Glottometrics 35
2016

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies.**

Herausgeber – Editors

External academic peers for Glottometrics

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 35 (2016), Lüdenscheid: RAM-Verlag, 2016. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 35 (2016) **ISSN 2625-8226**

Contents

Glottometrics 35, 2016, 1-5

German Compounds in the Texts of Technical Science

Ekaterina Shmidt[1]
Hanna Gnatchuk[2]

Abstract: The present investigation is engaged with a quantitative study of German compounds in the text of technical science. We have analyzed word classes for German compounds in Book "Wirtschaftsinformatik" by H. R. Hansen et al (2015). In such a way, 20 pages of the above-mentioned book have been studied with a sample of 221 German compounds. The data have been processed statistically. The results can be of great use for typological studies of compounds.

Keywords: German, compounds

1. Introduction: Some notes on German compounds

Compounds are considered to be the most productive ways of enriching the vocabulary of any language. According to Oguy (2003) "der Begriff *Zusammensetzung* bezeichnet sowohl einen spezifischen Typ der Wortbildungsprozesses (Komposition), der in der Verbindung einiger freier Morpheme besteht, als auch sein Ergebnis (Kompositum)". As far as the reasons for the appearance of compounds are concerned, they can be summarized as follows:

- The absence of the appropriate name for designating a notion or a thing;
- Language economy;
- The language unit gets older;
- The creation of euphemisms, metaphors, puns.

In the German morphology, one distinguishes three types of compounds: *determinative, coordinative copulative and independent compounds (unabhängige Zusammenrückungen)*. We are intended here to have a look at the above-mentioned types:

- *Determinativkomposita* (determinative compounds) consist of basic and specifying which are designated as "Grundwort" and "Bestimmungswort" (Oguy, 2003 : 200). By the way of illustration, let us take a German word *Großstadt*: *Stadt* = basic word, *groß* = specifying word. We can also take a more complex word – *Schreibwarengeschäft*. The first two components (*Schreib* + *Waren*) are specifying words. The third one (*Geschäft*) is a basic word which determines gender and word class of the whole compound. The most productive models of a determinative compound is Noun + Noun and Adjective + Noun. As far as the types of meanings are concerned, it is possible to differentiate *exocentric* and *endocentric* compounds. *The endocentric determinative compound*_means the notion formed by a sum of meanings of the components of a compound: Fensterbrett = "ein Brett am unteren Ende des Fensters". The exocentric compound "geht die Gesamtbedeutung über die Bedeutungen einzelner

[1] Jekaterina Shmidt, Alpen-Adria Universität, Institut für Romanistik, Universitätsstraße 65-67, 9020 Klagenfurt, Austria. Email: jekaterinaschmidt@gmail.com
[2] Hanna Gnatchuk, Universität Trier, Computational Linguistics and Digital Humanities, Universitätsring 15, Trier, Germany. Email: agnatchuk@gmail.com or s2hagnat@uni-trier.de

Bestandteile hinaus: ein Schlaukopf" which is applied to designate a person and his/her character (Oguy, 2003 : 201);
- *Kopulativkomposita* (copulative compounds) consist of equal constitutents of one word class: Dichter-Komponist, achtundvierzig, Strumpfhose. The most characteristic models are Noun + Noun (Strumpfhose) or Adjective + Adjective (taubstumm);
- "Zusammenrückungen" are represented by nominal word groups of different word classes and imperative sentences. Here the second constituent of a compound does not determine the word class of the whole compound. To this group we refer adverbs (infolge, zugrunde) and verbs (spazieren gehen, kennen lernen).

Moreover, it is worth introducing the notion of cohesion which deals with a linking element between two constituents of a compound. In the German language we have the following linking elements:

- *-e: Hund**e**fell, Gerät**e**technik;*
- *-er: Fächerkatalog, Rechnernetz, Rechnerstruktur, Speichertechnik;*
- *-i: Nachtigall;*
- *-o: Elektrotechnik, Neuroinformatik*
- *(e)s: Engelsgeduld, Arbeitsprozess, Forschungsvorhaben, Geisteswissenschaft;*
- *(e)n: Stellenwert, Datenschutz, Methodenlücke, Stellenmarkt, Weizenbau*
- *ens: Herzenslust*

On the whole, the linking element of a German compound may help us differentiate the meaning of a word: *Landsmann (Bauer)* and *Landmann (Heimatgenosse).* Moreover, Oguy outlines that syntactic properties of compounds can be of "binäre Struktur (vorwiegend zwei Konstituenten) und strukturelle Zweideutigkeit (z.B. Kinderfilmwoche als [*Kinder*[filmwoche] – Filmwoche für Kinder und [*Kinderfilm*[woche] – Woche des Kinderfilms" (Oguy, 2003 : 201).

2. A quantitative analysis of German compounds in terms of word classes of their constituents

The task of our analysis presupposes classifying German compounds in terms of their parts of speech in the text of technical science. Furthermore, we aim to find the frequencies of each model for German compounds as well as compare the results with English and Ukrainian compounds in this type of texts. Intending to do it, we have analyzed 29 pages of the book "Wirtschaftsinformatik" (2015) which belongs to the sphere of technical science.

As far as the procedure of our study is concerned, we have analyzed each page of the above-mentioned book (20 pages) where the German compounds have been written on each page. As a result, our sample includes 24 models (types of compounds according to the word classes of their constituents). The results are as follows:

1) **Noun + Noun:** *Schriftsteller, Grenzverschiebung, Geltungsanspruch, Informationsgesellschaft, Antrittsvorlesung, Spielkarten, Konsumartikel, Teilaspekt, Deutschland, Weltkrieg, Forschungsansatz, Programmiersprache, Gehaltsskalen, Bundesrepublik, Frankreich, Fachkraft, Datenverarbeitung, Fachrichtung, Studiengang, Ausbildungsprogramm, Kulturhoheit, Finanznote, Studienfach, Eröffnungsrede, Wirtschaftswissenschaftler, Fachbeirat, Forschungsprogramm, Arbeitsgruppe, Fachbereich, Wechselweise, Wechselwirkung, Wirtschaftsinformatik, Ingenieurinformatik, Datentechnik, Beschreibungsverfahren, Ingenieurseite, Informatik-Werk, Steuertechnik, Regeltechnik, Kommunikationsaspekt, Ausbildungsgang, Akademie-Definition,*

Sprachraum, Titeländerung, Grundlagen, Informationstätigkeit, Fachredaktion, Informationswissenschaft, Dokumentationswissenschaft, Informationsnutzer, Sprachschöpfung, Informatikstudium, Forschungsinstitut, Bibliothekwissenschaftler, Fragestellung, Forschungsfrage, Bibliothekskatalogen, Lieferdienst, Forschungsbereich, Anwendungsbereich, Gründungsphase, Handbedingung, Stellenwert, Teilaspekt, Forschungsvorhaben, Inhaltsverzeichnis, Schaltungsentwurf, Grundbegriff, Zeichenverarbeitung, Informationsbegriff, Geisteswissenschaft, Wissenschaftsklassifikation, Formierungsansatz, Automatisierungstechnik, Systemtheorie, Kodierungstheorie, Spieltheorie, Prozessautomatisierung, Entfaltungsmöglichkeit, Kooperationsmöglichkeit, Anwendungslücke, Anpassungsdruck, Abschlussarbeit, Informatikausbildung, Forschungsprogramm, Neuroinformatik, Datenschutz, Abgrenzungsentscheidung, Korrektheitsproblem, Programmieraufwand, Anwendungsproblem, Medizininformatik, Verwaltungsinformatik, Informatikfrage, Betriebswirtschaft, Naturwissenschaft, Informationsbegriff, Mustererkennung, Informatikmethode, Theorieverständnis, Methodenlücke, Klassifikationsproblem, Studienführer, Forschungsführer, Strukturwissenschaft, Ingenieurwissenschaft, Technikwissenschaft, Gestaltungwissenschaft, Prüfungsarbeit, Theoriebildung, Systemprogrammierung, Gerätetechnik, Stellenmarkt, Verfahrenstechnik, Schutzwall, Fächerkatalog, Rahmenrichtlinien, Volkswirtschaft, Gründungsphase, Notlösung, Computerindustrie, Rationalisierungstechniken, Produktionsbereich, Autoindustrie, Unternehmensberater, Jahrestagung, Diskussionsbeitrag, Kopfarbeit, Lösungsansatz, Rechnernetz, Hundefell, Netzwerk, Geschmacksfrage, Kooperationspartnerin, Berufsentscheidung, Rechnerstruktur, Unterstützungssystem, Staubsauger, Handwerk, Kunstfertigkeit, Engelsgeduld, Weizenbau, Restbestand, Informatikfirme, Arbeitsprozess, Rechnerunterstützung, Erkenntnisziel, Komplexitätstheorie, Modellierungsmöglichkeit, Modellbildung, Zustandsdiagramm, Computergrafik, Speichertechnik, Vermittlungstechnik, Präsentationstechnik, Datenmodelle, Lehrbereich, Lehrplan, Lehrbuch.

2) **Noun + Adjective**: *deutschsprachig, mittelfristig, forschungspolitisch, erfolgreich, wissenschaftspolitisch, informatikrelevant, verwaltungsrechtlich, maschinennahe, kampfbereit, wertfrei*;
3) **Noun + Noun + Noun:** *Bundesforschungsminister, Informatikstudiengang, Informatikfachbereich, Zeitschriftartikel, Fachzeitschrift, Mensch-Maschine-Kommunikation, Intelligenzformalisierungstechnik, Datenbanktechniken*, *Informatiklehrbuch, Stichwortgebe*;
4) **Adjective + Noun**: *Hochschule, Neugründung, Sowjet Union, Neubestimmung, Digitalrechner, Bereitstellung*;
5) **Adjective + Verb**: *vollziehen, hochqualifizieren, wahrnehmen, gleichberechtigen, bereitstellen;*
6) **Adjective + Noun + Noun**: *Halbleitertechnik, künstliche-Intelligenz-Forschung, Langfristziel;*
7) **Preposition + Noun + Noun**: *Nebenfachregelung, Überlebensstrategie, Übersichtband*;
8) **Adjective + Adjective**: *gleichzeitig, geistig-philosophisch, formallogisch*;
9) **Noun + Verb (Partizip 2)**: *teilnehmen, Kopfzerbrechen, Gerätefixiert*;
10) **Pronoun + Adjective**: *allgemein, alltäglich;*
11) **Pronoun + Noun + Noun**: *Alltagsgegenstand, Allmachtphantasie*;
12) **Preposition + Noun**: *Binnenreguliereung, Nebenfach*;
13) **Preposition + Adjectiv:** *außeruniversitär, kontraproduktiv;*
14) **Noun + Preposition + Noun**: *Grenzüberschreitung*;

15) **Noun + Pronoun + Noun**: *Berufsalltag;*
16) **Conjunctive + Adjective**: *wenngleich*;
17) **Noun + Noun + Noun + Noun:** *Fachliteraturdatenbank*;
18) **Preposition + Pronoun + Verb (Participle 2)**: *auseinanderliegend;*
19) **Pronoun + Noun:** *Selbstverständnis;*
20) **Numeral + Noun + Noun:** *drittmittelvorhaben;*
21) **Adverb + Adjective:** *vielfältig;*
22) **Preposition + Pronoun + Noun:** *auseinandersetzung;*
23) **Pronoun + Präposition + Noun:** *Selbstüberschätzung;*
24) **Preposition + Noun + Adjective:** *überarbeitungsbedürftig*

The results are given in Table 1.

Table 1
Rank-frequency distribution of the frequencies for German types of compounds in terms of parts of speech

Rank	**Models**	**Absolute frequencies**	**Computed values**
1.	Noun + Noun	159	158.9
2.	Noun + Adjective	10	13.35
3.	Noun + Noun + Noun	10	3.78
4.	Adjective + Noun	6	1.96
5.	Adjective + Verb	5	1.42
6.	Adjective + Noun + Noun	3	1.21
7.	Pronoun + Noun + Noun	3	1.12
8.	Adjective + Adjective	3	1.07
9.	Noun + Verb (Participle 2)	3	1.04
10.	Pronoun + Adjective	2	1.03
11.	Pronoun + Noun + Noun	2	1.02
12.	Preposition + Noun	2	1.01
13.	Preposition + Adjective	2	1.01
14.	Noun + Preposition + Noun	1	1.00
15.	Noun + Preposition + Noun	1	1.00
16.	Conjunctive + Adjective	1	1.00
17.	Noun + Noun + Noun + Noun	1	1.00
18.	Preposition + Pronoun + Verb	1	1.00
19.	Pronoun + Noun	1	1.00
20.	Numeral + Noun + Noun	1	1.00
21.	Adverb + Adjective	1	1.00
22.	Preposition + Pronoun + Noun	1	1.00
23.	Pronoun + Noun + Adjective	1	1.00
24.	Preposition + Noun + Adjective	1	1.00
Total		221	$R^2 = 0.996$ a = 157.8874 b = -3.6758

A statistical study of German compounds has shown that there are 24 models in the texts of technical science. It is worth mentioning a quantitative study of English compounds in the scientific texts (Gnatchuk, 2016) where 24 types (models) for English compounds have been found (West-Germanic languages). The analysis of Ukrainian compounds has shown only 11 types of compounds in the texts of computer science. In this case, it would be relevant to take into account family groups of each language. Similar to the frequencies of English compounds in the scientific style, the German model Noun + Noun has turned out to be the most productive one in comparison with others.

Since we have to do with simple ranking we suppose that the usual Zipfian power function may sufficiently capture the data. Instead of distribution we use a usual function. But since there is a long tail filled with 1, we displace the function and use

(1) $$\frac{dy}{y-1} = \frac{b}{x}dx$$

yielding

(2) $y = 1 + ax^b$.

The results of fitting are presented in the last column of Table 1. As can be seen, the construction of compounds is regulated by the same law as many other rankings of classes.

Bibliography

Gnatchuk, H. (2016). A Quantitative Analysis of English Compounds in Scientific Texts. *Glottometrics 33,* 1-7.

Hansen, H.R., Mendling, J., Neumann, G. (2015). *Wirtschaftsinformatik*. De Gruyter Studium.

Oguy, O. D. (2003). *Lexikologie der Gegenwärtigen Deutschen Sprache*. Winnyts'a: „Nova Knyha".

Glottometrics 35, 2016, 6-15

Persian Text Ranking Using Lexical Richness Indicators[1]

Tayebeh Mosavi Miangah[2]
Mohammad Javad Rezai[3]

Abstract
The adequacy of some quantitative parameters mainly based on frequency of lexical items (types and tokens) are to be demonstrated in this study through an experiment. The main purpose of the present article is to rank some Persian texts according to various indicators of vocabulary richness proposed in the state of the art literature. It is the first attempt towards a quantitative study of lexical characteristics of Persian texts to show the possible relationship between specific formal features of texts and vocabulary richness. The results show that journalistic texts in which repetition of certain words is inevitable are less rich in terms of vocabulary than poetry and literary texts, and that the type-token ratio and lambda indicators could well be able to distinguish genres in Persian language.

Keywords: *Lexical richness, Persian language, text ranking, type-token ratio, word frequency*

1. Introduction

One of the oldest sub-fields in quantitative linguistics is vocabulary richness measurement. As a virtue of a language, the concept tracing back to the Roman philosopher Cicero, is based on the notion of functional relationship between vocabulary and text length used in the process of text generation. This parameter demonstrates not only the size of active vocabulary at the disposal of the writer or the speaker, but also the way this is used in actual language usage. In fact, it is usually used in linguistics to for genre analysis and author attribution, that is, each person uses idiosyncratic and specific lexicon.

The present study is the first attempt in Persian linguistics at a quantitative study of various Persian text genres to be ranked based on various indicators of vocabulary richness proposed in the literature. In the first section of this paper we shed some theoretical light on various text analytic metrics including type-token ratio, h-point and lambda structure. In the section 2, we review some related literature on the field. Section 3 goes through methodology based on an experiment performed for text ranking according to various measures of vocabulary richness. In section 4, we summarize the findings and conclusion comes at the last section as usual.

