims.

Acknowledgment

This work was supported by the project Lingvistická a lexikostatistická analýza ve spolupráci lingvistiky, matematiky, biologie a psychologie, grant no. CZ.1.07/2.3.00/20.0161 which is financed by the European Social Fund and the National Budget of the Czech Republic.

References

Covington, M. A., McFall J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics 17(2), 94–100*.

Čech, R. (2015). Text length and the lambda frequency structure of the text. In: Mikros, G., Mačutek, J. (eds.), *Sequences in language and text* (accepted).

Guiraud, P. (1954). *Les catactères stitistiques du vocabulaire*. Paris: Presses Universitaires de France.

Guiraud, P. (1959). *Problèmes et methods de la statistique linguistique*. Dordrecht: Reidel.

Köhler, R., Galle, M. (1993). Dynamic Aspects of Text Characteristics. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative Text Analysis: 46–53*. Trier: WVT.

Kubát, M. (2013). Kvantitativní analýza žánrů v díle Karla Čapka. In *Lingvistika Praha 2013*. [online] WWW: <http://lingvistikapraha.ff.cuni.cz/sbornik>.

Kubát, M., Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics 20(4), 339–349.*

Milička, J. (2013). *MaWaTaTaRaD*. Praha. (Software)

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

Scott, M. (2013). *WordSmith Tools*. Liverpool: Lexical Analysis

Tešitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3, 103–120.*

Tuldava, J. (1977). O kvantitativnych charakteristikach bogatstva leksičeskogo sostava chudožestvennych tekstov. *Acta et Commentationes Universitatis Tartuensis 437*, 159–175.

Tuldava, J. (1995). On the relation between text length and vocabulary size. In: Tuldava, J. (ed.), *Methods in quantitative linguistics: 131–150.* Trier: WVT.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Part-of-speech concentration in Chinese

Lu Wang

1. Introduction

Words are considered to be lexical units with some grammatical formatives (Hřebíček 1992). This paper investigates a specific grammatical aspect of words, viz. its parts of speech. As the meaning of a word can diversify, its grammatical properties can also expand. Words with more than one part of speech (we call them polyfunctional words) exist in many languages, especially in analytic ones such as Chinese. This phenomenon is known also from English, where words such as *run, go* etc can be used as nouns, verbs or adjectives. Here we make a first attempt to study the "dominant" parts of speech (the most frequent part of speech) of polyfunctional words, test the relation with polyfunctionality (the number of parts of speech of a word), measure the degree of concentration and model the distributions. The most frequent part of speech of a given word is called "dominant" just in order to have a practical term, not because of other reasons such as semantic, grammatical, syntactic, contextual, stylistic, functional, systemic or other possible criteria. It would be, of course, worth-while to conduct other studies on the relations between such properties and the ones which we investigate in the present study.

2. Data and Method

2.1. Data

Our data is derived from the *People's Daily* (January 1998) corpus, which is a Chinese one-million word news corpus with word segmentation and part of speech tagging. All the words containing numbers or alphabetic characters are ignored; only words consisting of Chinese characters are taken into account.

2.2. Parts of speech in Chinese

Chinese words are classified into 12 parts of speech as presented in Table 1.The news corpus is processed according to *信息处理用现代汉语分词规范* (Information processing oriented modern Chinese word segmentation criterion). This segmentation system consists of 43 tags, including 12 generalized parts of speech (Table 1), 14 subclasses and 17 other components.

Table 1
Parts of speech in Chinese.

Parts of speech	**Examples**
Noun	书 book, 办法 method,朋友 friend
Verb	是 be, 来 come,相信 believe, 能够 can
Adjective	宽 wide, 猩红 scarlet, 花里胡哨, gaudy
Number	二 two, 几 several, 许多 many, 半百 fifty
Quantifier	厘米 centimeter, 次 time/occurrence, 天 day
Pronoun	我们 we/ us,这 this,哪里 where
Adverb	不 no/ not, 始终 always, 极其 extremely,总共 totally,反正 whatever
Auxiliary	的 used after an attribute to modify a noun; used at the end of a nominal structure to form a noun-phrase equivalent, 了 used after a verb or adjective to indicate the completion of an action or a change, 似的 similar/ as if
Conjunction	和 and, 然而 however, 即使 even if, 或者 or
Preposition	从 from/ since, 关于 about/ concerning, 按照 according to
Interjection	哦 oh, 哎哟 ouch, 好家伙 Good heavens!/ Good lord!
Onomatopoeia	汪 wow,滴答 tick tock, 噼里啪啦 pit-a-pat

We transform the subclasses to their corresponding generalized classes as shown in Table 2 with examples.

Table 2
Corpus tags and parts of speech.

Tag	**Meaning**	**Example**		**Part of speech**
a	adjective	重要/a 步伐/n	important steps	adjective
b	distinguish word	慢性/b 胃炎/n 女/b 司机/n	chronic gastritis female driver	
z	condition or state of affairs	短短/z 几/m 年/q	short years	
ad	adverbial adjective	积极/ad 谋求 /v	positively strive for	
an	nominal adjective	外交/n 和/c 安全/an	diplomacy and safety	
c	conjunction	合作/vn 与/c 伙伴/n	cooperation and fellows	conjunction

d	adverb	逐步/d 得到/v 缓解/vn	alleviate… progressively /step by step	adverb
e	interjection	啊/e，/w 那/r 金灿灿/z 的/u 麦穗/n	ah, the golden ears of wheat	interjection
f	noun of locality (up, down, left, right, north, south, inside, outside…)	军人/n 的/u 眼睛/n 里/f	inside the eyes of solders	noun
s	space/ place	海外/s 侨胞/n	overseas compatriots	
t	time	1997 年/t 3 月/t 19 日/t	March 19th , 1997	
nr	person name	张/nr 仁伟/nr	Mr. Zhang Renwei	
ns	place name	北京市/ns	Beijing	
nt	name entity	[世界/n 贸易/n 组织/n]nt	WTO	
nz	proper noun	诺贝尔奖/nz	Nobel Prize	
n	noun	城市/n	city	
m	number	10/m 公斤/q 一些/m 新/a 问题/n	10 kg some new problems	number
o	onomatopoeia	哈哈哈/o	hahaha	onomatopoeia
p	preposition	关于/p 地震/n 的/u 报告/n	the report on earthquake	preposition
q	quantifier	一/m 年/q	one year	quantifier
r	pronoun	这次/r 演出/v	this performance	pronoun
u	auxiliary	灵魂/n 的/u 共鸣/vn	sympathetic response of the souls	auxiliary
y	modal particle	只不过/d 做/v 了/u 我/r 应该/v 做/v 的/u 事/n 罢了/y。	I just did what I should do.	
v	verb	快速/d 增长/v	grow rapidly	verb
vd	adverbial verb	持续/vd 增长/v	grow continuously	
vn	nominal verb	规范/v 市场/n 管理/vn	standardize/ regulate market management	

The remaining 17 tags are: *h* (prefix component), *i* (idiom), *j* (abbreviation), *k* (suffix component), *l* (phraseology), *nx* (foreign alphabet), *x* (non-morpheme character), *w* (punctuation) and 9 morphemes (*Ag, Bg, Dg, Mg, Ng, Rg, Tg, Vg, Yg*).

2.3. Dominant part of speech

We extract the 1497 polyfunctional words from the corpus, obtain the parts of speech of a given word from the annotations in the corpus and calculate the frequencies of each part of speech separately. Some examples, where the number following a part of speech is the corresponding frequency, are

规定	noun	235	verb	205							
同时	conjunction	393	noun	184	adverb	103					
对	preposition	3697	quantifier	52	verb	45	adjective	6			
连	auxiliary	94	noun	23	adverb	23	verb	20	preposition	7

Obviously, the frequencies of the alternative parts of speech of a word are not distributed uniformly. The most frequent part of speech is called the dominant part of speech of the given word, and is underlined in the examples.

2.4. Repeat rate and relative repeat rate

In order to measure the concentration of the dominant part of speech, we adopt the repeat rate *R*:

$$R = \frac{1}{N^2} \sum_{i=1}^{S} f_i^2 ,$$

where *N* is the sum of the frequencies; *S* is the number of parts of speech; f_i is the frequency of an individual part of speech. The examples in 2.3 yield:

规定 $$R = \frac{235^2 + 205^2}{(235 + 205)^2} = 0.5023 ,$$

同时 $$R = \frac{393^2 + 184^2 + 103^2}{(393 + 184 + 103)^2} = 0.4302 ,$$

对 $$R = \frac{3697^2 + 52^2 + 45^2 + 6^2}{(3697 + 52 + 45 + 6)^2} = 0.9469 ,$$

连 $$R = \frac{94^2 + 23^2 + 23^2 + 20^2 + 7^2}{\left(94 + 23 + 23 + 20 + 7\right)^2} = 0.3709\,.$$

In general, the value of the repeat rate varies in the interval [1/*N*, 1], where 1/*N* stands for maximal dispersion, i.e. each part of speech of the given word has the same frequency of occurrence; and 1 means highest concentration, i.e. a monofunctional word, which is not taken into consideration in our study (Popescu, Altmann 2007). The polyfunctionality of our data varies from 2 to 5, which indicates that the lower limits of the intervals are not equivalent. To unify them, we use the relative repeat rate *RR* (Altmann 1988):

$$RR = \frac{1 - R}{1 - 1/n},$$

where *R* is the repeat rate; *n* is the polyfunctionality. Thus, in spite of varying polyfunctionality of the words, the interval should always be [0, 1]. However, we should distinguish the meaning of *RR* from *R*: 0 stands for the maximal concentration (i.e. monofunctional words), while 1 means that the frequencies of all the parts of speech are equal. The above examples yield:

规定 $$RR = \frac{1 - 0.5023}{1 - 1/2} = 0.9954\,,$$

同时 $$RR = \frac{1 - 0.4302}{1 - 1/3} = 0.8547\,,$$

对 $$RR = \frac{1 - 0.9469}{1 - 1/4} = 0.0708,$$

连 $$RR = \frac{1 - 0.3709}{1 - 1/5} = 0.7864\,.$$

3. Polyfunctionality and concentration

In this part, we study the relationship between polyfunctionality (the number of parts of speech a word carries) and their part of speech concentration. We calculate *RR* of the polyfunctional words, group the words with different polyfunctionality, and obtain the mean *RR* of each group. There is only one word*连* which carries the maximum of five parts of speech. We exclude its datum to avoid randomness. The data are presented in Table 3, where *x*[*i*] stands for polyfunctionality; and *f*[*i*] is the mean value of *RR*. As mentioned in 2.4, a smaller *RR* is correlated with higher concentration. Here, *RR* decreases with increasing polyfunctionality. Therefore, we can draw the conclusion: the more parts of speech a word can be used in, the stronger its frequency concentrates on one part of speech (on the average). As many linguistic properties are related in the form of a power function, we apply this model to our data and show the result in Table 3 and Figure 1. Although the goodness-of-fit indicates the success of modeling, at the present stage of research we did not make any conclusion, because (1) mathematically, it is very possible for a function with two parameters to fit 3 data points; (2) linguistically, there are no data from other languages to show that the power function is a widely applicable model. We will therefore consider the function just as a way to present the findings.

Table 3
Fitting a power function to the data.

x[i]	**f[i]**	**NP[i]**
2	0.6247	0.6194
3	0.5561	0.5698
4	0.5455	0.5369
a = 0.7146, b = -0.2061, R^2 = 0.9223		

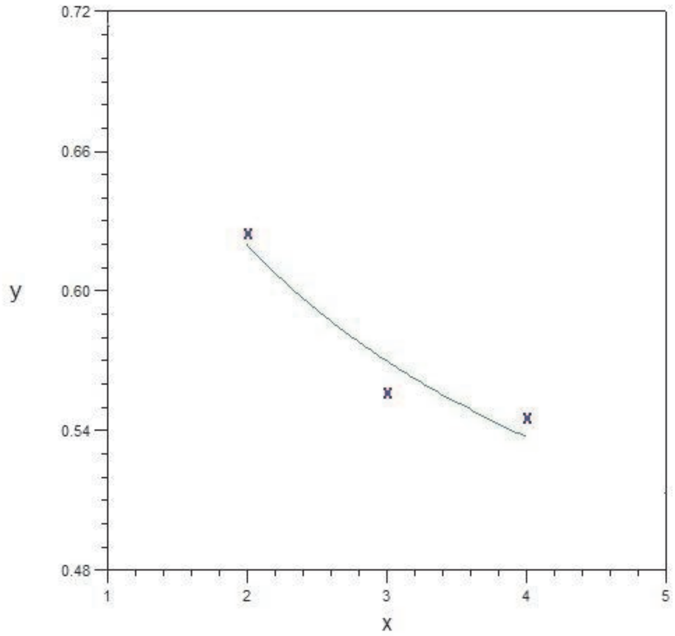


Figure 1. Fitting a power function to the data.

4. Part of speech concentration

First, we determine the dominant part of speech and calculate *RR* of all the polyfunctional words. Words with less than 10 occurrences and dominant parts of speech with less than 10 observations are not taken into consideration because of their statistical unreliability. Then, we calculate the mean value of *RR* for each dominant part of speech type. The results are listed in Table 4, arranged according to increasing *RR* (decreasing concentration) values. Interjection never plays the role of the dominant part of speech. Onomatopoeias were observed too rarely.

Table 4
RR of the dominant part of speech.

Dominant part of speech	*RR*
number	0.3814
pronoun	0.4176
auxiliary	0.4458
conjunction	0.4663
verb	0.4930
adjective	0.5097
preposition	0.5146
noun	0.5442
quantifier	0.5637
adverb	0.5744

Considering the examples in 2.3, *规定*,*同时* and *连* all belong to the noun category. We call them polyfunctional nouns. But their dominant parts of speech are not necessarily nouns; actually the dominant parts of speech can be noun, preposition and auxiliary respectively. Therefore, a question is raised: on which parts of speech and to what degree do polyfunctional nouns concentrate? To answer this question, we obtain the dominant parts of speech of polyfunctional nouns and calculate the mean *RR* of each dominant part of speech. The result is demonstrated in Table 5, as well as the other parts of speech. Unreliable data, as mentioned above, are omitted. Number, pronoun, auxiliary and conjunction concentrated only on themselves. Moreover, they are also the most concentrated parts of speech as shown in Table 4. Adjective has 3dominant parts of speech, but itself remains the highest concentrated class. Verb, noun, adverb, quantifier and preposition have more than one dominant part of speech and do not concentrate on themselves. If a polyfunctional preposition (or a polyfunctional quantifier) has also the function of the verb, the frequency of the verb will be higher. A polyfunctional verb, which belongs also to adverb, will lose its dominant position. However, adverb itself is not stable enough and will be less frequent than conjunction. Thus, we can draw a hierarchy from this concentration stratum (Figure 2). The arrow stands for the direction of stability. The part of speech on the left side is less frequent, while the one on the right side is more strongly concentrated. The part of speech with a cyclic arrow means that the category itself is the most dominating part of speech.

Table 5
Dominant parts of speech and *RR*.

Part of speech	**Dominant part of speech**	**RR**
number	number	0.3814
pronoun	pronoun	0.4176
auxiliary	auxiliary	0.4458
conjunction	conjunction	0.4663
preposition	verb preposition	0.5026 0.5146
quantifier	verb noun quantifier	0.4893 0.5171 0.5637
adjective	adjective noun verb adverb	0.5097 0.5848 0.5928 0.6336
adverb	conjunction verb adverb adjective noun	0.4348 0.4530 0.5744 0.6069 0.6616
noun	adjective verb noun adverb quantifier	0.4167 0.4840 0.5442 0.5585 0.5854
verb	adverb verb preposition noun adjective quantifier	0.4417 0.4930 0.5379 0.5418 0.5499 0.5762

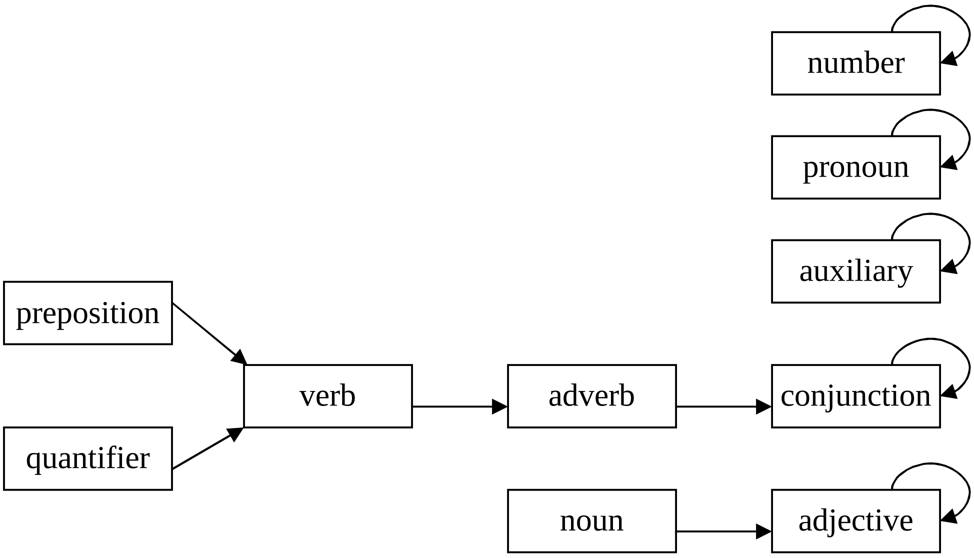


Figure 2. Hierarchy of the stability of concentration.

5. Dominant part of speech distribution

This part investigates the dominant part of speech distribution of each part of speech group. The problem of the data is: some words may have the same frequency for more than one part of speech as shown below:

一贯	adverb	19	adjective	19		
偏	adverb	29	adjective	29	verb	3
极端	noun	7	adverb	7	adjective	3

Consequently, we obtain the dominant parts of speech of adjectives as shown in the following list on the left side. As both parts of speech have the same chance to be the dominant part of speech, the frequency is evenly distributed to both parts of speech as shown by the list on the right side.

adjective	218		adjective	245.5
verb	104		verb	111.5
adverb	75		noun	89
noun	74		adverb	81.5
adjective / noun	**29**		number	3
adjective / verb	**14**		auxiliary	1
adjective / adverb	**11**		preposition	1
number	3		conjunction	1

auxiliary	1		onomatopoeia	0.5
preposition	1			
conjunction	1			
adverb / verb	**1**			
adverb / noun	**1**			
adjective / onomatopoeia	**1**			

To capture the distributions of all parts of speech (except for interjection which has only two data points), we found that the Popescu-Altmann function, which is a special case of the general model presented in Wimmer and Altmann (2005), is the best model. It is defined as

$$f(r) = 1 + ae^{-br}\ .$$

The constant 1 in the Popescu-Altmann function was added because there is no smaller frequency (Popescu, Altmann, Köhler 2009) than one. However, the smallest frequency in our data is 0.5, in the case where the frequency was distributed on two categories. We change the constant accordingly:

$$f(r) = 0.5 + ae^{-br}\ .$$

The fitting results are shown in Table 6 and Figure 3-13.

6. Conclusion

This study reveals three characteristics of polyfunctional words. First, the more parts of speech a word belongs to, the stronger it concentrates on one of the parts of speech. The relationship between polyfunctionality and relative repeat rate seems to abide by a power law. Second, part of speech concentration forms a hierarchy. Number, pronoun, auxiliary, conjunction and adjective are dominatingly concentrated on themselves. Verb and adverb are in the middle part, superior to preposition and quantifier but inferior to conjunction. Noun is second to adjective. Third, the dominant part of speech distributions of all the parts of speech (except for interjection which has only two data points) can be modeled by the Popescu-Altmann function.

Table 6
Fit of the Popescu-Altmann function to the data.

	verb			**noun**		
x[i]	dominant part of speech	f[i]	NP[i]	dominant part of speech	f[i]	NP[i]
1	verb	480	483.78	noun	405	428.76
2	noun	250.5	236.93	verb	317.5	245.79
3	adjective	112.5	116.17	adjective	100.5	141
4	adverb	36.5	57.09	quantifier	63	80.97
5	quantifier	28.5	28.18	adverb	53.5	46.59
6	preposition	26	14.04	number	9.5	26.9
7	conjunction	10	7.12	preposition	9	15.62
8	auxiliary	9	3.74	pronoun	6	9.16
9	number	7	2.08	auxiliary	5	5.46
10	pronoun	1	1.27	conjunction	2	3.34
11	onomatopoeia	1	0.87			
	a = 987.8559 b = 0.7149 R^2 = 0.9963			a = 747.7034 b = 0.5572 R^2 = 0.957		

Table 6 (continued).

	adverb			**quantifier**		
x[i]	dominant part of speech	f[i]	NP[i]	dominant part of speech	f[i]	NP[i]
1	adverb	170.5	167.25	quantifier	82.5	85.41
2	adjective	66.5	75.68	noun	52	46.67
3	noun	30.5	34.39	verb	33.5	25.61
4	verb	29	15.78	auxiliary	3	14.15
5	conjunction	15.5	7.38	number	3	7.92
6	number	9	3.6	adverb	3	4.53
7	preposition	5	1.9	preposition	2	2.69
8	auxiliary	3	1.13	adjective	2	1.69
9	quantifier	3	0.78	pronoun	1	1.14
10	pronoun	2	0.62			
	a = 369.8702 b = 0.7966 R^2 = 0.9837			a = 156.143 b = 0.6091 R^2 = 0.9641		

Table 6 (continued).

	adjective			**auxiliary**		
x[i]	dominant part of speech	f[i]	NP[i]	dominant part of speech	f[i]	NP[i]
1	adjective	245.5	239.4	auxiliary	16	15.93
2	verb	111.5	135.45	pronoun	3	3.56
3	noun	89	76.73	verb	2	1.11
4	adverb	81.5	43.56	quantifier	2	0.62
5	number	3	24.82	adverb	2	0.52
6	auxiliary	1	14.23	preposition	1	0.5
7	preposition	1	8.26	adjective	1	0.5
8	conjunction	1	4.88			
9	onomatopoeia	0.5	2.97			
	a = 422.9398 b = 0.5711 $R^2 = 0.9474$			a = 77.6453 b = 1.6154 $R^2 = 0.9674$		

Table 6 (continued).

	Pronoun			**preposition**		
x[i]	dominant part of speech	f[i]	NP[i]	dominant part of speech	f[i]	NP[i]
1	pronoun	13	12.84	preposition	32	32.52
2	adverb	4	4.66	verb	17	14.59
3	auxiliary	2	1.9	adverb	4	6.69
4	noun	2	0.97	conjunction	3	3.22
5	conjunction	2	0.65	noun	2	1.69
6	verb	1	0.55	auxiliary	1	1.02
7	preposition	1	0.51			
	a = 36.6267 b = 1.0873 $R^2 = 0.9658$			a = 72.8048 b = 0.8211 $R^2 = 0.9823$		

Table 6 (continued).

	conjunction			**number**		
x[i]	dominant part of speech	f[i]	NP[i]	dominant part of speech	f[i]	NP[i]
1	conjunction	27.5	27.66	number	23.5	23.72
2	adverb	15.5	15.64	noun	10.5	9.25
3	verb	11	8.94	verb	2	3.8
4	pronoun	3	5.21	adverb	2	1.74
5	preposition	3	3.12	preposition	1	0.96
6	noun	2	1.96			
7	conjunction	27.5	27.66			
	a = 48.6945 b = 0.5838 $R^2 = 0.9816$			a = 61.6064 b = 0.9754 $R^2 = 0.9866$		

Table 6 (continued).

onomatopoeia		
dominant part of speech	f[i]	NP[i]
verb	2	2.06
onomatopoeia	1.5	1.34
noun	1	0.95
adjective	0.5	0.74
a = 2.9030, b = 0.6198, $R^2 = 0.9274$		

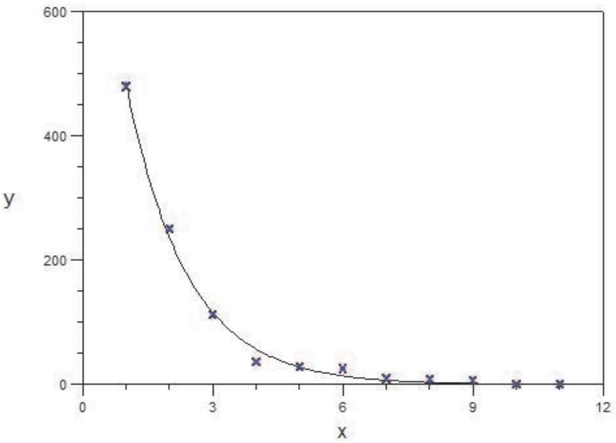


Figure 3. Fit of the Popescu-Altmann function to the data of verb.

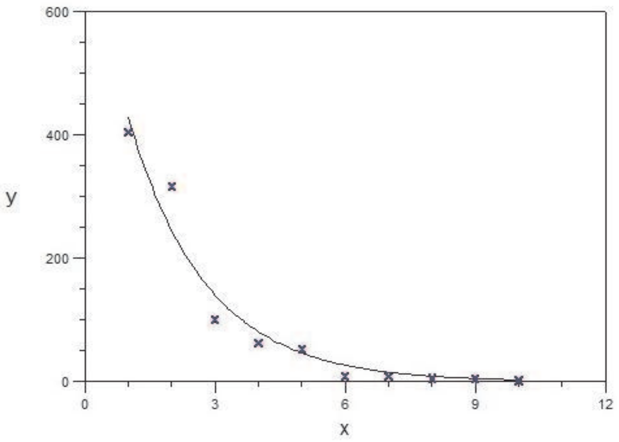


Figure 4. Fit of the Popescu-Altmann function to the data of noun.

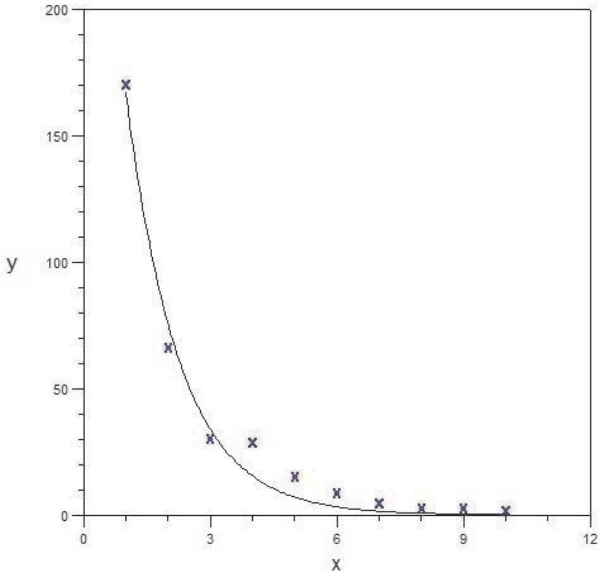


Figure 5. Fit of the Popescu-Altmann function to the data of adverb.

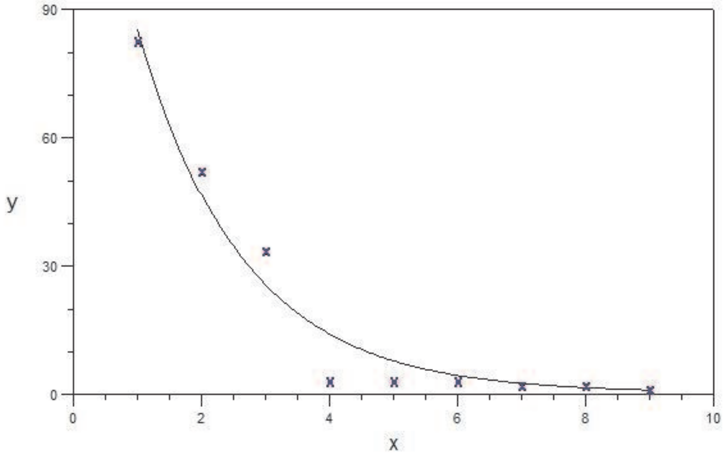


Figure 6. Fit of the Popescu-Altmann function to the data of quantifier.

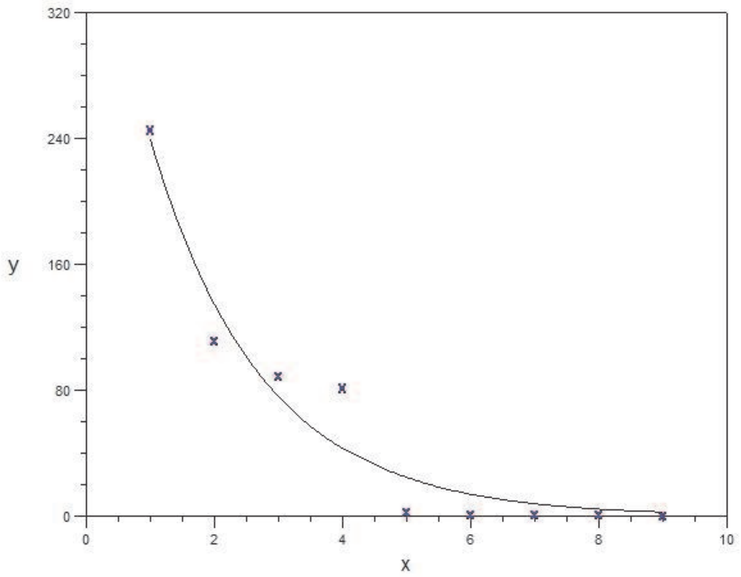


Figure 7. Fit of the Popescu-Altmann function to the data of adjective.

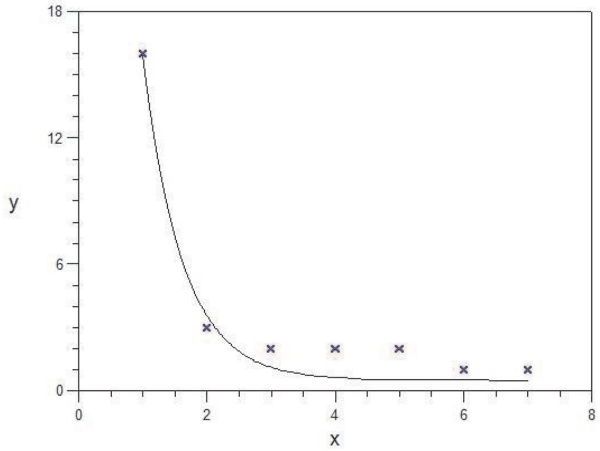


Figure 8. Fit of the Popescu-Altmann function to the data of auxiliary.

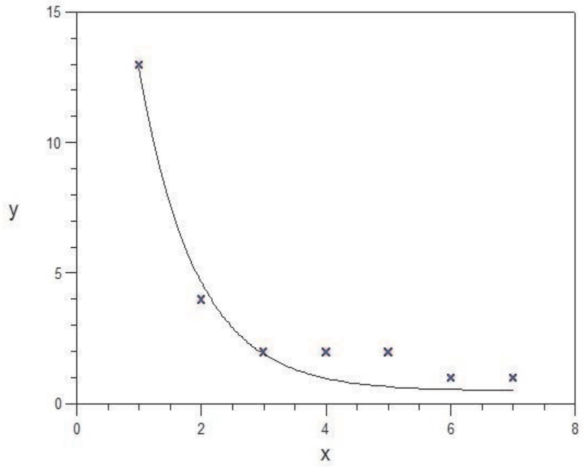


Figure 9. Fit of the Popescu-Altmann function to the data of pronoun.

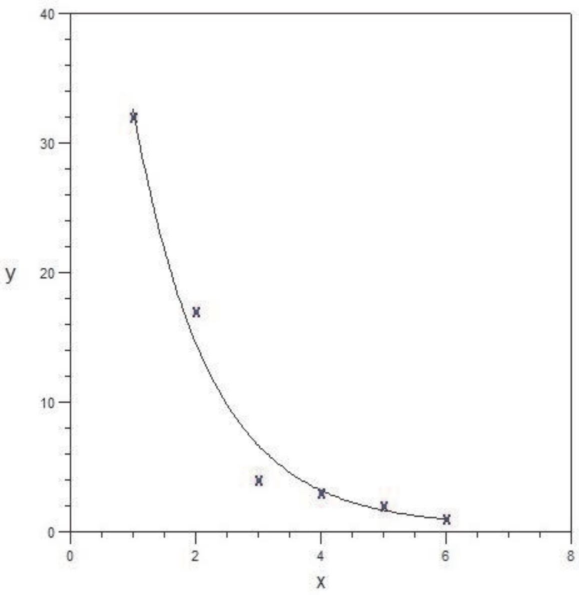


Figure 10. Fit of the Popescu-Altmann function to the data of preposition.

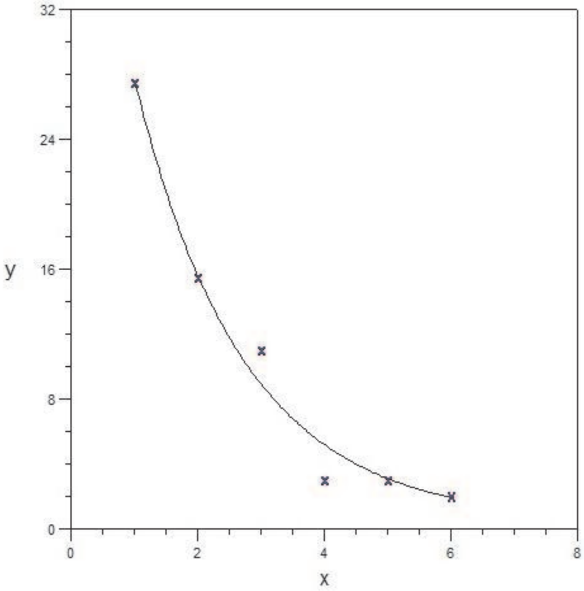


Figure 11. Fit of the Popescu-Altmann function to the data of conjunction.

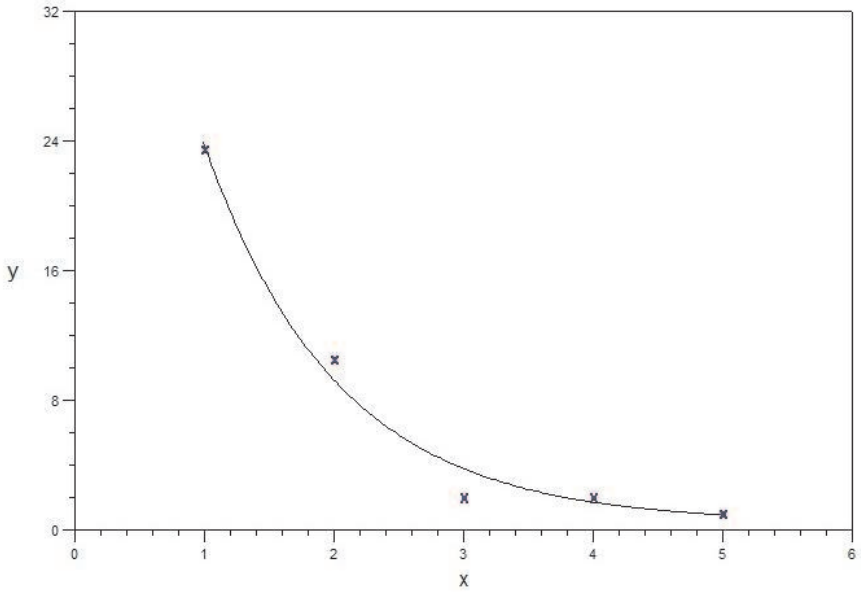


Figure 12. Fit of the Popescu-Altmann function to the data of number.