There are different approaches to vocabulary richness according to Wimmer and Altmann: a) Capturing the vocabulary richness by means of a measure, of an index which is the first step in most cases towards quantification, b) Capturing the unfolding of the vocabulary by a curve, e.g. Herdan, Tuldava, Köhler and Galle and several Russian scholars, c) Starting with the empirical distribution of words (types) occurring x-times (tokens) and deriving the theoretical distribution based upon combinatorial considerations and d) based

---

[1] This paper is dedicated to Prof. Dr. Reinhard Köhler in honor of his 65th birthday.

[2] Payame Noor University, Iran. mosavit@pnu.ac.ir
[3] Yazd University, Iran, mrezai@yazd.ac.ir

upon stochastic processes resulting in distributions, adopted by the guild of mathematicians like Brainerd, Gani, Haight, McNeil, Simon and others (Wimmer and Altmann, 1999:143).

Another criterion for measuring the lexical richness is the arc length indicator, which can be regarded as an elementary indicator of vocabulary richness. Popescu and his colleagues studied the behaviour of arc length computed from the ranked frequencies of some text units and tried to construct an indicator which is independent of text size N and which is useful for text characterization, text comparison and classification and even for language comparisons (Popescu, et al. 2013).

However, there are other measures for lexical richness to be discussed in this paper, namely type-token ratio, h-point indicator and lambda structure. Clearly, each of these approaches behaves differently depending on the type of language and text, length of the texts, the aim of the study and many other factors. As vocabulary richness is mainly used in stylometry, in this study we analyze genres in texts considering the most widely used measures of lexical richness applied to different Persian text types. The main purpose of the study is to discover whether such a parameter may be able to distinguish genres in Persian language.

1.1. Type-Token Ratio

Type-token ratio (TTR) is one of the oldest and easiest ways of vocabulary richness measurement. It is based on the simple ratio between the number of types and tokens in a text. The TTR value shows how much the vocabulary varies, that is, the more vocabulary variation in a text, the higher TTR (Kubát and Milička, 2013).

Token (text length) refers to the number of words in any form with any number of reoccurrence in a text; and the type (the number of different words) refers to any form of a word appearing in a string, regardless of the number of its frequencies in the text. Type-token ratio, as Wimmer puts it "is understood as the ratio of the number of different words to all words in text, or with other words, the ratio of vocabulary richness to text length. The problem developed probably in analogy to that of species frequency or species abundance in biology and has been imported in linguistics by statisticians who were active in both disciplines" (Wimmer, 2005: 361).

TTR is the basic measure of lexical richness which was first introduced by Chotlos (1944), and was formulated as follows:

$$TTR = \frac{V}{N}\ ,\ 0 < \text{TTR} < 1$$

in which V is vocabulary and N is the text length.

The information about the number of types and tokens (text length) of a text as well as the frequency of different words along with the rank distribution of the text vocabulary under study is to be taken into account as an index of lexical richness in every procedure measuring such a property of the texts. According to Panas, there are two interpretations of the TTR, namely (a) it is a characteristic of vocabulary richness of the text, and (b) it is a model of information flow in text (Panas, 2007).

From the quantitative point of view, the very basic property of a text is the frequency or the repeat rate of lexical units in the text. The repeat rate (which may be transformed in redundancy) in the data in text analysis has many applications in information systems. That is, the more often a lexical item appears in a text, the less amount of information the text under study will provide. In other words, based on information theory, more predictable linguistic units contain more redundancy (Weber 2005).

1.2. The *h*-point

Hirsch has originally introduced the concept of *h*-point, initially applied to scientometrics and bibliometrics (Hirsch, 2005), after which the concept was widely used in quantitative linguistics and particularly in word frequency studies (Popescu and Altmann 2006, Mačutek, Popescu and Altmann 2007). Studying the *h*-point is useful for descriptive linguistic purposes as well as cross linguistic studies. In a rank-frequency distribution of a list of some entries, the *h*-point is a fixed point at which the rank *r* and the frequency *f(r)* of the list are equal. It separates two different areas of this distribution, namely the synsemantic and the autosemantic branch. The formula for calculating the *h*-point proposed by Popescu and Altmann (2008:95) is given as follows:

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

In quantitative linguistics, *h*-point has many applications, the most important of which is language typology. In this field, the *h*-point may be interpreted as a sign of analytism, because in analytic languages the synthetic elements are replaced by synsemantics and hence the number of word forms is smaller. The *h*-point can also be used in determining the lexical richness of individual texts. In such an application, the area below the *h*-point may be considered the lexical richness measure of the given text (Zörnig et al. 2016). As Popescu points out, the applicability of the *h*-point as an indicator of text richness, however, depends on the text length, and the fact poses some limitations on exclusively applying such a measure for text richness studies without normalizing the measure by text length (Popescu et al. 2009).

1.3. Lambda structure

Lambda is the indicator of the structure of the language usage; it is not a direct indicator of the ability or talent of the authors. That is, to have a higher lambda value does not imply a better author. It is not a quality indicator, but may be expressed in relationship to some other qualitative characteristics of text

Moreover, lambda is something more than a vocabulary richness factor. It takes into account the extent of the vocabulary which is necessary for computing the arc length *L*, and at the same time, the difference between individual frequencies of the neighboring entries in the rank-frequency sequence. By computing lambda, one may study not only the rank-frequency structure of the text, but also the vocabulary richness of the text under study. Generally speaking, the greater the value of lambda, the greater is the vocabulary richness of the given text (Popescu, et al. 2011). The best formula of the lambda structure stabilizing the arc length and getting rid of text length is as follows:

$$\Lambda = \frac{L(\log_{10} N)}{N}$$

2. Literature review

Köhler and Gale investigated the dynamics of some features of a text, particularly a TTR measure. They believe that any dynamic property of a text is controlled by a linguistic law derivable from a theoretical model. A corresponding hypothesis for TTR follows from Altmann's proposal (1998:88), presenting an equation whose solution $T = L^a$ (in which T stands for types and L for tokens and $0 < a < 1$) may be used for prediction (Köhler and Gale, 1993).

In another study, the index of lexical richness has been investigated using the concept of elasticity of vocabulary with respect to text length. In this methodology, the existing indexes were derived as special cases. They believe that in order to explain the effect of the increase of text length on its vocabulary, it is correct to use the concept of elasticity; however, to explain the effect of the change in the text length on lexical richness, the elasticity of vocabulary with respect to text length as well as the *V-N* function (Vocabulary-Text length function) implied from the elasticity should be considered. So, by combining the analysis of *V-N* functions, they attempted to capture the idea of lexical richness (Panas, 2007).

Kubát and Milička proposed a new way of vocabulary richness measurement without any text length dependence. They applied their new method of vocabulary richness measure to genre and authorship analysis. Specifically, they used their new method for a genre analysis in texts written by the Czech writer Karel Čapek. They had two aims in their research: to propose a new way of vocabulary richness measure without any text size dependence, and to discover whether vocabulary richness is an advisable criterion for genre attribution (Kubát and Milička, 2013).

Popescu tried to use the *h*-index concept in ranking tasks. For this purpose he made use of three main classes of web text sources, namely the Bible, classical works, and Nobel lectures. The word distributions of these texts were produced using web available word frequency counters. Three main quantities describing the word distribution of the texts developed in detail were as follows: (a) text length or total word count representing the area under the (rank, frequency) word curve from the first rank up to the last rank, (b) *h*-index for words, indicating the "word distribution width", and (c) weight or percentage of the first *h* highly frequent words (*hfw*) out of the total word count. He believes that the *hfw* criterion seems as a consistent estimator of the ineffable style under which the text has been created. The results of sorting texts according to the above criteria showed that rankings by text length and by *h*-index are closely similar as expected. However, sorting the data by the third criterion (*hfw*), revealed top position of Bible texts, followed by classical texts and finally by Nobel lecture texts. He concluded that the *hfw* criterion is a consistent estimator of the ineffable style under which the texts are created. Based on his claim, a simple and objective measure can be used for evaluation of any type of text in a matter of seconds (Popescu, 2007).

Zornig and his colleagues studied lexical properties of different Serbian text types to reveal the characteristics and text parameters based on the frequency of word forms (relative frequencies, repeat rate, *h*-point and related indicators). They applied techniques of multivariate analysis (cluster analysis and multidimensional scaling — MDS) to classify the text types adequately and to illustrate the functioning of these techniques in detail by explicit calculations of all programming steps (Zörnig, et al. 2016).

3. The experiment

In order to evaluate the effectiveness of the three mentioned approaches, namely, TTR, *h*-point and lambda structure for lexical richness in text ranking, we collected some Persian texts from different genres and implemented the approaches on them.

3.1. Corpus used for this study

The corpus based on which our word frequency analysis and subsequent text ranking was performed consists of four different genres of Persian language, namely, poetry, religion, press or journalistic texts and literature.

- The poetry texts are extracted from the whole volume of Hafez Shirazi' sonnets collection (an Iranian poet of 1325/26–1389/90) including 8384 lines of 495 sonnets. The religious texts are extracted from the Persian version of the Holy Quran including 6345 sentences of 114 chapters. The Holy Quran is the main religious book of the Muslims all over the word.
- Journalistic texts are extracted from a selected part of Hamshahri Newspaper[2] of May 2015, including 13458 sentences.
- Literary texts are extracted from selected number of Persian short stories mainly taken from IranOnline[3], including 10860 sentences.

  The type-token information of the corpora used in this study has been demonstrated in Table 1.

Table 1
The type-token ratio of the used corpora

	Text type	Author/source	Type (v)	Token (n)	TTR
1	Poetry	Hafez	8383	60366	0.1388
2	Religious	Quran	10206	132066	0.0772
3	Journalistic	Hamshahri	19945	392058	0.0508
4	Literary	Short stories	20522	162452	0.1263

3.2. Methodology

For the purpose of this study, for each of the above mentioned selected texts from different genres the rank-frequency distribution of word forms was computed using a software developed by the author for the very purpose. Then, for each text the *h*-point, namely, the number for which $r = f(r)$ (i.e. rank = its frequency) was calculated. If there was no such number we took simply $h = r + 0.5$. The sum of relative frequencies or the cumulative relative frequency from 1 up to *h*-point, i.e. the distribution function up to *h* $F(r)$ for each text was computed separately. $F(r)$ shows the *h* coverage of the text. In this way, the Popescu indicator of vocabulary richness could be set up as follows:

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right)$$

[2] - http://www.hamshahrionline.ir/
[3] - http://www.iranonline.com/

where N is the text size (number of words in text) (cf. Tuzzi, Popescu, Altmann 2010:127). The lambda indicator for each text was computed in the following steps. In the first place, we computed the arc length between neighbouring frequencies expressing the frequency structuring of the text as follows:

$$L = \sum_{i=1}^{V-1} [(f_i - f_{i+1})^2 + 1]^{1/2},$$

where V is the size of the vocabulary (number of word-form types, the highest rank) and f_i are the individual frequencies. In the next step, using the arc length value gained, we computed for each text the lambda indicator as follows:

$$\Lambda = \frac{L(\log_{10} N)}{N}.$$

Using the above methodology, the h-point, the cumulative relative frequencies up to the h-point $F(h)$, the arc length, the Popescu-indicator of vocabulary richness R, as well as the lambda for each text is displayed in Table 2.

Table 2
The quantitative properties of the four Persian text genres

	Text type	**Author/source**	***H*-point**	***F*(*h*)**	***L***	***R***	***Λ***
1	Poetry	Hafez	83	0.4119	10720.2609	0.6451	0.8490
2	Religious	Quran	131.5	0.5736	15627.8181	0.5572	0.6059
3	Journalistic	Hamshahri	244	0.5374	33040.5528	0.6144	0.4713
4	Literary	Short stories	138	0.4361	25655.6180	0.6811	0.8229

4. Results

The statistics gained from the TTR part of the study indicated that the relationship between types and tokens in four text types under study varies from 0.1388 to 0.0508, belonging to poetry and press texts, respectively. Figure 1 depicts the TTR for these four text genres:

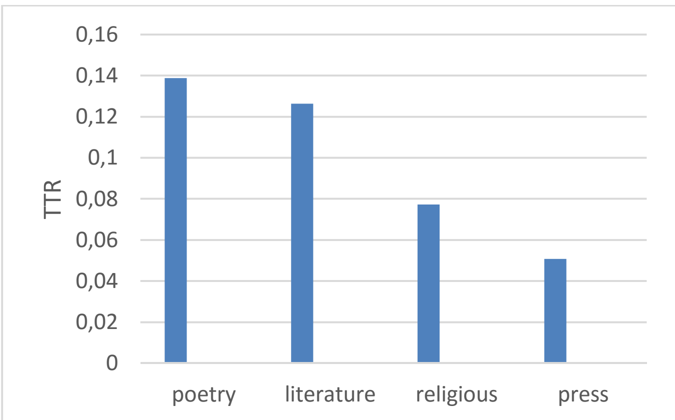

Fig. 1: TTR for four text genres

As the value of TTR varies between 0 and 1, a small value of TTR – that is, the one close to 0, indicates a lexicon decreasing its size with the growth of the text size, whereas, a large value of TTR, that is, the one closer to 1 indicates the growing size of distinct words as fast as the growth of the text size. In other words, in a text whose TTR is equal or very close to 1, almost all the words are unique. In contrast, in a text whose TTR is equal or close to 0, the most of the text words are repetitive or redundant. Type-token ratio for each of four texts have been calculated in 15 intervals to be able to scan the growth of TTR in different points. Figure 2 depicts the relationship between type and token, the corresponding numbers are found in Appendix 1.

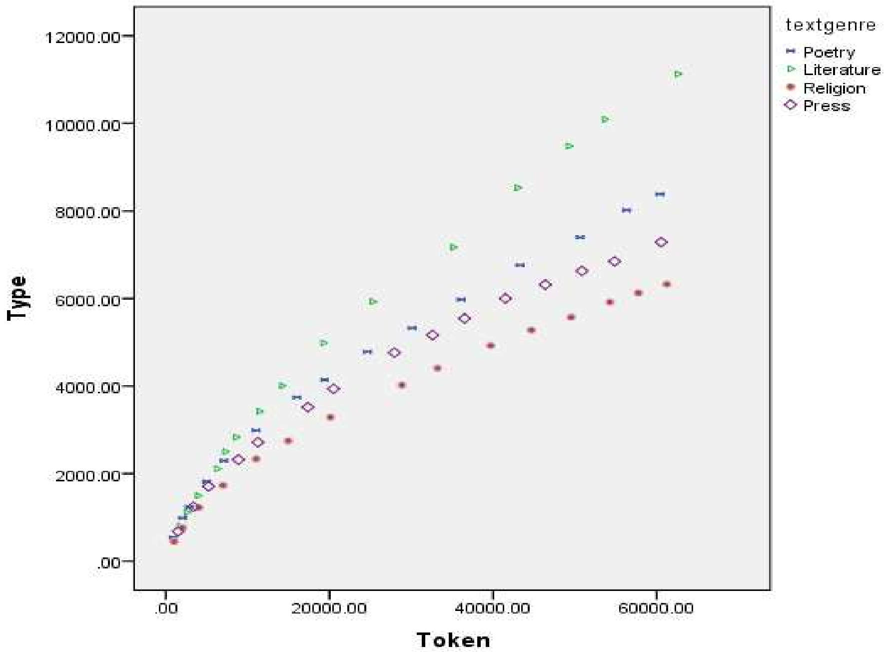

Fig 2. The relationship between type and token in 15 intervals of four text types

Based on the above figure, it is conjectured that the growth pattern of types of the poetry and literary texts are highly distinctive from the two other ones, namely, religious and journalistic texts. So, we can conclude that the lexical richness spectrum between these four text genres may be as follows:

Poetry > Literature > Religion > Press

These results are well compatible with the ranking the texts by lambda indicator in which the poetry texts stand at the top and journalistic texts at the bottom. However the ranking of the texts by these two indicators and the *h*-point indicator are very similar though not closely compatible, which is, somehow, due to difference in text lengths.

The vocabulary density is the ratio of text size and the vocabulary size *N/V* (*Buk and Rovenchak*, 2007). In this study the vocabulary density of the four text genres are calculated as shown in Table 3.

Table 3
The vocabulary density of the different text genres

	Text type	Author/source	VOC. DENSITY
1	Poetry	Hafez	7.2
2	Religious	Quran	12.94
3	Journalistic	Hamshahri	19.65
4	Literary	Short stories	7.91

As the table shows, the journalistic text has the highest vocabulary density comparing to other genres. This means that, in journalistic texts of Persian a new word appears in the text at every 19-20th word. In other words, journalistic texts are distinguished by a lower repeat rate as compared to other text genres.

5. Conclusion

The purpose of the present article is to rank some Persian texts according to various indicators of vocabulary richness proposed in the literature. It is the first attempt towards a quantitative study of lexical characteristics of Persian texts to show the possible relationship between specific formal features of texts and vocabulary richness. The results showed that journalistic texts in which repetition of certain words is inevitable are less rich in terms of vocabulary than poetry and literary texts, and that the type-token ratio and lambda indicators could well be able to distinguish genres in Persian language. It seems that the quantitative parameters like relative repeat rate, *h*-point, TTR, as well as lambda structure are appropriate measures for ranking and classification of text types of Persian language.

Further, the findings of this study was proved to be in line with the results gained from similar studies in other languages (e.g. Popescu, 2007; Panas, 2007). It is realistic to have more text genres to be investigated with more number of texts (written and oral) in each genre in order to prove the adequacy of the given quantitative parameters in different languages.

References

Altmann, G. (1988). *Wiederholungen in Texten. (Quantitative Linguistics 36).* Bochum: Studienverlag Brockmeyer.

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko’s novel Perekhresni stežky *(The Cross-Paths).* In: *Exact Methods in the Study of Language and Text.* Eds: Peter Grzybek & Reinhard Köhler. Berlin: Mouton de Gruyter, 39-48

Chotlos, J. W. (1944). Studies in Language Behaviour. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56, 75–111.

Devooght, J. (1957). Sur la loi de Zipf-Mandelbrot. *Academie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, 43, 244–251.

Herdan, G. (1964). *Quantitative Linguistics.* London: Butterworth.

Herdan, G. (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics.* s’Gravenhage: Mouton.

Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102 (46), 16569–16572.

Hirsch, Jorge E. (2005). An index to quantify an individual's scientific research output. In: *arXiv:physics*/0508025 v4 23 Aug 2005. http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf

Köhler, R., Gale, M. (1993). Dynamic Aspects of Text Characteristics. In L.Hřebíček, G. Altmann (eds.) *Quantitative Text Analysis: 46-53*. Trier: Wissenschaftlicher Verlag,

Kubát, M., Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics* 20(4), 339-349.

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

Panas, E.E. (2007). Indexes of lexical richness can be estimated consistently with knowledge of elasticities: some theoretical and empirical results. In: Peter Grzybek & Reinhard Köhler (eds): *Exact Methods in the Study of Language and Text: 523-534*. Berlin: Mouton de Gruyter.

Popescu, I.- I. (2007).Text ranking by the weight of highly frequent words. In: Peter Grzybek & Reinhard Köhler *(eds.): Exact Methods in the Study of Language and Text: 557-567*. Berlin: Mouton de Gruyter,.

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

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

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

Popescu, I.-I. et al. (2009). *Word Frequency Studies. (Quantitative Linguistics, 64).* Berlin, New York: Mouton De Gruyter.

Popescu, I.-I., Zörnig, P., Altmann, G. (2013). Arc length, vocabulary richness and text size. *Glottometrics*, 25, 43-53.

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

Tweedie, F.J., Baayen, R.H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. In: *Computers and the Humanities*, 32; 323–352.

Weber, S. (2005). Zusammenhänge [Interrelations]. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*: 214–226. Berlin, New York: de Gruyter.

Wimmer, G. (2005). *The type-token relation*. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook:* 361-368. Berlin, de Gruyter.

Wimmer, G., Altmann, G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics*, 6, 1–9.

Zornig, P., Kelih, E., Fucks, L. (2016). Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis. *Glottotheory*, 7(1). 41-66.

Appendix 1

Types and tokens belonging to four text types in 15 intervals

No.	Poetry		Literature		Religious		Press	
	type	*token*	*type*	*token*	*type*	*token*	*type*	*token*
1	536.00	869.00	803.00	1685.00	445.00	998.00	677.00	1489.00
2	982.00	2031.00	1112.00	2551.00	759.00	2003.00	1239.00	3353.00
3	1234.00	2853.00	1494.00	3909.00	1223.00	4037.00	1707.00	5214.00
4	1814.00	4974.00	2108.00	6242.00	1731.00	6986.00	2317.00	8857.00
5	2294.00	7093.00	2498.00	7228.00	2333.00	11034.00	2713.00	11250.00
6	2986.00	11021.00	2829.00	8556.00	2745.00	14930.00	3516.00	17363.00
7	3739.00	16017.00	3418.00	11402.00	3285.00	20112.00	3936.00	20488.00
8	4143.00	19399.00	4008.00	14188.00	4020.00	28845.00	4761.00	27938.00
9	4781.00	24632.00	4984.00	19214.00	4402.00	33216.00	5164.00	32589.00
10	5325.00	30079.00	5929.00	25266.00	4923.00	39702.00	5544.00	36505.00
11	5976.00	36073.00	7169.00	35114.00	5277.00	44672.00	6001.00	41499.00
12	6759.00	43277.00	8529.00	43005.00	5573.00	49530.00	6312.00	46380.00
13	7396.00	50637.00	9483.00	49219.00	5919.00	54292.00	6628.00	50852.00
14	8018.00	56327.00	10090.00	53627.00	6127.00	57759.00	6846.00	54860.00
15	8384.00	60367.00	11132.00	62547.00	6322.00	61234.00	7286.00	60545.00

Glottometrics 35, 2016, 16-21

Euphemisms in Political Speeches by USA Presidents

Lyubov Rimkeit-Vit[1]
Hanna Gnatchuk[2]

Abstract: The present study deals with the study of lexico-semantic groups (LSGs) of euphemisms in the political speeches of four USA presidents. The corpus of our study is represented by 16 political speeches. We have studied the public speeches by G. Bush Senior, B. Clinton, G. Bush Junior and B. Obama. The selected euphemisms have been classified into 14 lexico-semantic groups. The proposed classification has been captured by the power function

Keywords: *English, euphemism, US Presidents*

1. Introduction: General notes on euphemisms and their functions in the speech

Galperin (1981) considers that "euphemism is a word or phrase used to replace an unpleasant word or expression by a conventionally more acceptable one (Galperin, 1981 : 173)". A good example is illustrated in the newspaper *New Statesman and Nation* (June 15, 1957) mentioned in Book *Stylistics* (1981) by I. R. Galperin:

"The evolution over the years of a civilized mental health service has been marked by periodic changes in terminology. The mad-house became the lunatic asylum; the asylum made way for the *mental hospital* – even if the building remained the same. Idiots, imbeciles and the feeble-minded became low, medium and high grade mental defectives. All are now to be lumped together as patients of severely subnormal personality. The insane became persons of unsound mind, and are now to be mentally-ill patients. As each phrase develops the stigmata of popular prejudice, it is abandoned in favour of another, sometimes less precise than the old. Unimportant in themselves, these changes of name are the stigmapoints of progress." (June, 15, 1957):

Moreover, Galperin gives the example of a word "die" and its euphemisms *"to pass away, to expire, to be no more, to depart, to join the majority, to be gone, to kick the bucket, to give up the ghost, to go west"* (Galperin, 1981:173).

On the whole, the appearance of English euphemisms was traced in the XII – XV centuries and was indebted to the influence of the French culture. It is worth mentioning that Chaucer was the first person who used euphemisms in his "Canterbury Tales". Then the activity of purists gave way to increasing the euphemisms. In particular, they forbade using the word "God" in vain. In such a way. Shakespeares's expression "*Well, God give the spirit of persuasion"* was replaced by "*Well, maist thou have the spirit of persuasion*". In addition to it, theatres and dramas were forbidden because they were considered to be indecent. Thus, a new explanation of Shakespeare's works as well as censorship resulted in appearing a considerable number of euphemisms in the English language. The abundance of euphemisms flourished in the Victorian Age whereas wisdom and modesty played an important role for that period of time.

[1] Lyubov Rimkeit-Vit. BA. Alpen-Adria Universität, Institut für Anglistik und Amerikanistik, Universitätsstraße 65-67, Klagenfurt, Austria. Email: lyubavit@yahoo.com
[2] Hanna Gnatchuk, MA. Universität Trier, Computerlinguistik und Digital Humanities, Universitätsring 15, Trier, Germany. Email: agnatchuk@gmail.com

On the whole, euphemisms perform numerous functions in the speech. We shall take a look at *euphemistic, intentional, social and regulative, contacts-establishing, emotive and esthetic functions:*

1) The *euphemistic function* presupposes substituting a certain word for another one due to some fears, old traditions and conventions. We shall mention the example of the word "die" which is predominantly replaced by the politicians by "*to pass away, to expire, to be no more, to depart, to join the majority, to be gone, to kick the bucket, to give up the ghost, to go west"* (Galperin, 1981 : 173);
2) *Intentional function* is supposed to make a certain impression in the communicative situation. This function is of great importance to the speakers who tried to veil some unpleasant or negative facts. It is noteworthy mentioning that this function can manipulate a speaker in so far as he/she can present the information in better lights. For instance, T. Blair substituted a word "war" for *military action, military operation, armed intervention, conflict.*
3) *Social and regulative functions*. There can be certain situations where the nomination of certain things or phenomena are not acceptable. Here one must consider politeness. Focusing on a political speech, this politeness is called in the modern world "political correctness". In such a way, we can deal here with a considerable number of euphemisms. Let us consider the lexemes meaning age discrimination: *middlecence = the period from 40 to 65; the third age = the period from 65.* According to Peccei (1999), the direct nomination of a word "old people" can be replaced by *retired person* or *senior citizen* because the connotation of these euphemisms is associated with the semantics of "active, progressive, happy, strong" (Peccei, 1999: 107). Thus the euphemisms are used in these situations with the purpose of making an image of a perfect society regardless of age, financial backgrounds or gender;
4) *Contacts-establishing function*. This function helps a speaker establish contacts with the audience. In this case, a speaker must replace direct words for euphemisms in order to maintain a successful communication and avoid negative outcomes. For instance, it is better to use a euphemism "young offender" than "young criminal". In this situation the euphemisms can have a positive influence on the further youth life;
5) *Emotive function* is characteristic of the euphemisms with metaphorical or metonymical transferences. These euphemisms are concerned with an internal mood of a speaker. One can refer to this function the euphemisms of pity or sympathy to the children without parents. For example, we can apply the euphemisms "*to go through a prison gate"* to the people who have committed a crime;
6) *Esthetic function.* The usage of euphemisms can provide the speech with more delicate and elegant nominations of certain things or phenomena that evoke unpleasant emotions or feelings. Let us consider the situation told by a senator Tom Harkin. He said that one day Trumann's wife walked together with one lady around the White House. Having met the president, the lady gave her compliment to the beauty of the flowers. In answer to it, Trumann used a lexeme "manure". Although it would have been more appropriate to use a neutral word "fertilizer".

2. The analysis of euphemisms according their lexico-semantic groups in the political speeches of 4 USA presidents

Ginzburg (1979) is of the opinion that "the classification of vocabulary items into lexico-semantic groups is the study of hyponymic relations between words. By hyponymy is meant a semantic relationship of inclusion. Thus, e.g., **vehicle** includes **car, bus, taxi** and so on" (Ginzburg, 1979 : 53). In such a way, it is possible to deal with the lexico-semantic groups (LSGs) of **vehicles** (car, train, bus, trolley-bus, taxi, tram), **emotions** (happy, gay, satisfied, cheerful), **movements** (walk, hop, run, saunter).

The aim of our study consists of two steps: a) we intend to write out the euphemisms from 16 political speeches of four American Presidents; b) we aim to classify them into certain lexico-semantic groups as well as count their frequencies.

The material of our research is represented by 16 political speeches by 4 American presidents - *George Bush (Senior), Bill Clinton, George Bush (Junior) and Barack Obama*. The majority of the speeches have been taken from American President Speech Archive – Miller Center (http://millercenter.org/president/speeches):

George W. H. Bush (Senior): (political speeches dating back to 1992-1993)

1) *George H. W. Bush: Address at West Point. January 5, 1993;*
2) *George H. W. Bush: Remarks at Texas A & M University. December 15, 1992;*
3) *George H. W. Bush: Republican National Convention. August 20, 1992;*
4) *George H. W. Bush: State of the Union Address. January 28, 1992*

Bill Clinton: (political speeches dating back to 1993 – 2000)

1) *Bill Clinton. State of the Union Address. January 27, 1998;*
2) *Bill Clinton: First Inaugural Address. Washington D. C., January 21, 1993;*
3) *Bill Clinton: Farewell to DNC. Los Angeles, CA, August 14, 2000.*
4) *Bill Clinton: I Misled people. August 17, 1998.*

George W. Bush (Junior): (political speeches dating back to 2001-2002)

1) *George W. Bush: "A New Approach". Knoxville, Tennessee, June, 2000;*
2) *George W. Bush: "2000 RNC Presidential Acceptance". Philadelphia PA, August 3, 2000;*
3) *George W. Bush's 1st Inauguration Address. January 20, 2001;*
4) *George W. Bush: "Address to the Joint Session of the 107th Congress", February 27, 2001.*

Barack Obama: (political speeches dating back to 2004 – 2010)

1) *Barack Obama: Responsibly Ending the War in Iraq. 27 February, 2009;*
2) *Barack Obama: Official Announcement of Candidacy for US President. 10. February, 2007;*
3) *Barack Obama: Democratic National Convention Keynote Address. July 27, 2004;*
4) *Barack Obama: Oval Office Address to the Nation on BP Oil Spill Disaster. June 15, 2010.*

The procedure of the research and the discussion of the results. First of all, we have written out the euphemisms and classified them all into 14 lexico-semantic groups (LSGs). The results are as follows:

1) LSG of national or racial belonging (available in the speeches by Bill Clinton, G. Bush (Senior), Barack Obama):

Black-Africans = black-skinned people;
White South Africans = people from South Africa who are of European decent;
Celebrating our *diversity* = celebrating the availability of *different nations*
African-American families = black-skinned Americans
We're becoming more and more *diverse* = *different nations* inhabit America
Black-America = all black-skinned people
A white American = an American with a white skin
a black youth = a youth with a black skin
It does not matter whether you're ***black*** *or* ***white*** *or Hispanic or Asian or Native America"*
The most diverse nation on Earth = the people from all nations

2) LSG of abstract (or metaphorical) notions or actions (Barack Obama and Bill Clinton)

To *sit on opposite sides of the aisle = to belong to different parties, share opposite points of view;*
To struggle with bills = to have difficulties with paying the bills or hardly paying for everything;
To open wide the doors = to make something accessible;
The *lion's share of the credit = a considerable part of the credit*
The mystery of American renewal = the ceremony of inauguration
To say yes to democracy = to agree to join the NATO
A far-off storm = a problem
Temporarily left behind by the global marketplace – the Americans that have nothing to do with trade.
To lay a foundation = to start, to initiate;
The bedrock of this nation = hope, belief
To win the next *battle = to win the election*
The tyranny of oil = the dependence on oil
Now is the moment for this embark on a national mission to *unleash America's innovation = to refuse from fossil fuels*
Industry's watchdog = to control the consumption of oil industry;
I am in this race = I keep the policy of this kind

3) LSG of countries or states (G. Bush (Senior), Bill Clinton, G. Bush (Junior) Barack Obama)

The second world - previous socialist countries (the countries of the former Soviet Union);
The third world = the countries of Africa, Latin America, Oceania and Asia;
The Arab World = Arabic-speaking population and countries;
To form that "*more perfect union" = The United States of America*
The world's great power = the USA
A servant of freedom = American nation
A prosperous nation = the USA
Crossroad of a nation = Illinois
Magical place = USA
Land of Lincoln = Illinois

4) LSG of economic or social states (George Bush Senior, George Bush Junior)

Low-income people, family, housing = poor people, family, housing
Middle class = the social class of people being between working class and upper class

5) LSG of death (Bill Clinton, G. Bush (Senior), Barack Obama)

The people who have discharged their duty with honor and professionalism = the people who died in the interest of the USA;
To lose two patriots = to die;
Late father = dead father;
It lives in memories of your fellows, soldiers, sailors, airmen and Marines who *gave their lives = to die;*
They've both *passed away* now = *to die*

6) LSG of illnesses or physical disabilities (George Bush (Senior), George Bush (Junior) and Barack Obama)

The Americans with Disabilities = ill American people
The Disadvantaged = ill people
It does not matter whether you're black or white or Hispanic or Asian or Native American or young or old or rich or poor, abled, *disabled;*

7) LSG of (economic or political) periods of time (George Bush (Senior), Bill Clinton, George Bush (Junior) and Barack Obama)

The cold war = a rivalry between the USA and the previous Soviet Union in military, economic and political spheres;
The Great Depression = economic period of time
Passage = a new millennium;
The spring = the period of Clinton's presidency and policy
Service = the period of presidency of Bill Clinton
The most extraordinary *chapters* of service in the history of our nation = *periods of time*

8) LSG of people (Bill Clinton)

America's First Lady = Hilary Clinton;
She's been *a great first lady = Hilary Clinton*
We American have offered our most *precious resource* = *women and men* who will cooperate with the Iraqis

9) LSG of Age (George Bush Junior)

Elderly Americans = old American people
The seniors = the old;

10) LSG of military notions (George Bush Senior)

The women and men who proudly wear the uniforms of the USA = the people who serve in the army
A strong fighting force = army

11) LSG of political parties (Barack Obama)

Red states = the Republicans
Blue states = the Democrats

12) LSG of war (Barack Obama)

A tragic mistake = a war in Iraq

13) LSG of substance (George Bush Senior)

Nuclear nightmares = nuclear fuels

14) <u>LSG of close relationship</u> (Bill Clinton)

To have a relationship = to have an intimate relationship

The frequencies of lexico-semantic groups (LSGs) of euphemisms in 16 political speeches by four USA presidents illustrated in Table 1. As can be seen, even this very abstract level abides by a well known law, namely that of Zipf expressed by the power function $y = ax^b$. The computation is displayed in the last column of Table 1.