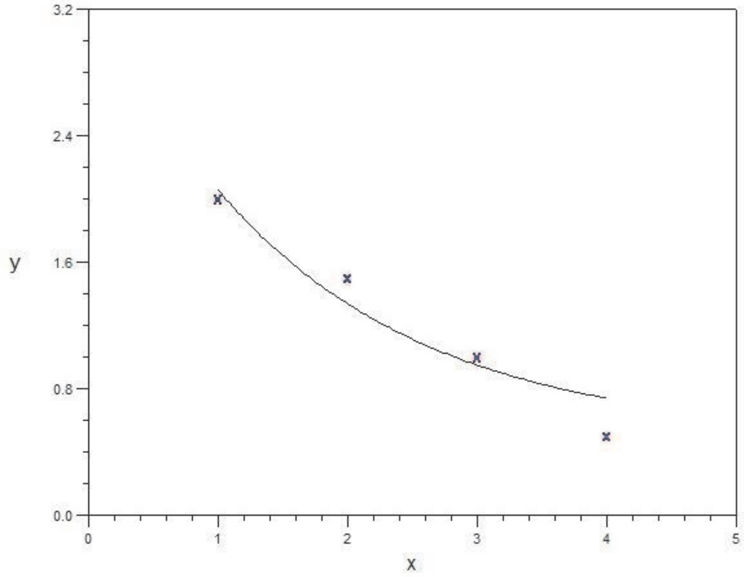


Figure 13. Fit of the Popescu-Altmann function to the data of onomatopoeia.

Acknowledgements

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

References

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

Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures.* Bochum: Brockmeyer.

Popescu, I-I., Altmann, G. (2007). On diversify of word frequencies and language typology. *Göttinger Beiträge zur Sprachwissenschaft 14, 83–91*.

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

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

A measure of lexical text compactness

Ján Mačutek, Gejza Wimmer

1. Introduction

This contribution presents a new text property which will be called lexical text compactness[1] (*LTC* hereafter). Two sentences in a text are considered linked if they contain the same word lemmas (in this paper, only content words: nouns, adjectives, verbs and adverbials, are taken into account[2]). We do not consider the numbers of the lemmas which occur in both of the two sentences – the sentences are either linked (if there is at least one lemma which occurs in both of them) or not (if such a lemma does not exist). The *LTC* is defined (cf. Section 2 for the formal definition) as the proportion of the number of pairs of linked sentences to the total number of pairs of sentences in a text – i.e., the more links among sentences, the higher the lexical compactness of a text. It will be shown that the number of the links can be modelled by the binomial distribution. A test for comparing *LTC*s in two texts will be introduced in Section 4.

2. Mathematical model

If a text consists of n sentences, there are $\binom{n}{2}$ pairs of sentences. Denote L the number of linked sentences (i.e., the number of pairs of sentences which contain the same noun, adjective, verb or adverbial, whereby word lemmas are taken into account, cf. Section 1). Then we define the lexical text compactness as

$$LTC = \frac{L}{\binom{n}{2}}. \tag{1}$$

L is an integer with possible values between 0 (no link) and $\binom{n}{2}$ (all sentences are mutually linked). The values can be considered random.

[1] We note that Popescu and Altmann (2008a) used the notion of text compactness, but they considered different properties of texts.

[2] Different parts of speech constitute different word lemmas in our paper, e.g., *pekný* (nice) and *pekne* (nicely) are two lemmas, not one. Of course, other approaches (i.e., merging all or some parts of speech into one lemma) are possible.

It is important to realize that the property “to be linked” is not transitive (i.e., if sentence A is linked with sentence B and sentence B with sentence C, sentences A and C do not have to be linked). We present a simple artificial example: let us have sentence A with words W and X, sentence B with words X and Y, and finally let sentence C consist of words Y and Z. Let there be no other words in the three sentences. Then, sentences A and B both contain word X, B and C both contain word Y, hence we have two pairs of linked sentences. On the other hand, sentences A and C are not linked.

Based on the non-transitivity justified above, (non-)existence of links between pairs of sentences can be considered independent of each other. For the sake of simplicity, let us suppose that probabilities of links do not differ throughout the text. Denote the one common probability of the link p. Under these conditions, the random variable L follows the binomial distribution with parameters $\binom{n}{2}$ and p, i.e.,

$$P(L=k)=\binom{\binom{n}{2}}{k}p^k(1-p)^{\binom{n}{2}-k}, \qquad k=0,1,\ldots,\binom{n}{2}. \tag{2}$$

We note that the assumption (or, at the very least, the possibility) of independence is crucial for the binomial distribution, which models the number of successes[3] in a sequence of independent binary experiments. As soon as links are transitive, independence is out of the question and the binomial distribution cannot be used as a model for the number of links.

3. Methodology and data

We will use two short Slovak journalistic texts (focused on football – a player briefly describes a match, and economy – investment policy in the Russian Far East, respectively) from the Slovak newspaper SME (www.sme.sk) as examples. We denote sentences S1,…, S13 those concerning sport and E1,…, E10 those concerning economy.

Text 1 (sport):

S1 Slovenský futbalový reprezentant Marek Hamšík asistoval pri oboch góloch Neapola v zápase 24. kola talianskej Serie A na pôde Sassuola (2:0).

[3] “Success” as a result of an experiment has a very broad meaning here, e.g., existence of a link between two sentences can be interpreted as a success.

S2 Zverenci trénera Rafaela Beniteza sa rehabilitovali za nečakanú jesennú domácu remízu 1:1 s nováčikom súťaže a v tabuľke sú tretí o šesť bodov pred Fiorentinou.
S3 K nováčikovi súťaže sme cestovali s jednoznačným cieľom získať tri body.
S4 Vedeli sme, že nečaká nás jednoduchý duel.
S5 Veď stačí si spomenúť na náš jesenný vzájomný duel, v ktorom sme proti Sassuolo v domácom prostredí iba remizovali.
S6 K zápasu sme pristúpili zodpovedne a nakoniec sme zaslúžene vyhrali.
S7 Dominovali sme hlavne v prvom dejstve.
S8 Získali sme tri body a zostali sme v tesnom závese za Rímom.
S9 Navyše Fiorentina prehrala, takže náš náskok pred štvrtými fialkami sa zvýšil, povedal pre svoju oficiálnu webovú stránku Hamšík, ktorý hral do 80. minúty a talianske média ho zaradili medzi najlepších hráčov na ihrisku.
S10 Hamšík priznal, že postupne sa dostáva späť do formy, akú mal pred zranením a nútenou pauzou.
S11 Boli to ťažké týždne.
S12 Mimo som bol dva mesiace, čo bolo najdlhšie počas mojej kariéry.
S13 Najdôležitejšie je, že v súčasnosti sa už cítim dobre, dodal slovenský stredopoliar.

Text 2 (economy)

E1 Ruský Ďaleký východ potrebuje do konca roka 2022 investície za približne 3 bilióny rubľov (62,30 miliardy eur), ak si región chce udržať 6% rast.
E2 Uviedol to minulý týždeň ruský minister pre rozvoj Ďalekého východu Alexander Galuška.
E3 Informovala o tom agentúra RIA Novosti.
E4 Ako povedal poslancom Štátnej dumy, dolnej komory ruského parlamentu, región takýto objem investícií nevyhnutne potrebuje, aby sa vytvorili vhodné podmienky pre investorov a prilákalo sa viac ľudí do oblasti.
E5 Očakávame, že každý rubeľ zo štátneho rozpočtu investovaný na Ďalekom východe umožní prilákať ďalších 20 rubľov súkromných investícií, povedal Galuška.
E6 Práve rozvoj Ďalekého východu sa stal jednou z priorít Kremľa v ostatných rokoch.
E7 Prezident Vladimir Putin vytvoril Ministerstvo pre rozvoj Ďalekého východu na začiatku jeho tretieho prezidentského obdobia v roku 2012.

E8 Minulý rok vláda schválila investičný program pre región, pričom pre východnú Sibír a Ďaleký východ vyčlenila každoročne minimálne 100 miliárd rubľov.

E9 Rozsiahly región Ďalekého východu, ktorý sa rozprestiera od Sibíri po Tichý oceán, už roky trpí v dôsledku zastaranej infraštruktúry, geografickej izolácie, extrémneho počasia a úbytku obyvateľstva.

E10 Ekonomiku nepodporujú ani viaceré tamojšie podniky ešte z éry Sovietskeho zväzu, ktoré nemajú šancu súperiť so zahraničnou konkurenciou.

Below we provide a list of linked sentences in the first text together with the words which link them. If two sentences share more than one lemma, only one of them is presented. In the parentheses, one finds an English translation, and, if a word form (in italics) differs from the respective lemma, also the grammatical number, gender, case or tense are specified[4]. We use the following abbreviations:

grammatical number	sing. = singular, pl. = plural,
grammatical gender	masc. = masculine, fem. = feminine,

cases	N = nominative, G = genitive, D = dative,
	A = accusative, L = locative, I = instrumental.

S1 – S5	Sassuolo (an Italian football team)
S1 – S6	zápas (match; *zápase* - sing. L, *zápasu* - sing. D)
S1 – S9	Hamšík (surname of the player)
S1 – S10	Hamšík (surname of the player)
S1 – S13	slovenský (Slovak)
S2 – S3	nováčik (newcomer, here means the team promoted from a lower league; *nováčikom* sing. I, *nováčikovi* sing. D)
S2 – S5	jesenný (autumn as adjective; *jesennú* sing. fem. A, *jesenný* sing. masc. A)
S2 – S8	bod (point; *bodov* pl. G, *body* pl. A)
S2 – S9	Fiorentina (an Italian football team; *Fiorentinou* sing. I, *Fiorentina* sing. N)
S3 – S8	získať (gain; *získať* infinitive, *získali* first person, pl., past tense)
S4 – S5	duel (duel, here meaning match)

Next, we present an analogous list of pairs of sentences which are linked in the second text.

[4] There are further grammatical categories in the Slovak language which are not considered in our examples.

E1 – E2	ruský (Russian)
E1 – E4	ruský (Russian; *ruský* sing. N, *ruského* sing. G)
E1 – E5	Ďaleký (far; *Ďaleký* sing. N, *Ďalekom* sing. L)
E1 – E6	Ďaleký (far; *Ďaleký* sing. N, *Ďalekého* sing. G)
E1 – E7	Ďaleký (far; *Ďaleký* sing. N, *Ďalekého* sing. G)
E1 – E8	Ďaleký (far)
E1 – E9	Ďaleký (far; *Ďaleký* sing. N, *Ďalekého* sing. G)
E2 – E4	ruský (Russian; *ruský* sing. N, *ruského* sing. G)
E2 – E5	Ďaleký (far; *Ďalekého* sing. N, *Ďalekom* sing. L)
E2 – E6	rozvoj (development)
E2 – E7	rozvoj (development)
E2 – E8	minulý (past as adjective)
E2 – E9	Ďaleký (far; *Ďalekého* sing. G)
E4 – E5	povedať (say; *povedal* third person, sing., past tense)
E4 – E7	vytvoriť (create; *vytvorili* third person, pl., past tense, *vytvoril* third person, sing., past tense)
E4 – E8	región (region)
E4 – E9	región (region)
E5 – E6	Ďaleký (far; *Ďalekom* sing. L, *Ďalekého* sing. G)
E5 – E7	Ďaleký (far; *Ďalekom* sing. L, *Ďalekého* sing. G)
E5 – E8	rubeľ (ruble; *rubeľ* sing. N, *rubľov* pl. G)
E5 – E9	Ďaleký (far; *Ďalekom* sing. L, *Ďalekého* sing. G)
E6 – E7	rozvoj (development)
E6 – E8	Ďaleký (far; *Ďalekého* sing. G, *Ďaleký* sing. A)
E6 – E9	Ďaleký (far; *Ďalekého* sing. G)
E7 – E8	Ďaleký (far; *Ďalekého* sing. G, *Ďaleký* sing. A)
E7 – E9	Ďaleký (far; *Ďalekého* sing. G)
E8 – E9	Ďaleký (far; *Ďaleký* sing. N, *Ďalekého* sing. G)

We thus have $n_1 = 13$ sentences and $L_1 = 11$ links in the first text, which results in the lexical text compactness (cf. (1)) $LTC_1 = \frac{11}{\binom{13}{2}} = \frac{11}{78} = 0.141$; in the other one there are $n_2 = 10$ sentences and $L_2 = 27$ links, with $LTC_2 = \frac{27}{\binom{10}{2}} = \frac{27}{45} = 0.6$.

4. Comparing LTCs in two texts

Within the context of this paper, to test the significance of the difference between *LTC*s of two texts means to test whether parameters p_1 and p_2 in (2) are significantly different. Testing the null hypothesis $p_1 = p_2$ is the same as constructing confidence intervals for each of the parameters. Then, if two intervals are disjoint (i.e., they contain no common points), the null hypothesis is rejected. Since we need two simultaneous two-sided confidence intervals with a common significance level α, we apply the Bonferroni correction (cf. Miller 1981). We divide α between the intervals (i.e., each of the two intervals will be constructed at the significance level $\frac{\alpha}{2}$), which leads to the $\frac{\alpha}{4}$-quantiles in the formulas below.

Many tests and confidence intervals for parameters of binomial distributions are based on different normal approximations (cf. Agresti 2013: 13-16; Johnson et al. 2005: 132-133). We only note that if *n* is very large, one can use, e.g., the interval

$$LTC \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{LTC(1-LTC)}{n}},$$

where $z_{1-\frac{\alpha}{2}}$ is the quantile of the standard normal distribution[5].

As our texts are short, we prefer another approach. For the binomial distribution (as well as for other discrete distributions), it is in general not possible to find an exact confidence interval at a given significance level $\frac{\alpha}{2}$. Approximate lower and upper limits (denoted hereafter p_{LL} and p_{UL}, respectively) for such an interval can be found (cf. Johnson et al. 2005: 130-131) as solutions of the equations

$$\text{(3)} \qquad \begin{aligned} \sum_{j=L}^{n} \binom{n}{j} p_{LL}^{j} (1-p_{LL})^{n-j} &= \frac{\alpha}{4}, \\ \sum_{j=0}^{L} \binom{n}{j} p_{UL}^{j} (1-p_{UL})^{n-j} &= \frac{\alpha}{4}, \end{aligned}$$

where *L* is the number of successes (in our case, the number of links in a text). The equations must be solved by computer[6]. The interval given by p_{LL} and p_{UL} from (3) is known as the Clopper-Pearson interval.

If we choose $\alpha = 0.05$ and apply (3) to the two texts from Section 3, we obtain

[5] If one needs two simultaneous confidence intervals, $z_{1-\frac{\alpha}{2}}$ must be replaced with $z_{1-\frac{\alpha}{4}}$.

[6] Statistical software R contains the command binom.test, which gives as output, among others, the limits of the interval.

$$p_1 \in (0.065; 0.253),$$
(4)
$$p_2 \in (0.423; 0.760),$$

which means that we reject the null hypothesis $p_1 = p_2$ (because the two intervals from (4) have no common points). Consequently, it can be claimed that the second text is, with respect to (1), lexically more compact than the first one.

5. Conclusion

We presented a measure of lexical text compactness, according to which a text is the more compact the more sentences share common words. The binomial distribution serves as a mathematical model. Two approaches (both using confidence intervals) to testing the difference between *LTC*s of two texts are discussed. The normal approximation of the binomial distribution is reliable for long texts (i.e., texts consisting of a large number of sentences). For short texts, the limits of the confidence intervals can be computed directly from formulas for the binomial distribution.

We emphasize once more that the binomial distribution assumes independence of the "experiments", which here means that only non-transitive properties (cf. Section 2) can be modelled in this way. The methodology introduced in this paper can be applied, e.g., if links are established among sentences with words from the same hrebs (Hřebíček 1997; Ziegler, Altmann 2002, Wimmer et al. 2003), or, if hrebs are linked when they contain the same words.

In our approach, two sentences are linked if they contain the same content word, which means that only words which occur in the text at least twice are considered. Hence, this paper complements, in a certain sense, the research focused on hapax legomena (e.g., Popescu, Altmann 2008b), which are irrelevant here.

Our methodology can be generalized or modified in several directions. Naturally, generalizations or modifications will require new mathematical models.

First, links can be defined not only among sentences, but also among other language units. It can be, e.g., more reasonable to link verses or stanzas in poetic texts, which can lack a sentence structure.

Second, we limited ourselves to binary modeling (i.e., either there is a link or not). It is possible to define multiple links if sentences (verses, stanzas,…) share more than one word.

Next, the distance between linked units can be taken into account. One can introduce weights of the links (e.g., the weights can be given by the difference of the sentence positions in the text – a link between the second and the fifth sentence would have weight 5 – 2 = 3).

Finally, text compactness (defined not only by sentences and words, but also by other units) is not an isolated property, but it can be embedded into the synergetic language model (cf. Köhler 2005). Tentatively, we allow ourselves to formulate three hypotheses: 1) the longer the sentence the more links it has; 2) sentences at the beginning of a text have more links than sentences at the end; 3) the higher the mean sentence length (in words) in a text, the higher the lexical text compactness.

Acknowledgments

The work was supported by the research grants ESF OPVK 2.3 - Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology (CZ.1 07/2.3.00/20.0161) (J. Mačutek), VEGA 2/0038/12 (G. Wimmer), and by the Slovak Research and Development Agency, grants APVV-0096-10, SK-AT-0025-12 (G. Wimmer).

References

Agresti, A. (2013). *Categorial Data Analysis.* Hoboken (NJ): Wiley.

Hřebíček, L. (1997). *Lectures on Text Theory.* Praha: Oriental Institute of the Academy of Sciences of the Czech Republic.

Johnson, N. L., Kemp, A. W., Kotz, S. (2005). *Univariate Discrete Distributions*. Hoboken (NJ): Wiley.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760–775.* Berlin, New York: de Gruyter.

Miller, R.G. (1981). *Simultaneous Statistical Inference.* Berlin, Heidelberg: Springer.

Popescu, I.-I., Altmann, G. (2008a). Autosemantic compactness of texts. In: Altmann, G., Zadorozhna, I., Matskulyak, Yu. (eds.), *Problems of General, Germanic and Slavic Linguistics: 472–480.* Chernivtsi: Books-XXI.

Popescu, I.-I., Altmann, G. (2008b). Hapax legomena and language typology. *Journal of Quantitative Linguistics 15, 370–378.*

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* Bratislava: Veda.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse.* Wien: Praesens.

The quantum of plurality. The relationship of singular and plural (and singularia and pluralia tantum) in Czech nouns

Jiří Mácha, Olga Richterová

1. Introduction

"In the opposition of two grammatical numbers, singular and plural, it is the plural which is clearly marked: i.e., it signifies plurality (either of individual referents, or of types), whereas the singular remains unmarked; it can denote both an individual referent as well as a class of such referents or an uncountable amount AND it can denote an abstract, uncountable notion too."[1]

This quote by Kroupová (1985) more or less summarizes the generally held view on the relationship of singular and plural number in Czech nouns. No matter that this quote will soon date thirty years back: although the descriptions of Czech available both to general public and to scholars devote a significant amount of attention and space to the classification of nouns that appear – exclusively or predominantly – either in singular or in plural forms (and are thus called singularia or pluralia tantum), they seem to avoid the question of proportion and of centrality with respect to the singular and plural as such.[2]

No description we could find would answer the question whether there are more singularia or pluralia tantum and the few remarks on the statistical distribution of the two realizations of the category of number[3] in Czech (see Bartoň et al. 2010) were not provided in general reference books. This seems to be similar to the situation in English or German linguistics or in contrastive Czech-English, Czech-German linguistics, which the authors are acquainted with (cf. Dušková 2003; Štícha 2003), yet one authoritative publication differs: (Biber et al. 1999). In this grammar book, the distribution of plural versus singular is

[1] „Je tedy v protikladu obou mluvnických čísel, singuláru a plurálu, plurál jasně příznakový, tj. označuje zásadně mnohost (ať už jednotlivin, nebo druhů), kdežto singulár je bezpříznakový — může označovat jak jednotlivinu, tak i třídu jednotlivin, popř. Nepočitatelné množství I nepočitatelný abstraktní pojem." (Kroupová 1985. Translation: authors.)

[2] This way of approaching the phenomenon in newer research complies with older papers by (Fiedlerová 1975; Jirsová 1981) or renowned grammar books of the past (Havránek, Jedlička 1960).

[3] The notion of 'dual', an archaic form referring to "twosomeness", can be dismissed and the few occurrences (although occurring in very frequent nouns) can be counted as regular plural forms.

discussed (p. 291) with respect to four genres: conversation (spoken language), fiction, news and academic texts. “The higher frequency of singular nouns agrees with their status as the unmarked form,” goes the text, also quoting a couple of examples that “occur in the [singular / plural] at least 75% of the time” (the three-quarter borderline is rather remarkable). The frequency per million words indicated for the four genres in English is a unique piece of information within the context of other grammar descriptions.

Lastly, in (Petr 1986: 52), there is a notion of “balance” between singular and plural in countable nouns. This finding is not returned to at all – either by questioning it or by further supporting it – in the (newer) literature quoted below. This gives the impression of an implicit symmetry fallacy: The tendency of humans, upon encountering a pair or a binary opposition, to expect an equal quantity, distribution, degree, etc. is not discouraged by explicitly stating that such a situation is not meant by “balance” between singular and plural forms.[4]

Having this as our starting point, we want to investigate the way singularity and plurality is expressed within the core representatives of Czech nouns. Our research is based on a statistical analysis of corpus data: As far as we know (and as the literature shows – see e.g. Petr 1986; Grepl et al. 1995; Cvrček et al. 2010), not only do the existing descriptions mostly neglect the above-stated questions, but they are also based mainly on the observations of linguists, accumulated over a large period of time yet relying on little real-use data. Therefore, following the data description in the central part of the present paper, the corpus data available to us will enable a more comprehensive approach and allow for a finer-grained description of the borderline cases (e.g. when the singular form of a lemma supposed to belong to pluralia tantum is used[5], which is considered extremely rare if not impossible by many descriptions). We will pay attention to these also in the end of the paper where we offer an explanation of the investigated problem area and provide paths for further research.

[4] Only in Nekula et al. (2002), the ratio of singular to plural forms in Czech nouns is provided as being 3:1.

[5] In SYN2010, we can find an illustration of such a case in a scientific text describing the ***gills*** **–** an individual *gill* is being referred to: <U žraloků a rejnoků má každá **žábra** vlastní otvor ústící do vody.>; however, the zero singular holds true for most pluralia tantum, e.g. the query [lemma="varhan.?"& tag="N..S.*"] delivers 0 hits, just as many other similar queries. *Žábry* (gills) and *varhany* (the organ) are arbitrarily selected examples from a long list of pluralia tantum found in several reference books. Another example of the singular occurring in an otherwise plural lemma is *pleticha* (an intrigue): <Když pak Vladimír Špidla odešel anebo byl odejit, na jeho místo se vyloupnul mistr zákulisní **pletichy** Stanislav Gross.> Here, the congruent adjective *zákulisní* (behind the scenes) proves the singularity of an otherwise plural lemma. One more example – *sourozenci* (siblings), [lemma="sourozenec" & tag="N..S.*"], occur in more than 200 cases, i.p.m. ca. 2, in the singular.

2. Continuity with Previous Research

The authors prepared a case-study[6] where inspiration was drawn from Eckhoff and Janda's new approach to investigating grammatical categories based on the identification of outliers within verbal paradigms (Janda 2013). The gist of this approach lies in the idea that an extreme value is a signal of a specific semantic behavior of an item.

Furthermore, the authors knew from previous research (cf. Bartoň et al. 2009) that not only the realizations of number, but also those of grammatical cases are not distributed symmetrically within Czech gender (genus) paradigms. Therefore, the authors replicated Janda's method and computed the quartiles for all case forms within every gender category.[7] This methodological procedure was applied to all high-frequency Czech nouns found in the latest balanced and referential corpus of Czech written texts, SYN2010. Their frequency of occurrence had to equal or exceed 300[8]. This threshold was set in order to filter out those lemmas whose frequency prevented the potentiality to realize all 14 case forms characteristic of Czech, which is highly inflected.[9]

3. Initial Assumptions

The subsequent analysis of extreme values showed that for each lemma, the proportion of singular to plural forms within its paradigm is usually scaled (sometimes giving preference to plural, and more often to singular forms), but in some cases a clear tendency to bipolarity (either plural, or singular) prevails. This observation confirms the pre-existing awareness of the existence of three classes of nouns: (i) nouns preferring only plural forms, (ii) nouns preferring only singular forms, and (iii) a vast group of nouns whose ratio of singular and plural

[6] Parts of this study were presented orally at the Lingvistika Praha 2013 Conference: http://lingvistikapraha.ff.cuni.cz/node/184 .

[7] First, the relative count of case forms in a given lemma was computed. Then the vectors of values for each case form within the given gender were construed. The upper border for extreme values was counted within these vectors as the value of Q3 + (value of interquartile range * 1.5). Every lemma (to be specific: every case form of a lemma) which exceeded this border was considered to be an outlier and thus having special semantics.

[8] This number was first set arbitrarily to avoid investigating low-frequency nouns which might be parts of phraseological units or show other distributional anomalies. Then it was checked and as its recall was acceptable, it was kept for the purposes of this paper.

[9] Actually, within the SYN2010 Corpus, only realizations of 12 case forms may be seen as truly inherent to the language system, because the occurrences of vocative forms – both in singular and plural realizations – are extremely rare. To illustrate this, we note that the two nominative forms realized in nouns are found in 8,198,410 occurrences, while the vocative forms are realized in 91,437 instances only, thus being 89 times less common.

forms does not show any clear-cut tendency – especially since we do not know what the average, benchmark ratio is.

The research question focused on setting borders among the mentioned groups by quantitative research and introducing new findings to the whole problem area.

4. Quantitative Analysis

4.1. Data Source and Extraction

First, the relative count of singular and plural forms for all Czech nouns in the SYN2010 Corpus was extracted[10]. To gain these results, the Perl programming language was used. Then, proper nouns[11] and lemmas with frequency lower than 300 were filtered out, because proper nouns show a different morphological behavior than common nouns. Most importantly, proper names tend to avoid plural forms[12] which might distort the overall ratio of singular and plural forms.

Following the extraction process, we first received a non-filtered and then a filtered list of lemmas containing information on the ratio of singular and plural forms and the absolute frequency for every given lemma. An example – Table 1 with the 20 most frequent lemmas – presenting a glimpse of the filtered list follows. The closest to the pluralia tantum group is the lemma *oko* (an eye), singularia tantum might be represented by *svět* (the world).

The non-filtered list contained 9,895 lemmas, and the filtered table contained 7,783 lemmas. This means that the filtered list represents ca. 78% of the non-filtered one, comprising the central core of Czech (appelative) nouns represented by the SYN2010 Corpus (and that ca. 22% of the list obtained from SYN2010 consisted of proper nouns). In the next step, the distributions of singular forms for both filtered and non-filtered data sets were compared. We created histograms[13], presented in Figure 1 and 2.

[10] SYN2010, above-characterized as a corpus of contemporary Czech, is annotated on the morphological level. This implies that final results may contain errors taken over from the automatic annotation. On the other hand, an extensive quantitative research based on morphological properties of lexical units would be impossible without such an annotation.

[11] First, we filtered out lemmas encoded with a capitalized initial letter. Because of the still-low precision, the second step was manual extraction of remaining proper names.

[12] Let us disregard such exceptions as "the little Beethovens" – www.thelittlebeethovens.com – where the proper noun has received a general meaning and can be used as a regular common noun.

[13] We used R software created by (R core team 2012) for the creation of graphs, charts and computation of all statistics mentioned in this paper. All charts were created by ggplot2 library (Wickham 2009).

The comparison of the charts in Figure 1 and 2 shows that a tendency for the preference of singular forms is clearly visible in both data sets. Although the shapes of both histograms resemble each other, a bootstrap version of the univariate Kolmogorov-Smirnov test (K-S test)[14] (Sekhon 2011) shows that there is a significant difference between both distributions of singular relative counts:

Bootstrap p-value: < 2.22e-16
Naive p-value: 0
Full Sample Statistic: 0.1207

Table 1
20 most frequent lemmas from the list of filtered nouns with frequency over 300. The table was sorted according to the absolute frequency of the given lemma.

lemma	**sg%**	**pl%**	**frequency**
rok	53	47	340483
člověk	32.6	67.4	222385
den	61.6	38.4	116831
doba	93.4	6.6	109465
místo	72.2	27.8	99420
dítě	31.5	68.5	92311
život	96.1	3.9	92237
město	84.3	15.7	89965
strana	75.2	24.8	85169
ruka	65.3	34.7	82730
země	70.8	29.2	82668
práce	85.8	14.2	82388
svět	97.3	2.7	81254
muž	61.7	38.3	77487
případ	76	24	75670
žena	61.9	38.1	73885
čas	91.9	8.1	73803
cesta	85	15	71682
oko	15.4	84.6	69298
hodina	32.8	67.2	68746

The initial assumption about the influence of proper nouns was confirmed by the K-S test. From this point on, only the filtered data set – with omitted proper nouns – is analyzed (cf. with Biber et al. 1999: 291, where proper nouns were also excluded from the analysis). The distribution of singular forms within the filtered data set is examined in Table 2 which shows the distribution of singular forms in individual quartiles. The median value 81.6% means that a half of all

[14] This method was chosen, because it allows us to compute the Kolmogorov-Smirnov test for tied data. We set the parameter of bootstraps to 7,000.

lemmas favors occurring in singular form in more then 80% of all their realizations.

Table 2
Quartile distribution of singular forms in filtered data set.

part of the dataset	0%	25%	50%	75%	100%
relative count of sg. forms	0	60	81.6	96.4	100

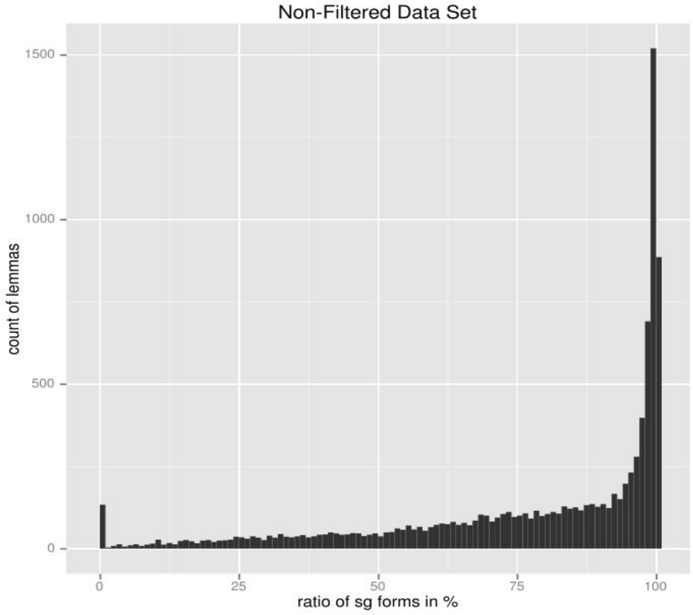


Figure 1. Non-filtered data set. x-axis represents the proportion of singular forms within all forms of a lemma while the y-axis shows the count of lemmas with the given proportion.

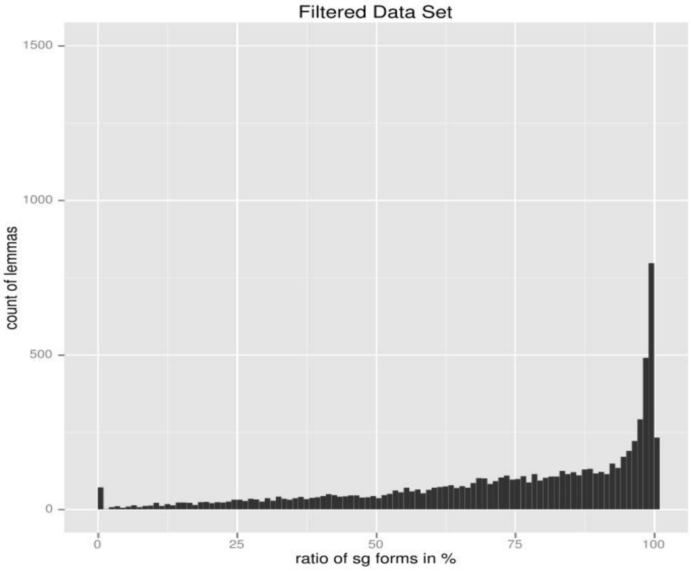


Figure 2. Filtered data set. x-axis represents the proportion of singular forms within all forms of a lemma while the y-axis shows the count of lemmas with the given proportion.

Table 2 leads to the conclusion that singular forms are absolutely dominant within the core of Czech common nouns. This is not a surprising finding (although the actual extent of the preference of the singular may be), but what is surprising is the consistent neglect of this fact by general reference books.

It is obvious that the above-mentioned method (Q3 + 1.5 IQR) for identifying of extreme values cannot be used in this case since the resulting value would exceed 100%. Therefore the decision was made to divide the distribution into 20 parts and inspect the first and last of them. The last vigintile of the vector of singular values is comprised of 2,201 lemmas (4,602,744 word forms), while the first vigintile covers 94 lemmas (267,079 word forms). The mean values of the frequencies found in both vigintiles are 2841.266 for the first and 2091.206 for the last vigintile. As the regression line in Figure 3 shows, the relation between the absolute frequency of occurrence of a lemma and its relative proportion of singular forms is not related. The Pearson's correlation coefficient computed for the absolute frequency and relative count of singular forms has value -0.03031023; Kendall's tau coefficient was -0.03677007 and Spearman's correlation coefficient was -0.05527446. Both Pearson's and rank based coefficients show that both observed variables are almost independent.[15]

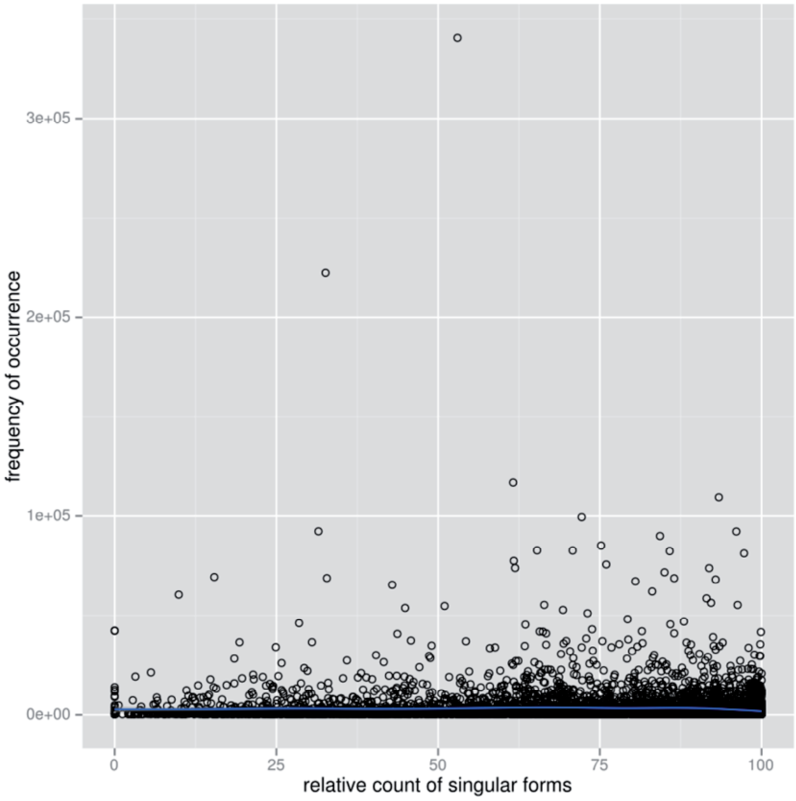


Figure 3. Relation of frequency and relative count of singular forms.