Table 1
LSGs of euphemisms in political speeches of 4 USA presidents

N	Lexico-semantic groups (LSGs)	Absolute frequencies	Power function
1	LSG of metaphorical notions	15	16.15
2	LSG of national or racial belonging	10	9.22
3	LSG of countries or states	10	6.64
4	LSG of (economic) periods of time	6	5.26
5	LSG of death	5	4.39
6	LSG of illnesses or physical disabilities	3	3.79
7	LSG of people	3	3.35
8	LSG of economic or social states	2	3.00
9	LSG of Age	2	2.73
10	LSG of military notions	2	2.51
11	LSG of political parties	2	2.32
12	LSG of war	1	2.16
13	LSG of substance	1	2.03
14	LSG of close relationships	1	1.91
		a = 16.1485, b = -0.8089,, R^2 = 0.9164	

Thus, we have found 63 euphemisms in 16 political speeches of 4 USA presidents which have been classified into 14 lexico-semantic groups: LSGs of *metaphorical notions*, *national or racial belonging, countries or states, periods of time, death, illnesses or physical disabilities, people, economic or social states, age, military notions, political parties, war*, *substance* and *close relationship*. Table 1 shows that the semantic groups of euphemisms denoting *metaphorical notions, national or rational belonging* as well as *countries or states* have the highest frequencies. The lowest frequencies are observed in LSGs of *economic or social states, age, military notions, political parties, war, substance and close relationships*.

The result testifies to the fact that Zipf's law holds true even in this abstract stylistic domain.

References

American President Speech Archive – Miller Center http://millercenter.org/president/speeches

Galperin I. R. (1981). *Stylistics.* The Third Edition. Moscow: Vysšaja škola.

Ginzburg R. S., Khidekel, S. S., Knyazeva, G. Y, Sankin, A. A. (1979). *A course in modern English Lexicology*. Moscow :Vysšaja škola.

Peccei, J. S. (1999). *Language and age. Language, society and power: An introduction: 99-116.* Ed. by L. Thomas, Sh. Wareing. L: N. Y.: Routledge.

Glottometrics 35, 2016, 22-27

The Impact of Code-switching on the Menzerath-Altmann Law

Lin Wang[1], *Radek Čech*[2]

Abstract. Based on the Chinese-English code-switching corpus and Modern Chinese corpus, the impact of code-switching on the Menzerath-Altmann law is observed. Specifically, the relationship between the sentence length and the clause length is analysed. Both code-switching and monolingual sentences abide by the Menzerath-Altmann law, however, differences are found in values of the determination coefficient R^2 and parameter b of the function expressing the law. As for the determination coefficient R^2, code-switching sentences evince worse fit of the model to the data then than the monolingual ones. Further, the lower value of b in the case of code-switching sentences expresses lower diversification (and a higher entropy) of the system.

Keywords: *Chinese-English code-switching, the Menzerath-Altmann law, sentence length, clause length*

1. Introduction

Code-switching is one of the important language contact phenomena. It refers to "language use that consists of material from two or more language varieties at any level from the discourse to the clause" (Jake and Myers-Scotton, 2009, p. 207). According to Myers-Scotton (2006), there are two types of code-switching: Classic Code-switching and Composite Code-switching. "Classic code-switching includes elements from two (or more) language varieties in the same clause, but only one of these varieties is the source of the morphosyntactic frame for the clause" (Myers-Scotton, 2006, p. 241). Composite code-switching is "bilingual speech in which even though most of the morphosyntactic structure comes from one of the participating languages, the other language contributes some of the abstract structure underlying surface forms in the clause" (Myers-Scotton, 2006, p. 242). The language or language variety that builds the morphosyntactic frame is the Matrix Language and the participating language is called the Embedded Language. The Morpheme Order Principle (Myers-Scotton, 2006) is applied to determine the Matrix Language: In mixed constituents which consist of at least one Embedded Language word and any number of Matrix Language ones, the surface word (and morpheme) order is the order of the Matrix Language.

Example (1) is a typical example of Classical Code-switching, in which Chinese is the Matrix language, and English is the Embedded Language and example (2) is a Composite Code-switching (Wang and Liu, 2013).

---

[1] School of Foreign Languages, Zhejiang Gongshang University, China.
[2] University of Ostrava, Czech Republic. Email 1 & 2: cechradek@gmail.com

(1) Xiao haizi dou xihuan cats and dogs
[Little kids all like cats and dogs]
Little kids all like cats and dogs

(2) Ta you contact wo.
[He have contact me]
He has contacted me.

In the clause (2), English contributes some of the abstract structure, that is, *you contact* is the Chinese word-for-word translation of English phrase *have contacted*. In Chinese, the aspectual marker *le* at the end of the sentence expresses the perfect tense instead of *you* preceding the verb, see the sentence (3).

(3) Ta lianxi wo le.
[He contact me ACCOMPLISHMENT]

From the theoretical point of view, there is an important question concerning the impact of code-switching on general language properties. One can find two different views on the phenomena; some researchers are persuaded that there are specific constraints for code-switching (Sankoff and Poplack, 1981; Woolford, 1983; Di Scilullo et al.,1986), while others deny this idea and consider the code-switching to be governed by the same constraints or principles underlying universal grammar (Mahootian, 1993; MacSwan, 1999, 2000; Chan, 2003).

The aim of the article is to analyze a relationship between code-switching and one of the very general language property which is expressed by the Menzerath-Altmann law (Cramer 2005; for syntax see Köhler 2012, chapter 4.1.3). According to the law, there is a systematic relationship between the length of language constructs (e.g., sentences) and their immediate constituents (e.g., clauses):

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

where y is the mean size of the immediate constituents, x is the size of the construct, and a, b, are parameters which seem to depend on the level of the units under study.

In this study, we focused on a relation between sentence length and clause length in two sorts of data: Chinese-English code-switching sentences and monolingual Chinese sentences. We assume that code-switching should affect the Menzerath-Altmann law because two different language systems are mixed in a language performance. Therefore, firstly we test the validity of the Menzerath-Altmann law in both samples. It is expected that the monolingual Chinese abides by the Menzerath-Altmann law. However, for Chinese-English code-switching sample there is no sure result – the potential impact must be tested empirically.

Four possible outcomes for Chinese-English code-switching data are:

a) They do not abide by the Menzerath-Altmann law;

b) They abide by the Menzerath-Altmann law, however, a fit between the data and the model is worse than for monolingual data;

c) They abide by the Menzerath-Altmann law with the same fit between the data and the model as for monolingual data, however, the parameters of the model differ;

d) There is no difference between code-switching and a monolingual sample, conesquently, code-switching has no impact on the Menzerath-Altmann law.

The article is organized as follows: in Section 2 the character of language material is described; methodology of the research is presented in Section 3; in Section 4 the results of the analysis are presented and interpreted; general findings and future study are discussed in Conclusion (Section 5).

2. Language material

Two sorts of data are used in our study: 100 Chinese-English code-switching sentences and 100 monolingual Chinese sentences. Chinese-English code-switching sentences were randomly selected from our self-built Chinese-English code-switching corpus. The corpus consists of 19,766 tokens, i.e. 16,267 (82.3%) Chinese tokens, 3,499 (17.7%) English tokens. Code-mixed data in the corpus are collected by audio-recording Chinese-English mixed discourses on mainland China and Hong Kong broadcasting or TV programs from June to September 2011. The mixed sentences in this paper are from the resource of the transcribed spoken language of TV programs. About 20% of the data come from interview programs and 80% from the entertainment news or social news. 100 mixed sentences include 2,385 tokens in total, including 2,119 Chinese tokens and 266 English ones. In this corpus most cases of code-switching are Classic code-switching, a few cases (only 11 of 773) are Composite code-switching; about 89% of code-switching are those with Chinese as the Matrix Language, and only about 11% with English.

Monolingual Chinese sentences are selected from the spoken part of Modern Chinese Corpus online (www.cncorpus.org). The Modern Chinese Corpus consists of 50 million tokens. The total 9,487 texts of the corpus are generally about the social sciences and the natural sciences, such as politics, economy, culture, law or psychology. The spoken Chinese sub-corpus is composed of 3 million tokens. Our data are chosen from the spoken texts after 1980s. In our sample, 100 monolingual Chinese sentences consist of 2,249 tokens in total.

3. Methodology

According to Hudson (2010), sentence can be defined as any string of words which are held together by syntactic relationships and which are not related to words outside that string. In other words, sentence boundaries are points where there are no syntactic relationships. In accordance to this approach, sentences are determined by referring to the transcription of spoken language in TV programs. Sentence length is calculated by the number of clauses in the sentence.

Lyon’s (1968) definition of clause is applied in this paper: a clause is a word sequence with subject and predicate and with pairs of words connected by syntactic relationships. Thus, the sentence can consist of one or more clauses. According to this definition, clauses can be segmented clearly in our sample. Clause length is calculated by the number of words in the clause.

We perform a manual count of 100 Chinese-English code-switching sentences and 100 monolingual Chinese sentences.

4. Results

The results, presented in Table 1 and 2 and Figure 1 and 2, show that both datasets follow the general tendency of the Menzerath-Altmann law, i.e. the longer the sentence the shorter the clause (in average). However, different behaviours of the two datasets are obvious at a first sight. Specifically, code-switching sentences (Table 1 and Figure 1) evince a deviation of the law in the area of the 2-clause and 4-clause sentences. The mean length of clauses in sentences with 2 clauses is shorter than the mean length of clauses in sentences with 3, and the mean length of clauses in sentences with 4 clauses is longer than the mean length of clauses in sentences with 3, which is in a contradiction to the prediction of the Menzerath-Altmann law. Further, the determination coefficient R^2, expressing the degree of correspondence between the empirical data and the model, is lower in the case of code-switching sentences than monolingual ones, which also indicates the deviation of the law. Finally, a comparison of parameter *b* of the function indicates different course of particular curves. The lower value of *b* in the case of code-switching sentences expresses lower diversification (and a higher entropy) of the system.

Table 1
Chinese-English code-switching *sentence l*engths and mean clause lengths

Sentence length (in clauses)	**Mean clause length (in words)**
1	9.69
2	6.38
3	6.71
4	6.9
5	6.06
6-7	5.64
a = 9.08, b = 0.25, R^2 = 0.76	

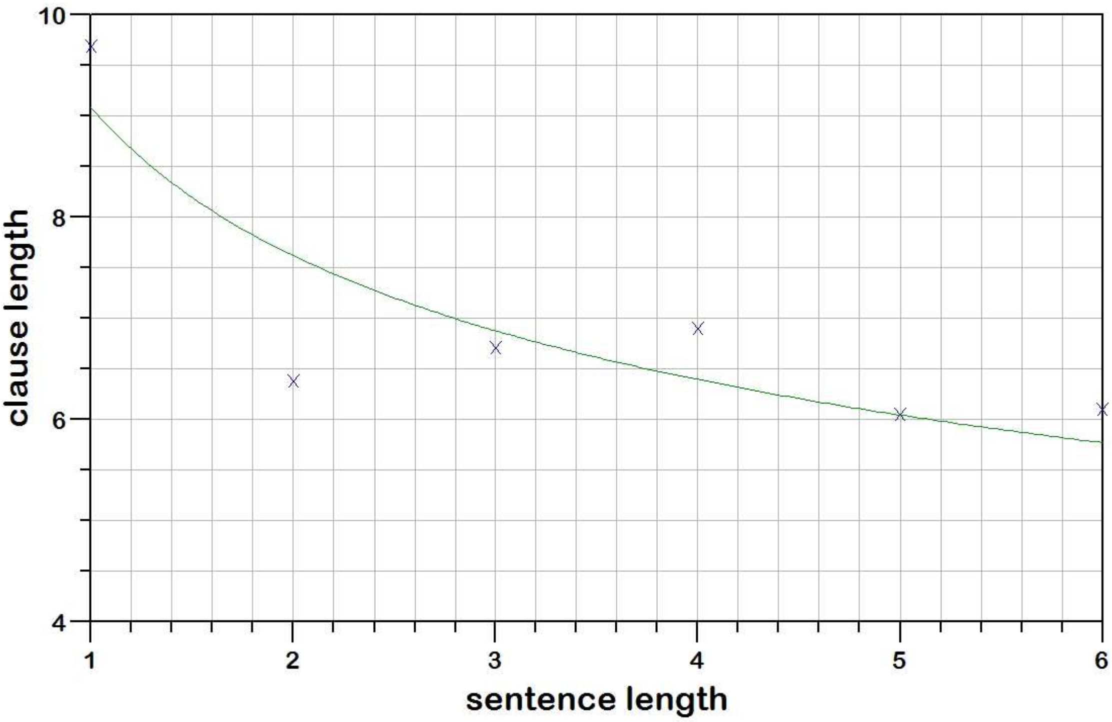

Figure 1. Fitting the Menzerath-Altmann law on the Chinese-English code-switching sentence level.

Table 2
Monolingual Chinese sentence lengths and mean clause lengths

Sentence length (in clauses)	Mean clause length (in words)
1	10.4
2	9.2
3	7.07
4	6.3
5	6.54
6-8	6.48
a = 10.53, b = 0.31, R^2 = 0.92	

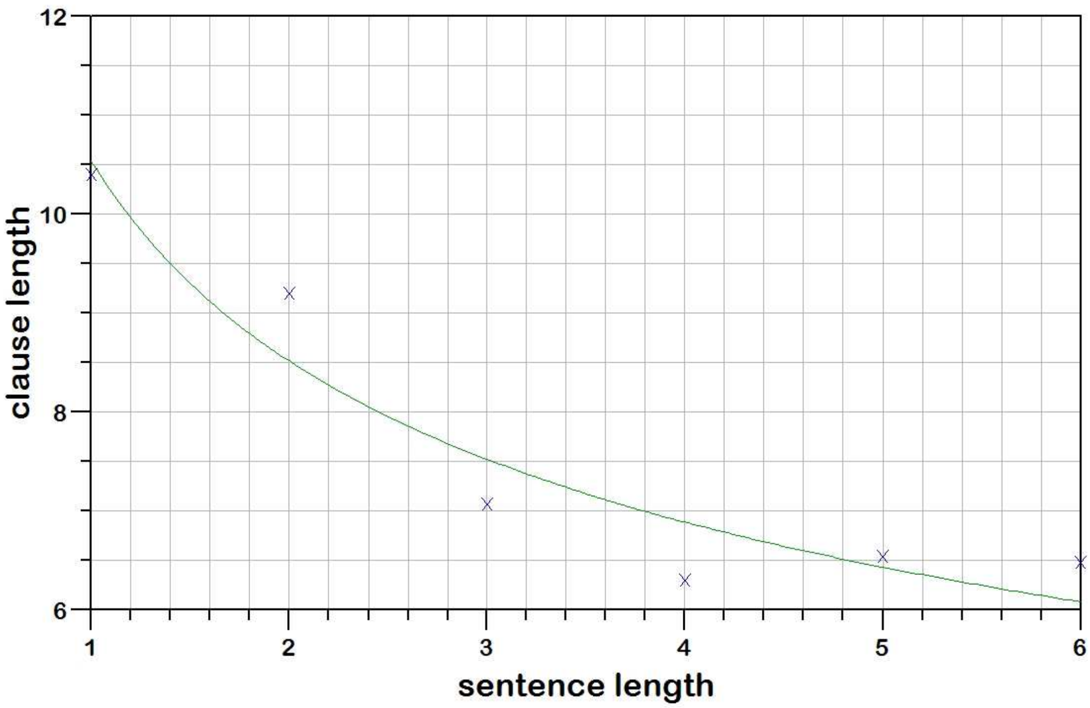

Figure 2. Fitting the Menzerath-Altmann law on the monolingual Chinese sentence level.

5. Conclusion

This study analyzes the impact of code-switching on the Menzerath-Altmann law. From the Chinese-English code-switching corpus and Modern Chinese corpus, 100 mixed sentences and 100 monolingual Chinese sentences have been randomly selected to test the Menzerath-Altmann law on the sentence level.

It is found that with regard to both Chinese-English code-switching and monolingual Chinese, the relation between the sentence length and the clause length generally abide by the Menzerath-Altmann law. However, a fit between the data and the model is worse in the case of code-switching sentences than for monolingual data. Moreover, the lower value of *b* indicates different behavior of code-switching sentences with regard to the Menzerath-Altmann law. Evidently, there is a boundary condition which could be captured by a further parameter or a slightly different function but it must be made for several languages at once.

It must be emphasized that this paper represents just a first step to study of the relationship between code-switching and the Menzerath-Altmann law. We are aware that only more detailed study can reveal the real impact of the phenomena under the study on the law; for instance, a proportion of words of Embedded Language in the sentence and their syntactic functions of should be taken in the account.

Acknowledgment

This work is supported by Zhejiang Gongshang University (Grant No. 1070KU115017).

References

Chan, B.H.-S. (2003). *Aspects of the Syntax, Pragmatics, and Production of Code-switching. Cantonese and English*. New York: PeterLang.

Cramer, I. M. (2005). Das Menzeratsche Gesetz. In: Köhler, R., Altmann, G. and Piotrowski, R. G. (eds), *Quantitative Linguistics. An International Handbook: 659–687*. Berlin; New York: de Gruyter.

Di Sciullo, A. M., Muysken P., Singh R. (1986). Government and code-mixing. *Journal of Linguistics 22, 1–24.*

Hudson, R. (2010). *Encyclopaedia of Word Grammar and English grammar*. Available at http://dickhudson.com/word-grammar/#books (15-06-2016).

Jake, J., Myers-Scotton, C. (2009). Which language? Participation potentials across lexical categories in code-switching. In: Isurin, L., Winford, D., Bot, K. (Eds.), *Multidisciplinary Approaches to Code-switching. 207–242*. Amsterdam/ Philadelphia: John Benjamins Publishing Company..

Köhler R. (2012) *Quantitative syntax analysis*. Berlin/Boston: Walter de Gruyter.

Lyons, J. (1968) *Introduction to Theoretical Linguistics*. Cambridge, New York, Melbourne: Cambridge University Press.

MacSwan, J. (1999). *A Minimalist Approach to Intra-sentential Code-switching*. New York: Routledge.

MacSwan, J. (2000). The architecture of the bilingual language faculty: evidence from intrasentential code switching. *Bilingualism: Language & Cognition 3, 37–54.*

Mahootian, S. (1993). *A Null Theory of Code-switching*. Doctoral Dissertation. Northwestern University, Evanston.

Myers-Scotton, C. (2006). *Multiple Voices: An Introduction to Bilingualism*. Victoria: Blackwell.

Sankoff, D., Poplack, S. (1981). A formal grammar for code-switching. *Papers in Linguistics*, 14, 3–46.

Woolford, E. (1983). Bilingual code-switching and syntactic theory. *Linguistic Inquiry 14, 520–536.*

Wang, L., Liu, H. (2013). Syntactic variations in Chinese-English code-switching. *Lingua 123, 58–73.*

Glottometrics 35, 2016, 28-37

The Meaning-Frequency Law in Zipfian Optimization Models of Communication

Ramon Ferrer-i-Cancho[1]

Abstract. According to Zipf's meaning-frequency law, words that are more frequent tend to have more meanings. Here it is shown that a linear dependency between the frequency of a form and its number of meanings is found in a family of models of Zipf's law for word frequencies. This is evidence for a weak version of the meaning-frequency law. Interestingly, that weak law (a) is not an inevitable of property of the assumptions of the family and (b) is found at least in the narrow regime where those models exhibit Zipf's law for word frequencies.

KEYWORDS: meaning-frequency relationship; Zipf's law; optimization of communication; linguistic universals

1. Introduction

The relationship between the frequency of a word and its number of meanings follows Zipf's law of meaning distribution: words that are more frequent tend to have more meanings (Zipf 1945; Baayen & Moscoso del Prado Martín 2005; Ilgen & Karaoglan 2007; Crossley et al. 2010; Hernández-Fernández et al. 2016). In his pioneering research, Zipf defined two laws where μ, the number of meaning of a word, is the response variable (Zipf 1945; Zipf 1949). One law where the predictor variable is i, the rank of a word (the most frequent word has rank 1, the 2nd most frequent word has rank 2 and so on), i.e.

$$\mu \propto i^{-\gamma}, \tag{1}$$

where $\gamma \approx 1/2$. Another law where the predictor variable is f, the frequency of a word,

$$\mu \propto f^{\delta}, \tag{2}$$

where δ is a constant satisfying $\delta \approx 1/2$. Zipf (1949) referred to Eq. 1 as the law of meaning distribution in his most famous book while he referred to Eq. 2 as the meaning-frequency relationship in a less popular article (Zipf, 1945). Eq. 1 and Eq. 2 describe, through different predictor variables, the qualitative tendency of the number of meanings of a word to increase as its frequency increases (assuming $\gamma > 0$ and $\delta > 0$).

Zipf (1945) derived Eq. 2 from Eq. 1 and Zipf's law for word frequencies (with rank as the predictor variable), i.e.

[1] Complexity and Quantitative Linguistics Lab. LARCA Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC). Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3. 08034 Barcelona, Catalonia (Spain). Phone: +34 934134028. E-mail: rferrericancho@cs.upc.edu .

$$f \propto i^{-\alpha}, \qquad (3)$$

where $\alpha \approx 1$ (Zipf 1945; Zipf 1949). Zipf's derivation of Eq. 2 is revisited in the Appendix. After Zipf's untimely decease, some researchers investigated Eq. 2 and provided support for it independently from Zipf's law for word frequencies. i.e. Eq. 2 (Baayen & Moscoso del Prado Martín 2005; Crossley et al. 2010; Hernández-Fernández et al. 2016) while others provided further empirical support for Eq. 1 (Ilgen & Karaoglan 2007).

μ is a measure of the polysemy of a word. If the mapping of words into meanings is regarded as a bipartite graph joining word vertices with meaning vertices (Ferrer-i-Cancho et al. 2005), μ defines the (semantic) degree of a word in that network.

The target of the preset article is the meaning-frequency law. Eq. 2 with $\delta \approx 1/2$ defines a strong version of the meaning-frequency law. A weak version of the meaning-frequency law can be defined simply as a positive correlation between μ and f. Notice that we are not assuming a Pearson correlation here, which is a measure of linear association (Conover 1999). Instead, we have in mind an association measure that can capture non-linear dependencies, e.g., Spearman rank correlation (Conover 1999; Zhou et al. 2003). The definition and use of a weak version of the law is justified for various reasons. First, looks of a pure power law of the form of Eq. 2 can be deceiving. This is a lesson of research on Menzerath-Altmann law in genomes (Ferrer-i-Cancho et al. 2013a), Heaps' law in texts (Font-Clos & Corral 2015) or the degree distribution of protein interaction and metabolic networks (Stumpf & Ingram 2005; Stumpf et al. 2005). Admittedly, Zipf's laws of meaning (Eq. 1 and 2) are not among the most investigated statistical laws of language and a mathematical argument illuminating the origins of both laws is not forthcoming. Thus, the basis for their current formulation is purely empirical. Second, the weak meaning-frequency law allows one to remain neutral about the actual dependency between μ and f. This neutral formulation has been adopted for research on Menzerath-Altmann's law in genomes (Ferrer-i-Cancho et al. 2013a) and Zipf's law of abbreviation in animal behavior (Ferrer-i-Cancho et al. 2013b and references therein). Third, the weak version allows for a unified approach to human language and the communicative behavior of other species. A positive correlation between frequency and behavioral context (a proxy for meaning) has been found in dolphin whistles (Ferrer-i-Cancho & McCowan 2009). Fourth, the weak version provides enough flexibility to allow for parsimonious models of language, models that reproduce more than one law of language at least qualitatively (Ferrer-i-Cancho, 2013).

In spite of the generality of the weak version of the law, its formulation imposes some hidden constraints. First, the fact that the law is an empirical law, imposes that only words that have non-zero probability matter for deciding if a theoretical model agrees with the law. Words that have zero probability are not observable. This second constraint might seem somewhat farfetched but Zipfian optimization models of communication can generate words with zero probability (Ferrer-i-Cancho & Díaz-Guilera 2007; Dickman et al. 2012; Prokopenko et al. 2010; Ferrer-i-Cancho 2013). Second, the definition of a proper correlation between μ and f (e.g., Pearson correlation, Spearman rank correlation) needs at least two different values of μ and at least two different values of f; otherwise, the variance of μ and that of f are undefined. To see it recall that the Pearson correlation between two variables X and Y (the ranks of μ and the ranks of f in case of the Spearman rank correlation between X and Y) is defined as

$$r = \frac{COV(X,Y)}{\sigma(X),\sigma(Y)}, \qquad (4)$$

where COV(X,Y) is the covariance of X and Y and $\sigma(X)$ and $\sigma(Y)$ are the standard deviation of X and Y, respectively. If $\sigma(X) = 0$ or $\sigma(Y) = 0$ the correlation is undefined.

If one wishes to determine if a theoretical model (e.g., Ferrer-i-Cancho & Solé 2003; Ferrer-i-Cancho 2005) agrees with the weak version of the law, frequencies must be replaced by probabilities. In that case, the definition of a proper correlation needs that there at least two different word probabilities. For the reasons explained above, words that have zero probability are excluded. Thus, the weak law cannot be defined properly in a communication system where only one word has non-zero probability or all words are equally likely. The weak law cannot be defined either in a system where only one word has non-zero degree or all words have the same degree.

Notice that the constraint of non-zero variance in the values of μ and f also concerns the strong meaning-frequency law (if f does not vary, Eq. 2 is not the only possibility for the relationship between μ and f).

Here we investigate if a family of Zipfian optimization models of communication (Ferrer-i-Cancho & Solé 2003; Ferrer-i-Cancho 2005) is able to reproduce some version of the meaning-frequency law. The family was conceived to investigate Zipf's law for word frequencies. It will be shown that those models yield Eq. 2 with $\delta = 1$ thus satisfying only the weak meaning-frequency relationship.

2. The family of Zipfian optimization models

The family of models departs from the assumption that there is a repertoire of V_S forms, $s_1,...,s_i,...,s_{V_S}$ and a repertoire of V_R meanings, $r_1,...,r_i,...,r_{V_R}$ that are associated through a binary matrix $A = \{a_{ij}\}$: $a_{ij} = 1$ if the s_i and r_j are associated ($a_{ij} = 0$ otherwise). The models of that family share also the assumption that the probability that a form s_i is employed to refer to meaning r_j is

$$p(s_i \mid r_j) = \frac{a_{ij}}{\omega_j}, \tag{5}$$

where

$$\omega_j = \sum_{i=1}^{V_S} a_{ij} \tag{6}$$

is the degree of the j-th meaning. The convention that $p(s_i|r_j) = 0$ when $\omega_j = 0$ is adopted. By definition, the marginal probability of s_i is

$$p(s_i) = \sum_{j=1}^{V_R} p(s_i, r_j) = \sum_{j=1}^{V_R} p(s_i \mid r_j) p(r_j). \tag{7}$$

The models of that family diverge by making further assumptions about $p(r_j)$. While one model (model B) assumes that $p(r_j)$ is given, e.g., $p(r_j) = 1/V_R$ assuming that no meaning is disconnected (Ferrer-i-Cancho & Solé 2003), another model (model A) assumes that (Ferrer-i-Cancho 2005)

$$p(r_j) = \frac{\omega_j}{M} \tag{8}$$

being M the total number of associations (links).

$$M = \sum_{j=1}^{V_R} \omega_j . \tag{9}$$

The Model A/B terminology is borrowed from Ferrer-i-Cancho & Díaz-Guilera (2007). Although the original Model B can easily be extended to allow for disconnected meanings, hereafter Model B with a ban for disconnected meanings is assumed for simplicity.

Applying the assumption of Eq. 5 and the convention on $p(s_i|r_j)$ above, Eq. 7 becomes

$$p(s_i) = \sum_{\substack{j=1 \\ \omega_j > 0}}^{V_R} \frac{a_{ij}}{\omega_j} p(r_j) . \tag{10}$$

For model A, Eq. 8 and 10 lead to

$$p(s_i) = \sum_{\substack{j=1 \\ \omega_j > 0}}^{V_R} \frac{a_{ij}}{\omega_j} \frac{\omega_j}{M} = \frac{1}{M} \sum_{j=1}^{V_R} a_{ij} = \frac{\mu_i}{M} , \tag{11}$$

where

$$\mu_i = \sum_{j=1}^{V_R} a_{ij} \tag{12}$$

is the degree of the i-th form. Eq. 11 indicates that form probability is proportional to form degree, i.e. model A satisfies Eq. 2 with $\delta = 1$. For model B, Eq. 10 and the assumption that $p(r_j) = 1/V_R$ (recall that no meaning can be disconnected) leads to

$$p(s_i) = \frac{1}{V_R} \sum_{j=1}^{V_R} \frac{a_{ij}}{\omega_j} . \tag{13}$$

The relationship between the probability of the i-th form and degree is not straightforward but it is possible to satisfy a weak version of the meaning-frequency law. Let us impose the following constraint on meaning degrees: $\omega_j = k$ with $0 < k \leq V_R$. This constraint transforms Eq. 13 into

$$p(s_i) = \frac{1}{kV_R} \sum_{j=1}^{V_R} a_{ij} = \frac{1}{kV_R} \mu_i . \tag{14}$$

Thus, the assumption of identical non-zero meaning degrees produces proportionality between the probability of the i-th form and its degree in Model B. Next section will show the utility of the case $k = 1$.

3. The weak meaning-frequency law is not inevitable

It has been shown that form probability is proportional to form degree directly in Model A and making further assumptions in Model B but this does not imply that those models are reproducing a weak meaning-frequency law. Some configurations of the matrix A where the weak law is missing will be shown next.

$H(S)$ is defined as the entropy of forms (S) and $I(S,R)$ is defined as the mutual information between forms (S) and meanings (R). The reader is referred to Ferrer-i-Cancho & Díaz-Guilera (2007) for definitions of those information theoretic measures.