[15] We would like to thank the anonymous reviewer who suggested employing one of the rank based correlation coefficients.

The dominance of singular forms leads us to the question if it is reasonable to define both singularia and pluralia tantum as equivalent categories; from a quantitative point of view, the superiority of the singular appears to be standard. Given the evidence that most lemmas prefer singular forms in more then 90% of all their occurrences, we decided to take a closer look at the category of pluralia tantum, as their role in the language system appears to differ from both remaining categories (ii) and (iii).

5. Qualitative analysis

5.1. Selecting pluralia tantum

What the histograms show very clearly is a small group of nouns which only occur in the plural – this group forms the very first vigintile of the whole data set. An analysis of these 94 pluralia tantum words began by comparing them with the detailed classification formulated by Jirsová[16] in Petr (1986: 51–2) and taken over by Karlík in Nekula et al. (2002). This comparison of the pre-existing semantic categorisation accompanied by a rich arsenal of examples and of the first vigintile revealed both a certain overlap and a surprising level of detail found in Petr (1986), but also rather profound lacunae in the existing description. To illustrate this: the lemma *dveře* (the door), identified as the second most frequent pluralium tantum by the data-driven approach, could not be found either in the category containing the lemma *vrata* (gate) or in any other sub-group. Similarly, very frequent and semantically varied lemmas such as *prostory* (premises), *poměry* (circumstances)*, lázně* (spa), *rozpaky* (to feel embarassed), or *odbory* (trade unions) were not included among the example representatives.

5.2. Comparing the existing categories

An overview of the semantic categories has to be split into a) those which tend to be realized in the plural, and b) those which belong to pluralia tantum (always occur in the plural). Among a), one finds: body parts, clothing, plant and food parts, names of food items, names of things (products), human and animal groups[17], some abstract nouns (with the semantics of "summarizing"). Among b), one finds both a very similar categorization and some strikingly different groups: body parts, clothing, illnesses, names of instruments and tools / equipment, names of holidays, festivities, (important) events and stretches of time, some mass nouns, names of food items, side products, remnants etc, names of written

[16] In the opening of Petr (1986: 5), Jirsová is mentioned as author of "background study" focused on genus and number of nouns.

[17] The original *názvy lidských a zvířecích kolektivů* was followed as closely as possible in the translation.

texts, names of games, some concrete nouns sharing the notion of being compounded.[18]

By comparing Jirsová's category "human and animal groups" and its representatives with its semantic partners in the filtered list, we discovered, that there was no overlap at all. Lemmas from this category were attributed to a), the larger group of "nouns in which plural forms vastly outnumber the singular" and "some of which only differ from pluralia tantum by the option to more easily denote a single item (out of the referred quantity)". Let us give a complete overview of the two lists in Table 3.

Table 3
Human and animal groups with zero overlap of representative lemmas

	Czech lemmas	**translation**
Petr (1986: 51)	manželé, rodiče, prarodiče, příbuzní, diváci, domorodci, krajané, účastníci, hosté, obyvatelé, svatebčané, zájemci, cestující, pozůstalí; plazi ap.	spouses, parents, grandparents, relatives, onlookers, aborigines, countrymen, participants, wedding guests, bidders/ applicants, passangers, the bereaved, reptiles etc.
Filtered list (SYN2010)	silničáři, spoluobčané/i, povstalci, demonstranté/i, obratlovci, pražané/i, stavbaři	road workers, fellow citizen, rebels, protesters, vertebrates, citizens of Prague, construction workers

In accord with Jirsová's observation, these lemmas do occur in singular forms, too (in 2.1–3.7 per cent of cases). But besides the *reptiles* (plazi) and the *vertebrates* (obratlovci), they do not even really share the same semantics. Jirsová gives either general terms for relatives or people living in a particular area or taking part in a special event. The corpus-based list is much more specific, reflecting the design of the corpus data (newspapers comprising about one third, just as fiction and academic texts do) and the topics discussed in the contemporary written discourse.

It is not that either of those lists would be better or worse (both are incomplete), but the difference is so striking that one cannot help but ask for an explanation. Is it possible that the discrepancy lies in the nature of linguistic introspection, which formed Jirsová's list and which probably largely draws on spoken language? Thus, the perspective of Jirsová seems to be that of events of family and local life which are regularly discussed in conversation.

[18] The Czech text can be also downloaded, it is available here: http://stream.avcr.cz/ujc/mluvnice-cestiny-2.pdf?0.903907053405419. (Petr 1986: 51-2)

The filtered list, on the other hand, shows which groups of people are generally referred to large bodies (*rebels*, *protesters*, *citizens*, *workers*), even though all of them consist of – often dissatisfied – individuals. The perspective reflects the view of the society in general: these nouns refer to human hordes partaking in public activities.

5.3. Borderline cases

All in all, 59 lemmas out of 94 (i.e. about two thirds) could not be found in the lists provided by Petr (1986) and taken over by Nekula et al. (2002). Out of these 59 lemmas, 21 occurred in singular forms, too. Since 7 of them are already listed in Table 3, we can name the remaining 14 lemmas: *kJ* (kilojoules), *muka* (troubles, pains), *odpadky* (garbage), *sacharidy* (carbohydrates)*, přijímačky* (entrance exams)*, nudličky* (noodles)*, běžky* (cross-country ski), *hektolitry* (hectolitres), *fakta* (facts, data), *lyže* (ski), *prostory* (premises, room), *zápisky* (notes)*, žampiony* (champignons), *zplodiny* (emissions).

This list shows the dilemmas faced by classification attempts: just as most people wear a pair of shoes, but a one-footed person needs to refer to his shoe only, most of the time it is sufficient to speak about "two skis" and only at rare moments, such as when an accident happens or a special technique is being described, one ski out of the pair needs to be referred to. And there is more to skiing: names of competitions, such as "Pražská / Kladecká / Zlatá lyže" distort the picture – this exemplifies very nicely that not all proper names can be eliminated that easily as a researcher might wish. Moreover, we also need to mention the possibly erroneous automatic annotation – all word forms of *přijímačky* with the morphological tag of the singular are actually realizations of the plural.[19] With *zplodiny* (emissions), this was the case for most (11) word forms tagged as singular, but 5 genuine singular forms could be found.

To stay with the word *zplodiny*: How are we to categorize a lemma with 367 occurrences in our corpus, out of which only 5 forms occur in the singular? The best answer seems to be: Investigate it more rigorously. Close analysis reveals a significant semantic divergence from the meaning of toxic gas substances[20]: in 3 cases out of 5, the authors of the utterances merely made use of the verbal root shared with the notion describing toxic emissions, and referred to toxic thoughts. So once again: How are we to categorize a lemma with 367

[19] We decided not to go deeper into the issue of form homonymy with respect to singular and plural, although the quality of automatic annotation would need to be tested, as this example aptly shows.

[20] To quote the hits: <Sovětský svaz je **zplodina** filozofického idealismu intelektuálů 19. století (...)> or < jeho osvícensky chladná ironie byla psychickou **zplodinou** racionalistické skepse>

occurrences, whose potential enables 3 metaphoric uses in singular form – and only 2 singular uses of the regular[21], core meaning?[22]

6. Summary of the findings and outlooks

Previous research that looked into the grammatical category of number did a good job in identifying the three important groups of nouns. However, it is quantitative analysis which is best-suited for (discovering and) defining the huge disproportion between the singularia and pluralia tantum groups. While the previous research relied mostly on introspective methods, corpus-based analysis combined with statistical testing strives for a methodology that is replicable and verifiable – and which can provide other professions with data to work with.[23] The example of *zplodiny* showed the deficiencies of an approach which overemphasizes classification and categorization. From a lexicographer's point of view, it may be precisely the information about the usage of the singular / plural word forms of the lemma *zplodiny* in the written language that s/he needs to know, not someone else's interpretation of whether it is a pluralium tantum or not. In the end, we must not forget that any classification is only an auxiliary tool in our attempts to understand the language system.

The investigation of the vigintile distribution of number forms clearly uses a mechanical criterion, yet it proved an optimal starting point for the follow-up, “manual” analysis which is necessary (and that not only because of the errors in automatic morphological annotation of corpora). Moreover, the top-level view (the so-called “bigger picture”) would not be possible without the – erroneous – morphological annotation. This paper's focus on the overall picture showed that the category of pluralia tantum differs a lot in quantity. Pluralia form ca. 1/70 of the ca. 7,000 lemmas representing the core of Czech nouns in SYN2010 Corpus, while singularia lie in a completely different range, comprising about 2/7 of frequent Czech nouns. Also, the comparison of non-filtered and filtered data sets proved the assumption that proper nouns would skew the results and it is reasonable to treat them as a specific category and investigate them in future

[21] To quote an example of the regular use in the singular: <Dochází při něm k neúplné oxidaci cukrů a pozvolnému uvolňování energie; **zplodinou** procesu je látka (etanol) ještě dosti bohatá na energii (...)>

[22] This issue borders also on the question of polysemy: we would like to thank the anonymous reviewer who pointed out that the behaviour of polysemous nouns and the distribution of singular / plural in individual meanings needs to be looked into as well.

[23] In the referenced publications, examples of singularia tantum are given in the dozens. However, our approach to their identification has detected ca. 2,200 of lemmas which can be considered to belong to this group. The important thing we would like to point out is the fact that we extracted, within the SYN2010 Corpus, two complete lists of frequent singularia and pluralia tantum represented within that corpus (we are aware that their completeness is a matter of interpretation, too).

work, on their own. The findings also show that the proportion of singular forms is not related to the frequency of a lemma.

There is a lot of room for further research: besides the lexicographic avenues mentioned above, there is need for a detailed semantic analysis of both distinct groups, including an analysis of the suffixes occurring within both groups, and finally, an analysis of the grades of “singularity” or “plurality” of a noun. Given the scope of this paper, the authors could only touch upon some of these topics, but they hope to have provided a large quantity of food for thought – as well as for future research

Acknowledgments
The study was conducted within the framework of the Charles University Programme of Development of Scientific Disciplines (PRVOUK), No. P11 Czech National Corpus.

References

Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., Petkevič, V. (2009). *Statistiky češtiny*. Praha: NLN.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Cvrček, V., Kodýtek, V., Kopřivová, M., Kovaříková, D., Sgall, P., Šulc, M. Táborský, J.,Volín, J., Waclawičová, M. (2010). *Mluvnice současné češtiny*. Praha: NLN.

Dušková, L., Knittlová, D., Peprník, J., Tárnyiková, J., Strnadová, Z. (2003). *Mluvnice současné angličtiny na pozadí češtiny*. Praha: Academia.

Fiedlerová, A. (1975). Nástin vývoje pomnožných jmen v češtině. *Slovo a slovesnost 36 (4), 266–285.*

Grepl, M., Hladká, Z., Jelínek, M., Karlík, P., Krčmová, M., Nekula, M., Rusínová, Z., Šlosar, D. (1995). *Příruční mluvnice češtiny*. Praha: NLN.

Havránek, B., Jedlička, A. (1960). *Stručná mluvnice česká*. Praha: SPN.

Janda, L.-A., Eckhoff, H.-M. (2013). Grammatical profiles and aspect in Old Church Slavonic. *Transactions of the Philological Society*.

Jirsová, A. (1981). Dynamika vztahu singuláru a plurálu u substantiv v češtině. *Slovo a slovesnost 42 (3), 193–199.*

Nekula, M. et al. (eds.) (2002). *Encyklopedický slovník češtiny*. Praha: NLN.

Kroupová, L. (1985). K pomnožným podstatným jménům v současné češtině. *Naše řeč 68 (2), 57–63*. Available at: http://nase-rec.ujc.cas.cz/archiv.php?art=6528

Petr, J. (1986). *Mluvnice češtiny 2, Tvarosloví*. Praha: Academia.

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for statistical Computing. Available at: http://www.R-project.org/.

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software 42(7), 1–52.*

SYN2010: Křen, M., Bartoň, T. – Cvrček, V., Hnátková, M., Jelínek, T., Kocek, J., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V., Skoumalová, H. (2010). *SYN2010: žánrově vyvážený korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Available at: http://www.korpus.cz.

Štícha, F. (2003). *Česko-německá srovnávací gramatika*. Praha: Argo.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. New York: Springer.

Palindrome-like structures in the *rongorongo* script

Tomi S. Melka

Intelligence loves patterns and balks at randomness.
(Hofstadter 1999 [1979]: 175)

1. Introduction

The amount of papers and studies about different aspects of the *rongorongo* (RR) script of Easter Island in the last ten years is on an unprecedented rise (see R. Wieczorek's 2013 webpage). The fact itself traces four major bearings, (a) it indicates to some extent the chronic disagreements over the nature and the decipherment of the script; (b) however, scholars may paradoxically reach a point of *a critical mass*,[1] a conduit driving out into a new quality: agreed and evidenced findings, (c) while the interpretation of any "mysterious" script is tricky,[2] proactive efforts may enable a plausible solution, if not a satisfactory one; and (d) it shows that for all the originality and exotic flair, *rongorongo* shares more than a few scribal features with other real-world scripts (Linear B, Egyptian hieroglyphics, Cuneiform, Zapotec, ancient Libyco-Berber writings, Maya glyphs, early Chinese writing, Naxi, Batak manuscripts, etc). Implications are important and they can be extended beyond the task of understanding and deciphering arrays of unknown signs coming from a small island of Oceania (see Barthel 1971; *Introduction* & *Conclusion* in Robinson 2002; Melka 2009b; Harris 2010: 8).

Over the years, very much has been written and said about *rongorongo* in several languages, and among the literature, the names of Routledge (1919), Métraux (1940), Ross (1940), Kudrjavtsev (1949), Butinov and Knorozov (1957 [1956]), and Barthel (1958), rise to prominence for their early scientific commitment. To this day, researchers owe a great intellectual debt to them.

Similarly, *rongorongo* has also been known to be deciphered, though the exact words contained in many of the tablets have yet to be translated into English, followed by independent and positive verifications.

General compilations of the historiography, description and taxonomy of RR are Barthel (1958) and Fischer (1997). Two other important sources of knowledge are the online editions of CEIPP (2005), and Wikipedia (2013a). As I wish not to repeat myself on the particulars and the deciphering problems of

[1] See especially Robinson (2002: 36).

[2] Along a sound methodology and credible arguments, small or large conceptual errors will likely persist.

rongorongo, readership may well follow Guy (1982, 1985, 1990, 2006), Pozdniakov (1996, 2011), Davletshin (2002, 2012a, 2014), Sproat (2003); Horley (2005, 2007, 2009, 2010), Harris (2010), Wieczorek (2011, 2014), Spaelti (2012), or Melka's (2009a, 2012, 2013, 2014) contributions.

The aim of this paper is to examine a limited number of specially inscribed structures in the classical script of Easter Island, *rongorongo*: the palindromic-like passages on tablets N*a*3, A*a*5 and their parallels. As allotted space is restrained in academic publications, I focus on this choice without claiming a full representativeness of the phenomenon. The subject of "palindromes," although partly treated and mentioned in the RR bibliography, has not seen to date a specifically designed study, at least formally published in paper or electronic media.

The current essay is structured in a simple manner. Sequences are visually and alphanumerically[3] presented in two main sections, compliant subsections and accompanied by supporting commentaries. A *Discussion* section gathers the mainstream opinions of RR scholars, and explores the domain of symmetry within and beyond *rongorongo*. *Conclusion* puts the finishing touch to the essay and leaves the door open for exchanging further ideas. As for descriptive and analytical conventions, they line up with those of former studies (see Barthel 1958; Guy 2006; Harris, Melka 2011a; Melka 2008, 2012, 2013).

2. Observed "*palindromes*" in some RR tablets

On the basis of a salient structural feature, Barthel (1958: 164) claimed that two RR strings (Fig. 1) have the property of a palindrome, a phrasal set (or a word) equally read forwards and backwards,

H*r*5: 4.64 - 44 - 4 - 49f - 4 - 44 - 4.64
N*a*3: 1.6 - 522 - 1.62 - 600 - 62.1 - 522 - 6.1

His following suggestion,

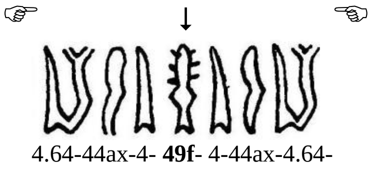
> "*Im zweiten Falle wird die Symmetrie noch dadurch unterstrichen, daß die Ligaturen ihre Handzeichen jeweils auf das Zentrum richten. Das ist die einzige Stelle in den Tafeltexten, bei der man sich zu einer ikonographischen Ausdeutung versucht fühlt*." [In both cases, the (mirror)-symmetry[4] becomes thereby emphasized, as the ligatures with hand-signs are oriented toward the center. This is the real placement in tablet-texts, in which someone should consider searching for an iconographic interpretation].

[3] We must keep in mind the fact that, so far, the real core of the repertory of RR signs is not even known with half certainty. Lacking a better notation, Barthel (1958) merely avoids confusion and more idiosyncratic tagging of the glyphs.

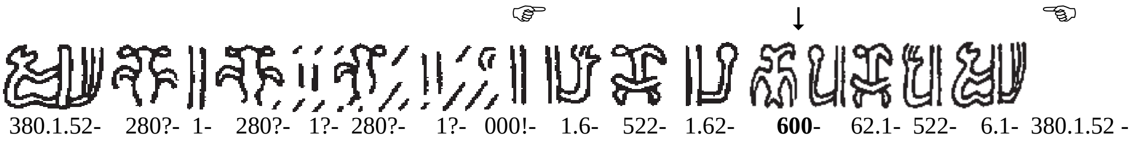
[4] The deployment of symmetry-related terms throughout the paper is essentially understood as "*… the total sum of pairwise correspondences in the complete pattern*" (Wagemans 1995: 25).

raises an interesting point in the *rongorongo* studies. By *iconographic interpretation* Barthel (1958) certainly meant a visual, non-phonetic approach with regard to these constructs.[5]

"*Large Santiago*" tablet, H*r*5 (# 1)
Barthel's (1958) tracings



"*Small Vienna*" tablet, N*a*3 (# 2)
CEIPP (2005) after Barthel's (1958) tracings (see also Haberlandt 1886, Tafel X, Fig. 2a; and Horley 2010, p, 53, Figure 9)



Fischer's (1997) tracings

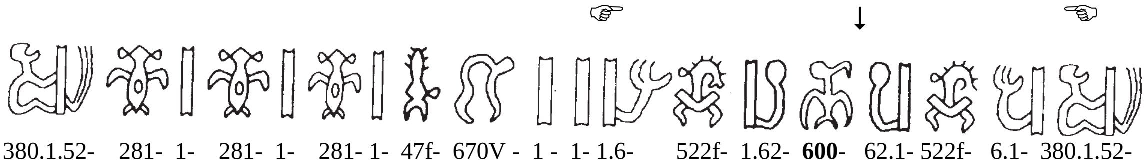

☞
↓
☜
380.1.52- 281- 1- 281- 1- 281- 1- 47f- 670V - 1 - 1- 1.6- 522f- 1.62- **600**- 62.1- 522f- 6.1- 380.1.52-

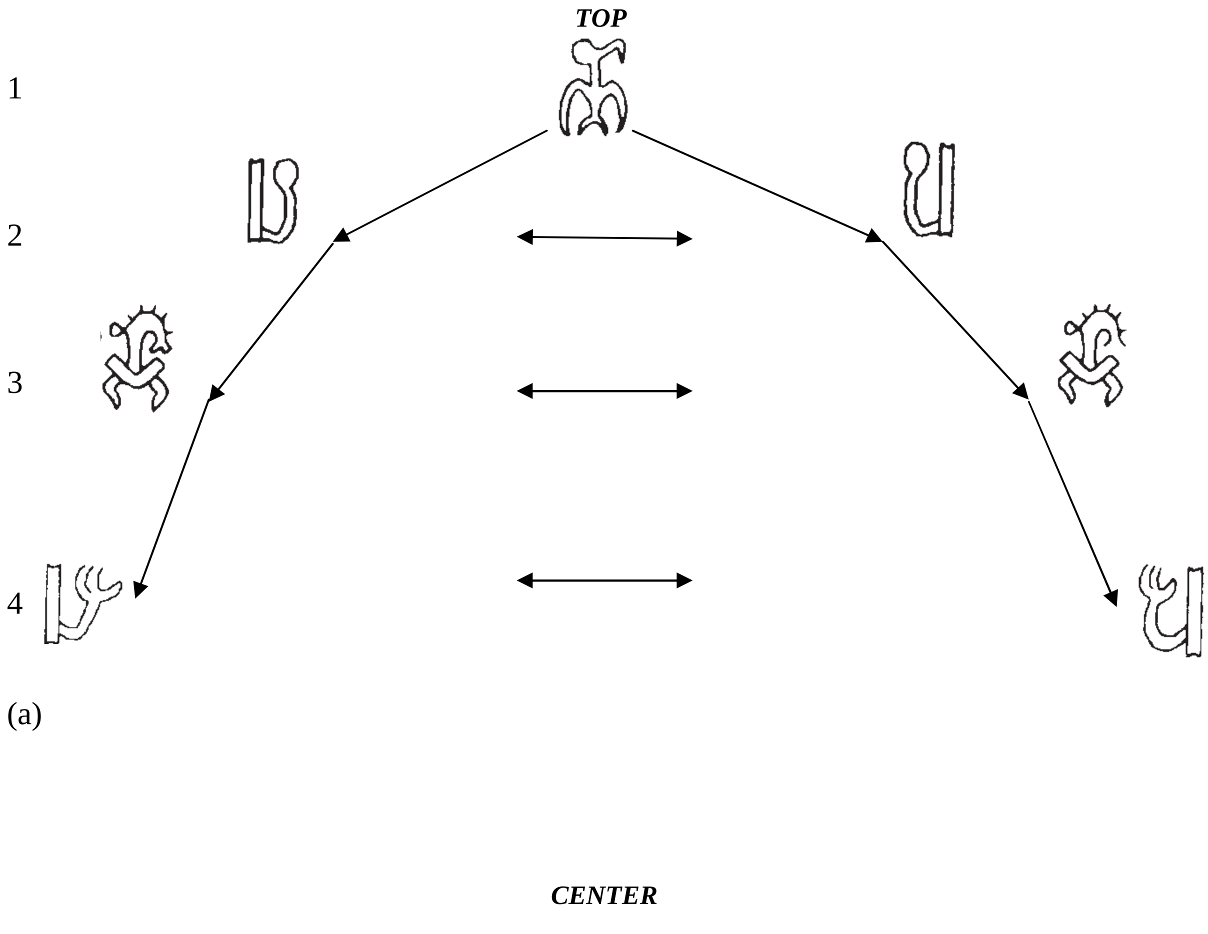
Figure 1. Illustration of the palindromic-like segment N*a*3, with Barthel's (1958) numbering applied to Fischer's (1997) improved drawings. Granted the iconography, N*a*3 is very likely a mini-scene, and not a mini-text. H*r*5, as a matter of fact, is the "core" of a longer palindrome confirmed in the *Great Tradition*'s items (Wieczorek 2011: 34, Fig. 4; see also Kudrjavtsev 1949; Sproat 2003; and Ávila Fuebtealba 2007).

2.1. The N*a*3-palindrome

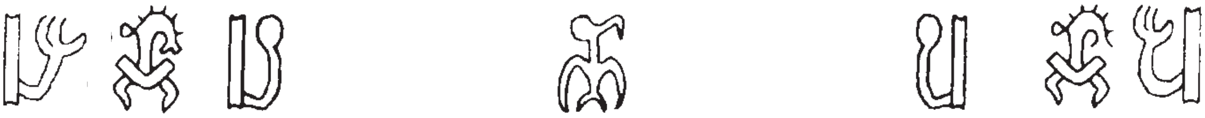
Sequence /1.6-522-1.62-600-62.1-522-6.1/ is first published in Barthel (1958). Later, it was reconfirmed and filed by several researchers, including the author (2004). The N*a*3-structure (Figure 1, # 2) calls for scrutiny. The Janus-faced sequence is an *isolate* in the present form, although a much reduced parallel is found on E*v*4, /1.6.6-200-1.62-522/ (Horley 2010, Figure 9). Two other passages

[5] Fischer (1997: 224) offers a similar glimpse, "*Certain series of glyphs appear to orient themselves toward a glyphic centre, suggesting an iconographic reading of the passage*."

(cf. C*b*1 and S*a*1) are distinguished by such considerable variations that raise suspicions about their functional similarity. N*a*3 appears incised in an astonishing equipoise with the progressive-regressive "reading," most probably, visually-based.[6] Borrowing from the domains of algebra and geometry, we realize the one-to-one correspondence between the two glyphic sub-sets (i.e. bijection) and any property (be that, semantic / linguistic / emblematic) preserved in the components of left half, is also true of the right half's components. One wonders if there is any obvious advantage in choosing the left or the right isomorph in reading N*a*3. The palindrome, as it keeps its structure-preserving mappings, does not show to be entirely self-contained. Rather, it is related to what is implicit in the preceding sub-string /281-1-281-1-281-1-47f-670V-1-1/. On their part, both sub-strings are placed in between a delimiter-type /380.1.52/, an expansion of the /380.1/ (cf. Barthel 1958), meaning, they point at a *coherent portion of text.*



[6] To evaluate a potential solution to the observed "palindrome" (or to other ones), we might consider the rhetorical devices of *chiasmus* and *antimetabole* (A. Davletshin, personal communication, 2012). These figures of speech are frequently used in the literature and the oratory tradition of many people, e.g. ancient Greeks and Romans, Hebrews, etc. For more, see *Section 4. Discussion*.



(b)

Figure 2. “Binary” tree of the palindrome-like sequence in N*a*3 depicts its structural elegance and sparsity. Discussed glyphs are placed at the vertices. The four (4) levels or heights of the pyramid-shaped sketch (a) hypothetically designate Old Easter Island personages, metaphorical or real, according to their social influence and projected power. The assumption consents Purse, Campbell (2013: 377), “*When writing develops in contexts of iconography, the vertical axis* (*in column formats,* my note) *corresponds to the orientation of dominant figures in scenes*.”

With glyphs strung out in a straight line (Figure 2b), the “frigate bird”-glyph takes up the center of the palindrome. Otherwise, in constructing a binary tree (see Figure 2a), glyph /600/ subsumes the apex position in the spatial hierarchy. Whether intuitively or consciously, pinpointing the central axis of a palindrome is necessary, but not everything. Under normal circumstances, when reading a palindrome from left-to-right, it is almost impossible to locate the “middle” until the complete word has been read. Beneath the “frigate bird” glyph are listed the figures of “barbed” *ariki* /522f/ (Figure 2a). In either level, /600/ and /522f/ are flanked or followed by the “staff”-glyph /1/, perhaps drawn after the celebrated *kohau* of the *rongorongo* men, or the long staff *ua*, another authority mark in Old Rapanui (Orliac, Orliac 2008, Figure 103 and 104). A different view is that /1.6/ (/6.1/) may work as some sort of “*graphical frame*” as in the section C*a*5-C*a*9 (Pozdniakov 2011: 58; see also Ávila Fuentealba 2007, on a broader context). I am disposed to consider /1.62/ (/62.1/) in tune with the earlier suggested function, and slightly comparable to a silent “incipit.” Both “staff”-like compounds draw glyphs appearing to imply *preeminence* and *power*, i.e. /600/ and /522f/. If evinced, no grammatical constructions, or clauses, seem to have been inserted between /600/ and /522f/. Neither is clear if a subject follows a predicate here (or *vice versa*),[7] dismissing the language-specific bond.

As the frigate bird was considered *messenger of the gods* (see Barthel 1971: 1173; Mordo 2002: 124; Craig 2004: 65), or related to the Miru lineage group (Barthel 1963: 378), it is realistic to think that /600/, shaped after real-life *makohe*,[8] is invested with utmost significance, and plausibly stands for the Old supreme deity Makemake. The different meanings of glyph /600/ may be

[7] Rapanui language is of a VSO syntactic order (du Feu 1996: 9-10).

[8] The ultimate mirroring effect would have been the carving of a double-headed bird, similar to the Christian Orthodox emblem featured in banners and flags.

indebted to the polyvalence spotted in Rapanui language, a branch of the Polynesian family. Sign /600/ is one of the most-frequent of the RR corpus, and we may suppose that the high recurrence may have motivated its logographic use in a number of environments. Beyond this context, a striking parallel comes from Cuneiform writings of the Achaemenids. In the cuneiform system for writing the Old Persian there are in addition logographs for the common recurring words *king*, *god*, *country*, and *earth* (see Lunde 2009: 23). The symmetric alignment *more than a quirk in grammar* seems to fill *a visual role* or hint at *some graphical punning*. Perhaps the "palindrome" was appealing to the Old Rapanui intellect and conveniently adjusted to *otium cum dignitate* [leisure with dignity] (Filloy 2005).

Alternatively, as the "frigate bird" /600/ occupies the top hierarchy, I surmise that it is glorified as a champion and likewise as a guardian of the divine and earthly order. From an archetypical worldwide perspective, peaks and tops of natural objects (e.g. hills, mountains) or of artificial ones (e.g. shrines, temples, pyramids) match with kingship, heaven and divinity-like concepts.

2.2. Other "*palindrome*"-like structures

Routine inspections have shown so far that RR "palindromes," although not carved at an alarming degree, are not that rare after all (see Barthel 1958; Ávila Fuentealba 2007; Horley 2007: 25, 2011: 104-105; Pozdniakov 2011; Wieczorek 2011; Melka, Harris 2011; Davletshin 2014). Based on their analysis strands, the identified sequences seem not to share a unifying style, or a standard level of orthography, given that space and individual scribal factors are frequently at play. Some glyphs are quite stylized and geometric-like, e.g. /2/; /4/; /25/; /64/, others remarkably *lifelike*, e.g. /200/; /600/, and their core- and flank-structures show variety in terms of morphology. In what follows, additional palindromic-style "mirror reflections" and quasi-symmetrical structures are merely reported, with no comments on the details. Without conveying neglect at their expense, I'd point out that another paper would beg a serious analysis in the near future.

A*a*2-3, "*Tahua*" tablet, see Davletshin (2014, Figure 2C'),

A*a*6, "*Tahua*" tablet *vs*. R*a*2, "*Small Washington*" tablet

A*b*3, "*Tahua*" tablet

B*v*3, "*Aruku Kurenga*" tablet, see Horley (2011: 104, Figure 11)

B*v*9, "*Aruku Kurenga*" tablet, see Horley (2007: 25, Figure 2)

H*v*4, "*Large Santiago*" tablet and P*v*6, see Horley (2005: 111, Figure 13)

P*v*8 and P*v*10 “*Large St. Petersburg*” tablet {P*v*10 ~ H*v*9}, see Horley
(2005: 112, Figure 15)

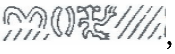
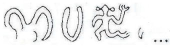
S*a*3-4, “*Large Washington*” tablet

Some rulings about their perceived function are included at this point: Fischer (1997: 224), opts for an iconographic reading when *two figures appear to be exchanging glances as a recurring situation or plot*; Horley (2005: 111, Fig. 13) in mentioning mirror-reflections in H*v*4 and P*v*6, states that they probably serve for ornamental purposes only. In turn, de Laat (2009: 218, Fig. 98, 4-6), admits that *mirroring of two glyphs that appear next to each other is a very common feature*, adding that the attribute *could only have been appreciated by the reader*. Therefore, these cases *contain another extra-dimension* in the RR script. For my part, I surmise that mirror objects with a vertical axis (be RR glyph sequences, or any other design with such a property) are clearer than any other object for the human neuro-optical system, and hence easier to recognize or retrieve (see, for instance, Barlow, Reeves 1979, on the detection of mirror symmetry in random dot displays; Washburn, Crowe 1988: 23, …*all peoples use symmetry as a diagnostic feature in the perception of form*, and also *kaomoji* (face-mark) constructions in Twitter messages where symmetry, configurationally, plays an additional role in delivering and identifying them, Bedrick et al. 2012: 59).

3. On four parallel sequences found on Tablets “*Tahua*” and “*Aruku Kurenga*”

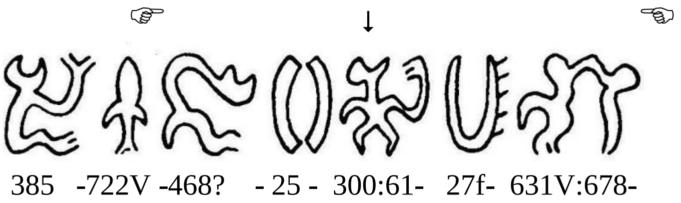
3.1. Brief synopsis on A*a*5

Based on the *mirroring text* in A*a*5, each additional sequence is co-referenced, compared to it and is discussed in turn (see Figure 3; and Ávila Fuentealba’s 2007: 51, Fig. 49, fine-grained identifications). For convenience and optimization purposes, A*a*5 – the longest common string – is regarded as the *original occurrence* during the analysis. Anyway, before facts are known, there are no prior expectations as to this claim. So, to demonstrate that A*a*5 is indeed the “prototype” and not some episodic event, plus to put to a serious test the validity of the comments, further substantial evidence is needed.[9]

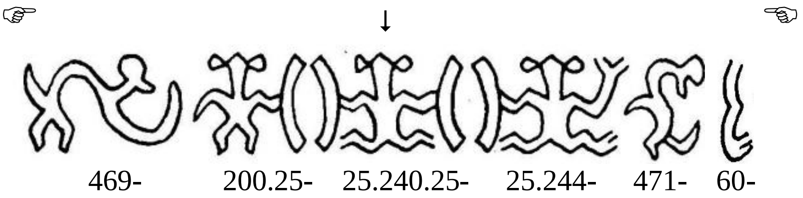
[9] A possible meaningful substring of the *original sequence* occurs in R*a*1, coded after T. Barthel as /493?-25-226-?-?/. The following portion is obliterated either in Barthel (1958) , or Fischer’s (1997: 466)  tracings, hence definite statements are refrained. Nevertheless, a computer-enhanced analysis or multispectral imaging techniques applied to Ross (1940: Plate 1) and Kjellgren’s (2001: 76; Plate 48) photo of “*Small Washington*” tablet (or to the artifact itself), might yield a positive identification. The restoration of the damaged portion may be also made on the basis of

The actual RR corpus is what we have left, being the outcome of dire human action and time itself (Melka 2009a), with related remarks set by default. Missing the entire supposed corpus (Eyraud 1866; Routledge 1919) is *a little*, as preaching from outside. Cross-checking is thereby an additional asset as it allows coordinating the analysis and drawing more persuasive arguments. The fully enacted form (A*a*5) also grounds the idea that sequence's cohesion is ensured or justified by the anthropomorphic figure – notated as /244/ by Barthel (1958) – plus the rounded sides of glyph /25/ . Sequence's boundaries are set at /469/ (*the prefacing glyph*) and /471/-/60/ (*the terminal glyphs*), in analogy with parallels in A*a*1, A*b*7 and B*v*8. Glyph /60/ is not a "dummy" appendage, rather than an integral part of the strings, possibly simulating or balancing the twisted projection perceived in /469/. A*a*5 includes most likely *redundancy*, which in any case means some kind of *repetition*. While struggling with orthography, the *rongorongo* scribe expanded the "phrase" or the intended message until he was content that he had conveyed the meaning, by making sure also that the information could be extracted (see Cherry 1966 [1957], p. 120, for general conventions).