If $H(S)$ is minimized, it is well known that then only one form has non-zero probability and non-zero degree (Ferrer-i-Cancho & Díaz-Guilera 2007). Then the variance of form probabilities is zero (recall that form probabilities that are zero are irrelevant for the weak version of the meaning-frequency law) and thus the correlation between form probabilities and semantic degree is not defined (recall Eq. 4). The same problem happens if $I(S,R)$ is maximized. Then the optimal solutions are those where all forms that have non-zero probability have the same degree or the same probability (Ferrer-i-Cancho 2013; Ferrer-i-Cancho & Díaz-Guilera 2007) and thus their variance is zero again.

4. A weak meaning-frequency law is possible

4.1. Possible in globally optimal configurations

The meaning-frequency law is possible (at least) in the global minima of $H(S|R)$, the conditional entropy of forms when meanings are given. The minima of $H(S|R)$ are characterized by $\omega_j \in \{0,1\}$ for model A (Ferrer-i-Cancho 2013; Ferrer-i-Cancho & Díaz-Guilera 2007) and $\omega_j = 1$ for model B (Ferrer-i-Cancho & Díaz-Guilera 2007; Prokopenko et al. 2010; Dickman et al. 2012). Those minima allow for an arbitrary number of words with non-zero probability/degree (Ferrer-i-Cancho & Díaz-Guilera 2007; Trosso 2008; Prokopenko et al. 2010; Dickman et al. 2012), a requirement of both the strong and weak meaning-frequency law.

The family of models assumes that languages minimize a linear combination of $H(S)$ and $I(S,R)$, i.e.

$$\Omega(\lambda) = -\lambda I(S,R) + (1-\lambda)H(S) \quad (15)$$

with $0 \leq \lambda \leq 1$.

Eq. 15 is equivalent to (Ferrer-i-Cancho 2005; Ferrer-i-Cancho & Díaz-Guilera 2007)

$$\Omega(\lambda) = (1-2\lambda)H(S) + \lambda H(S \mid R). \quad (16)$$

It is not surprising that the mapping of forms into meanings exhibits the principle contrast, the tendency of different forms to contrast in meaning (Clark 1987): the global minima of both $H(S)$ and $H(S|R)$ in Model A and B share $\omega_j \leq 1$ when $V_S \leq V_R$ (Ferrer-i-Cancho & Díaz-Guilera 2007; Ferrer-i-Cancho 2013).

The global minima of $\Omega(\lambda)$ split the range of variation of λ into three domains (Ferrer-i-Cancho & Díaz-Guilera 2007):

- $0 \leq \lambda < 1/2$ where only $H(S)$ is minimized. The weak meaning-frequency law is impossible (Section 3).
- $\lambda = 1/2$ where only $H(S|R)$ is minimized. The weak meaning-frequency law is possible (this section).
- $1/2 < \lambda \leq 1$ where only $I(S,R)$ is maximized. The weak meaning-frequency law is impossible (Section 3).

4.2. Possible in suboptimal configuration

The appearance of the weak meaning-frequency law is easier when the global minima are not reached. Indeed, those models generate a distribution of forms resembling Zipf's law for word frequencies when λ equals λ^*, a critical value of λ when $\Omega(\lambda)$ is optimized by means of an evolutionary algorithm based on a Monte Carlo method at zero temperature (Ferrer-i-Cancho & Solé 2003; Ferrer-i-Cancho 2005; Prokopenko et al. 2010). λ^* is typically a value below 1/2 but close to 1/2, i.e.

$$\lambda = 1/2 - \varepsilon \tag{17}$$

being ε a small positive quantity, e.g., $\varepsilon = 0.1$ (Ferrer-i-Cancho 2005; Ferrer-i-Cancho & Solé 2003; Prokopenko et al 2010). When $\lambda = \lambda^*$, there is enough variability in form probabilities and their degree to reproduce the meaning-frequency law (Ferrer-i-Cancho & Solé 2003; Ferrer-i-Cancho 2005; Prokopenko et al 2010), a requirement of both the strong and weak meaning-frequency law.

4.3. Where a weak meaning-frequency law is found

For model A, a weak meaning-frequency law in the minima of $H(S|R)$ (equivalent to $\Omega(1/2)$) or the suboptimal configurations appearing for $\lambda = \lambda^*$ is not only a possibility but a fact thanks to Eq. 11. For model B, some further reasoning is needed to turn apossibility into a fact. The global minima of $H(S|R)$, i.e. $\omega_j = 1$ with $M>0$, imply $k = 1$ in Eq. 14, which gives

$$p(s_i) = \frac{1}{V_R}\mu_i . \tag{18}$$

Being the probability of the i-th form proportional to its degree, a weak meaning-frequency law is expected in general in the minima of $H(S|R)$ for model B due to the variability of form degrees of these minima: configurations where (a) all forms have the same degree or (b) only one form is connected are unlikely (Trosso 2008; Prokopenko et al. 2010; Dickman et al. 2012). A weak law is also expected in the suboptimal configurations that are obtained for $\lambda = \lambda^*$. This implies that the minimization of $\Omega(\lambda)$ in Eq. 16 is being dominated by the minimization of $H(S|R)$: while the weight of $H(S)$ is small, i.e.

$$(1\text{-}2)\lambda^* = 1\text{-}2(1/2\text{-}\varepsilon) = \varepsilon, \tag{19}$$

the weight of $H(S|R)$ is relatively large, i.e.

$$\lambda = \lambda^* = 1/2 - \varepsilon, \tag{20}$$

thanks to Eq. 17. The fact that $H(S|R)$ is much stronger than $H(S)$ is critical for the emergence of the weak meaning-frequency law. The point is that the minimization of $H(S)$ implies the minimization of $H(S|R)$. On the one hand, this is positive for the emergence of the weak law because their minimization promotes in both cases $\omega_j \in \{0,1\}$. On the other hand, this is negative for the emergence of the weak law because we have shown that the minima of $H(S)$ turn the weak meaning-frequency law impossible (Section 3) and the fact that $H(S|R) \leq H(S)$, implies that, if $H(S)$ is minimum, i.e. $H(S) = 0$, then $H(S|R)$ is also minimum, i.e. $H(S|R) = 0$ (Ferrer-i-Cancho & Díaz-Guilera 2007). Being $H(S|R)$ much stronger than $H(S)$ a weak law is expected. Additional support for the arguments comes from the presence of Zipf's law for word frequencies for $\lambda = \lambda^*$ (Ferrer-i-Cancho, 2005; Ferrer-i-Cancho & Solé, 2003). If the minimization of $H(S)$ was dominating, instead of a distribution of this kind one would find one form (or a few forms) taking all probability. This is not what happens (Ferrer-i-Cancho 2005; Ferrer-i-Cancho & Solé 2003; Propopenko et al. 2010).

It is important to note that an inverse-factorial distribution has been derived for $\lambda = \lambda^*$ in Model B (Prokopento et al. 2010) and that this distribution differs from the traditional power-law that is typically used to approximate Zipf's law (Eq. 3). The inverse factorial should be considered as a candidate for in empirical research on Zipf's law (e.g., Li et al. 2010; Font-Clos et al 2013; Gerlach & Altmann 2013).

5. Discussion

We have shown some conditions where a weak meaning-frequency law, i.e. Eq. 2 with $\delta = 1$, appears in a family of Zipfian optimization models although the law it is not an inevitable property of the probabilistic definitions. Interestingly, that weak meaning-frequency law emerges (at least) in the narrow range where the models are argued to exhibit Zipf's law for word frequencies. Tentatively, those findings do not imply that a weak meaning-frequency law emerges only under very special circumstances. Suppose that m=$V_S V_R$. The binary association matrix A allows one to produce 2^m different mappings of words into meanings. The proportion of mappings (configurations of A) where a Spearman rank correlation is defined and has a positive sign could be large. That should be the subject of future research.

Randomness may facilitate the emergence of the weak law. For instance, consider configurations of the matrix A where the weak law is not found because the correlation is undefined (e.g., those where $H(S)$ is minimum or $I(S,R)$ is maximum). Producing a few random mutations in those configurations, it might be possible to obtain a variance of non-zero probabilities or non-zero degrees that is greater than zero and thus the correlation is defined (recall Eq. 3). Although the correlation is defined, the variation in form probability or form degree may be still too small with regard to real language.

There are many models of Zipf's law for word frequencies (Piantadosi 2014) but as far as we know the family of models reviewed here is the only that illuminates the origin of synchronic properties of language such as the principle of contrast and also dynamic properties such as the tendency of children to attach new words to unlinked meanings (Ferrer-i-Cancho 2013). It is tempting to conclude that the prediction of a linear relationship between number of meanings and frequency instead of the actual power law dependency of Eq. 2 is a reason to abandon this kind of models (i.e. the current family or variants stemming from it). Although the disagreement between the models examined so far and reality is a serious limitation (and thus we encourage future research), we cannot miss an important point: modern model selection is based on a compromise between parsimony and quality of fit (Burnham & Anderson 2002). To our knowledge, generative models for Eq. 2 are not forthcoming and the predictions of current models of Zipf's law beyond word frequencies are unknown, unexplored or simply impossible (Piantadosi 2014). There is at least one exception: the family of optimization models reviewed here, which is able to shed light on various statistical patterns qualitatively but in one shot from minimal assumptions.

The virtue of that family is not only its parsimonious approach to various laws but also its capacity to unify synchrony (patterns of language such as Zipf's law for word frequencies, a weak meaning-frequency law, the principle of contrast) with diachrony/ontogeny (through the vocabulary learning bias above). The science of the future must be unifying (Morin 1990). Theoretical linguistics cannot be an exception (Alday 2015; Ferrer-i-Cancho 2015).

Appendix

In his seminal work, Zipf derived Eq. 2 from Eq. 1 (law of meaning distribution) with $\gamma = 1/2$ and Eq. 2 (Zipf's law for word frequencies) with $\alpha = 1$. Here a more general and detailed derivation of Eq. 2 is provided. Notice that Eqs. 1 and 3 give, respectively,

$$i \propto \mu^{-\frac{1}{\gamma}} \tag{21}$$

and

$$i \propto f^{-\frac{1}{\alpha}}. \tag{22}$$

Combining Eqs. 21 and 22 it is obtained

$$\mu \propto f^{\delta} \tag{23}$$

with

$$\delta = \frac{\gamma}{\alpha}. \tag{24}$$

Therefore, $\gamma = 1/2$ and $\alpha = 1$ predict $\delta = 1/2$ as shown originally by Zipf (1945). Interestingly, Eq. 24 shows that Ilgen & Karaoglan's (2007) assumption, namely that $\gamma = \delta$, is only valid if $\alpha = 1$.

ACKNOWLEDGEMENTS

We are grateful to A. Corral for helpful comments on the manuscript. We also thank P. Alday and S. Piantadosi for helpful discussions. This work was supported by the grant BASMATI (TIN2011-27479-C04-03) from the Spanish Ministry of Science and Innovation, the grant APCOM (TIN2014-57226-P) from MINECO (Ministerio de Economía y Competitividad) and the grant 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya).

REFERENCES

Alday, P. (2015). Be careful when assuming the obvious. Commentary on "The placement of the head that minimizes online memory: a complex systems approach". *Language Dynamics and Change 5(1), 138-146.*

Clark, E. (1987). The principle of contrast: a constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition: 1-33*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crossley, S., Salsbury, T. & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language and Learning 60(3), 573-605.*

Baayen, H. & Moscoso del Prado Martín (2005). Semantic density and past-tense formation in three Germanic languages. *Language 81, 666-698.*

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd Ed. New York: Springer-Verlag.

Conover, W.J. (1999). *Practical nonparametric statistics*. New York, Wiley.

Dickman, R., Moloney, N.R., Altmann, E.G. (2012). Analysis of an information-theoretic model for communication. *Journal of Statistical Mechanics: Theory and Experiment, P12022*.

Ferrer-i-Cancho, R. (2005). Zipf's law from a communicative phase transition. *European Physical Journal B 47, 449-457.*

Ferrer-i-Cancho, R. (2013). The optimality of attaching unlinked labels to unlinked meanings. http://arxiv.org/abs/1310.5884

Ferrer-i-Cancho, (2015). Reply to the commentary: "Be careful when assuming the obvious", by P. Alday. *Language Dynamics and Change 5 (1), 147-155.*

Ferrer-i-Cancho, R., Riordan, O. & Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London Series B 272, 561-565.*

Ferrer-i-Cancho, R. & Solé, R.V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences USA 100, 788-791.*

Ferrer-i-Cancho, R. & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics, P06009*.

Ferrer-i-Cancho, R. & McCowan, B. (2009). A law of word meaning in dolphin whistle types. *Entropy 11 (4), 688-701.*

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G. & Baixeries, J. (2013a). The challenges of statistical patterns of language: the case of Menzerath's law in genomes. *Complexity 18 (3), 11–17.*

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M.J. & Semple, S. (2013b). Compression as a universal principle of animal behavior. *Cognitive Science 37 (8), 1565–1578.*

Font-Clos, F., Boleda, G. & Corral, A. (2013). A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics 15 (9), 093033*.

Font-Clos, F. & Corral, A. (2015). Log-log convexity of type-token growth in Zipf's systems. *Physical Review Letters 114, 237801.*

Gerlach, M. & Altmann, E.G. (2013). Stochastic model for the vocabulary growth in natural languages. *Physical Review X 3, 021006.*

Hernández-Fernández, A., Casas, B., Ferrer-i-Cancho, R. & Baixeries, J. (2016).Testing the robustness of laws of polysemy and brevity versus frequency. In: *4th International Conference on Statistical Language and Speech Processing (SLSP 2016)*. P. Král and C. Martín-Vide (eds.). *Lecture Notes in Computer Science 9918: 19–29*.

Ilgen, B. & Karaoglan, B. (2007). Investigation of Zipf's law-of-meaning on Turkish corpora. In: *22nd International Symposium on Computer and Information Sciences (ISCIS 2007), 1-6.*

Li, W., Miramontes, P. & Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy 2010, 12, 1743-1764.*

Morin, E. (1990). *Introduction à la pensée complexe*. Paris, ESF.

Piantadosi, S. (2014). Zipf's law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review, 21 (5), 1112-1130.*

Prokopenko, M., Ay, N., Obst, O. & Polani, D. (2010). Phase transitions in least-effort communications. *Journal of Statistical Mechanics: Theory and Experiment, P11025.*

Stumpf, M., P. Ingram, I. Nouvel & Wiuf, C. (2005). Statistical model selection methods applied to biological network data. *Transactions in Computational Systems Biology 3, 65-77.*

Stumpf, M.P.H. & Ingram, P.J. (2005). Probability models for degree distributions of protein interaction networks. *Europhysics Letters 71, 152.*

Trosso, A. (2008). La legge di Zipf ed il principio del minimo sforzo (Zipf's law and the principle of least effort). March-April 2008, University of Turin (Italy).

Zipf, G.K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology 33, 251-266.*

Zhou, K., Tuncali, K. & Silverman, S. G. (2003). Correlation and simpler linear regression. *Radiology, 227, 617–628.*

Glottometrics 35, 2016, 38-48

Activity in Italian Presidential Speeches

Peter Zörnig[1], *Gabriel Altmann*

Abstract. We analyze the activity of New-Year-Speeches of Italian presidents evaluated over the period from 1949 to 2012. The activity is measured in terms of Busemann´s indicator. The results are used both to compare the speeches of a given president and to describe the alteration over the analyzed 64 years. Some possible interpretations of the formal analysis are outlined.