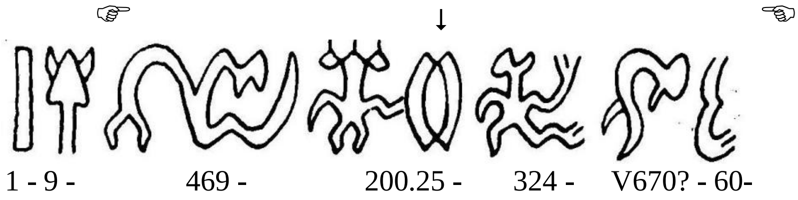
A*a*1



A*a*5



A*b*7



the symmetry. Thus, the remaining glyphs in R*a*1 could be diagnostic of the leftward "isomorph" /468?-25-300:61…/ present in A*a*1.

B*v*8 ☞ ↓ ☜

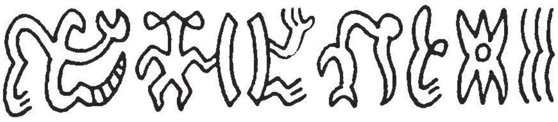


Figure 3. Although not perfect cognates due to inconsistent orthography, sequences A*a*1, A*a*5, A*b*7, and B*v*8, appear motivated by similar-looking glyphs and are far from being disjoint (see Sproat 2003; Ávila Funtealba 2007: 51, Fig. 49; Horley 2009: 260, Figure 7). Sequence repetition across the tablets, with slight or strong variation, appears to have been an established routine among Old *rongorongo* scribes or local schools.

A case in point comes from ancient Egyptian. Syntactic redundancy implies additions to a text; something more is said or written than is strictly necessary to convey the message. DeFrancis (1989: 162, after Budge 1963: 38) illustrates the idea, see Figure 4.

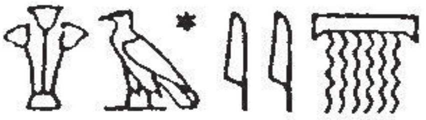


Figure 4

Specifically, the first symbol is a biconsonantal approximating the sound *ha* –the /*a*/ actually stands for a sort of coughed consonantal sound called a *glottal stop*–. The second symbol is a uni-consonantal representing the same sound /*a*/ and functioning as a phonetic complement reinforcing the sound /*a*/ in the preceding symbol. The third double "*reed*" glyph represents a sound somewhat like the "y" of English <*hay*>, conventionally written as /*i*/. The fourth is a semantic determinative referring to *rain* / *waters* (Ritner 1996: 79, Table 4.4). All four taken together represent the word <ḥai> "*rain*" (DeFrancis 1989). To Western eyes, these added elements are quite *unnecessary* since the Greco-Roman alphabetic writing system ingrained in learned people's mind tends in theory to make matters much simpler.

3.2. Search and compare - A*a*1

The greater part of numbering in A*a*1[10] is amended due to misidentification and empty slots (see Barthel 1958: p. 44, *Transkription der Tafel „Tahua“* [Tablet’s “*Tahua*” transcription]). /385/, a /380/-series figure with an upward “forked hand” /64/ almost certainly serves as a delimiter. Next, I am disinclined to view /722V/ as an impulsive *contamination*, i.e. an extraneous glyph brought in from elsewhere by the scribe. Its phonetic relation to the rest of the string is uncertain, though it may be suggested that semantically complements the palindrome’s content,[11] i.e. it stands for a meaningful subsidiary of /468?-25-300:61-27f-631V:678/. The core structure is condensed into /25-300:61-27f/ , an alleged variation of /25.240.25/ in A*a*5. Therefore, it may be presumed that rightward glyph /27f/ swaps with /25/.[12] Despite different numbering, the visual likeness beteween the two glyphs is recognizable, especially along the smoothed edges. The palindrome effect is just reversed: only *one* “anthropomorphous” is left instead of the *two* or *three* rendered in A*b*7, B*v*8 and A*a*5. These apparent alterations have been formerly described by researchers, see e.g. Métraux (1957: 184) “…*the same engraved symbols recur, with numerous variations of detail, on the majority of these objects*…”

Now, let’s check the glyphs that constrain (or sandwich) the core. “Opening” /468?/ is strikingly twisted, in the vein of the other “break-dancing” glyphs in the parallel sequences: /492/ on B*v*8; or the variants /469/ on A*a*5 and on A*b*7.[13] The “gaping mouth” orientation, downward or upwardly set, does not seem to be of importance and is open to scribal whim. The slightly different shapes would also point to the fact that the RR sign repertoire was not stabilized yet, with scribes experimenting with and rejoicing in exploring graphic and visual possibilities. Final glyph /631V:678/ is, in effect, a composite, bonding two bird-shaped elements. The “long beak” of the /600/-series alternates with the “head-and-body” glyph of /400/-series; given the case,

[10] For segmentation, see Ávila Fuentealba. (2007: 34, Figura 29, Tahua Aa, 07).

[11] At any rate, as there is no successful correlation between the spelling system and any assumed phonologic unit, it is hard to deem /722V/ as a violation in line with the other iterative sequences.

[12] Barthel (1958: 279, Footnote 10) likened the shape of this glyph with a bivalve mollusk, the mussel. As for translation, he assigned the nominal value, “pure,” prayer, invocation.

[13] Many glyphs in the RR corpus offer formal similarity to the so-called complex “break-dancing” glyphs, e.g. /461V/ in Ia7; /484/ in Bv2; /493/ [60.490.64] in Bv10; the upturned /493/ [61?.490.61] in G*v*1, –to cite a few–. Their spectrum distribution across the corpus is virtually rare and they are characterized by asymmetry. But none of the RR glyphs can be compared to the sheer and formidable complexity of certain Maya glyphs, see Lounsbury (1991 [1989]: 232-233, Fig. 10),

we may again suggest forthwith *allography*. Switching of the /400/-series with /600/-series is long known in RR upon observing manifold strings where these structural relationships are confirmed (cf. Barthel 1958: 155, Paraphrase 1 (g); 1963: 378-379; Fischer 1997: 389; Melka 2007-2008; Horley 2009: 258; Wieczorek 2011).

The "opening" and "closing" glyphs appear to serve as "brackets" in a sign-group, that is, the sign-group begins and ends with the same or a closely similar sign. The *bracketing* is observed in other real-world scripts, see e.g. Brice (1976: 37-38, Fig. 2.6) on the Cretan Linear A, and the Iranian proto-Elamite.

3.3. Search and compare - A*b*7

Another abbreviated parallel sequence is carved on A*b*7 (see Figure 3, and Horley 2009: 260, Figure 7). To begin with, bi-gram /1.9/ most likely performs the function of the repeated delimiter /1.9:5/ , as detected on "*Tahua*" (Barthel 1958;[14] Horley 2007: 27; Ávila Fuentealba 2007: 133-134, 136). Broadly speaking, although implicit RR "words" are not clearly separated[15] by white space (as in English), or by some special sign (as in the ancient *Meroïtic*, Smith 2010: 1), or as in the Etruscan *Tabula Cortonensis* (Robinson 2002: 180), identification of "word" or "paragraph" delimiters offers realistic hope for correct text segmentation. To which purpose, I'd like to further illustrate with another case from a different context and age. William Caxton, the publishing pioneer in England, set a precedent by dividing in 1485 the manuscript "*Le mort d'Arthur*" of Thomas Malory into eight chapters or "books," thus making it more palatable or readable (Olmert 1995 [1992]: 126). Textual division and indexing, whether on *the word*, *paragraph* or *the chapter level*, democratizes and considerably helps the organization and retrieval of information for the concerned community.

Returning to *rongorongo*, the "centripetal" anthropomorph /25.200.25/ of A*a*5 is absent, as demonstrated in Figure 3. The omission has a double significance, (a) it may create a problem if we believe in a lineup of syllabic units regularly underlying the two sequences, and (b) the pattern is re-used and compacted at will –occasionally due to economy in respect to both material and carving effort, and quite often due to personal interpretation of what must have been "good looking" *rongorongo* in pre-missionary times (before 1864).

Examining the curving outline of the "opening" glyph /469/ , the figure appears quite contorted, or exaggerated as opposed to its form on A*a*5. The configuration of /V670?/ is somewhat impaired, although it must surrogate

[14] Barthel (1958) thought of this repetitive intercalation in "*Tahua*" tablet as a *refrain*.

[15] In pre-modern societies that did not divide words and/or sentences, *writing was done by specialists who were completely immersed in writing conventions and who could therefore dispense with such aids to legibility*, i.e. sundry word dividers (see Gaur 1987: 55-56.)

the "closing" glyph /471/. Next, it seems that glyph /324/ complies with what /244/ is for in A*a*5. The *mirroring effect* at string A*b*7 is clearly in remission, and barely recognizable by the casual observer. Measured up to the full cognate in A*a*5, the A*b*7-sequence is not *ad litteram* carved, being visibly reduced. Providing a fixed "syllable-structure" in many *rongorongo* sequences, one must say then the script is exceptionally cumbersome, where syntax is *persistently breached* or *disordered* to the point that "syllable" positioning looses relevance. Similarly, given the size of *Tahua*'s text, carved in a-bit-less-than-a-meter-long piece of ash wood (*Fraxinus excelsior*): 91 cm (Fischer 1997: 409), or 91.4 cm (Orliac, Orliac 2008: 246, 248-249) and its fine state of preservation, Geiseler's (1995 [1883]: 55) report[16] (see also Métraux 1940: 398), although original, is quite improbable. Considering that *Tahua*'s scribe was involved in devising a crypto-system by occasionally or cyclically shifting glyphs across the sequences after a *CAESAR cipher-like* in order to hand the long object to a messenger to send it elsewhere would defeat its purpose. Text A with a *highest quality of inscription*, appears to be among the most revered and ceremonial RR artifacts (Fischer 1997: 411-412).

Furthermore, assuming the same anonymous person carved text A (see e.g. Fischer 1997: 394), an explanation for such "under-spelling" or alteration on A*b*7 could be that after doing more than ¾ of the tablet[17] (91 centimeters long; *c*.1825 glyphs, cf. Barthel 1958: 15; Fischer 1997: 409), the scribe's hand – and focus – perhaps got shaky to some extent, opting for economical and *not* very symmetric solutions. A long handwriting process is known to affect the stability and aesthetics of penmanship, resulting now and then in under-achievement and lack of scribal uniformity. We meet again the problem of how to distinguish with greater or lesser certainty between *artistic preferences*;[18] *functional differences related to text pragmatics*; *tool-writing medium interaction* and *fatigue error* of a given scribe across the RR corpus (see e.g. Melka 2012). However, despite spelling "deficiencies," it does not mean that A*b*7 is a corrupt passage, and hence less suited to be retrieved by the skills of a trained mind. To flesh out any missing piece of information, the chanter most likely relied on his background

[16] "*Mr. Salmon made further inquiries and he learned from an old chief that the tablets were used to send brief messages concerning important matters to chiefs in other villages when one did not wish to give these orally to a messenger…*"

[17] This is true *only* if irrefutably confirmed that "*Tahua*"'s side *a* is the *recto*, and side *b*, the *obverse* of the inscription (see Fischer 1997:. 411). Pozdniakov (1996: 299), based on the distribution of repetitive patterns on "*Tahua*" (see also Ávila Fuentealba 2007: 66-71) suggests two possible reading orders, A*a*1-A*a*8-A*b*1-A*b*8 and Ab1-A*b*8-A*a*1-A*a*8. Another concern relates to the chronology of "writing": text A is relatively a large sample, and if the inscription was done in one uninterrupted session, or in several sessions of carving under different conditions, this remains to be seen. The temporal structure of a text certainly allows for stylistic modifications.

[18] These are usually (but not absolutely) one-off scribal attempts, observed in different text portions.

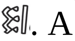
knowledge, on tactile qualities of the artifact, plus the nearby glyphic context of "*Tahua*." Thus, no matter how odd *rongorongo* orthography is, the eye derived comfort and speed from a conventional recurrence of the same visual impression for the same glyphs in all the writing the chanter read –specifically, A*a*5 vs. A*b*7–, to paraphrase Pulgram (1976 [1966]: 27). *Writing* never was a faultless mapping of the speech sounds,[19] much less of an Old system that resorts to a mixed bag of linguistic and pre-linguistic devices such as proto-Cuneiform, Zapotec, Maya glyphs, Cuneiform, or Egyptian hieroglyphs. The statement echoes Springer Bunk (2003 [2001]: 177), as she analyzes the Old Libyan-Berber inscriptions in Canary Islands, and similarly Urcid (2005: 10), while reflecting on the *inconsistencies and oddities from which traditional, antiquated* (i.e. *ancient*, my note) *orthographies suffer*. Even Finnish orthography recognized as highly regular, leaves out a great deal of phonetic information (Marshall Unger, DeFrancis 1995: 56, note 7; see also Harris 2010: 16).

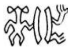
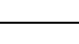
A possible meeting point of RR is with Old Egyptian, a system known for encoding extensively meaning-plus-sounds. Meltzer (1980: 53) suggests that Egyptian hieroglyphs entail *the mnemonic principle*. Seeing matters through his lens, the script does not offer a complete and accurate representation of the language utterances, rather than providing enough information to bring a phonetic or a semantic item unambiguously to mind. Another author, A. Gaur discusses several real-world memory aids, such as *Lukasa* of the Luba people of Zaïre, the *quipu* of the Incan Peru, genealogical staffs from Aotearoa (modern New Zealand), the Aztec "writing," the picture-writing of North American Indians, etc. In this respect, the option that *at even more sophisticated level*, *memory aids can be integral elements of an already* (partly) *phonetic script* is not ruled out (Gaur 1987 [1984]: 28). At any event, to guard against misapprehension, I need to specify that external shared attributes do not mean full equation between Egyptian hieroglyphics and *rongorongo*. Likewise, *no* insinuation is made thereof on *rongorongo*'s nature: that it stands for a pre-linguistic system totally free of speech. In the end, I should say that *minor or major variation* is common among several of the parallel passages in *rongorongo* texts (see also Barthel 1958: 167), and A*b*7 proves again the rule.

3.4. Search and compare - B*v*8

Weighed against the "original" sequence (A*a*5), the B*v*8 would seem to be *the most compressed glyphic form* or an *ellipsis*. The assessment does not suggest that "*Aruku Kurenga*"'s scribe strayed from the writing rules, rather than taking enough liberty to practice there some "short-hand" *rongorongo* due to familiarity with the topic (see also Harris 2010: 18, in a similar estimate). For a heuristic comparison, as an abbreviated instance of the so-called *speed writing*, we may evoke the Latin "*etc*" instead of the full-blown "*et cetera*." Horley (2009: 260,

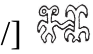
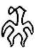
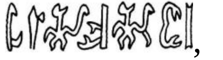
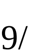
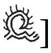
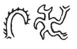
[19] Recall e.g. the "notorious" case of English orthography (see Rogers 2005: 190-195; Fengxiang, Altmann 2008).

Figure 7) reports a somewhat longer form of the sequence, counting the "prefacing" glyphs /165-54-1/ . A faintly different segmentation shows a *clef*-related sequence in the form of a partial duplication /54-165/ - /54-1/ . Glyph /165/ is a "*staff with two appended hooks*" in contrast to a "*stripped staff*" /1/ in the second mini-sequence. Codifying we obtain /A-B.^{x} $_{x}$ / - /A-B/ which slightly consorts with Barthel's (1958: 152) formula "a-b-a-c."

Compound  is notated in Barthel (1958) as /200.171/. As a matter of fact, part of this agglomerate can be deconstructed in the following components, /200.25.6.60/, with glyph /25/ –bound to the nucleus position. The deconstruction becomes plausible, if compared to the other reformatted sequences, in particular, when testing out /25.244/ in A*a*5 and /25-324/ in A*b*7. "Body" of /200/ and /300/ glyph-series (cf. A*a*5 and A*b*7) is omitted in B*v*8, with "*lower limb*"-glyph /60/ and a "hand"-glyph /6/ conjoined to the right half of /25/ (see also Ávila Fuentealba 2007: 51). This type of glyphic interaction has been reported earlier: Pozdniakov (1996: 296, Fig. 4c) describes several glyphic fusions and deletions of a similar nature. All in all, interchanging "hand"-glyph /6/ and "forked hand"-glyph /64/ hint at complying with a related function in many RR contexts (see Barthel 1958: 155;. Pozdniakov 1996: 295-296; Guy 2006: 62, Melka 2014). In the same way, series /200/ and /300/ often indicate allography, i.e. substitutable glyphs (see e.g. Barthel 1958: 273; Pozdniakov 1996: 295). Inclusion of final glyph /60/ in the sequence owes to the fact that a similar element is soldered in the "body" of /492/, unmarked however in Barthel (1958). At this juncture, the equation /492/ * /670-60/ is quite suggestive, both standing together for dividing boundaries within text B. The point worth noting here is that *mirroring symmetry*[20] is roughly visible: were not for a sharp sight and comparison of parallel passages, let alone having been formally instructed in *rongorongo*, we must tell that glyph-balance and -reversibility in B*v*8 are upset. Even if clear symmetry is running low, the *bare bones of the message* (or of *the formulaic chant*, for that matter) are present and able to be seen, that is, /200.25.6.60/ . "*Aruku Kurenga*"'s scribe applied the principle of *pars pro toto* by which *an item is rendered by a prominent part of a whole*;[21] explicitly, via "hand"-glyph and "lower limb"-glyph he delivers the complete human-like forms of /244/ (A*a*5) and /324/ (A*b*7). This graphic convention is not unusual in *rongorongo*, see e. g. Guy (2006: 59) about string /60-10.8/ standing for /275.8/  in C*a*12; Guy (2006: 57; Figure 3) on "masks," "head" and "arm"-

[20] *Mirroring symmetry*, in the sense as applied on the "original" sequence A*a*5.

[21] See Prem (1992: 54). *Pars pro toto*, structurally likening the metonymy, is common in graphic and symbolic systems, for instance, in Aztec (Prem 1992: 54); in Olmec iconography (Houston 2004: 284); in Mayan glyphs (Zender 2005: 1); or in the case of Egyptian hieroglyphs where a simple eye-ball, for instance, was a used as variation on the Eye of Horus glyph (Kinnaer 2009).

elision in comparing P*r*4 and H*r*5; and H*r*1 and P*r*1; Ávila Fuentealba (2007: 6, Fig. 5), with regard to a full juxtaposition /510.390/ [i.e. /513:548.79:546/]  in S*a*4, reduced through *pars pro toto* to /515/ [i.e. /513:79/] in E*v*6. The "sockets & nose," plus the double "curves" have rid the need of the "bodies," present in S*a*4; Melka (2008) and Horley (2009: 254, Fig. 3) on a reduplication in B*r*4-5, /200*-63s-300.63-79.1-300.63s-79-1/ , where the complete glyph of the "*double serpentine gaping-mouth*" /79/  is shortened before in the guise of two "*claws*" or "*gaping mouths*," attached to the stick-like sign in glyph /13/ (according to Barthel 1958) or /79.1/ after deconstruction; Horley (2010: 53, Fig. 9) on glyph /59/ [/59f:60/ ] in R*a*5, in which the lower "limb" /60/ surrogates the whole "body" glyph /324/ of the juxtaposition /59-324/  in E*v*6.

Evidence collected on the scrutiny of A*a*1, A*a*5, A*b*7, B*v*8, and R*a*1 points to the fact that Barthel's (1958) coding is repeatedly ambiguous and misleading (Pozdniakov 1996: 297; Fischer 1997: 218-219; Guy 2006: 59; Ávila Fuentealba 2007: 107; Davletshin 2012a: 246). For all its virtues as a scientific effort in cataloguing the mixture of RR signs, it misjudges simple glyphs and their spelling variants, complex glyphs and their components, parallel sequences, their relationships and syntactic order, by formalizing unneeded *chaos* in the system.

Short of a better transliteration system (*v. supra*, Footnote 3), the criterion for establishing parallels is met and improved if we visually and structurally compare these sequences. The conclusion echoes in quite the same way a few observations based on other sequences (cf. Melka 2012).

Notwithstanding the *distinctive features* and *tablet size* where the contrasted sequences occur – "*Tahua*"'s text is engraved on a yard-long European or North American oar; "*Aruku Kurenga*" on a 41 cm oblong piece of wood, with rounded edges and one end shaped as a handle; "*Small Washington*" on a 24 cm long object with one flat edge and a bulging opposite edge (see Fischer 1997) – I am of the mind that RR scribes were rehearsing, or pursuing an optimal choice, or some kind of concordance among these available strings, to the detriment of strict linearity of the same glyphs. If a "spoken sentence" was rendered in the sequences, then, it was not completely rendered (at least, in three of them), rather than omitting some fractions. I find inevitable to suggest that the senior RR scribe[22] availed himself of the proficiency of the Old Pascuan language and rested upon previously known indigenous traditions. He might have equally resorted to the specific tablet context, or to its physical characteristics (v. *supra*) to connect the missing semantic and/or phonetic units.

Chadwick (2000 [1958]: 131) has some astute lines when tackling the decipherment of Linear B. Although the scripts are far removed from each other, it gives us an approximation about the *rongorongo* sequences, "*It may also seem*

[22] If we account for the expertise applied in the carving of "*Tahua*" and "*Aruku Kurenga*" tablets, the term "*senior scribe*," although speculative, is most likely *correct*.

odd that such a useful invention as writing should be confined to such humdrum uses. Why should not letters, histories or even poems have been written down? The clumsiness of the script imposes a limitation; we may question how far a document in linear B would be readily intelligible to someone who had no knowledge of the circumstance of its writing. It is rather like shorthand; the man who wrote it will have little difficulty in reading it back. But a total stranger might well be puzzled, unless he knew what the contents were likely to be."

4. Discussion

At the outset, I should say that the RR scribes did not wander off track when inscribed, or intended through a metonymical relation to inscribe symmetries in several glyph-sequences. Such sequences, appearing to correspond to *full* or *partial palindromes* are generally and formally pleasing and raise one important question: *to what purpose served the same, or almost the same measurable order and what was the motivation behind the use*?

Leaving aside the assumption about its *notational* or *mnemonic* nature, several researchers believe that RR is *a functional script*, if not regularly sound-based, then partly sound-based: *logo-syllabic*, or *an emergent form of writing* with a nascent phoneticism (Butinov, Knorozov 1957 [1956]; Barthel 1958; Pozdniakov 1996; Macri 1996; Fischer 1997; Guy 1985, 2006; Horley 2005; Harris, Melka 2011a; Davletshin 2014). Following the rationale, it is reasonable to look for correlates in Rapanui or in other Polynesian languages. As long as we find in the data credible palindromes of different length, related to the Old RR genres (see Routledge 1919; Barthel 1958; Fischer 1997), the situation is potentially explicable. On the contrary, other alternatives should be pursued. P. Horley's (2007: 25) opinion properly covers both ends of the current discussion,

"*Until similar constructions* (*palindromes*, my note) *could*
be found in Rapa Nui lore, it seems to be safer to exclude
such fragments from the corpus intended for statistical analysis."

The cited *analysis* is based on the assumption of phonetic elements in RR.

If we turn to *antimetabole* for a possible explanation (Davletshin 2012b), then we have a quasi-symmetric format in A*a*5, applicable thereof; whereas A*a*1, A*b*7 and B*v*8 parallels appear misconfigured and do not fit the figurative device. Possibly, by long practice, RR scribes could have attained to some readiness in recognizing contextual clues in the strings, and in reciting the full form by filling missing pieces. Referring to the Scriptures (Genesis 9:6) we extract the full-size parable

Z 3 2 1 1 2 3
"Who *sheds the blood* of *a man*; by a man shall *his blood* be *shed*..."

Applying tentatively some RR-like variation and permutation in the cited *anti-metabole*, we obtain,

<blood of man, by a man shall blood shed>

<shed blood, man bloodshed>.

Despite the absent bits, there are still meaningful linkages among the lexical components, and most likely, a Bible connoisseur would have not much trouble in elaborating and reconstructing the original sequence.

Similarly, it is proper at this point referring to Greber's (1996) description, "*Largely a visual phenomenon, the palindrome epitomizes the spatiality of language and scripture...* (*script*, my note)." Spatiality is also interpreted as a evenhanded distribution of characters or figures in a specific text to emphasize *harmony* and *liveliness*. Upon which, from the perspective of a phonetic Latin-based script, palindromes -and ambigrams- (…*a special*, *restricted case of anagram*, see Greber 1996), playful and ludic as they are,[23] are known to depict balance and symmetry aiming at *perfection*, *aesthetic pleasure*, *circularity* and *eternity* (cf. Borgmann 1985; Greber 1996; Darvas 2007).

Rongorongo is still an unknown script; hence we cannot suppose nor gain comfort from what might be real and/ or probable in alphabetic systems, e.g. English, e.g. <*racecar*>; Spanish, e.g. <*reconocer*> [recognize, admit]; or Welsh, <*llad dafad dall*> [kill a blind sheep], should automatically translate into the Old Pascuan script. Before we can say much about the possibility of phonetic palindromes, a comparison among such structures present in speech-bound and speech-free expressions is potentially of assistance. In this vein, more work can be done to quantitatively express the measure of symmetry; to ascribe it to these glyphs and state the distribution of symmetry of the glyph *types*.[24]

Palindromes exist sporadically in the Pascuan language. Given the characteristics of a Polynesian language with a syllable structure (C)V,[25] we expect them to begin and uniformly end with the same vowel, e.g. <*ana*> [cave]; <*epe*> [earlobe]; or to be doublings, <*áka-áka*> [bungle, do badly]; <okooko> [take all] (see Englert 1978 [1948]). Indeed, additional palindromic words are present in Rapanui (see Englert 1978 [1948]; Fuentes 1960), or in other Polynesian languages, e.g. <*aninanina*> [giddy, frivolous] in Māori; <*anonanona*> [ant] in Hawaiian (cf. Borgmann 1985), *Ele*'*ele* [in Polynesian mythology, *the first woman*], *Eleele* [a Hawaiian town]; in Mangarevan, <*akaka*> [an unusually low tide], <*omoomo*> [suck], <*ekakekake*> [waves of the sea coming and going back on the beach] (Grant 1986), etc. However, it seems quite likely that Polynesian speakers would not have been aware of them as "*palindromes*." That only becomes obvious when one transcribes them at the segmental level, which

[23] See e.g. Borgmann (1965), and Filloy (2005).

[24] I wish to thank Altmann (2010) for the tip.

[25] See du Feu (1996: 186); and Fischer (1997: 204) on phonotactic restricttions in Rapanui language.

requires a theory of segmental (phonemic) transcription that ancient Polynesians evidently lack. For instance, one may imagine a speaker of Hawaiian missing that *anonanona* is a palindrome. Palindromes at the syllable or word level would presumably have been more recognizable,[26] considering also the fact that wood carving (i.e. inscribing) as a process, generally commands more time and concentration than the pure verbalization of thoughts (i.e. speaking). Since RR may suggest (at best) mixed elements, that is, intermittent syllables, several word-signs, complements and/or semasiograms, any prominent palindrome (e.g. N*a*3) perhaps indicates a purposely devised structure by the Old literate as speaker of a Polynesian language. Whether for fast and better access, for prestigious and aesthetic display, or for "throwing" a pun in-between the textual passages, such a salient feature calls for attention.

But as yet, the consulted literature (Englert 1978 [1948]; Fuentes 1960) has not yielded long, meaningful sentences constructed in a palindromic-like fashion and fitting the assumed ritualistic and cult practices "written" on the RR tablets. For that reason, the visual aspect may well gain importance. The idea is in line with Pozdniakov (2011: 58) that such *mirrored graphemes* (see N*a*3, Figure 1, #2), *most possibly*, *do not have any phonetic basis*. Still, it does not mean that they are devoid of significance. D. Hofstadter in examining the common aspect of repetition and symmetry in the works of three intellectuals (S. Bach; M. C. Escher, and K. Gödel) insinuates that isomorphic properties *lead to symbolic associations of meaning, which in turn are parts of larger formal systems* (Washburn, Crowe 1988: 10). As for RR, the pattern symmetries may represent cultural relationships and concepts, metaphorically embedded (see e. g. Washburn 2004b: 54).

We should realize that if *visual enhancement* of the text was intended (Horley 2011), the implicated symmetry may be thought of as tradition-bound descriptive "pictures" in a running RR text, supposing the remaining text encodes speech units. As we statistically compute RR in search of phonetic mappings, we must behold and reposition ourselves with regard to these specific constructions. Skewed results could occur, adding more doubt to the *rongorongo* "puzzle." Now, we may also toss aside with skepticism the comments on the *visual nature* of the palindromes in RR, since they are open to personal "readings," spurring increased disagreement and a penchant for the mnemonic-like model in *rongorongo*. The only real way to handle the *phonetic* and/or the *visual* content of palindromes is testing them out all across the extant corpus and let researchers have their say. Yet, in the face of uncertainty we need patience, as we accordingly refine and update the investigation. Such a fact leaves the RR discipline *in a flux*, with many hypotheses vying with each other and claiming feasibility (Métraux 1940; Butinov, Knorozov 1957 [1956]; Barthel 1958; Guy 1990; Fischer 1995; Pozdniakov 1996; Horley 2005, 2011; Pozdniakov, Pozdniakov 2007; de Laat 2009; Wieczorek 2011, 2014; Harris, Melka 2011a; Davletshin 2002, 2012, 2014; Melka 2012, 2013).

[26] Thanks to Sproat (2013) for the observation.

Verifying with laser precision a mirrored scribal choice, whether motivated by *a functional* and/or an *aesthetic necessity* (Pozdniakov 2011) *vs*. a real palindrome motivated by mere whim, would be quite mind-boggling. The statement rings true, especially for non-deciphered scripts with limited data, e.g. *rongorongo,* Isthmian, Zapotec, proto-Biblian, etc (cf. Melka 2009a). It can be concluded that palindromic reversions are not exclusive to Easter Island's classic script, see Brice (1976: 38-40) regarding the *Cretan Linear A* and *Iranian proto-Elamite,* and Urcid (2005: 12), on the Old Zapotec. The comparative framework is in any case exploratory, and we wonder at this point about the intersecting similarities in these distanced Old scripts. Evidently, there are no inherited or linguistic affiliations among Cretan Linear A, proto-Elamite, Zapotec and *rongorongo*.

5. Conclusion

The palindromic-like structures, inserted at irregular intervals within the inscription by RR practitioners with no regard to the side of the tablet, are noteworthy accomplishments. As humans are very sensitive in general to *balance* and *symmetry* features (Livi 2006; du Sautoy 2008; Sparavigna 2008), these formations are distinguishable from the rest of other RR text chunks and passages.

The described interactions in these structures (palindromes, palindromic-like and half-palindromes) are more than often the result of a permutation, and/or of an elision, if we dismiss unintended error, distractions, and/or occasional weariness. Such effects appear not to distort their conceptual and semantic arrangement, and may be explicated in terms of aesthetic choices, reshaping and of a malleable orthography, i.e. of the scribes' privilege by "trading off" and deleting glyphs (or parts of them) while fine-tuning the inscription. In my opinion, the recorded information whether tightly "packed" or almost wholesale, does not mean *flawed literacy* or *a secret encryption*, rather than an idiosyncratic way to present it. For some aspects that could not be retrieved or predicted by language alone, the RR scribes had to consider the artifact being chanted, the surrounding conditions, the religious and social relations in which they were immersed. Thus, from the perspective of the "full-time" scribes of *tablets A* and *B,* or of other co-fellows, the visual and/or "kinetic"-like features set within the context made the sequences *legible* and *identifiable* as parallels. It may also be inferred that by practice and repetition the scribes bettered and explored the expressive potential of the script. Given the time, certain scribal norms would have been consolidated and stabilized among the *surplus of graphic variants* (Melka 2009a: 119), giving way even to a more cursive form of writing. As the *rongorongo* phenomenon was disrupted in the sixties of the 19th century due to several factors (Melka 2012), this must remain however an educated guess.

It is acknowledged that under different sampling conditions (a smaller or a larger corpus), hypothesis might have turned different. However, considering the randomness of the extant corpus both in terms of time-depth and provenance (see Fischer 1997; Melka 2009a), any particular palindrome suggests that its realization was *not* a short-lived event, neither a fortuitous shred of evidence, rather than a diagnostic feature of the script. If these patterns of use are unproblematic for outsiders to discern, then, for the qualified Old Rapanui scribe would have been quite a cinch. Palindromes offer further support that *rongo-rongo* was not immune neither to scribal puns, nor to a remarkable aesthetic and calligraphic creativity.

Without alleging full commonality among them, comparison with other sign systems serves better the assessment of *rongorongo*. As each system has characteristics influenced by their specific socio-cultural grounds, geography and chronology, matters cannot be entirely generalized. Evaluating the discussed RR glyphs only on their outer shape is fraternizing with bias. The picture-like quality does not mean per se *pictography* in each and all the contexts. Whether the amount of speech is *pervasive*, *marginal* or *nil*, this is another issue. Suffice it to say that phonetic signs may transmit imagery with special vigor, or *vice versa*, apparently script-like signs (but that are not), illustrate purely emblematic or decorative sets (see Sproat 2013). As of now, there is not a clear-cut line in *a priori*[27] discriminating systems that encode speech *vs*. systems that do not.

Mirror-like patterns and texts are neither a shift away in the Old Pascuan script or exclusive of it: they are evident in several linguistic and non-linguistic systems and designs, dating back to pre-history and historical times (see examples in Borgmann 1965: 71; Brice 1976: 39-40; Barlow, Reeves 1979; Hofstadter 1999 [1979]; Makkay 1984: 143, Fig. 17; Washburn, Crowe 1988; Farmer et al. 2004; Filloy 2005; Urcid 2005: 12; Sparavigna 2008: Fig, 1; LDL 2008-2012; Bedrick et al. 2012: 57; Saltveit 2012; Wikipedia 2013b). Along with playfulness, a primary formative element in human culture (Huizinga 1955 [1938]), they seem to comply with a human need for better data perception and organization.

Acknowledgements

I am indebted to predecessors who seriously took up the study of *rongorongo* (and not only). To surrender to expediency, special thanks are given to the following people for exciting ideas: G. Altmann, F. Ávila Fuentealba, A. Davletshin, R.W. Sproat, and R.M. Wieczorek.

References

Ávila Fuentealba, F. (2007). *Ensayo de Estudio Visual de las Tablillas Rongorongo.* Unpublished Manuscript: Temuco, Chile.

[27] By this, we understand the signs’ shape, combinatorial properties of the signs, and their text distribution.

Barlow, H.B., **Reeves, B.C.** (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research* 19, *783-793*. Available at:
http://www.sciencedirect.com/science/article/pii/0042698979901548 (accessed May 23, 2012).

Barthel, T.S. (1958). *Grundlagen zur Entzifferung der Osterinselschrift*. (Abhandlungen aus dem Gebiet der Auslandskunde 64, Reihe B.) Hamburg: Cram, de Gruyter & Co.

Barthel, T.S. (1963). Rongorongo-Studien (Forschungen und Fortschritte bei der Weiteren Entzifferung der Osterinselschrift). *Anthropos 58, 372-436.*

Barthel, T.S. (1971). Pre-Contact Writing in Oceania. In: Sebeok, T. et al. (eds.) *Linguistics in Oceania: Current Trends in Linguistics*. *Volume 8: 1165-1186*. The Hague-Paris: Mouton.

Barthel, T.S. (1993). Perspectives and Directions of the Classical Rapanui Script. In: Fischer, S. R. (ed.) *Easter Island Studies*: *Contributions to the History of Rapanui in Memory of William T. Mulloy*. *Oxbow Monograph 32: 174-176*. Oxford: Oxbow Books.

Bedrick, S., **Beckley**, **R.**, **Roark B.**, **Sproat**. **R.** (2012). Robust *kaomoji* detection in Twitter. In: *Workshop on Language and Social Media*. Montreal, Canada.
Available at: http://aclweb.org/anthology-new/W/W12/W12-2107.pdf (accessed June 20, 2013).

Bible. (n. d.) *Genesis*: *Chapter 9*, *Verse 6*. Available at:
http://bible.cc/genesis/9-6.htm (accessed June 22, 2013).

Borgmann, D.A. (1965). *Language on Vacation*: *An Olio of Orthographical Oddities*. New York: Charles Scribner's Sons.

Borgmann, D.A. (1985). *The Majestic Palindrome*. Available at:
http://digitalcommons.butler.edu/cgi/viewcontent.cgi?article=2943&
context=wordways (accessed March 20, 2012).

Brice, W.C. (1976). The Principles of non-Phonetic Writing. In: Haas, W. (ed.), *Writing without Letters*. *Mount Follick series*. *Volume Four: 29-44*. Manchester: Manchester University Press.

Butinov, N.A., **Knorozov**, **Y.V.** (1957). Preliminary Report on the Study of the Written Language of Easter Island. *Journal of the Polynesian Society 66(1), 5-17.*

CEIPP, Centre (formerly Cercle) d'Études sur l'Île de Pâques et la Polynésie [Study Centre (formerly Circle) of Easter Island and Polynesia] (2005).
Available at:
http://www.rongorongo.org/ (accessed January 23, 2008).

Chadwick, J. (2000 [1958]). *The Decipherment of Linear B*. Cambridge: The Press Syndicate of the Cambridge University.

Cherry, C. (1966 [1957]). *On Human Communication*: *A Review, a Survey and a Criticism*. Cambridge, MA: The MIT Press.

Craig, R. D. (2004). *Handbook of Polynesian Mythology*. Santa Barbara, CA: ABC-Clio.

Daniels, P. T., **Bright**, **W**. (eds.) (1996). *The World's Writing System*. Oxford, NY: Oxford University Press.

Darvas, G. (2007). *The Interpretation of the Concept of Symmetry in Everyday Life*, *Science and Art*: *Symmetry*, *Invariance*, *Harmony*. Basel: Birkhäuser Verlag AG.

Davletshin, A. (2002). *Names in the Kohau Rongorongo Script*. Paper presented as From Kohau Rongorongo Tablets to Rapanui Social Organization: From Rapanui Social Organization to Kohau Rongorongo Script at the 2nd International Conference "Hierarchy and Power in the History of Civilizations," Saint Petersburg, Russia, July 4-7, 2002.

Davletshin, A. (2012a). Numerals and Phonetic Complements in the '*Kohau rongorongo*' script of Easter Island. *Journal of the Polynesian Society 121 (3), 243-274.*

Davletshin, A. (2012b). *Literary approach to structure and content of* Kohau Rongorongo *texts of Easter Island*. 8th International Conference on Easter Island and the Pacific. Santa Rosa, CA, 8-13 July, 2012.

Davletshin, A. (2014). Word-signs and Sign groups in the *Kohau Rongorongo* Script of Easter Island. In: Conrich, I., Mueckler, H. (eds.), *Easter Island*: *Cultural and Historical Perspectives*. *Proceedings of the Symposium on "Easter Island*: *Cultural and Historical Perspectives"*, 19 November 2010. Chilean Embassy, London. Wien, Zürich: LIT Verlag GmbH & Co. KG, forthcoming.

DeFrancis, J. (1989). *Visible Speech*: *The Diverse Oneness of Writing Systems*. Honolulu, HI: University of Hawai'i Press.

de Laat, M. (2009). *Words out of Wood*: *Proposals for the Decipherment of the EasterIsland Script*. Delf, The Netherlands: Eburon Academic Publishers.

du Sautoy, M. (2008). *Symmetry*: *A Journey into the Patterns of Nature*. New York: HarperCollins.

Englert, S. (1978 [1948]). *Idioma Rapanui*: *Gramática y Diccionario del Antiguo Idioma de la Isla de Pascua*. Santiago de Chile: Ediciones de la Universidad de Chile.

Eyraud, E. (1866). Lettre du Frère Eugène Eyraud, au T. R. P. Supérieur Général de la Congrégation des Sacrés-Coeurs de Jesús et de Marie. Valparaíso, décembre 1864. *Annales de la Propagation de la Foi 38, 52-71, 124-138*. Lyon.

Farmer, S., **Sproat, R.**, **Witzel, M.** (2004). The Collapse of the Indus-Script Thesis: The Myth of a Literate Harappan Civilization. *Electronic Journal of Vedic Studies (EJVS) 11(2), 19-57.*

Fengxiang, F. Altmann, G. (2008). Graphemic Represenation of English Phonemes. In: Altmann, G., Fengxiang, F. (eds.), *Analyses of Script*: *Properties of Characters and Writing Systems*. *Quantitative Linguistics 63: 23-57* . Berlin, New York: Mouton de Gruyter.

Filloy, J. (2005). *Karcino*: *Tratado de Palindromía*. Buenos Aires: El Cuenco de Plata.

Fischer, S. R. (ed.) (1993). *Easter Island Studies*: *Contributions to the History of Rapanui in Memory of WilliamT. Mulloy*. *Oxbow Monograph 32*. Oxford, UK: Oxbow Books.

Fischer, S. R. (1995). Preliminary Evidence for Cosmogonic Texts in Rapanui's Rongorongo Inscriptions. *Journal of the Polynesian Society 104 (3), 303-321.*

Fischer, S. R. (1997). *Rongorongo, the Easter Island Script*: *History, Traditions, Texts*. Oxford, NY: Oxford University Press.

Fuentes, J. (1960). *Diccionario y Gramática de la Lengua de la Isla de Pascua*. Santiago de Chile: Editorial Andrés Bello.

Gaur, A. (1987 [1984]). *A History of Writing*. London: The British Library.

Geiseler, W. (1995 [1883]). *Geiseler's Easter Island Report*: *an 1880s Anthropological Account*. *Kapitänleutnant Geiseler, Die Osterinsel: eine Stätte prähistorischer Kultur in der Südsee (Beiheft zum Marine-Verordnungsblatt Nr. 44)*. Berlin. Translated by W. Ayres and G.S. Ayres. Social Science Research Institute, UH. Honolulu: University of Hawai‘i at Manoa.

Grant, J. (1986). *The Palindromes of Mangareva*. Available at: http://digitalcommons.butler.edu/cgi/viewcontent.cgi?article=3215 &context=wordways (accessed February 6, 2014)

Greber, E. (1996). *A Chronotope of Revolution*: *The Palindrome from the Perspective of Cultural Semiotics*. Available at: http://www.palindromist.org/chronotype (accessed January 14, 2012).

Guy, J.B.M. (1985). On a Fragment of the "*Tahua*" Tablet. *Journal of the Polynesian Society 94, 367-387.*

Guy, J.B.M. (1990). On the Lunar Calendar of tablet 'Mamari.' *Journal de la Société des Océanistes 91 (2), 135-149.*

Guy, J.B.M. (2006). General Properties of the *Rongorongo* Writing. *Rapa Nui Journal 20 (1), 53-66.*

Haas, W. (ed.) (1976). *Writing without Letters*. *Mount Follick series. Volume Four*. Manchester: Manchester University Press.

Haberlandt, M. (1886). Über Schrifttafeln von der Osterinsel. *Mitteilungen der Anthropologischen Gesellschaft in Wien 16, 97-102.*

Harris, M. (2010). *Corpus linguistics as a method for the decipherment of* rongorongo. Mres in Applied Linguistics. London: Birkbeck University. Available at: http://www.academia.edu/243385/Corpus_linguistics_as_a_method_for_ the decipherment_of_*rongorongo*_Mres_Dissertation_(accessed June 20, 2012).

Harris, M., Melka, T.S. (2011a). The Rongorongo script: On a listed sequence in the *recto* [*verso*, repaired] of tablet '*Mamari*.' *Journal of Quantitative Linguistics 18 (2), 122-173.*

Hofstadter, D.R. (1999 [1979]). *Gödel, Bach, Escher*: *The Eternal Golden Braid*. New York: Basic Book, Inc.

Horley, P. (2005). Allographic Variations and Statistical Analysis of the *Rongorongo* Script. *Rapa Nui Journal 19 (2), 107-116.*

Horley, P. (2007). Structural Analysis of *Rongorongo* Inscriptions. *Rapa Nui Journal 21 (1), 25-32.*

Horley, P. (2009). *Rongorongo* Script: Carving Techniques and Scribal Corrections. *Journal de la Société des Océanistes 129 (2), 249-261*.

Horley, P. (2010). *Rongorongo* Tablet Keiti. *Rapa Nui Journal 24 (1), 45-56.*

Horley, P. (2011). Lunar Calendar in *rongorongo* Texts and Rock Art of Easter Island. *Journal de la Societé des Océanistes 132 (1), 87-107*.

Houston, S. D. (2004). Writing in Early Mesoamerica. In: Houston, S.D (ed.), *The First Writing*: *Script Invention as History and Process: 274-309*. Cambridge, UK: Cambridge University Press.

Huizinga, J. (1955). *Homo ludens* [Man, the Player]: *A Study of the Play-element in Culture*. Boston: Beacon Press.

Kinnaer, J. (2009). *The Written Language*: *So-called* "*Ptolemaic*" *writing*. Available at:
http://www.ancient-egypt.org/index.html (accessed June 4, 2011).

Kjellgren, E. (2001). *Splendid Isolation*: *Art of Easter Island*. With contributions by J.A. van Tilburg and A.L. Kaepler. New York: The Metropolitan Museum of Art; New Haven, London: Yale University Press.

Kudrjavtsev, B.G. (Кудрявцев Б. Г.) (1949). Письменность острова Пасхи. *Сборник музея антропологии и этнографии 11, 175-221* [Pis'mennost Ostrova Paschi. *Sbornik Muzeja Antropologii i Etnografii 11, 175-221*].

LDL (2008-2012). *15 Inspiring Ambigram Logos*. Available at:
http://www.logodesignlove.com/ambigram-logos (accessed February 3, 2012).

Livi, M. (2006). *The Equation That Couldn't Be Solved*: *How Mathematical Genius Discovered the Language of Symmetry*. New York: Simon & Schuster.

Lounsbury, F.G. (1989). The Ancient Writing of Middle America. In: Senner, W.M (ed.), *The Origins of Writing: 203-237*. Lincoln. NE: University of Nebraska Press.

Lunde, P. (ed.) (2009). *The Book of Codes*: *Understanding the World of Hidden Messages*. Berkeley, Los Angeles: University of California Press.

Macri, M.J. (1996). *RongoRongo* of Easter Island. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 183-188*. New York, Oxford: Oxford University Press.

Makkay, J. (1984). *Early Stamp Seals in Early South-East Europe*. Budapest: Akadémiai Kiadó.

Marshall Unger, J., DeFrancis, J. (1995). Logographic and Semasiographic Writing Systems: A Critique of Sampson’s Classification. In: Taylor, I., Olson, D.R. (eds.), *Scripts and Literacy: 45-58*. Dordrecht: Kluwer Academic Publishers.

Melka, T.S. (2007-2008). Structural and Distributional Analysis of the *rongo-rongo* Text in Tablet “*Mamari*.” Manuscript in the collection of the author.

Melka, T.S. (2008). On some common glyphic features of tablet ‘*Aruku Kurenga*’ and other *rongorongo* tablets. Unpublished manuscript in the collection of the author.

Melka, T.S. (2009a). The Corpus Problem in the *rongorongo* Studies. *Glotto-theory*: *International Journal of Linguistics 2 (1), 111-136.*

Melka, T.S. (2009b). Linearity, Calligraphy and Syntax in the *Rongorongo* Script. *Glottotheory*: *International Journal of Linguistics 2 (2), 70-96*.

Melka, T. S. (2012). On a “kinetic”-like sequence in *rongorongo* Tablet “*Mamari*.” *Writing Systems Research*, iFirst. Available at: http://www.tandfonline.com/doi/full/10.1080/17586801.2012.742005

Melka, T.S. (2013). “*Harmonic*”-like Sequences in the *rongorongo* Script. *Glottotheory*: *International Journal of Linguistics 4 (2), 115-139*.

Melka, T. S. (2014). Distinctive Sequences in *rongorongo* Texts C, G, and B. In: Conrich, I., Mueckler, H. (eds.), *Easter Island*: *Cultural and Historical Perspectives*. *Proceedings of the Symposium on “Easter Island*: *Cultural and Historical Perspectives”*, 19 November 2010. Chilean Embassy, London. Wien, Zürich: LIT Verlag GmbH & Co. KG, forthcoming.

Melka, T.S., Harris, M. (2011). Possible List Content in the *rongorongo* Inscriptions. Manuscript in the collection of the authors.

Meltzer, E.S. (1980). Remarks on Ancient Egyptian Writing with Emphasis on its Mnemonic Aspect. In: Kolers, P.A., Wrolstad, M.E., Bouma, H. (eds.), *Processing of Visible Language: 43-66*. New York, London: Plenum Press.

Métraux, A. (1940). *Ethnology of Easter Island*. *Bernice P. Bishop Museum Bulletin 160*. Honolulu, HI: Bernice P. Bishop Museum Press.

Métraux, A. (1957). *Easter Island*: *A Stone-Age Civilization of the Pacific*. Trans. from French by Michael Bullock. New York: Oxford University Press.

Mordo, C. (2002). *Easter Island*. Auckland, New Zealand: Firefly Books Ltd.

Olmert, M. (1995 [1992]). *The Smithsonian Book of Books*. New York – Avenel, New Jersey: Wings Books / Random House, Inc.

Orliac, M. ,**Orliac, C**. (2008). *Trésors de l’île de Pâques* / *Treasures of Easter Island*. Collection de la Congrégation des Sacrés-Coeurs de Jésus et de Marie SS.CC. Genève: Éditions D, Paris: Éditions Louise Leiris.

Pozdniakov, K. (1996). Les Bases du Déchiffrement de l’Écriture de l’Ile de Pâques. *Journal de la Societé des Océanistes 103 (2), 289-303*.

Pozdniakov, I., Pozdniakov, K. (2007). Rapanui Writing and the Rapanui Language: Preliminary Results of a Statistical Analysis. *Forum for Anthropology and Culture 3, 3-36.*

Pozdniakov, K. (2011). Tablet *Keiti* and calendar-like structures in Rapanui script. *Journal de la Societé des Océanistes 132, 39-74.*

Prem, H.J. (1992). Aztec Writing. In: Reifler Bricker, V., Andrews, P.A. (eds.), *Epigraphy*: *Supplement to the Handbook of Middle America Indians*. *Vol. 5: 53-69* . Austin, TX: University of Texas Press.

Pulgram, E. (1976 [1966]). The Typologies of Writing-systems. In: Haas, W. (ed.), *Writing without Letters*. *Mont Follick series. Vol. 4: 1-29*. Manchester, UK: Manchester University Press.

Purse, L., Campbell, L. (2013). *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Ritner, R.K. (1996). Egyptian Writing. Section 4. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 73-84.* Oxford, NY: Oxford University Press.

Robinson, A. (2002). *Lost Languages*: *The Enigma of the World's Undeciphered Scripts*. New York: McGraw-Hill.

Rogers, H. (2005). *Writing Systems*: *A Linguistic Approach*. Oxford, UK: Blackwell Publishing Ltd.

Ross, A.S.C. (1940). The Easter Island tablet Atua-mata-riri. *Journal of the Polynesian Society 49 (196), 556-563.*

Routledge, K. (Mrs. Scoresby). (1919). *The Mystery of Easter Island*: *The Story of an Expedition*. London, Aylesbury: Hazell, Watson and Viney LD.

Saltveit, M. (2012). *The Palindromist*: *For People who WRITE and Read Palindromes*. Available at:
http://www.palindromist.org/ (accessed March 17, 2012).

Smith, R. (2009). Constructing Word Similarities in Meroitic as an Aid to Decipherment. *British Museum Studies in Ancient Egypt and Sudan 12, 1-10.*

Spaelti, P. (2012). *Philip Spaelti's Rongorongo pages.*
Available at:
http://kohaumotu.org/Rongorongo/ (accessed September 19, 2013).

Sparavigna, A.C. (2008). *Symmetries in Images on Ancient Seals*.
Available at:
http://arxiv.org/ftp/arxiv/papers/0809/0809.3566.pdf (accessed September 12, 2013).

Springer Bunk, R.A. (2003 [2001]). *Origen y Uso de la Escritura Líbico-Bereber en Canarias*. Segunda Edición. Arafo-Tenerife, España: Litografía Romero S.L.

Sproat, R. W. (2003). *Approximate String matches in the RR Corpus*.
Available at:
http://www.cslu.ogi.edu/~sproatr/ror/ (accessed September 5, 2013).

Sproat, R.W. (2013). *Corpora and Statistical Analysis of Non-Linguistic Symbol Systems*. Available at: http://rws.xoba.com/monograph.pdf (accessed December 4, 2013).

Urcid Serrano, J. (2005). *Zapotec Writing*: *Knowledge, Power and Memory in Ancient Oaxaca*. Available at: http://www.famsi.org/zapotecwriting/zapotec_text.pdf (accessed February 9, 2013).

Wagemans, J. (1995). Detection of visual symmetries. *Spatial Vision 9, 9-32*. Available at: https://lirias.kuleuven.be/bitstream/123456789/126480/1/Wagemans+ Spatial + Vision +1995.pdf (accessed November 22, 2013).

Washburn, D.K., Crowe, D.W. (eds.) (1988). *Symmetries of Culture*: *Theory and Practice of Plane Pattern Analysis*. Seattle: University of Washington Press.

Washburn, D.K. (ed.) (2004a). *Embedded Symmetries*: *Natural and Cultural*. Albuquerque: University of New Mexico Press; published in cooperation with the Amerind Foundation.

Washburn, **D.K.** (2004b). The Genesis of Realistic and Patterned Representa--tion. In: Washburn, D.K. (ed.), *Embedded Symmetries*: *Natural and Cultural: 47-59*. Albuquerque: University of New Mexico Press; published in cooperation with the Amerind Foundation.

Wieczorek, R.M. (2011). The Double-Body Glyphs and Palaeographic Chronology in the Rongorongo Script. *Rapa Nui Journal 25 (2), 31-40*.

Wieczorek, R.M. (2013). *Rafal Wieczorek - rongorongo page*: *rongorongo literature*. Available at: http://flint.sdu.dk/index.php?page=rongorongo-literature (accessed December 14, 2013).

Wieczorek, R.M. (2014). Putative duplication glyph in the rongorongo script. *Cryptologia*, forthcoming.

Wikipedia (in English). (2013a). *Rongorongo*. Available at: http://en.wikipedia.org/wiki/Rongorongo (accessed November 11, 2013).

Wikipedia (in English). (2013b). *Triple spiral*. Available at: http://en.wikipedia.org/wiki/Triple_spiral, (accessed November 11, 2013).

Zender, M. (2005). Teasing the Turtle from its Shell: AHK and MAHK in Maya Writing. *The PARI Journal 6 (3), 1-14*. Available at: http://www.mesoweb.com/pari/publications/journal/603/Turtle_e.pdf (accessed December 8, 2013).

Distribution of the Menzerath's law on the syllable level in Greek texts

Georgios Mikros, Jiří Milička

1. Introduction

Since 1980, when Gabriel Altmann published his famous article on the Menzerath's Law (Altmann 1980), the law has been corroborated on many linguistic levels and even non-linguistic material inspiring generations of linguists. Both Menzerath's Law and Altmann's equation assume that the relation between construct length and the constituent length is a monotonic decreasing function like the function depicted in Fig. 1.

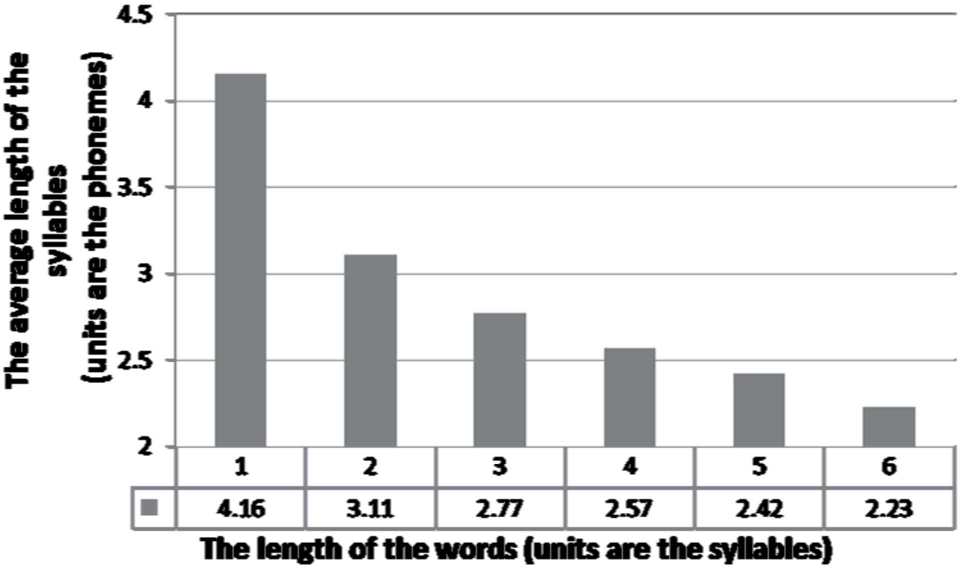


Figure 1. Menzerath's Law on the word-syllable-phoneme level in a German text (Source: Altmann et al. 1989: 38, Table 5.1a).

Counterexamples to this law can be found, for example in the relation between the length of a Greek word and its corresponding syllables in this text (Fig. 2). But as the single counterexample could be only a random fluctuation from a trend (especially here, the text contains only 4 word tokens with 10 syllables), the following hypothesis needs to be tested:

H1: *On average, the function measured on a sample of texts is monotonly decreasing.*

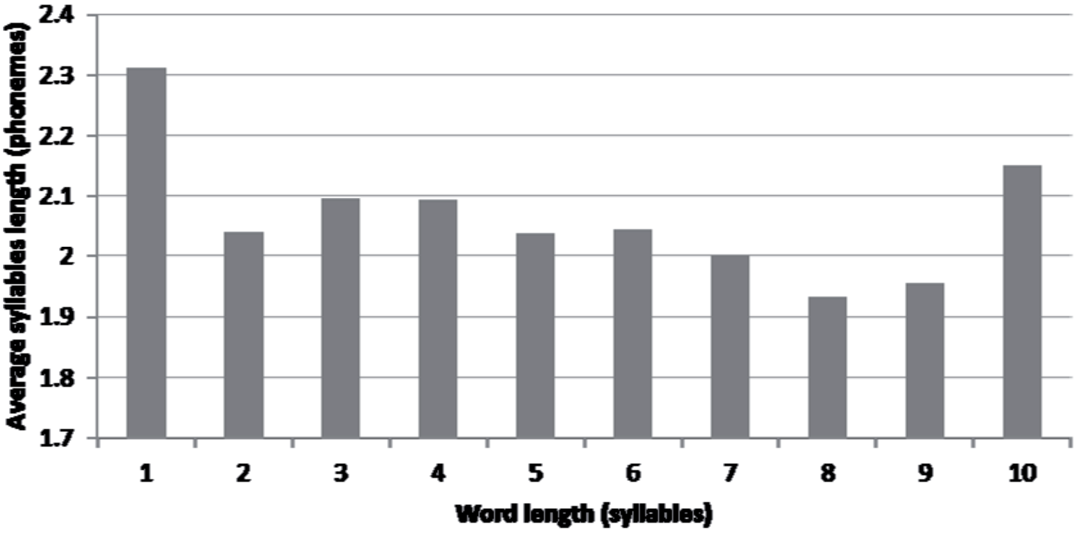


Figure 2. Menzerath's Law on the word-syllable-phoneme level in a Greek text[1].

2. Material and methods

Modern Greek is one of the least quantitatively studied modern European languages. Zipf's law and some basic quantitative characteristics in various linguistic elements have been measured in Hatzigeorgiu et al. (2001) and Mikros et al. (2005). However, we don't have any published research focusing on Menzerath's law and its application to Modern Greek data[2]. This is the focus of the present research and in the following sections, we will describe the corpus, tools and methods used in this study.

2.1. The HNC corpus

The corpus used is the Hellenic National Corpus (HNC). HNC was developed by the Institute for Language and Speech Processing (ILSP) (Hatzigeorgiu et al. 2000) and is an ongoing effort. It currently contains more than 50,000 written MG texts, published from 1976 on, totaling 47 million words.

[1] The specific text is a scientific monograph belonging to the broader discipline of Epistemology (Title: "The peel of apricot", 2000, Ed. Ellinika Grammata: 526).

[2] The only relevant research that we are aware of is the unpublished MA thesis of Zenetzi and Papachristos (2013). However, in this study Menzerath-Altmann's law has been studied in each text of a Modern Greek corpus separately using a very different methodology from the methodology that most quantitative linguistic studies use. A regression model was applied to the corpus data (mean word length in syllables per text and mean syllable length in letters per text) yielding a fit which confirmed a weak correlation of the investigated relationship.

Texts in HNC are classified according to PAROLE standards (PAROLE 1995), which follow the TEI (Sperberg-McQueen, Burnard 1994) and EAGLES (EAGLES 1994) guidelines. Texts are classified as regards to Medium, Genre, Topic, Detailed Genre, Detailed Topic and bibliographical information. As far as Medium is concerned, texts are classified into four categories, according to their source. The current percentage of words for each one of the four categories can be seen in Table 1.

Table1
Percentage of words according to text medium.

Text Medium	Percentage
Newspaper	61.29%
Miscellaneous	23.08%
Book	9.41%
Magazine	5.89%
Internet	0.32%

We used a “clear” version of HNC texts, with all their metadata removed from the text and stored in an excel file with pointers to the related text files.

3. Tools

In order to test the application of the Menzerath law on Modern Greek (MG) we had to develop a number of tools that would permit the computational processing of the HNC. The first of these tools is the MGphontranscriber, a specialized PERL script for converting texts written using standard MG spelling to broad phonemic transcriptions. The script is based on 99 letter to phoneme rules implemented using appropriate regular expressions. The regular expressions apply sequentially to the text input and are ordered for producing the appropriate MG phonemic representations. The output of the tool was validated against human MG phonemic transcriptions and after some fine tuning produced 100% correct outputs. In the last preprocessing step, the tool tokenized the input removing all punctuation marks and putting each token in a separate line (vertical text representation).

The output of MGphontranscriber was used as input to a separate PERL script that calculated the basic variables of Menzerath’s law, i.e. the length of words (measured in phonemes) and their number of syllables (measured as the number of vowels in each word).

The resulting data have been processed by Menzerath.exe software[3] package that was designed to measure Menzerath's Law for many kinds of pre-processed data. The bootstrap resampling has been done by the Bootstrapper software[4].

A simple bootstrap method was used i.e. taking a sample of *N* values from the original data and resampling from them to form a new sample that is also of size *N*. The bootstrap sample is taken from the original one using sampling with replacement so it is not identical with the original one. This process is repeated in our case 100,000 to 1,000,000 times, and for each of these bootstrap samples its mean is computed. Confidence intervals for the sample mean are estimated from the histogram of the bootstrap means.

4. Results

The Menzerath's law for the phoneme-syllable-word level has been measured on a sample of 45,691 Greek texts. The distribution of the results is depicted in Fig. 3.

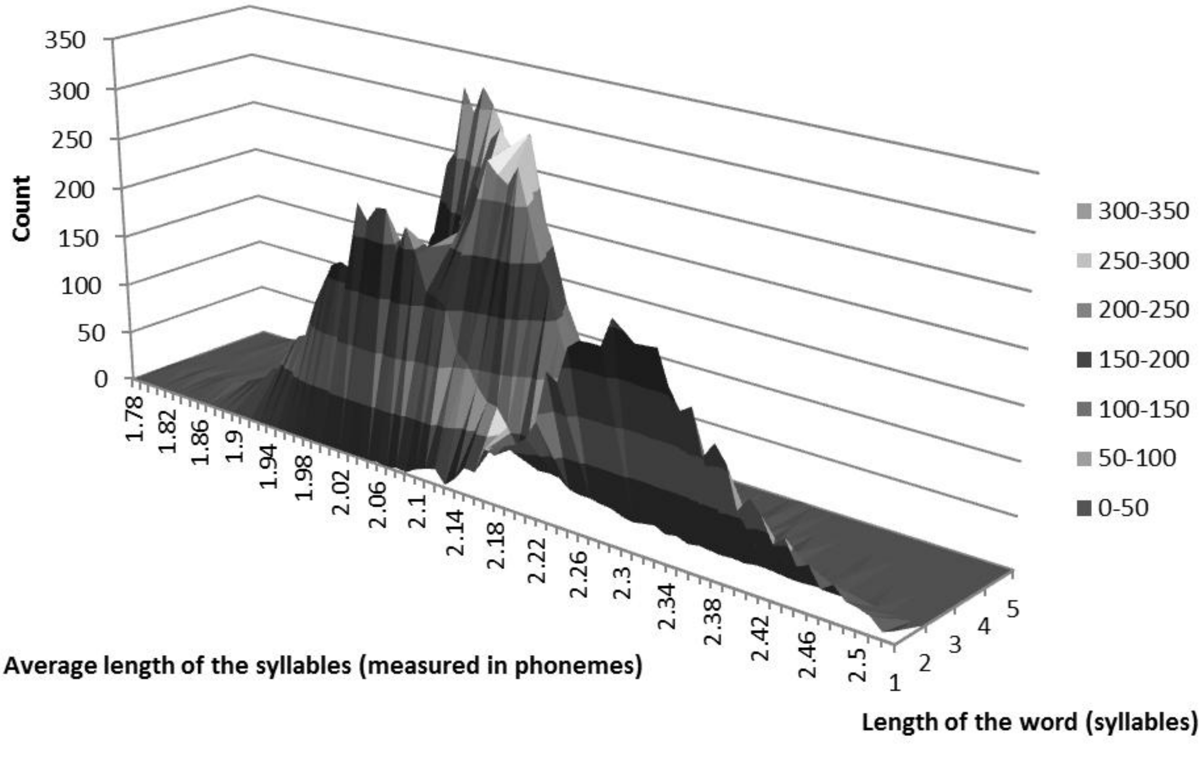


Figure 3. The distribution of results for words that contains 1–5 syllables.

[3] Available at http://milicka.cz/en/menzerath

[4] Available at http://milicka.cz/en/bootstrapper

From the distribution we bootstrapped confidence intervals[5] for the average values. Fig. 4 displays the resulting distribution produced by 100,000 bootstrap samples.

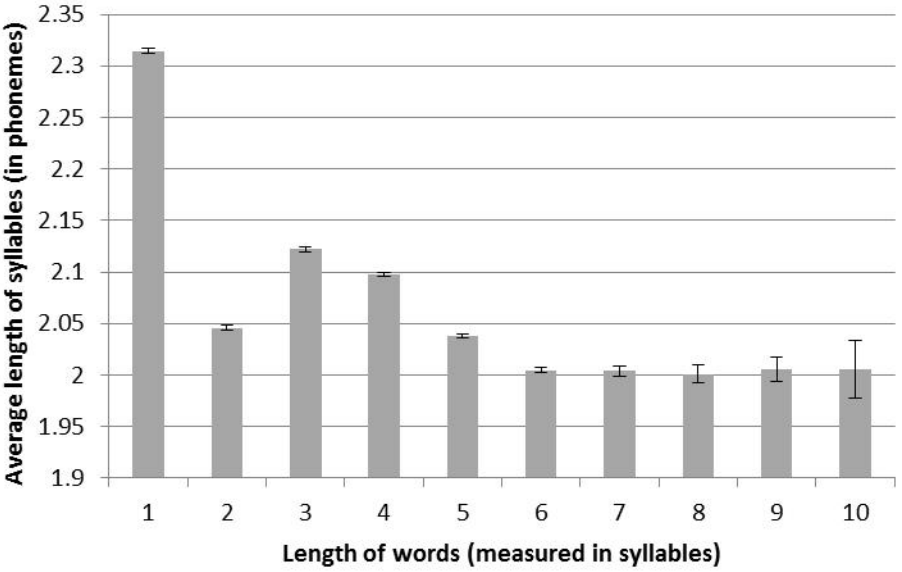


Figure 4. The Menzerath's law on the sample of texts. The error bars stands for the 95% confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11 – 20) were omitted.

Fig. 4 shows a break in the Menzerath-Altman (MA) law in the area of the 2-syllable words. The average length of syllables in words with 2 syllables is shorter than the average length of syllables in words with 1, 3 and 4 syllables making the function non-monotonic.

In order to investigate this phenomenon in detail on the single-text level we selected randomly 2 texts (#07439 & #35855) from the corpus and produced their word length histograms displayed below (Fig. 5). In Fig. 5 we see that the 2-syllable words exhibit a left-skewed distribution lowering their mean length and consequently producing bad fits to the MA law.

Another way to investigate further the differences between the various word lengths is to examine their relative ratios. More specifically, for each text, the ratios between values of neighbouring columns were calculated, i.e. average length of the syllables in words with n-syllables were divided by the average length of the syllables in words with n+1 syllables for each text. The distribution of the ratios is displayed in Fig. 6.

Again, from the distribution we bootstrapped confidence intervals (con-fideence level 95%) for the average values, see Fig. 7 (the confidence intervals

[5] For an introduction to the method see Efron (1979).

are marked by the error bars). The number of bootstrap samples was 100,000. This measurement brings results equivalent to the results that are depicted in Fig. 4.

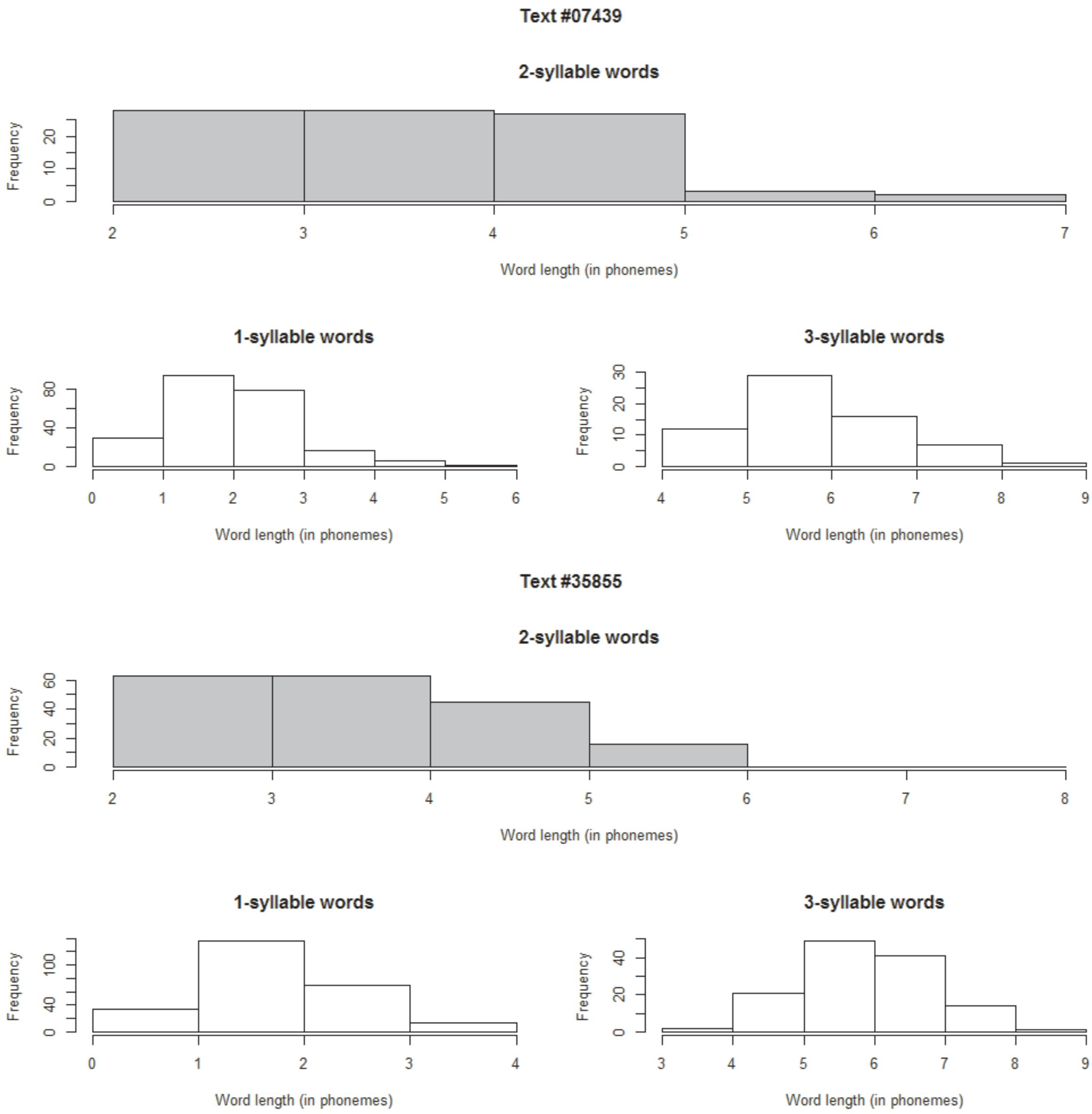


Figure 5. Distribution of word length (measured in phonemes) in 1, 2 and 3-syllable words in two random selected texts (#07439 & #35855).