Keywords: *Busemann's indicator, classification of speeches, text comparison and development, Italian language.*

1 Introduction

The mood of a text can be evaluated in many different ways. One may analyze the theme, the structure of sentences, the use of words, repetitions, conceptual inertia, etc. The number of aspects is constantly growing, it is practically infinite. A possible approach is to study the activity measured by the modified Busemann´s indicator

$$Q = \frac{V}{A+V}, \quad (1)$$

where V and A denote the number of verbs and adjectives in the text (Busemann 1925; Popescu et al. 2014; Zörnig et al. 2015, 4 ff., Popescu et al. 2015). One can say that basically verbs and adjectives indicate activity and descriptiveness, respectively. The expression (1) is the only "activity measure" for texts or speeches encountered in the linguistic literature. It should be clear that it is a very rudimentary tentative to capture the much more complex reality. Complications arise, for example, from the fact that verbs may express various "degrees of activity" – some of them being not active at all, e.g. "be", "have", "sleep" in English – and in some languages verbs do not differ formally from adjectives or can be used in the same grammatical positions (e.g. in Hungarian and Indonesian). Moreover, adverbs could also be taken into account. But adjectives and verbs are predicates of the first degree of nouns, and considering adverbs would complicate the calculation of the activity indicator.

It would be adequate to invent a kind of scaling of activity and take into account the adverbs which modify the given activity, but such a procedure requires analyzing all verbs and all adverbs that modify them and should be performed by psycholinguists. For this purpose test persons would be necessary since no scaling of this kind has been performed up to now. Restricted investigations concerning the orientation of space in language (expressed mostly by prepositions, adverbs and affixes) were performed for Nimboran and Slovak (cf. Altmann, Dömötör, Riška 1968a,b).

However, such a procedure is only of limited practical use. Anyway, software exists that may discern adjectives from verbs so that the indicator (1) can be calculated mechanically. Thus our studies of the present article are based on this specific measure of

[1] Univ. Brasilia (Brasilia). Email: peter@unb.br

activity. In the following section we analyze characteristics of the random variable Q in (1). In Section 3 we compute the Busemann´s indicator and the corresponding characteristics of the New-Year-Speeches of Italian presidents. A data analysis and tentative interpretations are presented in Section 4, followed by some final remarks in Section 5.

2 Characteristics of Busemann´s indicator

In order to model the distribution of the random variable V in (1), we restrict the text to a sequence of verbs and adjectives, simply ignoring other parts of speech that may occur, obtaining thereby the topical universe. Assuming that in the construction of this sequence a verb is chosen with probability p and an adjective with probability q = 1-p (cf. e.g. Altmann, Köhler 2015), the resulting number V of verbs is binomially distributed, i.e.

$$(2) \qquad P(V=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ for } k = 0,1,\ldots n,$$

where n is the number of verbs and adjectives, i.e. $n = V + A$.

From elementary probability theory (see e.g. Zörnig 2016, Sections 5.3, 8.1, 8.2) it is known that the expectation and variance of the random variable V are given by

$$(3) \qquad E(V) = np, \quad Var(V) = np(1-p)\ ,$$

implying

$$(4) \quad E(Q) = E\left(\frac{V}{n}\right)=\frac{1}{n}E(V)=p\,, \quad \mathrm{Var}(Q) = Var\left(\frac{V}{n}\right)=\frac{1}{n^2}Var(V)=\frac{p(1-p)}{n}$$

for the activity defined by (1). When the sequence of verbs and adjectives is sufficiently long, the activity Q is approximately normally distributed with the parameters given in (4). In columns three and four of Table 1 we present the number of adjectives and verbs in the analyzed speeches of Italian presidents. The fifths column shows the corresponding observed value of the Busemann's indicator Q. In order to decide whether the value is large or small, we calculate the squared deviation

$$(5) \quad X^2 = \frac{(V-A)^2}{V+A}=\frac{(2V-n)^2}{n}$$

between the numbers of adjectives and verbs. The distribution of X^2 is relatively complicated. By means of simulation one can show that it is "similar" to a chi-square distribution with one degree of freedom if p is "close to 0.5" and similar to a normal distribution otherwise.

Based on the value of the random variable X^2 we consider the following five activity classes:

SA = significantly active (V > A, X^2 > 3.84)
AC = active (V > A, X^2 < 3.84)

(6) NE = neutral ($X^2 < 0.5$)
DE = descriptive ($V < A$, $X^2 < 3.84$)
SD = significantly descriptive ($V < A$, $X^2 > 3.84$)

The constant 3.84 is derived from the chi-square distribution with one degree of freedom, having the probability density function $f(x) = \frac{e^{-x/2}}{\sqrt{2\pi x}}$. A variable *Y* following that distribution satisfies $P(Y > 3.84) = \int_{3.84}^{\infty} \frac{e^{-x/2}}{\sqrt{2\pi x}} dx \approx 0.05$. The threshold 0.5 to delimit the neutral region has been chosen arbitrarily.

3. Data collection regarding the activities in presidential speeches

The statistic X^2 and the respective class of the presidential speech are indicated in columns 6 and 7 of Table 1. The last column contains the variance of *Q* calculated according to (4), assuming that *p* is given by the observed ratio *V/n* of the respective speech·

Table 1
Activity in speeches of Italian presidents

President	Speech	*A*	*V*	*Q*	X^2	Class	Var(*Q*)
Einaudi	1949	30	33	0.5238	0.14	NE	0.00395926
	1950	15	20	0.5714	0.71	AC	0.00699708
	1951	41	34	0.4533	0.65	DE	0.00330430
	1952	27	28	0.5091	0.02	NE	0.00454395
	1953	34	24	0.4138	1.72	DE	0.00418221
	1954	43	36	0.4557	0.62	DE	0.00313971
Gronchi	1955	64	51	0.4435	1.47	DE	0.00214613
	1956	88	87	0.4971	0.01	NE	0.00142853
	1957	170	126	0.4257	6.54	SD	0.00082593
	1958	131	127	0.4922	0.06	NE	0.00096876
	1959	92	80	0.4651	0.84	DE	0.00144641
	1960	112	107	0.4886	0.11	NE	0.00114096
	1961	184	162	0.4682	1.40	DE	0.00071962
Segni	1962	120	83	0.4089	6.74	SD	0.00119062
	1963	170	131	0.4352	5.05	SD	0.00081662
Saragat	1964	85	64	0.4295	2.96	DE	0.00164452
	1965	141	138	0.4946	0.03	NE	0.00089595
	1966	185	144	0.4377	5.11	DE	0.00074808
	1967	167	145	0.4647	1.55	DE	0.00079730
	1968	176	134	0.4323	5.69	DE	0.00079165
	1969	222	232	0.5110	0.22	AC	0.00055039
	1970	272	257	0.4858	0.43	NE	0.00047221
Leone	1971	37	35	0.4861	0.06	NE	0.00346954
	1972	134	111	0.4531	2.16	DE	0.00101142
	1973	174	205	0.5409	2,54	AC	0.00065522

	1974	120	139	0.5367	1.39	AC	0.00096006
	1975	200	191	0.4885	0.21	NE	0.00063905
	1976	196	211	0.5184	0.55	AC	0.00061342
	1977	216	262	0.5481	4.43	SA	0.00051817
Pertini	1978	156	283	0.6446	36.74	AS	0.00052182
	1979	280	442	0.6122	36.35	SA	0.00032883
	1980	164	244	0.5980	15.69	SA	0.00058919
	1981	330	571	0.6337	64.46	SA	0.00025762
	1982	322	495	0.6059	36.63	SA	0.00029228
	1983	452	760	0.6271	78.27	SA	0.00019295
	1984	163	269	0.6269	26.01	SA	0.00054386
Cossiga	1985	404	289	0.4170	19.08	SD	0.00035082
	1986	215	187	0.4652	1.95	DE	0.00061887
	1987	349	248	0.4154	17.09	SD	0.00040678
	1988	369	311	0.4574	4.95	SD	0.00036497
	1989	303	231	0.4326	9.71	SD	0.00045965
	1990	533	396	0.4263	20.20	SD	0.00026325
	1991	64	57	0.4711	0.40	NE	0.00205920
Scalfaro	1992	360	472	0.5673	15.08	SA	0.00029504
	1993	387	469	0.5479	7.86	SA	0.00028938
	1994	482	590	0.5504	10.88	SA	0.00023084
	1995	523	741	0.5862	37.60	SA	0.00020470
	1996	326	313	0.4898	0.26	NE	0.00039107
	1997	521	1048	0.6679	177.01	SA	0.00014136
	1998	415	775	0.6513	108.91	SA	0.00019086
Ciampi	1999	276	291	0.5132	0.40	NE	0.00044061
	2000	272	291	0.5169	0.64	AC	0.00044354
	2001	260	338	0.5652	10.17	AC	0.00041095
	2001	301	312	0.5090	0.20	NE	0.00040770
	2003	211	231	0.5226	0.90	AC	0.00056445
	2004	264	268	0.5038	0.03	NE	0.00046990
	2005	164	181	0.5246	0.84	AC	0.00072288
Napolitano	2006	285	356	0.5554	7.86	SA	0.00038523
	2007	241	274	0.5320	2.11	AC	0.00048344
	2008	219	281	0.5620	7.69	SA	0.00049231
	2009	268	374	0.5826	17.50	SA	0.00037879
	2010	336	354	0.5130	0.47	NE	0.00036207
	2011	330	341	0.5082	0.18	NE	0.00037248
	2012	325	398	0.5505	7.37	SA	0.00034226

4. Data analysis and preliminary interpretations

Analyzing the activities in Table 1, we may distinguish between "internal" and "external" characteristics, the first of which refers to properties related to one and the same legislative period, the latter examining the alteration of presidential features over the time. In the first case we may e.g. ask whether the above defined activity of a president varies in the course of his mandate or which textual characteristics are typical of a politician. The End-of-Year speeches are of course dictated by the particular political, economic circumstances etc. but depend also on the attitude of the writer. It is well known that presidents have their writers, but the texts on which the speeches are based, are official documents reflecting the president´s opinion and may be used for analysis.

In previous studies it has already been demonstrated that the topic of presidential speeches depends only on individual choices where no clear time-effect can be observed. However, this does not mean that the choice of subject is completely random and independent of the geopolitical situation. One might consider the possibility that a certain regularity underlies the temporal evolution that could be detected by means of quantitative analyses. Moreover, clear temporal patterns have been observed e.g. in discourses of prime ministers, presidents of chambers and of the confederation of Italian industries.

In this paper we aim to extract any available information from the quantitative analyses.

4.1. Activity alteration over time

In a first attempt to express internal characteristics related to the legislative periods of presidents, we present the mean values of the random variables Q and X^2 in Table 2. The last column contains the *activity vectors*, whose components express the numbers of speech classes of the form SA, AC, NE, DE, SD, respectively (see (6)), which were presented by the same president.

Table 2
Mean activities and deviations in the presidential speeches

Legislation period	President	Mean activity	Mean squared deviation	Activity vectors [SA,AC,NE,DE,SD]
1: 1949-54	Einaudi	0.4879	0.64	(0,1,2,3,0)
2: 1955-61	Gronchi	0.4686	1.49	(0,0,3,3,1)
3: 1962-63	Segni	0.4220	5.90	(0,0,0,0,2)
4: 1964-70	Saragat	0.4651	2.28	(0,1,2,4,0)
5: 1971-77	Leone	0.5103	1.62	(1,3,2,1,0)
6: 1978-84	Pertini	0.6206	42.02	(7,0,0,0,0)
7: 1985-91	Cossiga	0.4407	10.48	(0,0,1,1,5)
8: 1992-98	Scalfaro	0.5801	51.09	(6,0,1,0,0)
9: 1999-05	Ciampi	0.5222	1.88	(0,4,3,0,0)
10:2006-12	Napolitano	0.5434	6.17	(4,1,2,0,0)

The alteration of the mean activity of the presidents is illustrated graphically in Figure 1.

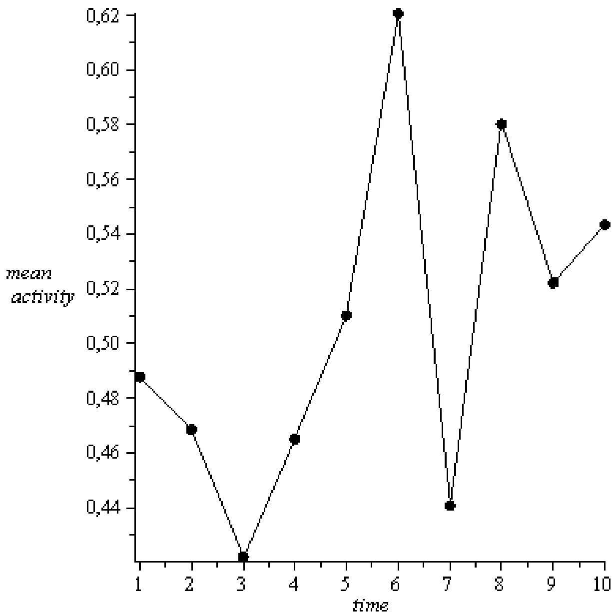

Figure 1. Mean activities of legislative periods

According to the above remarks no temporal trend can be discovered for this oscillating curve. However, one can identify the peaks corresponding to the presidents Pertini (1978-1984) and Scalfaro (1992-1998). These two presidents are characterized by the highest textual activities. In accordance with other analyses, these politicians exerted a great influence and differed from the other presidents. While Pertini often held extemporaneous speeches without reading from the template, Scalfaro was known for long speeches with an easy-to-read preachy language. The minima of the activity curve correspond to the politicians Segni (1962-1963) and Cossiga (1985-1991). The former gave only two speeches since he early resigned from his office and the latter announced his position a year ahead of schedule (in the last speech he informed his decision to resign from his office).

In summary one can say that the four mentioned statesmen are easily detected in the formal analysis as atypical personalities, since they represent the extrema of the curve in Fig. 1. Interpreting the activity indicator one might guess that high values indicate a great influence and a comprehensible language, while low values may indicate missing animation and career frustration.

4.2. Jumps in the course of time

In order to discover "jumps" i.e. abrupt changes between subsequent New-Year-Speeches, one may test the differences in the activity levels of subsequent years. We apply an asymptotic two-sided normal test in order to check for significant differences in subsequent activities. One compares the Q-values of consecutive years, Q_1 and Q_2, considering the test statistic

$$(6) \qquad U = |Y|, \qquad \text{where } Y = \frac{Q_1 - Q_2}{\sqrt{Var(Q_1) + Var(Q_2)}}.$$

Thereby *Y* is a reduced random variable, having expectation 0 and variance 1. Suppose that the activity is approximately normally distributed, then *Y* has the standard normal distribution.

It holds P(Y > 1.96) = 0.025, hence P(U > 1.96) = P(Y > 1.96) + P(Y < -1.96) = 2(0.025) = 0.05. That is, if U > 1.96 we can identify the transition as a "jump" for a significance level of 5%.

The *U*-values are summarized in Table 3. As can be seen, not all transitions from one president to his follower are jumps, and on the other hand jumps may occur between subsequent years of the same legislation period. The significant values (> 1.96) are indicated in bold face. The last value of a legislation period represents an activity change between two consecutive presidents.

Table 3
Activity changes in the New-Year speeches

President	Year	U	President	Year	U
Einaudi	1949-1950	0.45	Pertini	1978-1979	1.11
	1950-1951	1.16		1979-1980	0.47
	1951-1952	0.63		1980-1981	1.23
	1952-1953	1.02		1981-1982	1.19
	1953-1954	0.49		1982-1983	0.96
	1954-1955	0.17		1983-1984	0.16
				1984-1985	**6.88**
Gronchi	1955-1956	0.90	Cossiga	1985-1986	1.55
	1956-1957	1.51		1986-1987	1.55
	1957-1958	1.57		1987-1988	1.51
	1958-1959	0.55		1988-1989	0.86
	1959-1960	0.46		1989-1990	0.24
	1960-1961	0.47		1990-1991	0,93
	1961-1962	1.36		1991-1992	**1.98**
Segni	1962-1963	0.59	Scalfaro	1992-1993	0.80
	1963-1964	0.11		1993-1994	0.11
				1994-1995	1.74
				1995-1996	**3.99**
				1996-1997	**7.72**
				1997-1998	0.92
				1998-1999	**5.49**
Saragat	1964-1965	1.29	Ciampi	1999-2000	0.12
	1965-1966	1.40		2000-2001	1.65
	1966-1967	0.69		2001-2002	**1.97**
	1967-1968	0.81		2002-2003	0.44
	1968-1969	**2.15**		2003-2004	0.59
	1969-1970	0.79		2004-2005	0.60
	1970-1971	0.005		2005-2006	0.92
Leone	1971-1972	0.49	Napolitano	2006-2007	0.79
	1972-1973	**2.15**		2007-2008	0.96
	1973-1974	0.10		2008-2009	0.70
	1974-1975	1.21		2009-2010	**2.55**
	1975-1976	0.85		2010-2011	0.18
	1976-1977	0.88		2011-2012	1.58
	1977-1978	**2.99**			

One can see from Table 3 that there are four significant jumps between succeeding presidents (out of nine possible), and six jumps "within" a legislative period. In particular, the highest jumps occurred between 1984 and 1985 and between 1996 and 1997. One could ask again, whether significant activity jumps correspond to important historical data. A definite answer can only be given by historians specialized in Italian politics. The study of this question would be a step towards an interdisciplinary research.