On average, short words are the most frequent ones (for references to this phenomena see Strauss, Altmann 2006). Simple theoretical consideration leads us to the suspicion that the abnormality observed in the short words can be caused or accompanied by some non-trivial abnormality in the most frequent words. Let us measure the Menzerath's law for the sets of types (or dictionaries) instead of the real texts. This should neutralize the impact of the most frequent words (which are somewhat special because the set of the most frequent words contains synsemantic words and many proper names).

The Fig. 8 compares the values from Fig. 4 that were measured on word tokens with the values that were measured on the lists of types (for each text, a list of word types was extracted).

As we can see, the Menzerath's law on the phoneme-syllable-word level measured on the list of types that were extracted from the Greek texts is in accordance with the theory (monotonically decreasing). At the same time we can observe that the values for mono-, di- and tri- syllabic words that were measured on the lists of types are significantly different from the values that were measured on the word tokens (original texts).

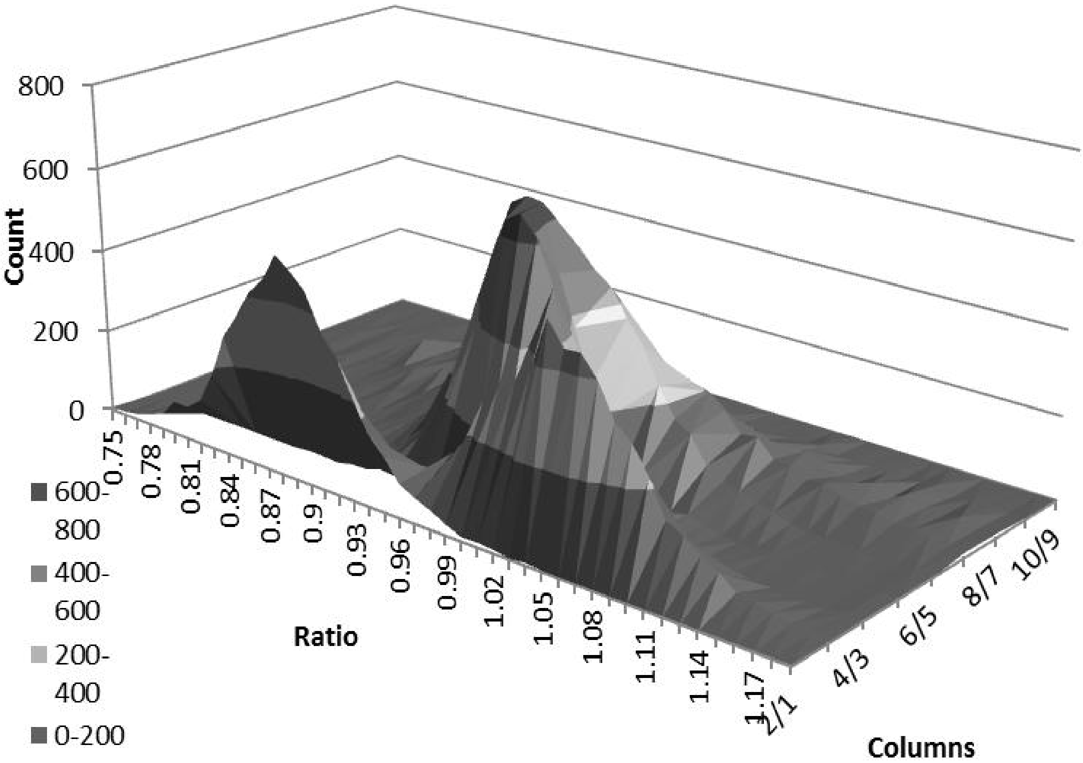


Figure 6. The distribution of ratios between neighbouring columns for columns 1–6.

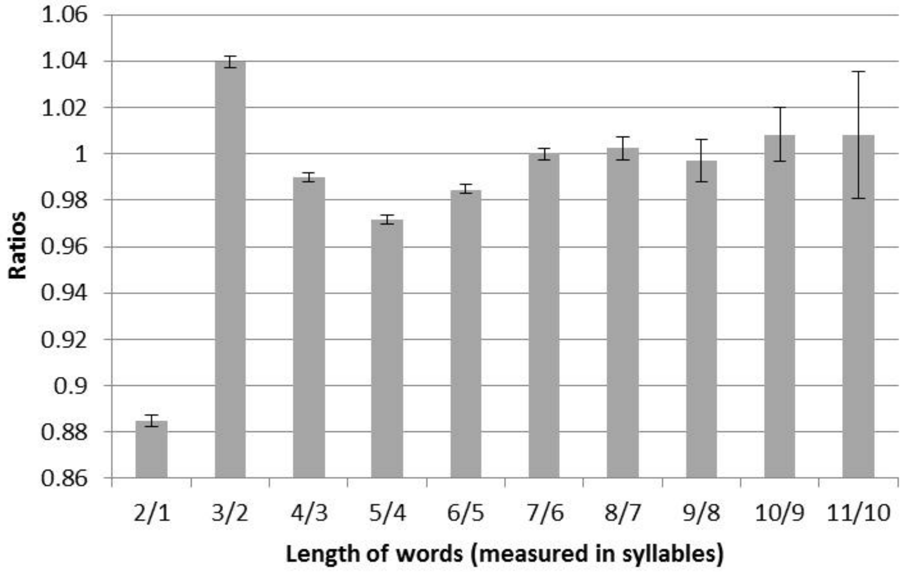


Figure 7. The Menzerath's law on the sample of texts. The error bars stands for the 95% confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11–20 syllables) were omitted.

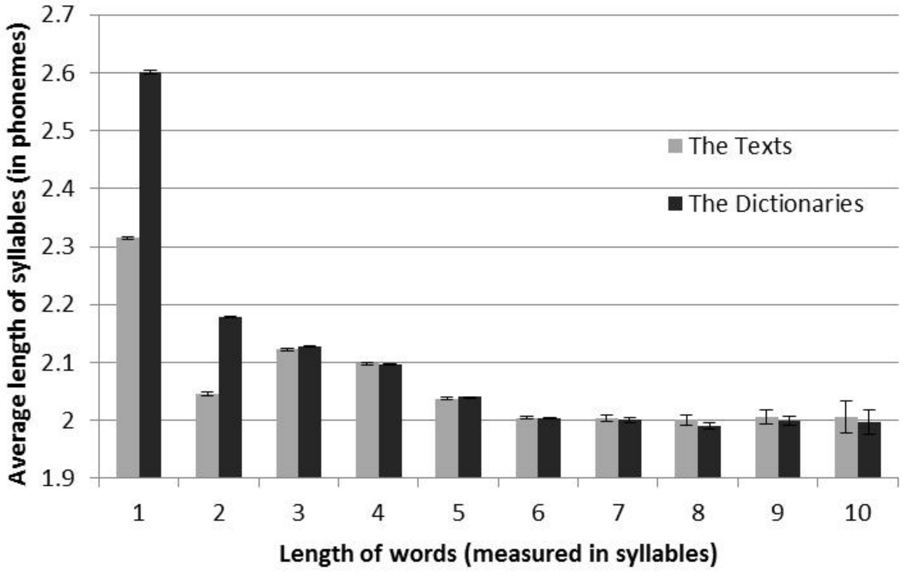


Figure 8. The Menzerath's law measured on the sample of texts and on the sample of lists of types (dictionaries) that were extracted from these texts. The error bars stands for the 95% confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11–20 syllables) were omitted.

5. Conclusion

Examining a large corpus of Greek texts we found that the average length of syllables in the disyllabic words is lower than the average length of the syllable in monosyllabic words and lower than the average length of syllables in tri-syllabic words. This peculiar phenomenon seems to be ubiquitous in Greek texts and our future research is directed to a wider explanatory framework including both the phonological properties of the Greek syllables contained in two-syllable words, their frequency distribution and possible historical explanations.

Word frequency seems to be a factor that can partially explain the deviant behaviour of the two-syllable words. Data that were measured on lists of types (extracted from each text separately) are significantly different from the data that were measured on the raw texts, and the type data follow Menzerath's law.

We are aware that some well-taken objections can be raised, e.g.:

1. The sample was not really randomly chosen from the population of Greek texts (even if the population is limited for the texts written in some time period).
2. The validity of the Menzerath's Law on the phoneme-syllable-word level is sometimes considered to be a side effect of the validity of the law on the phoneme-morpheme-word level and thus less interesting than the latter one.

This is the first attempt to study systematically Menzerath's law in a large Modern Greek corpus. Despite the limitations of this study outlined above, we hope that this study will to draw attention to this problem and to foster the discussion about the proposed methodology.

Acknowledgments

This work was supported by the project “Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology”, grant no. CZ.1.07/2.3.00/20.0161 which is financed by the European Social Fund and the National Budget of the Czech Republic.

References

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

Altmann, G., Schwibbe, H. (eds.) (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Georg Olms Verlag.

Andres, J. (2010). On a Conjecture about the fractal structure of language. *Journal of Quantitative Linguistics 17, 101–122.*

EAGLES (1994). Corpus encoding: Draft. Technical report, EAGLES. Document EAG-CSG/IR-T21.

Geršič, S., Altmann. G. (1980). Laut-Silbe-Wort und das Menzerathsche Gesetz. *Frankfurter Phonetische Beiträge 3, 115-128.*

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari E., Papageorgiou, H., Demiros, I. (2000). Design and implementation of the online ILSP Greek Corpus. In: Gavrilidou, M. et al. (eds.), *Proceedings of the LREC 2000 Conference: 1737-1742*. Athens.

Hatzigeorgiu, N., Mikros, G., Carayannis, G. (2001). Word length, word frequencies, and Zipf’s law in the Greek language. *Journal of Quantitative Linguistics 8*, 175–185.

Hřebíček, L. (1990). The constants of Menzerath-Altmann’s Law. In: Hammerl, R. (ed.), *Glottometrika 12: 61–71.* Bochum: Brockmeyer.

Hřebíček, L. (1992). *Text in comunication: Supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1994). Fractals in language. *Journal of Quantitative Linguistics 1,* 82–86.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7, 1–26.*

Menzerath, P. (1928). Über einige phonetische probleme. In: *Actes du premier Congrès International de Linguistes*. Leiden: Sijthoff

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümler.

Mikros, G., Hatzigeorgiu, N., Carayannis, G. (2005). Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics 12, 167–184.*

PAROLE (1995). *Design and composition of reusable har-monised written language reference corpora for euro-pean languages. Technical report*. PAROLE Consortium. MLAP, WP 4 - Task 1.1, *63–386*.

Sperberg-McQueen C. M., Burnard, L. (1994). *Guide-lines for electronic text encoding and interchange: Tei-p3. Technical report*. Chicago and Oxford: ACH-ACL-ALLC Text Coding Initiative.

Strauss, U., Altmann, G. Word Length and Frequency. In: *Laws in Quantitative Linguistics*. [Available at http://lql.uni-trier.de/index.php/Word_length_and_frequency] [cit. 2013/11/25].

Zenetzi, K. Papachristos, D. (2013). *Mathematical – empirical methods in Greek language. A study of the Menzerath-Altmann Law*. Unpublished MSc. Thesis, Postgraduate Program “Technoglossia VI”, National and Kapodistrian University of Athens and National Technical University of Athens, Athens.

The choice of postpositions of the subject and the ellipsis of the subject in Japanese

Haruko Sanada

1. Background of the study: Former studies on the valency theory and verbs in Japanese

The study of the valency of verbs in Japanese linguistics has developed in theoretical linguistics as an aspect of syntactic/grammatical studies, aiming at introducing it into the teaching of Japanese for foreigners or at developing an automatic translation system. Ishiwata and Ogino (1983a) and Ishiwata (1999) mentioned how the valency theory[1] (Tesnière 1959, 1988) had been introduced and developed in Japanese linguistics.

However, studies tended to be theoretical rather than empirical in Japanese linguistics. There are notable conceptual papers in this field: Nitta (1973, 1974), Ishiwata (1981, 1983, 1998, 1999, 2010, 2011), Maruyama (1990, 2010, 2011), Ikehara et al. (1997, 1999), National Language Research Institute (1997), Koizumi (2000, 2007), Ogino et al. (2003, 2005, 2005), and Kimura (2007).

Ishiwata already pointed out the importance of quantitative studies for the valency as "it is necessary and important to quantitatively investigate how many cases individual verbs take, or which cases are taken by which verbs" (1999: 132).

Miyajima has performed quantitative studies on valency since the 1960's. Concerning the necessity of the study from the point of view of both the frequency of use and the extent of use, he mentioned that: "It is necessary to state how frequently individual verbs are used with which case, not only whether it is correct that individual verbs are used with some case" (Miyajima 1994b: 465); and "the problem which case is more frequently governed by the verb can be considered the pragmatic aspect of the case government, and the problem how many cases are governed by the verb in the sentence can be considered the syntagmatic aspect of the case government" (Miyajima 1994a: 438).

Empirical studies of valency have not progressed because of the difficulty to provide corresponding mass data, e.g. a syntactically analysed and tagged corpus of present-day Japanese, or an adequate software system for sentence structure analysis (parser). It is difficult to treat such problems using a frequency

[1] In the present study, we discuss the verbs and their complements though the valency theory has been enlarged from the verb to the adjective and the noun (Sommerfeldt, Schreiber 1974, 1977).

table, texts with Key Word In Context (KWICK), or searching for a system of neighbouring words like collocations in the sentence. Ogino et al. (2005) said: "Studies on the valency have developed to the theoretical determination of the proper case employed by individual verbs. However, there is no study showing the entire picture of use or ellipsis of valency in texts" (2005: 63).

Not only in Japanese linguistics but also in general linguistics, the empirical study of valency is not sufficiently developed. Köhler (2012) provided quantitative research on distributions of the valency in corpus of various languages, and Helbig and Schenkel (1969, 1983) did the same in a German valency dictionary.

2. The aim of the project and its position in the studies on valency

In the present and future studies we aim at building a bridge between theoretical and empirical research concerning valency grammar. Referring to empirical studies, i.e. Miyajima (National Language Research Institute 1964, Miyajima 1994a, 1994b), Ishiwata (1999), or Ogino et al. (2005), and Köhler (2012), we would like to study the following hypotheses:

1. In the text, no verb takes uniformly all complements which are theoretically defined in the valency frame, and the frequency of complements or ellipses depends on the cases and postpositions.
2. The categories of obligatory or non-obligatory complements (or the degree of graduation) do not correlate with the observed frequency of complements.
3. In the text, there are combinations of more or less compatible cases, namely, combinations of cases (including postpositions) with a strong connection or a weak connection with the verb, though theoretically the cases are defined equally.
4. The number of complements which a verb takes in the sentence depends on the verb; some verbs always display more complements than other ones, though the valency dictionary shows always the maximum number of complements.
5. From the point of view of a collocation study, verbs occur with high or low frequent combinations of complements.

3. Outline of the study

In the study on distributions of Japanese postpositions (Sanada 2012), the frequency of valency for 6 verbs was observed employing two corpuses and a text, and the following results were found:

1. The proportions of postpositions for each verb in the three texts have a similar tendency.
2. The proportions of postpositions or of the ellipsis of the case seem to have a similar tendency for individual verbs.

Referring to our last study (Sanada 2012), we consider what is a condition of the choice of postpositions of the subject, or another choice, i.e. the ellipsis of the subject.

The postposition "ga" is mainly used for a subject in Japanese (Example 1), but in many cases another postposition "wa" for a theme is used instead (Example 2). An ellipsis of the subject is also often observed (Example 3).

Example 1:
Beikoku dewa 18 sai miman no kodomo wo motu josei no uchi *61% ga* hataraite iru.
[In America, *61% of women* who have children younger than 18 years old have a job.]

Example 2:
Kami kara atae rareta sakka de, *watashi wa* tomi wo eta.
[*I* gained wealth by the talent of football which is gifted by the god]

Example 3:
Yujin no shokai de Putnam sha no William Targ ni atta.
[*I* met William Targ of Putnam who was introduced by a friend]

The usages of "ga" and "wa" differ from each other (Noda 1996). For example, at the beginning of a story, "wa" is employed to show a theme and a subject, followed by a sentence with "ga" (Example 4).

Example 4:
Mukashi mukashi, aru tokoro ni *ojiisan to obasan ga* imashita. *Ojiisan wa* yama he shibakarini…
[Once upon a time there were *an old man and an old woman*. *The man* goes firewood gathering to the hills...]

However, in some cases both postpositions or an ellipsis of the subject are possible, and writers choose one among three (Other postpositions are also possible (Example 5) but they have a smaller frequency) considering a balance of redundancy and the detailed-ness of the whole text.

Example 5:
Seidoku dewa jissei *kinri no* ugoku haba ga, rengin no kin'yu chosetsu ni yotte ittei no han'inai ni osamerarete iru.

[In West Germany the range where *the market interest rate* can move is settled by the German Federal Bank.]

We do not find a linguistic rule yet for such unconscious choice among these three alternatives when there is no constraint of the context. However, it is already known that a linguistic balance obeys quantitative laws (Köhler 2012). Linguistic rules for the unconscious choice would be applicable in natural language processing, e.g. a machine translation from English to Japanese because Japanese sentences written by the native speaker are achieved with such choices while a subject is obligatory for most English sentences. In the field of natural language processing they presume words or sentences based on frequency data of a huge number of examples, i.e. words used more frequently must have greater probabilities of occurrence if a text is generated by a program. We are studying not only the conceptual aspect but also the empirical aspect of the choice of postpositions or an ellipsis of the subject.

Employing the valency database (Ogino et al. 2003), we have made statistical tests for the distributions of postpositions of the subject "ga" and "wa" and an ellipsis of the subject of the same 6 verbs as employed in our last study (Sanada 2012): "au" (meet), "hataraku" (work), "yaburu" (tear, break), "umareru" (be born, arise), "ugoku" (move), and "ataeru" (give). We tentatively chose these 6 verbs because of the following reasons (Sanada 2012): they do not use the auxiliary verb, and have no homonyms; they have relatively little ambiguity of meaning or usage; they do not have very low frequency in the given text. The frequencies of the subject case, an ellipsis of the subject, and the number of occurrences of the verb are shown in Table 1. The frequency of Type (1) in Table 1 includes "ga", "wa" (emphasis), "mo" (also), derivative forms like "dake ga" (only) or "made ga" (even) and some other postpositions. The number of examples of nouns without postposition, i.e. an ellipsis of the postposition, is also included.

In the present study the following relationships are tested with verbs shown in Table 1 to investigate whether the distributions are significant.

(1) An ellipsis of the subject and an objective case or a dative case.
(2) An ellipsis of the subject and the type of verbs with respect to the valency.
(3) The choice of postpositions or an ellipsis of the subject.

The valency data base (Ogino et al. 2003) categorized postpositions into the following 11 groups: (1) "wa" (emphasis, theme), (2) "ga" (subject), (3) "wo" (direct object), (4) "ni" (indirect object, direction), (5) "e" (direction), (6) "kara" (from), (7) "yori" (from, than), (8) "made" (until, to), (9) "de" (by means of, with), (10) "to" (with, and), and (11) others.

Table 1
Distribution of subjective case for 6 verbs in the valency database (made from Ogino et al. 2003)

Type	"au" (meet)	"hataraku" (work)	"yaburu" (tear, break)	"umareru" (be born, arise)	"ugoku" (move)	"ataeru" (give)
(1) Subjective case: exists	95	165	24	28	165	75
(2) Ellipsis of the subject	144	110	27	1	35	95
(3) Dependent clause whose subject is in the main clause	4	61	0	3	15	4
Total	243	336	51	32	215	174

The categorizing was done by hand, and according to the context, postpositions are placed in the different groups, e.g. "wa" (subject) and "no" (subject) are placed with "ga" (subject); "ga" (direct object) is placed with "wo" (direct object); and "mo" (also) is placed with the proper group. With this database, we can analyze the surface case and the deep case, e.g. how many postposition "ga" occurred in the database, and how many of them are classified into Group (2) (subject) or into Group (3) (direct object).

In this paper we refer to Group (2) (subject) as the subjective case, to Group (3) (direct object) as the objective case "wo", and to Group (4) (indirect object), as the dative case "ni".

4. The ellipsis of the subject and an objective case or a dative case

We found that the proportions of postpositions for each verb in the three texts have a similar tendency (Sanada 2012). For example, the verb "au" (meet) is followed by postpositions of the object "to" (with) or "ni" (object). Both postpositions are possible, however, the frequency of "ni" is higher than "to" in the three sources. It can be interpreted that a subject is “assumable” (i.e. the reader can assume there is a subject) if a context has two persons as in the sentence "A meets B" and an object is shown to the reader. In Japanese the subject is not obligatory, and the reader might feel there is prolixity and the text is not natural if a subject which can be assumable is often given.

In this section we investigate the relationship between frequencies of the subject or ellipsis of the subject and frequencies of an objective case or a dative case. We made the Chi-square tests at the 5% significance level for two groups with/without subject with/without an objective case or a dative case for the following three individual verbs[2]: We also employ the extended Fisher's exact test offered by the statistical software package *R*, if some of the expected values were smaller than 5.

(1) The verb "au" (meet) with the subjective case and the dative case "ni",
(2) The verb "ugoku" (move) with the subjective case and the dative case "ni",
(3) The verb "ataeru" (give) with the subjective case and the dative case "ni", and
(4) The verb "ataeru" (give) with the subjective case and the objective case "wo".

The hypothesis H_0 assumes that there is no difference between two groups with or without subject with respect to distributions of with or without an objective case or a dative case. The results of tests are shown with frequency data in Tables 2 to 5. The total number excludes examples of Type (3) (Dependent clause whose subject is in the main clause) of Table 1. Some verbs also exclude examples of the dependent clause whose dative or objective case is in the main clause. The frequency of the dative case "ni" from Table 2 to Table 5 includes "wa" (emphasis), and derivative forms like "ni wa" (emphasis) or "ni mo" (also).

Table 2
Subjective case and dative case of the verb "au" (meet)

	Dative case "ni": exists	Dative case "ni": does not exist	Total
Subjective case: exists	39	54	93
Subjective case: does not exist	64	64	128
Total	103	118	221[3]
$\chi^2_0 = 1.408 < \chi^2_{0.05}$ (3.84) n.s. Hypothesis (H_0) is not rejected.			

[2] The verbs "hataraku" (work), "umareru" (be born, arise), and "yaburu" (tear, break) are not tested because they have 0 in the cells of the tables.
[3] The total number of verbs excludes 18 examples of the dependent clause whose dative case is in the main clause.

Table 3
Subjective case and dative case of the verb "ugoku" (move)

	Dative case "ni": exists	Dative case "ni": does not exist	Total
Subjective case: exists	35	130	165
Subjective case: does not exist	9	26	35
Total	44	156	200
$\chi^2_0 = 0.341 < \chi^2_{0.05}$ (3.84) n.s. Hypothesis (H_0) is not rejected			

Table 4
Subjective case and dative case of the verb "ataeru" (give)

	Dative case "ni": exists	Dative case "ni": does not exist	Total
Subjective case: exists	43	31	74
Subjective case: does not exist	52	40	92
Total	95	71	166[4]
$\chi^2_0 = 0.004 < \chi^2_{0.05}$ (3.84) n.s. Hypothesis (H_0) is not rejected			

Table 5
Subjective case and objective case of the verb "ataeru" (give)

	Objective case "wo": exists	Objective case "wo": does not exist	Total
Subjective case: exists	60	2	62
Subjective case: does not exist	91	1	92
Total	151	3	154[5]
$\chi^2_0 = 0.136 < \chi^2_{0.05}$ (3.84) n.s. Fisher's $P = 0.565 > 0.05$ Hypothesis (H_0) is not rejected			

[4] The total number of verbs excludes 4 examples of the dependent clause whose dative case is in the main clause.

[5] The total number of verbs excludes 16 examples of the dependent clause whose objective case is in the main clause.

In all the above cases, the null hypotheses are not rejected, and there is no significant difference between the two groups: a verb with or without subject and with or without an objective case or a dative case in the text. It means that an ellipsis of the subject does not depend on the frequency of an objective case or a dative case even if an objective case or a dative case contributes to a presumption of the subject.

5. An ellipsis of the subject and the type of verbs from the point of view of valency

In this section, we consider whether the type of verb with respect to the valency exerts an influence on the ellipsis of the subject. In the previous section we investigated the relationship between the subject and the object using their frequencies. However, verbs need not have an object in the text though they theoretically have complements.

The valency patterns for the 6 verbs in Table 6 show the theoretical possibility of complements. In Japanese, "au" (meet) which employs the postposition "ni" for a direct object is grammatically defined as an intransitive verb because only verbs with "wo" are defined as transitive verbs. For this problem, we cannot simply categorize verbs as intransitive or transitive. We have to consider the type of verbs on the basis of the similarity of the behavior of the verbs.

Table 6
Valency patterns of 6 verbs (Ishiwata, Ogino 1983b)

Verb	Valency pattern
"au" (meet)	N[hum] "ga" +N[div] "ni"/ N[div] "to" + V
"hataraku" (work)	N[abs] "ga" / N[hum] "ga" +V
"hataraku" (work) (in particular, commit a wrongdoing or a crime)	N[hum] "ga" +N[abs] "wo" +V
"yaburu" (tear, break)	N[hum] "ga" +N[con] "wo" +V
"umareru" (be born, arise)	N[ani] "ga" / N[con] "ga" +V
"ugoku" (move)	N[con] "ga" / N[hum] "ga" +N[loc] "kara" +N[loc] "ni" / N[loc] "e" +V
"ataeru" (give)	N[hum] "ga" / N[hum] "kara" +N[con] "wo" +N[hum] "ni" +V
[hum]=human, [div]=divers, [abs]=abstract, [con]=concrete, [ani]=animal, [loc]=locality, ga (subject), wo (direct object), ni (indirect/ direct object, direction), to (together, with), kara (from), e (direction)	

Employing the valency database (Ogino et al. 2003), we compared a pair of verbs out of 6 verbs with the presence or ellipsis of the subject using the Chi-square tests at the 5% significance level, and considered the characteristics of verbs which have no significant difference. The H_0-hypothesis underlying the Chi-square test is: the distributions of the presence or the ellipsis of the subject with two verbs are not different. For example, Table 7 shows the distributions of the subjective case in verbs "ugoku" (move) and "ataeru" (give) and the result of the Chi-square test. The hypothesis is rejected, hence the employment of the subject in these two verbs is significantly different.

Table 7
Distributions of the subjective case in verbs "ugoku" (move) and "ataeru" (give)

	Subjective case: exists	Subjective case: does not exist	Total
"ugoku" (move)	165	35	200
"ataeru" (give)	75	95	170
Total	240	130	370
$\chi^2_0 = 59.40^* > \chi^2_{0.05}$ (3.84) Hypothesis (H_0) is rejected.			

The results of tests for 30 pairs of 6 verbs are shown with their Chi-square statistics in Table 8. The mark (**) shows that the pair is significantly different, and "n.s." shows that the pair is not different. The critical value of the Chi-square statistics at 5% significance level with 1 degree of freedom is 3.84.

Pairs of verbs "au" (meet), "yaburu" (tear, break), and "ataeru" (give) are not significantly different. These three verbs all take an objective complement. The pair "umareru" (be born, arise) and "ugoku" (move) are also not significantly different, and both take no objective complement. Therefore it could be supposed that the characteristics of verbs to take an objective complement exerts an influence on the tendency of the subject to be present or to be omitted.

The verb "hataraku" (work) and "yaburu" (tear, break) display no statistical difference. According to the valency patterns in Table 6, the verb "hataraku" (work) with an object means to commit a crime as a limited usage. However, we found a few examples with the meaning of “labor” and taking an object in the valency database (Ogino et al. 2003)[6] and in another corpus[7]. It can be inter-

[6] "Kono kuni dewa *zangyo bun ijo wo* josei ga *hatawaite* (...)" (Women *work more than their overtime payment* in this country) (RID: JCO0038954) (Ogino et al. 2003).

[7] "1 ka getsu no natsu no vacance no tame ni *nokori 11 ka getsu wo hataraku*, nado to iwareru furansu jin de aru ga, (...)" (French people are said that for one month of their summer vacation they *work the other 11 months of the year*) (BCCWJ)

preted that the verb "hataraku" (work) is intermediate between intransitive and transitive. This problem should be a task for the future.

Table 8
Chi square statistics of pairs of verbs for distributions of the subjective case

	"au" (meet)	"hataraku" (work)	"yaburu" (tear, break)	"umareru" (be born, arise)	"ugoku" (move)
"au" (meet)					
"hataraku" (work)	20.98**				
"yaburu" (tear, break)	0.93 n.s.	2.96 n.s.			
"umareru" (be born, arise)	32.68**	14.24**	19.03**		
"ugoku" (move)	82.41**	27.64**	27.44**	3.05 n.s.	
"ataeru" (give)	0.78 n.s.	10.67**	0.14 n.s.	26.37**	59.40**

6. The choice of postpositions or an ellipsis of the subject

The postpositions "ga" and "wa" are frequently used for the subjective case though there are some examples with other postpositions, e.g. "no" or "mo", etc. An ellipsis of the subject is also often seen in Japanese. In this section we investigated whether the distributions of the three possibilities - postposition "wa", "ga", and an ellipsis of the subject - have similar tendency in texts.

As many studies already pointed out, the usages of "ga" and "wa" are different from each other and we choose one depending on the context. Also in some cases we find less significant difference of the context for the two choices, or we choose an ellipsis of the subject avoid prolixity. The text is altered so as to be “natural” Japanese – a repetition of the same postposition makes a text monotonous, too many ellipses make a text unclear, and no ellipsis makes a text be wordy.

We employed the valency database (Ogino et al. 2003) and an essay (Miki 1941, 1995), and performed the Chi-square tests at the 5% significance level for 6 verbs separately. The frequency in the essay is very small, and we also employ the extended Fisher's exact test offered by the statistical software package *R*, if some of the expected values were smaller than 5. The hypothesis (H_0) is: the distributions of two texts are not different (two-sided test). The results of the tests using the frequency data are shown in Tables 9 to 14.

Table 9
Distributions of the frequency of the subjective case for the verb "au" (meet)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	62	30	144	236
Essay	2	1	4	7
Total	64	31	148	243
$\chi^2_0 = 0.04 < \chi^2_{0.05}$ (5.99) n.s. Fisher's $P = 1.00 > 0.05$ Hypothesis (H_0) is not rejected.				

Table 10
Distributions of the frequency of the subjective case for the verb "hataraku" (work)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	72	71	110	253
Essay	7	4	6	17
Total	79	75	116	270
$\chi^2_0 = 0.125 < \chi^2_{0.05}$ (5.99) n.s. Fisher's $P = 0.59 > 0.05$ Hypothesis (H_0) is not rejected.				

Table 11
Distributions of the frequency of the subjective case for the verb "yaburu" (tear, break)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	8	15	27	50
Essay	1	1	6	8
Total	9	16	33	58
$\chi^2_0 = 0.135 < \chi^2_{0.05}$ (5.99) n.s. Fisher's $P = 0.67 > 0.05$ Hypothesis (H_0) is not rejected.				

Table 12
Distributions of the frequency of the subjective case for the verb "umareru" (be born, arise)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	10	13	1	24
Essay	3	1	6	10
Total	13	14	7	34
$\chi^2_0 = 14.28^* > \chi^2_{0.05}$ (5.99). Fisher's $P = 0.001 < 0.05$ Hypothesis (H_0) is rejected.				

Table 13
Distributions of the frequency of the subjective case for the verb "ugoku" (move)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	73	77	35	185
Essay	1	1	17	19
Total	74	78	52	204
$\chi^2_0 = 045.16^* > \chi^2_{0.05}$ (5.99). Fisher's $P = 9.77e\text{-}10 < 0.05$ Hypothesis (H_0) is rejected.				

Table 14
Distributions of the frequency of the subjective case for the verb "ataeru" (give)

	"wa"	"ga"	Ellipsis of the subject	Total
Valency Database	34	34	95	163
Essay	4	3	12	19
Total	38	37	107	182
$\chi^2_0 = 0.28 < \chi^2_{0.05}$ (5.99) n.s. Fisher's $P = 0.94 > 0.05$ Hypothesis (H_0) is not rejected.				

For 4 of 6 verbs, the hypotheses are not rejected, and the distributions of 3 cases in two texts display no difference. It can be assumed that the distributions of these 3 cases may depend on the individual verb and not on the context.

7. Discussions and future problems

The present study explores the linguistic balance of Japanese. In this paper, we studied frequencies of postpositions of the subject and the ellipsis of the subject employing the valency database, and considered the choice of postpositions or an ellipsis. The following three points are tested and discussed.

(1) An ellipsis of the subject and an objective case or a dative case. There is no significant difference between two groups, i.e. verbs with the subject or verbs without subject, and they both do not depend on the occurrence of the objective case or the dative case.

(2) An ellipsis of the subject and the type of verbs with respect to the valency. As there are significant groups of verbs in the valency database, it could be supposed that the characteristics of verbs to take an objective complement exert an influence on the tendency of the subject to be present or to be omitted.

(3) The choice of postpositions or an ellipsis of the subject. It can be assumed that the distributions of cases - "ga", "wa" and the ellipsis of the subject - may depend on the individual verb and not on the context.

We examined whether a noun for the subject could be supplied from other parts of the same sentence if the subject is omitted. The frequency of the ellipsis of the subject for 6 verbs are employed from Table 1 and the result is shown in Table 15. In most of the cases, the noun cannot be supplied from other parts of the same sentence. It can be assumed that in most of the cases, the omitted subject must be supplied from other sentences in the reader's consciousness. However, this problem must be one of our future tasks.

As future tasks we should investigate various aspects of the valency of Japanese employing more verbs and more linguistic materials.

Table 15

Distribution of the ellipsis of the subject for 6 verbs

Type: from where an omitted subject can be supplied.	"au" (meet)	"hataraku" (work)	"yaburu" (tear, break)	"umareru" (be born, arise)	"ugoku" (move)	"ataeru" (give)
In the same sentence	15	31	7	1	12	12
Not in the same sentence	129	79	20	0	23	83
Total: Ellipsis of the subject	144	110	27	1	35	95

Remarks

This study is based on data from Sanada (2013) and revised. The study is partly supported by the Grant-in-Aid for Scientific Research (No. 23520567) of Japan Society for the Promotion of Science (JSPS) in 2012.

References

Helbig, G., Schenkel, W. (1969, 1983). *Wörterbuch zur Valenz und Distribution deutscher Verben.* Leipzig: Bibliographisches Institut. (Reprint of De Gruyter)

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y., NTT Communication Science Laboratories (eds.) (1997). *Nihongo Goi Taikei* (Japanese Lexicon), vol.1-5. Tokyo: Iwanami Shoten.

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y., NTT Communication Science Laboratories (eds.) (1999). *Nihongo Goi Taikei* (Japanese Lexicon): *CD-ROM Edition.* Tokyo: Iwanami Shoten.

Ishiwata, T. (1981). Review: K.-E. Sommerfeldt, Ketsugoka wo Koryo Shita Goi no Bunrui (The classification of lexical items according to their valence). *Keiryo Kokugogaku* (Mathematical Linguistics), *13*(2), 97–102.

Ishiwata, T. (1983). Meishi no Bunrui (A classification of nouns). In: Mizutani, Sh. (ed.), *Bunpo to Imi 1* (Grammar and meaning 1). Tokyo: Asakura Shoten, *273–274*.

Ishiwata, T. (1998). Shinsokaku to Doshi no Imi (Deep case and meanings of verbs). In: *Kokubun Mejiro* (Japanese literature Mejiro). Japan Women's University, vol. *37, 1–8*.

Ishiwata, T. (1999). *Gendai Gengo Riron to Kaku* (Modern linguistic theories and the case). Tokyo: Hitsuji Shobo.

Ishiwata, T. (2010). Book Review: Thomas Herbst / Susen Schuller, *Introduction to Syntactic Analysis. A Valency approach.* Gerhard Helbig, Theoretische und praktische Aspekte eines Valenzmodells in Gerhart Helbig (Hrsg.) *Beiträge zur Valenztheorie*. In *Keiryo Kokugogaku* (Mathematical Linguistics) *27-5, 194–195*.

Ishiwata, T. (2011). Book Review: Gizella Boszak, *Realisierung der valenzbestimmten Korrelate des Deutschen.* K. Fischer, E. Fobbe, S. Schierholz (Hrsg.) *Valenz und Deutsch als Fremdsprache.* In *Keiryo Kokugogaku* (Mathematical Linguistics) *28-3,* 106–109.

Ishiwata, T., Ogino, T. (1983a). Ketsugoka kara Mita Nihon Bunpo (Japanese grammar focused on the valency). In: Mizutani, Sh. (ed.), *Bunpo to Imi 1* (Grammar and meaning 1). Tokyo: Asakura Shoten, 81–134.

Ishiwata, T., Ogino, T. (1983b). Nihongo Yogen no Ketsugoka (Valency of declinable words in Japanese). In: Mizutani, Sh. (ed.), *Bunpo to Imi 1* (Grammar and meaning 1). Tokyo: Asakura Shoten, 226–272.

Kimura, M. (2007). Kakubunpo Ketsugoka Bunpo to Colocation (Case Grammar, Valency Grammar and colocation). *Nihongogaku* (Japanese linguistics), *26(12)*, 28–36.

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

Koizumi, T., Funaki, M., Honda, Ky., Nitta, Y., Tsukamoto, H. (2000). *Nihongo Kihon Doshi Yoho Jiten* (Dictionary of usage of basic verbs in Japanese). Tokyo: Taishukan Shoten.

Koizumi, T. (2007). *Nihongo no Kaku to Bunkei: Ketsugoka Riron ni Motoduku Sinteian* (Japanese cases and sentence patterns: a new theory based on the Valancy Grammar). Tokyo: Taishukan Shoten.

Maruyama, N. (1990). Kakujoshi to Kaku to Ketsugoka (Case Particles, Cases, and Valency). In: *Keiryo Kokugogaku* (Mathematical Linguistics), *17*(*4*), *169–192.*

Maruyama, N. (2010). Joshi Ni wo Tomonau Yakuwari Seibun: Corpus ni Motoduku Bunseki (Role Constituents with the Particle Ni: A Corpus Analysis). *Nihongo Bunpo* (Journal of Japanese Grammar), *10*(*1*), *71–87.*

Maruyama, N. (2011). Doshi no Kaku Joho: Kokugo Jiten no Kijutsu to Corpus (Case frames of verbs: descriptions in the Japanese dictionary and corpora). *Nihon Bungaku* (Japanese literature), *107*, *227–245.*

Miyazima, T. (1994a). Kaku Shihai no Ryotei Sokumen (Quantitative aspect of the case government). In: Miyajima, T. (ed.) *Goiron Kenkyu* (Study of morphology). Tokyo: Mugi Shobo, *437–461.*

Miyazima, T. (1994b). Ido Doshi to Kaku, Zenchishi (Motion verbs, cases and prepositions). In: Miyajima, T. (ed.) *Goiron Kenkyu* (Study of morphology). Tokyo: Mugi Shobo, *463–474.*

Miyazima, T. (1994c). Kakari no Ichi (Order between elements preceding the predicate). In: Miyajima, T. (ed.) *Goiron Kenkyu* (Study of morphology). Tokyo: Mugi Shobo, *514–521.*

National Language Research Institute. (1964). *Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki* (Vocabulary and Chinese Characters in Ninety Magazines of Today: vol.3: Analysis of Results). Tokyo: Shuei Shuppan.

National Language Research Institute. (1997). *Nihongo ni Okeru Hyosokaku to Shinsokaku no Taio Kankei* (Cases and Japanese postpositions).Tokyo: Sanseido.

Nitta, Y. (1973). Doshi no Kaku Shihai (Case government of verbs). *Kokugogaku Kenkyu* (Study of Japanese Linguistics), *12*, *54–64.*

Nitta, Y. (1974). Nihongo Ketsugoka Bunpo Josetsu: Doshi Syntax no Hitotsu no Model (An Introduction to Valence-Grammar of Japanese). *Kokugogaku* (Japanese Linguistics), *98*, *93–112.*

Noda, H. (1996). *"Wa" to "ga"* (Post positions "wa" and "ga"). Tokyo: Kuroshio Shuppan.

Ogino, T., Kimura, M., Yoshida, N., Kobayashi, M., Isahara, H. (2005). Nihongo Doshi no Ketsugoka Kijutsu ni Hitsuyo na Imi Bunrui Komoku to Imi Bunrui Komokusu (Semantic categories critical to verb valence and the distribution of the numbers of the categories). *Keiryo Kokugogaku* (Mathematical Linguistics), *25-1*, *1–31*.

Ogino, T., Kobayashi, M., Isahara, H. (2003). *Nihongo Doshi no Ketsugoka* (Verb valency in Japanese). Tokyo: Sanseido.

Ogino, T., Ueda, Y., Kobayashi, M., Isahara, H. (2005). Kopasu Deta ni Motoduku Kakujoshi Kumiawase Reberu ni Okeru Ketsugoka no Jittai to Doonihyoki Hantei deno Riyo (Valence behavior at the level of particle combinations based on corpus data and its contribution to homophone distinction). *Journal of Natural Language Processing*, *12*(*4*), *21–54*.

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

Sanada, H. (2013). Shukaku wo Shimesu Joshi no Hindo Bunpu to Shukaku Shoryaku no Mondai: Ketsugoka Riron wo Sanko ni Shita Keiryoteki Kenkyu (Distributions of frequencies of postpositions for the subject and the ellipsis of the subject: a quantitative analysis employing the valency theory). In: *Gakugei Kokugo Kokubungaku* (Journal of Japanese linguistics and literature of Tokyo Gakugei University), *45*, *1–14*.

Sommerfeldt, K.-E., Schreiber, H. (1974). *Wörterbuch zur Valenz und Distribution deutscher Adjektive.* Leipzig: Bibliographisches Institut.

Sommerfeldt, K.-E., Schreiber, H. (1977). *Wörterbuch zur Valenz und Distribution der Substantive.* Leipzig : Bibliographisches Institut.

Tesnière, L. (1959, 1988). *Éléments de Syntaxe Structurale. 2nd edition.* Paris: Klincksieck.

Software and corpus

Graduate Schools of Informatics in Kyoto University, NTT Communication Science Laboratories. Morphological analyzer: *MeCab*, version 0.97.
(http://mecab.googlecode.com/svn/trunk/mecab/ doc/index.html)

Kudo, T. Sentence structure analyzer: *Cabocha.* ůok
(http://code.google.com/p/cabocha/)

Miki, K. (1941, 1995). Jinseiron Note (Essay on the life). Tokyo: Sogensha.
(Digital edition included in Shinchosha 1995).

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

National Institute for Japanese Language and Linguistics. *Balanced Corpus of Contemporary Written Japanese (BCCWJ).* (http://www.ninjal.ac.jp/kotonoha/)

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

The volume of Ottoman lexical influence on Romanian

Kamil Stachowski

0. Rationale

This paper continues the pursuit of two of the goals set by quantitative linguists. On the one hand, it wants to help implement Luděk Hřebíček's wish (1990: 371), and demonstrate one way of how quantitative and qualitative methods and results can be used jointly to gain a kind of insights that either of them would find difficult to obtain on its own. On the other hand, it will advance the work of K.-H. Best (2005) by further exploring Turkic influence on the languages of Europe, and the possibilities and limitations of the so-called Piotrovskij-Altmann law. In this sense, this paper is a continuation of Stachowski (2013), where a similar approach was applied to Turkic glosses in Hungarian and Polish.

In particular, we investigate the vocabulary of Oriental origin that Romanian acquired as a result of the Ottoman presence in the Balkans. The problem is an extensive and complex one, and the current paper does not intend to offer a full analysis. It will be limited to establishing just the overall, general quantitative characteristic of the said vocabulary.

I will: 1. introduce the historical background, its linguistic results, and the assumptions made in this paper, 2. explore selected quantitative aspects of Ottoman influence on Romanian as a whole (2.1), and on the individual regions where Romanian is spoken (2.2), and lastly 3. summarize the conclusions.

1. Introduction

This section provides background information without which the analytic part of the paper may be unclear (1.1–1.2), and introduces Suciu (2009, 2010) as the source of the material discussed in it, together with the adopted assumptions and abbreviations (1.2–1.4). In particular, the note on the terms “Ottoman” and “Romanian” in 1.3 should not be skipped.

1.1 Political history

Romanian is spoken in a territory that encompasses several regions whose histories are often separate, complicated, and marked by long periods of conflicting foreign interferences which had resulted in an unclear political situation (see e.g. Ágoston, Masters 2009; Pop, Bolovan 2006; Sallanz 2005; Sugar 1977; see also Fig. 1 below).

Wallachia, i.e. Muntenia and Oltenia taken together, is home to one of the first Romanian principalities, founded in 1330 by Basarab I, probably of Turkic (Cuman) origin himself. The country fell again under Hungarian suzerainty soon afterwards, and in 1394, it sent tribute to the Ottomans, an act that the Porte considered an acknowledgment of vassalage. The Ottoman grip on Wallachia tightened further in 1476, after the death of the infamous Vlad the Impaler, but the region was never incorporated into the Empire as a province. The next two hundred years were not free of turmoil; finally, in 1715, the so-called Phanariot epoch began, when voivodes were appointed by the Ottomans from among the Greeks of Istanbul. In the years 1716–39, Oltenia, the western part of the region, was under Austrian rule, and in the second half of the 18th century, the entire country was occupied more than once by Russians and by the Habsburgs. In 1829, Wallachia was together with Moldavia under Russian military rule, in 1859 both elected the same ruling prince, thus forming a personal union, and in 1878 they were united with Dobruja giving rise to the sovereign Kingdom of Romania.

Moldavia, before becoming an independent state in 1359, remained mostly under Hungarian control and was repeatedly invaded, among others, by Turkic-speaking Cumans and Tatars. In 1377 it sent taxes to the Ottomans for the first time. Nevertheless, the country had managed to retain a degree of independence until the late 15th and early 16th century, and never became a province of the Ottoman Empire. The two hundred years before that time, and the following two hundred, were a period of almost constant attempts to regain full independence, with Moldavia’s own forces, or by various alliances, and of the consequent Ottoman retaliations. In 1601–18, the country recognized Polish suzerainty. Beginning from 1711, its voivodes were chosen from among the Phanariots. In 1775, the Habsburgs gained control over Bukovina, i.e. the northwestern part of the region, and in 1812, its eastern part, Bessarabia, was seized by the Russian Empire. The remainder came under Russian control after 1829, and became part of independent Romania in 1878.

Dobruja came under Bulgarian control in late 12th century, and was repeatedly invaded, by, among others the Romanian Principality of Wallachia, and by the Turkic Cumans and Pechengs, until it was incorporated into the Ottoman Empire in 1419 to remain one of its provinces until 1878 when it was united with Moldavia and Wallachia as a part of the newly created sovereign Kingdom of Romania. For almost half a millennium, the region was inhabited primarily by Turks and Tatars while Bulgarians and Romanians were in a minority. In 1878, the population of the Constanța County was composed of Tatars (38%), Romanians (23%), Turks (18%), Bulgars (13%), and other nations (8%). These proportions changed dramatically and rapidly, and by 1930, the same county was mainly inhabited by Romanians (66%) while the Turkic element constituted less than 10% of the total population, still about twice as many as it does presently.

Conquered in late 9th and early 10th century, Transylvania remained under Hungarian control, albeit with an increasingly high degree of autonomy, until the defeat at Mohács in 1526 when it effectively became independent, and an object of struggle between Austrians, Hungarians, and Ottomans. The country began to pay tribute to the Porte in 1543, and remained her vassal until the late 17th century. It was occupied by the Habsburgs since 1687, but the Ottoman Empire did not concede the loss until 1699. After World War I, Transylvania proclaimed unification with Romania; the act was internationally recognized in 1920.

Three regions have for some time been considered parts of Transylvania. In the southwest, Banat passed from Hungary to the Ottoman Empire in 1552, and from the Empire to the Habsburgs in 1718. After World War I, it was divided between Hungary, Yugoslavia, and Romania which obtained the largest, eastern part of the land. In the west, Crişana was divided between Transylvania and Austria until 1660 when it was seized by the Ottomans for twenty-six years before being incorporated into the Habsburg Empire until 1918. In the northwest, Maramureş shared the fate of the core part of Transylvania in the 16th and 17th century. It fell gradually under Habsburg control, was fully annexed in 1732, and remained Austrian till the end of World War I.

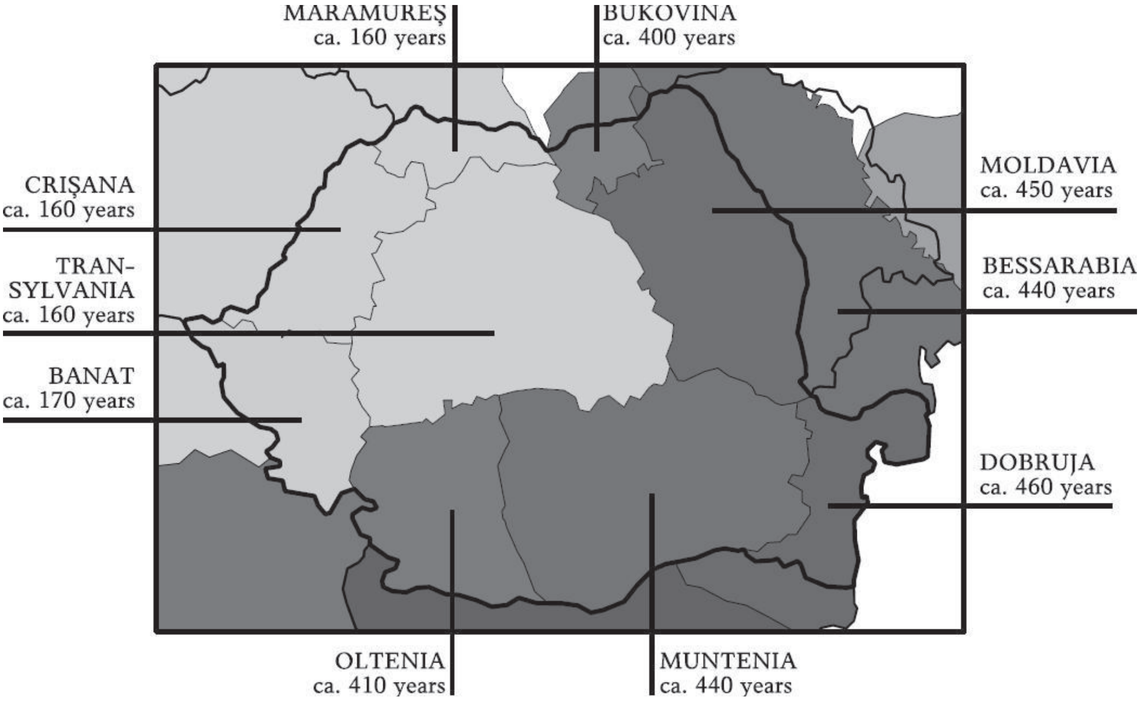


Figure 1. The approximate duration of Ottoman authority over historical regions where Romanian has been spoken, and over the neighbouring lands. Modern borders are added for reference.

1.2 Linguistic results

Remarkably, Romanians have carried their national and linguistic identity through the turbulent times, but traces of old partitions can still be seen in the language today. Dialectal division of Romanian has been the object of a long debate, see e.g. Ursan (2008: 77f) which primarily focused on the status of the Transylvanian variety, while the dialects of Banat, Moldavia, and Muntenia appear to be relatively clearly marked.

There is also a subtler result of this situation, which will be perhaps best illustrated by the word *taban*, which Suciu (2010: 709) defines so:[1]

TABÁN^{2} s[ubstantiv] n[eutru] (reg[ional]; 1879–); pl[ural] *-uri* şi *-e*.
1. (Tehn[ică], Agric[ultură]; Dobr[ogea]) ‘Patin de la charrue’: **tabán** 1884–1885. ***2.*** (Constr[ucţii]; Ban[at]) ‘Socle d’un four’: **tăván** sec[olul] XX/1. ***3.*** (Constr.; Mold[ova], S[ud] Transilv[ania]) ‘Poutre qui forme la base d’une construction’: **tabán**, **taván** sec. XX/1. ***4.*** (Constr.; Mold., Munt[enia], Olt[enia]; şi la pl., cu sens de s[in]g[ular]; în Mold. determinat prin *de jos*) ‘Plancher’: **taván** sec. XX/1. ***5.*** (*P[rin] ext[ensie]*; Constr.; rar) ‘Terrasse (d’une maison paysanne)’: **tabán** 1939. ***6.*** (Constr.; Mold.) ‘Planche longue et mince; latte’: **tabán** 1879, **taván** 1895, **tabá** sec. XX/1, **tabón** mijl[ocul] sec[olului] XX. ***7.*** (Vestim[entaţie]; Mold.) ‘Semelle intérieure, semelle première’: **tabán** 1900. ***8.*** (*P. anal[ogie]*; Vestim.; Mold., rar) ‘Bordure de pelisse d’un vêtement’: **tabán** 1924. ***9.*** (Med[icină]; V[est] Munt.) ‘Enflure du sabot des bêtes’: **taván** 1967. ***10.*** (Agric.; Dobr.) ‘Cadre de bois auquel on attelle les chevaux dans l’aire de battage’: **tabán** sec. XX/2.
– Din t[ur]c[ă(-osmanlie)] **taban** „idem (1, 2, 3, 4, 6)” şi, în general, ‘semelle; plante du pied; socle, support, base, partie inférieure’ […] probabil şi *„idem (10)”, **taban[astarı]** „idem (7)” […], **[iç] taban** „idem (7)” […], **taban [kabartısı** sau **şişi]** „idem (9)” […] – Cf. s[ârbo]cr[oată] *taban* „talpa piciorului; talpa plugului; fund de brazdă” [‘foot, ploughshare, bottom of the furrow’], n[eo]gr[ească] *ταμπάνι* „grindă, stâlp” [‘beam, pole, pillar’], arom[ână] *tăbáne* „grindă” [‘beam’], alb[aneză] *tabán* „talpa piciorului; talpa încălţămintei, pingea; creştetul capului” [‘foot, sole, crown’] […]

It seems easy to count loanwords based on just the head of the entry. On closer inspection, however, it quickly becomes apparent that the results so obtained may be debatable and, in fact, severely underestimated. Of course, Suciu (2009, 2010) is also aware of these difficulties, and discusses them in more detail in the first volume of his work (2009: 58f).

On the Romanian side, Dobr. *taban* ‘ploughshare’ (meaning 1) is, for example, quite unlikely to have resulted from the same act of borrowing as e.g.

[1] For readability, certain parts of the originally rather long entry have been omitted here without notice. In particular, sources have not been copied past the date of the first attestation, and Romanian definitions of meanings have been left out and only the French translations preserved. Where these were missing, English ones have been added in square brackets. Abbreviations are only explained where they appear for the first time, and their meaning is not always given in full (e.g. “Tehn.” stands for ‘tehnică, termen tehnic *(technique)*’, but here it is only expanded to “Tehn[ică]”).

Ban. *tăvan* ‘oven base’ (meaning 2), or even Dobr. *taban* ‘wooden block to which horses are attached during threshing’ (meaning 10), i.e. the same phonetic shape used in the same region but with a rather different semantics. On the Turkic side, *taban* ‘sole, heel, base, floor, &c.’ does not necessarily have to be considered the same etymon as the phrases *taban kabartısı* or *taban şişi* ‘swelling of *taban*’ which yielded western Munt. *tavan* ‘swelling of the inner hoof in cattle’ (meaning 9), and where *taban* is merely an attribute.

Meanings 5 and 8 are marked by Suciu as Romanian innovations (indicated by the qualifiers “P[rin] anal[ogie]” ‘by analogy’ and “P[rin] ext[ensie]” ‘by extension’, see Suciu (2010: 8)), highlighting the possibility of internal borrowing between Romanian dialects. Cases where semantic or phonetic change did not occur in the process may be virtually indistinguishable from words borrowed directly from Turkish. Also, note the presence of Alb. *taban* ‘foot, sole, crown’ and SCr. *taban* ‘foot, ploughshare, bottom of the furrow’, both of which might well have mediated the word from Turkish to Romanian. The eventual Turkic source may also have been Gagauz or Tatar in some cases, especially in 20th century borrowings in Dobruja and Moldavia (e.g. *amlă* ‘sitting board at the end of a fishing boat’, attested in Dobruja since 1971, or *saman* ‘straw’, attested in eastern Moldavia since 1978).

Thus, the actual number of Romanian words is open to discussion, and so is the number of etymons that they stem from, and the routes that they have taken. Technically, words borrowed via Serbian or any other language – including Romanian dialects – are not *Turkish* loanwords. Leschber (2011) discusses in more detail almost seventy Romanian words with such complicated etymologies, and several interesting cases can also be found investigated in Mitu (2002).

1.3 Treatment of the material

Since etymology is often unable to resolve the three issues listed above, and the answers are necessary before a quantitative analysis can begin, simplifications had to be made.

The material will be considered first as a whole, and then as a collection of independent, regional wholes. In the first case, Suciu’s judgement will be followed and every entry counted as one loanword. In the second, entries will be split, and each individual meaning with a regional qualifier treated as a separate borrowing – unless it is believed by Suciu to have evolved on Romanian ground. For example, *taban* above will count as one loanword in Romanian in the first part of the analysis, while in the second, as one borrowing in Banat (meaning 2), one in Oltenia (4), one in southern Transylvania (3), two in Dobruja (1, 10), two in Muntenia (4, 9), and four in Moldavia (3, 4, 6, and 7), and meanings 5 and 8 will be ignored as Romanian innovations. In the second part, also entries (meanings) without regional qualifiers will be omitted even when they are marked as literary. The beginnings of the Romanian literary language reach back to the 16th century (Gheție, Mareş 1985: 451) but it cannot be expected to have already been universally and fully adopted in the remote past.

This unequal treatment has been dictated by that the first part will be considering the heterogeneous sum of regional varieties and the literary language, while the second part will be dealing with each in isolation as if with a separate language.

The limitations mentioned in 1.2 and the methodological choice made above will also affect use of terms:

- “Romanian” will denote ‘Daco-Romanian, as a whole or any of its (sub)dialects’,
- “Ottoman” will be used to refer to ‘any Turkic language from which word(s) have penetrated into the Balkans as a result of the presence of the Ottoman Empire in the region’, i.e. not so much to the nation and the high language of its elites as to the political entity it ran. (See e.g. the ethnic composition of Dobruja mentioned in 1.1.)

The material analyzed in this paper has been primarily extracted from Suciu (2010). The datation of loanwords is also given in a table in Suciu (2009: 619f), and it was consulted where the notation was unclear (e.g. s.v. *bairac-agă*). More importantly, however, these series of dates are not always identical. The differences are discussed in 2.1.1, but essentially of no consequence for the main points of the paper.

1.4 Piotrovskij-Altmann law

First suggested by R.G. Piotrovskij in 1960, the so-called Piotrovskij-Altmann law has since evolved in Altmann (1983) and other works into a convenient formula for quantitative modelling of change in language (eq. 1).

(1) $$p(t)=\frac{1}{1+ae^{-bt+Ct^2}}.$$

It has three variants, that describe: (1.) a reversible (nonmonotonic) change ($C \neq 0$), or (2.) a non-reversible (monotonic) change, which can be either represented (2a.) in absolute terms (C = 0), or (2b.) as a fraction of the complete metamorphosis (C = 0, c = 1). For a more detailed introduction, see Strauss, Altmann ([2013]), and Stachowski (2013: 109f).

Variant 1. will be used as is and commented on in 2.1.1. Variant 2a. will be used in a variation with linguistically more meaningful coefficients (eq. 2, see Stachowski (2013: 110)). Variant 2b. has no application in our case as there is no fixed maximal number of loanwords that a language can possibly absorb:

(2) $$p(t)=\frac{1}{1+ae^{-b(t-A)}}.$$

2. Analysis

This section is composed of two parts, one where the entire Romanian territory is considered en bloc (2.1), and the other where the varieties of specific regions are analysed in isolation (2.2). The material is not treated equally in both parts, see 1.3.

2.1 Romanian as a whole

In this subsection, the Romanian literary language is treated as one with all of its regional and stylistic varieties. Each entry in Suciu (2010) counts as one loanword, regardless of whether it is literary or dialectal, or whether it appears in all regions with the same or different meanings and phonetics. More assumptions apply, see 1.3.

Suciu (2010) collects as many as 2776 words. Unlike most dictionaries, he does not limit himself to the first attestations but also provides the approximate time when the given word ceased to be used in Romanian. This allows the Piotrovskij-Altmann law (see 1.4) to be tested on a new kind of data, and significantly enriches the overall quantitative picture of Ottoman influence on Romanian. In 2.1.1, its total volume is discussed, and in 2.1.2, the longevity of loans from different periods.

2.1.1 Chronology

The Piotrovskij-Altmann law was fitted for three series: the number of loanwords newly appearing during the given century (cumulatively, symbolized ● in Fig. 2, fitted to eq. 2), the number of disappearing loanwords (also cumulatively, symbolized ■, also fitted to eq. 2), and the number of loanwords that were in use at the given time (symbolized ▲, fitted to equation 1). The results have proven to be very precise; see Figure 2 and Table 1.

The last of these series requires additional attention. It was mentioned in 1.3 that Suciu (2009; the monographic volume) and Suciu (2010; the dictionary volume) provide dates that do not match up. Suciu (2010) ignores uncertain attestations and does not inform us about gaps, e.g. *părpăți* ‘Oriental-style white woolen trousers, wrinkled, loose at the top and tight at the bottom, worn by Romanian peasants’ is marked as “probably used” in the first half of the 19th century in Suciu (2009), and as “1884–” in Suciu (2010), while e.g. *hamailâu* ‘amulet, talisman’ is marked as first attested in the second half of the 16th century and then throughout the 19th century in Suciu (2009), and as “1594–sf[ârşitul] sec[olului] XIX” in Suciu (2010). In some cases, the nature of the inconsistency is less clear, e.g. *pazar* ‘market, fair; buying and selling, transactions, haggling’ is marked in Suciu (2009) as attested in the second half of the 15th century, and

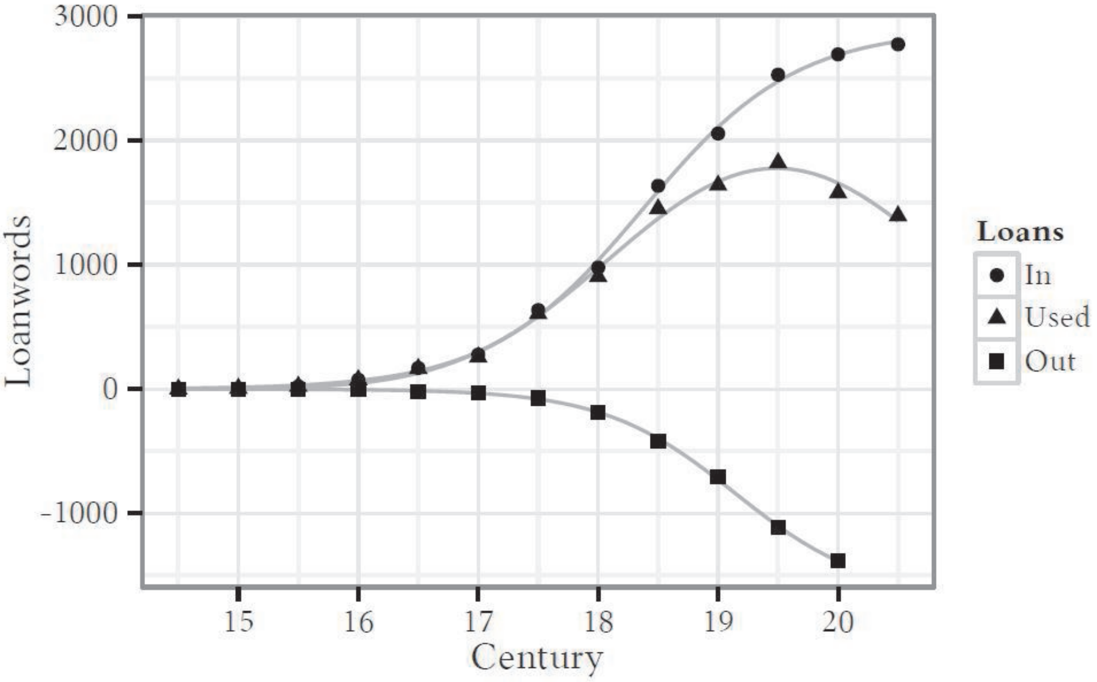


Figure 2. Fitting of the Piotrovskij-Altmann equation to the Ottoman-Romanian data. The three series are: loanwords coming into use (●, cumulatively), in use at the time (▲), and going out of use (■, cumulatively, and multiplied by -1 for graphical clarity). The middle series is based on Suciu (2010). Century numbers were decreased by 19 to enable fitting; see the main text. Fitted with *R*.

then from the first half of the 19th century on, and as “1500–” in Suciu (2010). Also, Suciu (2010: 826f) adds sixteen new words and withdraws one. Both volumes distinguish between words that were normally used in the given period, and those that were seen as dated or obsolete. In the material extracted here from Suciu (2010), both are treated as still in use. Suciu (2009), on the other hand, provides two series of summary dates, one that includes uncertain attestations and such archaic words, and one that does not.

All three series can be accurately modelled with eq. 1, as is shown in Fig. 3. However, in order to achieve the match, century numbers had to be decreased. This is not unacceptable in itself as dates are a pure convention, but it was my arbitrary decision to decrease them by precisely 19. In the case of eq. 2 (non-reversible change, the “In” and “Out” series in Fig. 2), the meanings of coefficients are clear: *A* represents time (the location of the curve on the horizontal axis),[2] b the intensity (the slope of the curve), and c the strength (the height of the curve). With eq. 1 and $C \neq 0$ (reversible change), however, their function cannot be so clearly defined. They hinge on by how much the dates have been decreased, and their variation is also much greater (compare Tables 1 and 3).

Table 1
The empirical (actual) and theoretical (fitted) number of Ottoman loanwords in Romanian. Coefficients for loanwords in use at the time (▲) refer to eq. 1; the other two series to eq. 2. The R^2 coefficient of determination represents essentially the proportion of variability in the dependent variable that is accounted for by the model.

Century	● In (cumul.)		▲ Used		■ Out (cumul.)	
	empirical	theoretical	empirical	theoretical	empirical	theoretical
14.5	1	6.6	1	0.9	0	0.4
15.0	6	14.4	6	3.8	0	0.9
15.5	22	31.5	22	14.5	0	2.2
16.0	75	68.2	75	46.8	4	5.5
16.5	170	145.4	166	128.7	20	13.5
17.0	278	300.8	258	299.4	29	33.3
17.5	636	587.1	607	587.5	73	80.3
18.0	979	1036.9	906	973.5	185	186.4
18.5	1637	1593.9	1452	1373.3	416	397.6
19.0	2058	2111.0	1642	1671.9	709	730.7
19.5	2531	2477.4	1822	1778.2	1116	1102.1
20.0	2695	2690.4	1579	1660.2	1381	1385.2
20.5	2776	2800.1	1395	1353.5	NA	NA
	A = 18.373		a = 5.011		A = 19.141	
	b = 1.571		b = 0.306		b = 1.821	
	c = 2899.267		c = 10050		c = 1674.76	
			C = 0.3149			
	R^2 = 0.999		R^2 = 0.996		R^2 = 0.999	

[2] Note that A is in the same units as the original data. In this paper, time is given in centuries, so that e.g. A = 18.37 corresponds to the year 1737 rather than *1837.

In their present form, the coefficients do not in themselves add to the linguistic value of the Piotrovskij-Altmann law in its “reversible change” variant. It is not clear to me how the formula can be improved. Possibly, it might be beneficial to view nonmonotonic change as a sequence of two monotonic changes, and do away with eq. 1 altogether, concentrating only on eq. 2. One might also want to note, as a separate issue, that all the three series presented in Fig. 3 have in fact a rather sharper peak than the fitted curves suggest. Perhaps just a coincidence, the same phenomenon can be observed in the data collected and analyzed by Imsiepen (1983), unaffected by smoothing in Best, Beöthy, Altmann (1990).

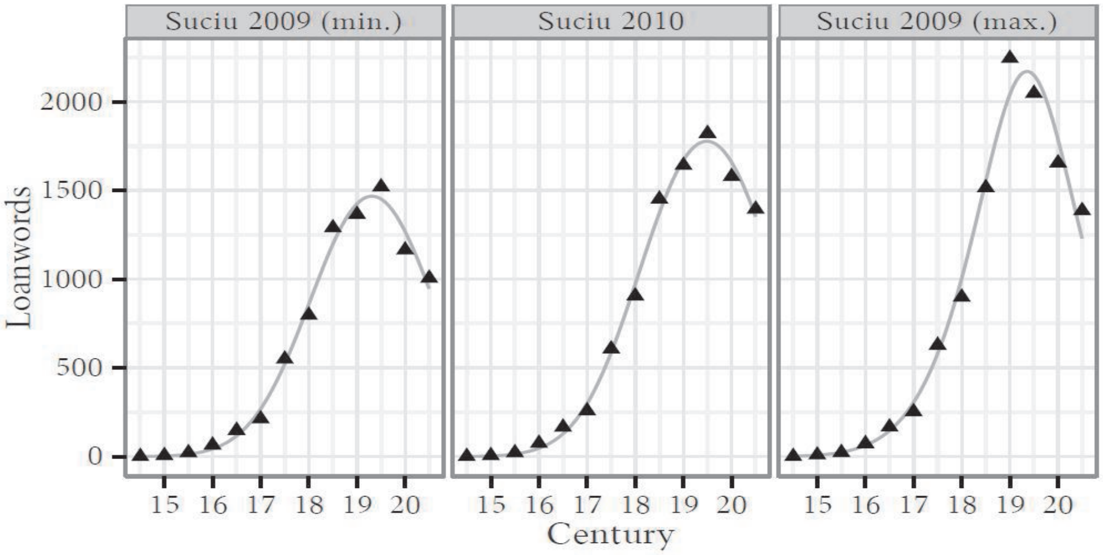


Figure 3. The number of Ottoman loanwords in use in Romanian at different times, and the fitted Piotrovskij-Altmann curves (eq. 1). The three series are a result of different selection criteria. Century numbers have been decreased by 19 in all cases, see the main text. Fitted with *R*.

Table 2
The coefficients used in the fitting of eq. 1 to the number of Ottoman loanwords in use in Romanian at different times. The three series are a result of different selection criteria. Century numbers have been decreased by 19 in all cases, see the main text. Fitted with *R*.

	Suciu (2009; min.)	Suciu (2010)	Suciu (2009; max.)
a	14.74	5.011	-2.1880
b	0.2043	0.306	0.1849
c	22430	10050	-2426.9160
C	0.3283	0.3149	0.2597
R^2	0.990	0.996	0.987

The very precise fitting obtained here is, no doubt, due in some degree to the fact that many words are only dated with the accuracy of half a century. This forced binning of the data has ironed out many possible outliers and improved the goodness of fit but, at the same time, obscured the correlation with historical events; see also Stachowski (2013: 111f). Naturally, the material being essentially the sum of regional varieties and the literary language, precise correspondence could not have been expected in any case.

What can be observed is that the influx does not gain impetus until the first half of the 17th century and once acquired, maintains it till the beginning of the second half of the 19th century – which is also when the outflux, after having visibly accelerated about a hundred and fifty years earlier, begins to surpass the influx, and the number of loanwords in use starts finally to drop.

A historical explanation can easily be found for the closing date: severely weakened by the end of the 18th century, the Ottoman Empire looses its authority over most of Romanian lands in the first half of the 19th century, and in 1878, they unite and form the independent Kingdom of Romania.

The starting date is less obvious. One may perhaps speculate about the intensification of Ottoman control in that period, but another factor might also need to be taken into account. The oldest surviving document in Romanian is dated 1521, and the 16th century is generally when the literary language begins to form, see e.g. Gheție, Mareş (1985: 450f). The increase of the number of recorded loanwords does not seem to correlate with one particular event in the political history but it does coincide with the rise of writing in Romanian. It is difficult to estimate to what extent this has affected our results, but the issue certainly deserves attention. A similar phenomenon can be observed in the regional data, see 2.2.1 below.

2.1.2 Longevity

Thanks to the fact that Suciu (2009, 2010) provides not only the dates of first attestation, but also the approximate time of disappearance of Ottoman loanwords from Romanian, we are allowed a rare chance to inspect their longevity. Suciu himself devotes much thought to this question in respective chapters in (2009: 109f), but he does not seem to proceed to make general statements.

Fig. 4 might create the impression that, with the exception of the 20th century, the later the borrowing took place, the more short-lived it has proven to be. A detailed examination of these words, however, fails to reveal clearly why it should be so.

Let us divide Fig. 4 into three parts: the first stretching from the very beginning to the first half of the 18th century, the second from the second half of the 18th to the second half of the 19th century, and the last encompassing only the first half of the 20th century.

Politically, the first part covers the period of almost undisputable domination of the Porte over most or all of the Romanian regions. Similarly to what was said in 2.1.1 above, one might wish to explain the visible rise in the number of short-lived loanwords during that period by speculating about tightening of the Ottoman grip. But again, if the history of Romanian writing is factored in, together with the fact that the oldest surviving purely Romanian documents are not older than the 16th century, another explanation becomes available, namely that the distribution of longevity of borrowings might have been in fact relatively stable throughout the period, only the oldest examples were simply either never attested or the documents that contained them did not survive till our times.

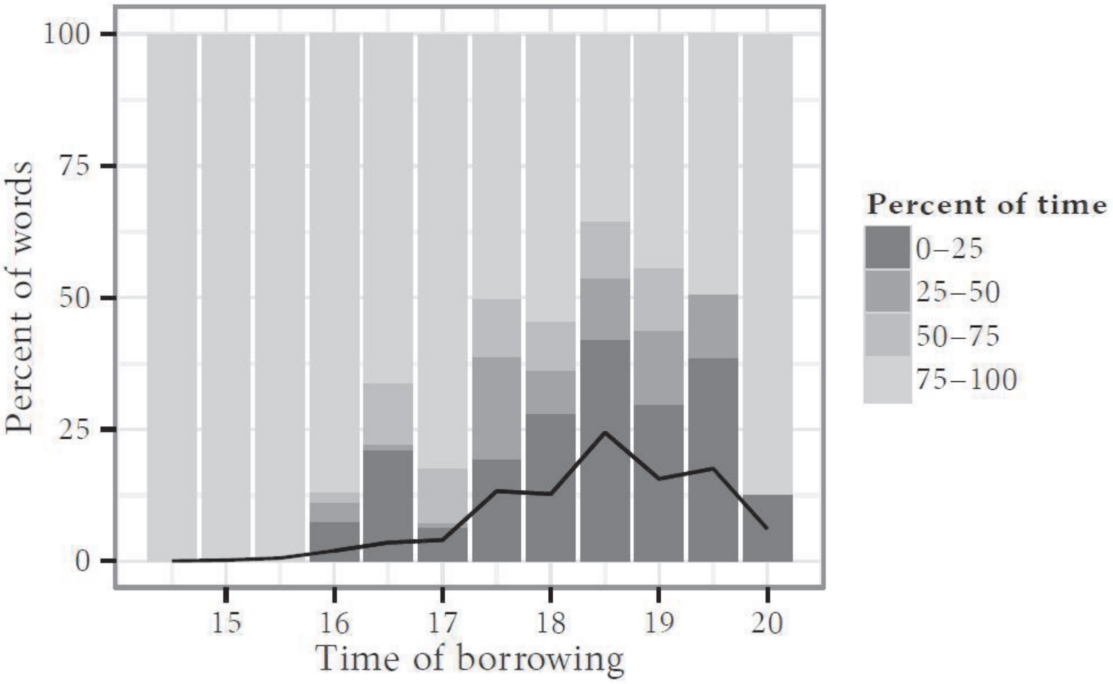


Figure 4. Longevity of Ottoman loanwords in Romanian. Bars represent the percent of words borrowed at the given time that have survived in the language for a specific percent of the time till nowadays. For example, all of the words borrowed in the second half of the 14th century (the bar at 14.5) have remained in Romanian for 75–100% of the six centuries that have passed since their appearance; at the same time, more than 27% of words borrowed in the second half of the 18th century have disappeared after 0–25% of the two centuries that have since passed. Words marked as *historic* by E. Suciu are not treated in a special way. The line shows what percentage of the total number of loanwords comes from the given time. For example, almost a quarter of all Ottoman loanwords that were ever in Romanian were borrowed in the second half of the 18th century.

The second period is the time of the fight for independence, and of much increased volume of writing. Possibly, many loanwords were simply the need of the hour, that happened to have appeared at the right time to become recorded.

Finally, in the last period the representation of longevity as a percentage somewhat breaks down. With the best available accuracy of half a century, words borrowed in the first half of the 20th century could have either survived for 0% of the maximum time, i.e. disappeared in the same period, or 100% of the maximum time, i.e. still have existed in the second half of the century. The visible drop in the former category is surely a result of regaining independence in 1878.

It remains to be observed that the absolute number of loans in the given period correlates quite well with the number of the most short-lived and that of the most long-lived borrowings (Spearman's ρ = 0.979 and 0.986, respectively), but not so well with the two middle categories (ρ = 0.856 and 0.614). The relation appears to be logarithmic rather than linear in nature, but requires a more detailed examination before conclusions can be drawn or explanations suggested.

2.2 Regional varieties

In this subsection the material collected in Suciu (2010) is treated as belonging to separate varieties. (The term *dialect* is deliberately not used here as Suciu's qualifiers refer to historical regions rather than the still disputed dialectal division of Romanian, see e.g. Ursan (2008).) Individual meanings are considered independent words (see 1.3 for justification), and only the words with regional qualifiers are taken into account. Also the words marked as "literary" are discarded because they are often additionaly labelled as being used "today mostly in …" which suggests that even when known in other regions, they do not strictly belong to the regional speech.

By such criteria, Suciu (2010) contains as many as 1513 words from six regions: Banat, Dobruja, Moldavia, Muntenia, Oltenia, and Transylvania. Occasionally, locations beyond the core of the modern range of Romanian are marked. They will not be discussed here, and the six words that are not attested outside of these areas will be ignored (northwestern Bulgaria: *chimur*, *peringi*, northern Serbia: *pai*, *tere*, southern Ukraine: *otac*, *tarciniu*). Additional geographical specifications are also sometimes given, such as "eastern", "northwestern", &c. They will not be taken into account here except in "northern Moldavia" which denotes 'southern Ukraine' and "eastern Moldavia" which stands for the 'Republic of Moldova', see Suciu (2010: 6). The former will be ignored, as was stated before, and the latter will be treated as a separate region for historical reasons (independent from Romania since 1947).

Regional collections are not nearly as rich as the general Romanian material. Apparently, the historic data are particularly incomplete, as will be seen in 2.2.1. For this reason, longevity of borrowings will not be discussed, and the subsequent parts will be limited to only those loanwords that are still in use today.

2.2.1 Chronology

The general Romanian material was analyzed with respect to both the time of appearance and that of disappearance of loanwords (2.1.1). The regional data, however, do not appear to be sufficiently representative.

It can be seen from Fig. 5 that the Piotrovskij-Altmann law can very accurately model the influx of Ottoman loanwords into regional varieties of Romanian. The correlation with historic events, however, is below expectations.

The rise of the rate of influx begins before suzerainty in Banat and Transylvania, and inversely in the remaining five regions. But in all the regions, it can be said to gain impetus in the first half of the 16th century, i.e. together with the growth of writing in Romanian. Similarly then to the general material discussed in 2.1 above, the historic data appear to reflect more the volume of Romanian writing than that of Ottoman influence.

Data on the decrease of the rate of influx come from newer and better attested times, and appear to be more representative. In almost all the regions, the period of rapid growth ends with the second half of the 19th century, i.e. just when Romania gained independence. Only Dobruja is an exception, most probably because Turkic-speaking peoples remained a sizeable part of its population well into the 20th century. If predictions can be made based on the Piotrovskij-Altmann law, a significant drop in the rate of influx is not to be expected before mid-22nd century. Intuitively, a more imminent decrease would seem more likely.

Banat and Transylvania were under Ottoman authority for a relatively short time. The former was fully incorporated into the Empire, the latter managed to maintain a degree of autonomy, but the influx of Ottoman words proceeded in both at a very similar, staid pace. In all the other regions, a sharper spike can be observed in the second half of the 19th century. Supposedly, it was caused primarily by B.P. Hasdeu’s extensive dialectological questionnaire started in 1884 (the manuscript of the responses comprised eighteen volumes, Suciu (2010: 17) s.v. *H*), L. Şăineanu’s *Influența orientală asupra limbei şi culturei române* (1900), and by other publications prompted by the general growth of interest in folklore at the time. The political situation and fight for independence might have contributed to the zeitgeist but they would be an unlikely cause of the increased influx of Ottoman loanwords.

Overall, the regional data are strongly influenced by the volume of Romanian writing at the given time. Pre-16th century parts do not at all seem complete, but their accuracy appears to improve with time, and they might be expected to have reached a fair degree of representativeness with the end of the 19th century.

For material ranging from the second half of the 14th to the second half of the 20th century, this is quite late. The number of loanwords going out of use, therefore, and the number of loanwords currently in use at different times, will not be discussed here.

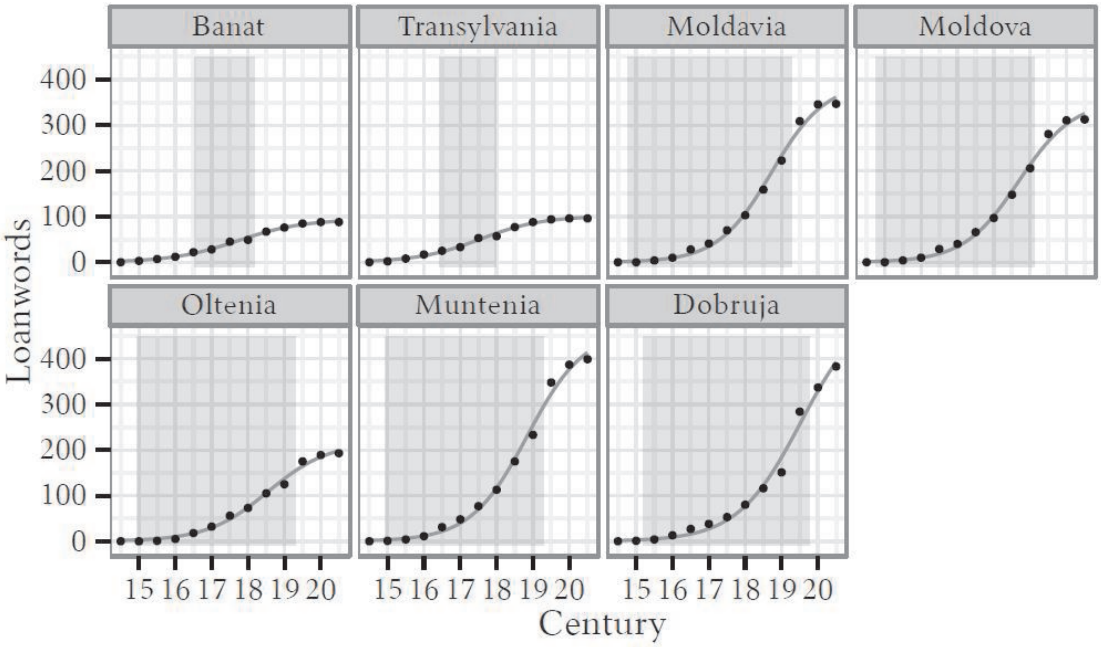


Figure 5. The influx of Ottoman loanwords into regional varieties of Romanian, and the fitted Piotrovskij-Altmann curves (eq. 2). Regions are ordered to resemble the geographical distribution rather than alphabetically. Grey rectangles show the duration of Ottoman authority over respective regions; short episodes of dependency to other countries are not marked (e.g. Austrian rule in Oltenia in the years 1716–39, see 1.1). Fitted with *R*.

Table 3
The coefficients used in the fitting of eq. 2 to the influx of Ottoman loanwords into regional varieties of Romanian. Fitted with *R*.

	A	*b*	*c*	R^2
Banat	17.717	1.114	94.045	0.995
Dobruja	19.532	1.145	522.093	0.989
Moldavia	18.726	1.353	394.402	0.995
Moldova	18.687	1.324	355.273	0.995
Muntenia	18.849	1.281	463.288	0.994
Oltenia	18.554	1.187	217.97	0.994
Transylvania	17.553	1.166	101.446	0.994

2.2.2 Duration and volume of influence

One might be tempted to link the volume of influence to the duration of Ottoman authority over individual regions, and the time that has elapsed since its end. Historical data not being fully representative (see 2.2.1), let us limit the test to those words only that are still in use today.

Settling on the exact number of loanwords was shown in 1.2 to be a neither obvious nor an entirely objective procedure, and so is determining the precise number of years a region was, and then was not, under Ottoman authority. Some were incorporated into the Empire; others retained varying degrees of autonomy; in some the Turkic presence was considerably more visible at certain times than in others. The brief summary in 1.1 cannot do justice to the complex history of the region.

For the purpose of the experiment, the numbers shown in Fig. 6c were used. Two formulas were tested, yielding the following relations (*l* stands for ‘loanwords in use today’, t_O for ‘duration of Ottoman authority’, and t_n for ‘time since the end of Ottoman authority’; fitted with *R*):

(3a) $l = 0.4384t_0 - 0.6382t_n + 205.1582$,

(3b) $l = 0.5020\left(t_0 - t_n\right) + 154.881$.

The accuracy is effectively identical in the two cases, $R^2 = 0.784$. This is a disappointingly low result. As it happens, the correlation between the mutually independent t_O and t_n dwarfs them both (Pearson’s $\rho = -0.9628736$).

It is possible that the problem lies in precision. Muntenia and Oltenia are often regarded as merely the two parts that together constitute Wallachia. This is a perfectly justified practice from the historical point of view, and when applied to our case, it yields the following relations:

(4a) $l = 0.6824t_0 - 0.3436t_n + 82.5277$,

(4b) $l = 0.57398\left(t_0 - t_n\right) + 167.72428$,

which have a considerably improved accuracy of $R^2 = 0.908$ for eq. 4a, and 0.907 for 4b. This effect must be akin to what is observed when fitting the Piotrovskij-Altmann law to grouped, and to ungrouped data (see 2.1.1).

2.2.3 Similarity of regions

The duration and extent of Ottoman authority differed considerably from region to region. The unification of the entire Romanian territory happened in several steps over almost a century, and was followed by the loss of Moldova and parts of Bukovina and Dobruja in 1947. It is to be expected, then, that some regions will be more similar to one another than to others. As in 2.2.2 above, the discussion will be limited to words that are currently still in use.

Perhaps the most conspicuous feature in Fig. 7 is that the similarities appear to more reflect historical divisions than geographical proximity, which suggests a relatively low level of communication between the regional varieties even during the past century when they were parts of one state.
Three partitions are especially visible: Transylvania with Banat, Moldavia with Moldova, and the remaining four regions which formed the Kingdom of Romania in the years 1878–1913.

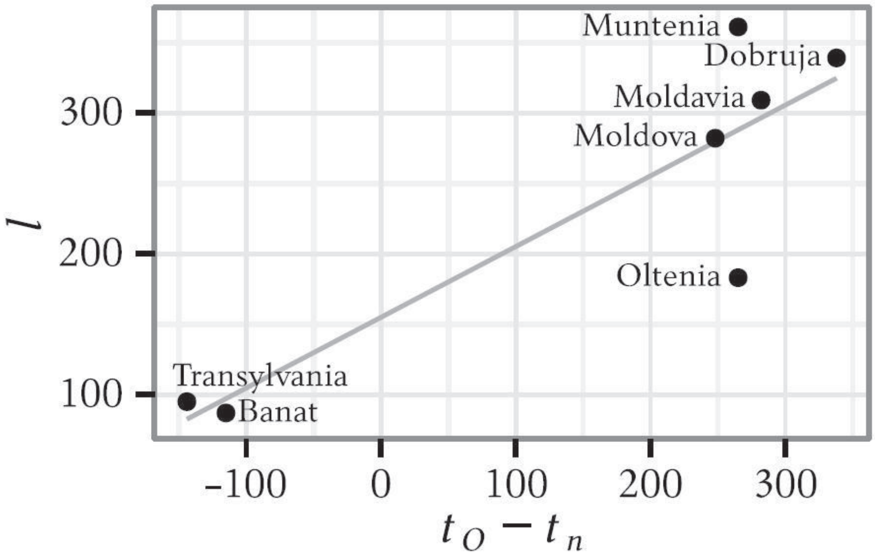


Figure 6a. Muntenia and Oltenia considered separately (eq. 3b). See Table 4.

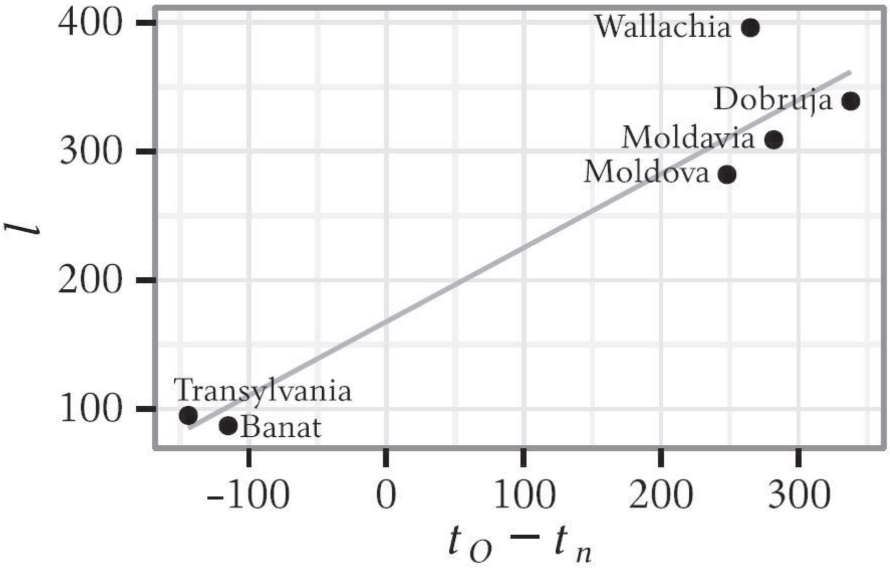


Figure 6b. Muntenia and Oltenia considered together (eq. 4b). See Table 4.

Banat is more alienated than any other region. Its stock of Ottoman loanwords only resembles that of Transylvania, but not at all even that of the neighbouring Oltenia. This could be explained by its peripheral location, but Transylvania which lies in the very centre of the country also does not really seem to have assimilated its stock to Moldavia's or Wallachia's, or acted as a middleman between the two. Banat and Transylvania share a large part of their history, and were only officially tied to the other regions after World War I, although they have been populated in a considerable proportion by Romanians for at least several centuries. Note also that the separateness of the Banatian dialect does not raise doubts, while the Transylvanian variety has eluded clear categorization for a very long time (see 1.2).

Table 4
The relation between the number of Ottoman loanwords in use today in different regions (l), the time these regions remained under Ottoman authority (t_O), and the time that has since passed (t_n), with reference to the year 1999, as Suciu's data covers the period till the end of the 20th century). See also Figures 6a and 6b.

	l	t_O	t_n
Banat	87	166	281
Dobruja	339	459	121
Moldavia	309	452	170
Moldova	282	435	187
Muntenia	361	435	170
Oltenia	183	435	170
Transylvania	95	156	300
Wallachia	396	435	170

Moldavia and Moldova were considered one region for some time. In fact, Suciu (2010) refers to the latter as 'eastern Moldavia'. In the years 1947–91, however, it was part of the USSR and this proved sufficient to kindle a national consciousness. In the sphere of Ottoman loanwords, nonetheless, the effect appears to have been minimal. The stock is very similar in the two regions, and has relatively many similarities to the other parts of the first Kingdom of Romania. Only Moldova and Dobruja are further part, due perhaps to the geographical wedge of Budjak.

Muntenia and Oltenia are commonly seen as parts of one region, therefore their relative proximity does not surprise, and neither does the fact the the stock of the eastern one of the two, Muntenia's, is more similar to the stock of Dobruja. Oltenia's similarity to Moldavia and Moldova is perhaps slightly more unusual as they do not have a shared border. The mediation must have occurred via Muntenia. Geographical proximity appears to have only played a more important role in the case of Dobruja.

Table 5
The similarity of the stocks of Ottoman loanwords between Romanian regions.
The index used is the Jaccard distance, $\left|\frac{A \cap B}{A \cup B}\right|$,
where 1 denotes identity, and 0 no similarity at all.

	Banat	Dobruja	Moldavia	Moldova	Muntenia	Oltenia	Tran-sylvania
Banat	1						
Dobruja	0.052	1					
Moldavia	0.073	0.232	1				
Moldova	0.076	0.190	0.870	1			
Muntenia	0.067	0.296	0.376	0.331	1		
Oltenia	0.098	0.165	0.224	0.214	0.374	1	
Transylvania	0.422	0.061	0.107	0.102	0.096	0.121	1
all	0.104	0.262	0.418	0.388	0.380	0.219	0.115

3 Conclusions

Although unusually rich, the material collected in Suciu (2009, 2010) has not proven to be sufficiently representative for a quantitative diachronic analysis. Especially older parts of the collection depend heavily on the volume of writing in Romanian that has survived to our times. The problem affects the general Romanian and the regional data equally (2.1.1–2.1.2, 2.2.1). The issue is not of major importance for the present paper, but it highlights a possible weakness in some of the results obtained so far for the Piotrovskij-Altmann law.

Remarks were also made in 2.1.1 about the "reversible" variant of the law. It was pointed out that its coefficients lack clear linguistic meaning, and that the resulting curve fails to reflect what might be a characteristic spike in nonmonotonic phenomena in language.

Regardless of these shortcomings, the discussion in 2.1.2 appers to suggest that the total number of loanwords from a certain period depends primarily on those borrowings which are about to disappear very quickly, and those which are

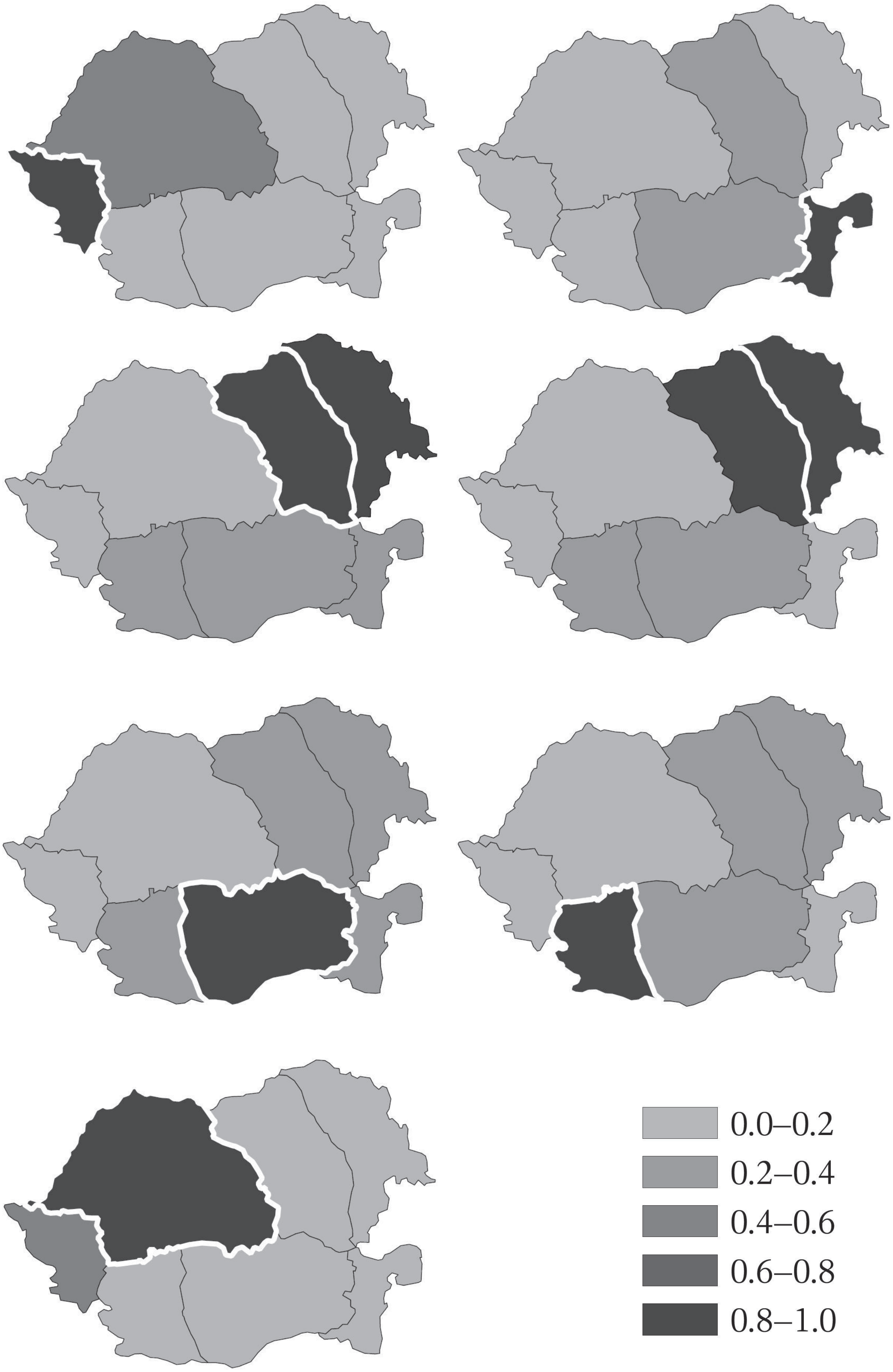


Figure 7. The similarity of the stock of Ottoman loanwords in the highlighted, and the remaining regions of Romania. The index used is the Jaccard distance, , where 1 denotes identity, and 0 no similarity at all.

to survive for a rather long time. The medium-lived ones seem to be the least numerous group, and more randomly distributed across the centuries.

The number of surviving loanwords was shown in 2.2.2 to bear some correlation with the duration of Ottoman authority, and the time that has elapsed since its end. Two regions, Muntenia and Oltenia, did not quite conform to the tendency set by the other five. When they were considered together, however, as is often the historically justified practice, the accuracy of the fitting significantly improved.

Lastly, the stock of Ottoman loanwords in different regions was compared in 2.2.3 to reveal that the old political partitions have apparently had a greater impact on transmission of borrowed vocabulary than did geographical proximity. The present paper does by no means exhaust the topic or the material contained in Suciu (2009, 2010). This would require a much larger enterprise. Here, just an overall quantitative characteristic has been presented, together with some theoretical observations, and hopefully a way of combining qualitative and quantitative research and findings to obtain a new kind of insights.

Abbreviations

Alb. = Albanian | **Ban.** = Banatian | **Dobr.** = Dobrujan | **Munt.** = Muntenian | **SCr.** = Serbo-Croat

References

Ágoston, G., Masters, B. (eds.) (2009). *Encyclopedia of the Ottoman Empire*. New York: Facts On File.

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: In: Best, K.–H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte: 59–90*. Göttingen: Edition Herodot.

Best, K.-H. (2005). Turzismen im Deutschen.*Glottometrics 11, 56–63*.

Best, K.-H., Beöthy, E., Altmann, G. (1990). Ein methodischer Beitrag zum Piotrowski-Gesetz. *Glottometrika 12, 115–124*.

Best, K.–H., Kohlhase, J. (eds.), (1983). *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte*. Göttingen: Edition Herodot.

Gheție, I., Mareş, A. (1985). *Originile scrisului în limba română*, Bucureşti, Editura Ştiințifică şi Enciclopedică.

Hřebíček, L. (1990). Quantitative studies. In: Hazai, G. (ed.), *Handbuch der türkischen Sprachwissenschaft 1: 371–387*. Budapest: Akadémiai Kiadó.

Imsiepen, U. (1983). Die *e*-Epithese bei starken Verben im Deutschen. In: Best, K.–H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung. Theoretische*

Beiträge, statistische Analysen und Arbeitsberichte: 119–141. Göttingen: Edition Herodot.

Leschber, C. (2011). Lehnwege einiger Orientalismen und Wörter eurasischer Herkunft im Rumänischen und den sonstigen Balkansprachen. *Studia Etymologica Cracoviensia 16, 33–61*.

Mitu, M. (2002). Orientalizmy leksykalne w języku polskim i rumuńskim. In: Rusek, J., Boryś, W., Bednarczuk, L. (eds.), *Dzieje Słowian w świetle leksyki: 301–306*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Pop, I.-A., Bolovan, I. (eds.) (2006). *History of Romania. Compendium*. Cluj-Napoca: Romanian Cultural Institute. Center for Transylvanian Studies.

Sallanz, J. (ed.) (2005). *Die Dobrudscha: Ethnische Minderheiten – Kulturlandschaft – Transformation. Ergebnisse eines Geländekurses des Instituts für Geographie der Universität Potsdam im Südosten Rumäniens* (= *Praxis Kultur- und Sozialgeographie* 35). Potsdam: Universitätsverlag Potsdam.

Stachowski, K. (2013). The influx rate of Turkic glosses in Hungarian and Polish post-mediaeval texts. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3, dedicated to Karl-Heinz Best on the occasion of his 70th birthday: 100–116*. Lüdenscheid: RAM-Verlag.

Strauss, U., Altmann G., *Change in language*. In: Laws in Quantitative Linnguistics. http://lql.uni-trier.de/index.php/Change_in_language [July 2013]

Suciu, E. (2009). *Influența turcă asupra limbii române, 1: Studiu monographic*. Bucureşti: Editura Academiei Române.

Suciu, E. (2010). *Influența turcă asupra limbii române,* 2: *Dicționarul cuvintelor româneşti de origine turcă*. Bucureşti: Editura Academiei Române.

Sugar, P. F. (1977). *Southeastern Europe under Ottoman Rule, 1354–1804*. Washington: University of Washington Press.

Ursan, V., (2008). Despre configurația dialectală a dacoromânei actuale. *Transilvania 1, 77–85*.

Addresses of authors

Altmann Gabriel
Lüdenscheid, Germany
ram-verlag@t-online.de

Andreev Sergey
Smolensk State University, Smolensk, Russia
smol.an@mail.ru

Andres Jan
Palacký University, Olomouc, Czech Republic
jan.andres@upol.cz

Buk Solomija
Ivan Franko National University of Lviv, Lviv, Ukraine
solomija@gmail.com

Čech Radek
University of Ostrava, Ostrava, Czech Republic
cechradek@gmail.com

Fan Fengxiang
Dalian Maritime University, Dalian, China
fanfengxiang@yahoo.com

Huang Wei
Beijing Language and Culture University, Beijing, China
huangwei@blcu.edu.cn

Kelih Emmerich
University of Vienna, Vienna, Austria
emmerich.kelih@univie.ac.at

Köhler Reinhard
Universität Trier, Trier, Germany
koehler@uni-trier.de

Kubát Miroslav
Palacký University, Olomouc, Czech Republic
miroslav.kubat@gmail.com

Liu Haitao
Zhejiang University, Hangzhou, China
lhtzju@gmail.com

Mačutek Ján
Palacký University, Olomouc, Czech Republic
jmacutek@yahoo.com

Mácha Jiří
Charles University, Prague, Czech Republic
jiri.macha.84@gmail.com

Melka Tomi S.
Parkland College, Champaign (IL), USA
tmelka@gmail.com

Mikros George K.
National and Kapodistrian University of Athens, Athens, Greece
gmikros@gmail.com

Milička Jiří
Palacký University, Olomouc, Czech Republic
Charles University, Prague, Czech Republic
milicka@centrum.cz

Richterová Olga
Charles University, Prague, Czech Republic
richterova.olga@gmail.com

Rovenchak Andrij
Ivan Franko National University of Lviv, Lviv, Ukraine
andrij@ktf.franko.lviv.ua

Sanada, Haruko
Saitama Gakuen University, Saitama, Japan
h_sanada@nifty.com

Stachowski Kamil
Jagiellonian University, Cracow, Poland
kamil.stachowski@gmail.com

Su Hong
Dalian Maritime University, Dalian, China
suzihesuhong@126.com

Uhlířová Ludmila
Czech Academy of Sciences, Prague
lidauhlirova@seznam.cz
uhlirova@ujc.cas.cz

Wang Lu
Universität Trier, Trier
wanglu-chn@hotmail.com

Wimmer Gejza
Matej Bel University, Banská Bystrica
Slovak Academy of Sciences, Bratislava
wimmer@mat.savba.sk

Zhou Pianpian
Dalian Maritime University, Dalian
zhoubianer@163.com