The activity changes are illustrated graphically in Figure 2.

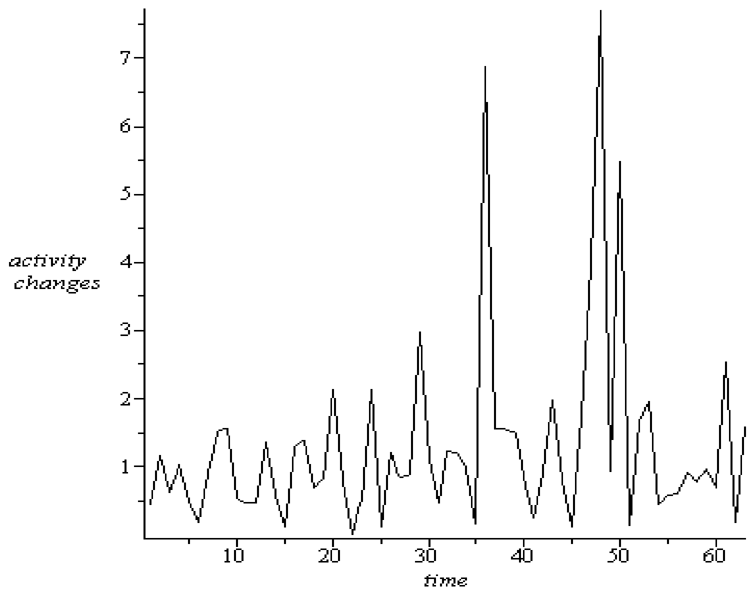

Figure 2. Activity differences between the New-Year-Speeches

4.3. Activity vectors

In the above studies we have considered activity in terms of a singular value of the Busemann´s indicator. We now characterize presidents or legislation periods by means of the activity vectors considered in Table 2. Using this concept, one may compare the presidents in different ways. One possibility is to determine the angles between the vectors. We make use of the elementary relation

$$\text{(7)} \qquad \cos(\alpha) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

where $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_n)$ are n-dimensional vectors and α denotes the angle in radians between x and y. By $x \cdot y = \sum_{i=1}^{n} x_i y_i$ and $\|x\| = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{n} x_i^2}$ we denote the scalar product between x and y and the norm of x, respectively.

In our application we have $n = 5$, since there are five activity vectors in Table 2. The angle α between the vectors x and y is now given as

(8) $$\alpha = \arccos\left(\frac{x \cdot y}{\|x\| \cdot \|y\|}\right).$$

Its value can vary between 0 and π, corresponding to 0 and 180 degrees, i.e. identical and opposed directions. Perpendicular vectors form an angle of $\pi/2$ radians or 90 degrees, respectively. In the following we analyze the direction change between the activity vectors of consecutive presidents. A large angle between these vectors might indicate a course change in the political strategy. For the vectors E = (0,1,2,3,0) and G = (0,0,3,3,1) corresponding to Einaudi and Gronchi we get e.g.

$$\cos(\alpha) = \frac{0\cdot 0 + 1\cdot 0 + 2\cdot 3 + 3\cdot 3 + 0\cdot 1}{\sqrt{0^2+1^2+2^2+3^2+0^2}\cdot\sqrt{0^2+0^2+3^2+3^2+1^2}} = \frac{15}{\sqrt{14}\cdot\sqrt{19}} = 0.9197.$$

Hence the angle between E and G is α = arcos(0.9197) = 0.4035 radians or $\alpha \cdot \frac{180}{\pi} \approx 23$ degrees. Performing these calculations for all consecutive presidents we obtain the results in Table 4.

Table 4
Angles between activity vectors of consecutive presidents

Subsequent presidents	Angle	
	in radians	in degrees
Einaudi – Gronchi	0.4035	23
Gronchi – Segni	1.3393	77
Segni – Saragat	1.5708	90
Saragat – Leone	0.9023	52
Leone – Pertini	1.3096	75
Pertini – Cossiga	1.5708	90
Cossiga – Scalfaro	1.5392	88
Scalfaro – Ciampi	1.4720	84
Ciampi – Napolitano	1.1192	64

One can observe that every change of a president caused a considerable change in the direction of the activity vector. The smallest change occurred when Einaudi was replaced by Gronchi. The largest occurred in the transitions from Segni to Saragat, Pertini to Cossiga and Cossiga to Scalfaro. Once more it would be interesting to check by political studies whether large angles correspond to certain changes in the style of governing

4.4. Reduced activity vectors

One possibility to simplify and sketch the activity vectors graphically consists of restricting them to the two extreme classes SA (significantly active) and SD (significantly descriptive). The *reduced activity vector* obtained thereby can be represented as a point in the plane (see Fig. 3). The dotted diagonal line divides the plane into a more active part (above) and a more descriptive part (below). It is interesting to observe that no president has combined high activity with high descriptiveness, i.e. if one of the president´s speeches is significantly active, none of his other speeches is significantly descriptive and – vice versa - if one of the president´s speeches is significantly descriptive, none of his other speeches is significantly active.

Geometrically interpreted, this means that the reduced activity vectors of all presidents are all located on one of the axes in Fig. 3. One could possibly use this graphic to group the presidents in personalities which tend to be more active, more descriptive or neutral in their speech (with respect to the above defined activity vector).

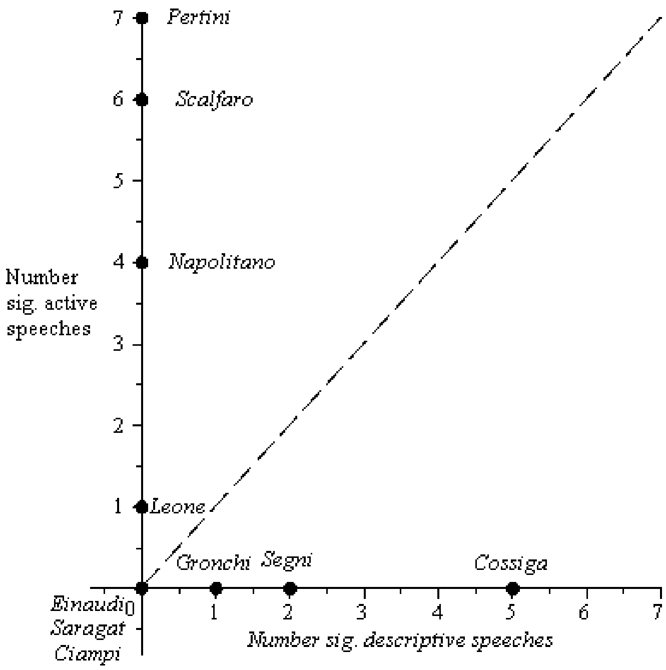

Figure 3. Reduced activity vectors

5. Conclusions

As already indicated in the introduction, “activity” in a text or speech can be defined and measured in different ways. The only concrete computable activity indictor available in the linguistic literature seems to be Buemann´s indicator, which we have used as the basis of our calculations.

It is obvious that other definitions of activity might yield different results. Only some individual analyses of textual activity have been performed in linguistics since Busemann´s “classical” paper of 1925 (cf. Altmann 1978; Wimmer et al. 2003). Therefore it is interesting to apply this concept to an extensive data material like the Italian presidential New-Year-Speeches. The analysis can be used as a basis for further investigations. The interpretations are of preliminary type. It should be clear that only intensive comparisons with political, sociological, economical or other important facts may lead to useful and reliable interpretations.

Acknowledgement

We would like to thank Arjuna Tuzzi for providing us with the files of formalized Italian presidential speeches.

References

Altmann, G. (1978). Zur Anwendung der Quotienten in der Textanalyse. *Glottometrika 1, 91-106.* Bochum: Brockmeyer.

Altmann, G., Dömötör, Z., Riška, A. (1968a). Darstellung des Raumes im System der slowakischen Präpositionen. *Jazykovedný časopis 19, 25-48.*

Altmann, G., Dömötör, Z., Riška, A. (1968b). The partition of space in Nimboran. *Beiträge zur Linguistik und Informationsverarbeitung 21, 56-71.*

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

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena: Fischer.

Popescu, I.-I., Čech, R., Altmann, G. (2014). Descriptivity in special texts. *Glottometrics 29, 70-80.*

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative analysis of poetic texts.* Berlin/Boston: de Gruyter Mouton.

Wimmer, G., Altmann, G., Hřebiček, L., Ondrejovič, S., Wimmerová, S. (2ßß3). *Úvod do analýzy textov.* Bratislava: Veda.

Zörnig, P. (2016). *Probability Theory and Statistical Applications: A profound treatise for self-study.* Berlin/New York: De Gruyter.

Zörnig, P., Stachowski, K., Mosavi Miangah, T., Mohanty, P., Kelih, E., Chen, R., Altmann, G. (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* Lüdenscheid: RAM-Verlag.

Glottometrics 35, 2016, 49-63

An Optimization Model of Global Language Complexity

Germán Coloma[1]

Abstract. In this paper we develop a theoretical model of global language complexity, based on a constrained optimization approach. We assume that language is a system that chooses different levels of complexity for its different domains (i.e., phonology, morphology, syntax, vocabulary) in order to minimize a global complexity function subject to an expressivity constraint (which also depends on non-linguistic variables related to geographic, phylogenetic and demographic factors). The model is illustrated with the aid of a dataset based on a short text translated into 50 languages, for which global complexity is measured using a version of Kolmogorov complexity. That dataset is used to run simultaneous-equation regressions, which represent different relationships between language complexity measures.

Keywords*: language complexity, optimization, Kolmogorov complexity, simultaneous-equation regression.*

1. Introduction

The literature about global language complexity is relatively vast and diverse. On one hand, there is a considerable amount of theoretical literature that has dealt with topics such as the definition of language complexity (e.g., Kusters 2003, Miestamo 2008, Culicover 2013) and its determinants (e.g., McWhorther 2001, Balasubrahmanyan & Naranan 2002, Hawkins 2004, Trudgill 2009). On the other hand, there is a good deal of empirical work that has either analyzed the relationship between complexity measures (e.g., Nettle 1995, Fenk-Oczlon & Fenk 2005, Shosted 2006, Sinnemäki 2008) or the relationship between those measures and other (non-linguistic) variables (e.g., Hay & Bauer 2007, Atkinson 2011).

The theoretical literature has also developed models assuming that language is a system, and that its behavior is guided by a hidden process which tries to achieve some desired objective. Among the main contributions to that literature we can mention Beckner et al. (2009), which states that language is a complex adaptive system whose structures emerge from interrelated patterns of experience, social interactions and cognitive mechanisms. Another group of studies in a similar line are the ones related to the concept of "synergetic linguistics" (e.g., Köhler 2005), for which language is a self-organizing and self-regulating system whose properties come from the interaction of several constitutive, forming and control requirements.

Part of that theoretical literature has explored the possibility of explaining the behavior of the language system through an optimization model (e.g., Ke, Ogura & Wang 2003, Ferrer-i-Cancho 2014, Futrell, Mahowald & Gibson 2015). However, we have not found any example from that literature in which the model used is directly related to complexity minimization, and this is probably the main contribution of the current article. The model that we develop here is, nevertheless, well known in other

[1] CEMA University, Buenos Aires, Argentina. Email: gcoloma@cema.edu.ar

social sciences such as economics (e.g., Chiang & Wainwright 2005), where cost minimization is a standard approach.

The main challenge for using a model like this is probably the fact that global language complexity is a rather indefinable concept, and it is therefore very hard to measure. The theoretical literature that has sought to find results related to its determinants (e.g., Hawkins 2004, Culicover 2013) has in general ended up with the conclusion that language complexity had better be studied using concepts that are applicable to specific situations (e.g., markedness, economy, efficiency). In the empirical literature, however, there is a measure derived from information theory that could represent the global complexity of a text. That measure is Kolmogorov complexity (Kolmogorov 1963), and it has been used by some authors in different linguistic settings (e.g., Juola 2008, Ehret & Szmrecsányi 2015).

Kolmogorov complexity can be defined as the length of the smallest algorithm required to generate a certain string of characters (Li & Vitányi 1997). Although in general it is formally incomputable, it can be approximated by the size of a compressed text file that comes from another (original) file. The ratio between the sizes of the two files, therefore, can be seen as an empirical measure of the global complexity of the text to which those files refer to, since the possibility of compressing the original file into a smaller one is directly linked to a series of characteristics (e.g., letter inventory, letter repetition, morpheme repetition, word repetition, clause length) that signal the complexity of the text.

In the following pages we will develop a model in which we assume that global language complexity is measurable (for example, by computing the Kolmogorov complexity of a representative text) and that it depends on several partial complexity variables (which can also be measured). We will also assume that those partial complexity variables are somehow "chosen" by the language under analysis in order to minimize global complexity, but that they are also influenced by non-linguistic variables related to phylogenetic, geographic and demographic factors. Those factors can also be important to determine language "expressivity", i.e., the capacity of a language to discriminate between possible alternative referents for a certain expression (Kirby et al. 2015). That expressivity will also depend on the different language domains involved in the production, transmission and decoding of linguistic messages (e.g., phonology, morphology, syntax and vocabulary).

Our model will be illustrated with an example based on data from a short text for which we have translations to 50 different languages. With those translations we compute several complexity measures (including Kolmogorov complexity) and build a dataset in which those measures are seen as the variables of the empirical version of our model. As the languages belong to different families and regions, and are spoken by different numbers of people, we can make use of that diversity to build three additional (categorical) variables. With all that we proceed to estimate the parameters implicit in our theoretical model, using a statistical procedure of simultaneous-equation regressions known as "three-stage least squares" (Zellner & Theil 1962).

2. Theoretical model

Let us assume that we can measure the global complexity of a language by a numerical continuous variable "g". Let us suppose, moreover, that the value of that variable is an increasing function "C" of several partial complexity variables related to different language domains (e.g., phonology, morphology, syntax, vocabulary). Let us now assume that those partial complexity variables are themselves numerical and continuous, and can be associated to "g" in the following way:

$$g = C(p, m, s, v) \qquad (1)$$

where "p", "m", "s" and "v" may represent, for example, the phonological, morphological, syntactic and lexical complexity of language.

In a context like this, global complexity can be seen as a measure of the effort that speakers have to exert in order to use the language under analysis. Therefore, the smaller the value of "g", the less costly a language is to be used by its speakers. But as language has to express meanings associated to its different components (i.e., to its words, clauses, texts, etc.), then its partial complexity levels can also be positively associated to its expressivity (through a function "E", which will be increasing in "p", "m", "s" and "v").

Following the ideas that appear in the literature about language as a complex adaptive system, we can think of the process of language evolution and transmission as an attempt to choose optimal levels for "p", "m", "s" and "v", which simultaneously minimize "C" and maximize "E". But this trade-off between opposing objectives can be influenced by other variables, such as phylogenetic, geographic and demographic factors ("pg", "gg", "dg"). One possible way to introduce those factors is to suppose that they operate as a determinant of the level of expressivity that a language must possess, through a restriction "R" (which integrates them into a single function). If that is the case, we can think of an "expressivity constraint" that can be written in the following way:

$$R(pg, gg, dg) = E(p, m, s, v) \qquad (2) .$$

If "R" is a constraint for the level of "E", and its determinants are exogenous to the language system, then our problem of choosing the optimal levels of "p", "m", "s" and "v" is somehow simplified, since it can be converted into one where we minimize "g" subject to the constraint stated in (2). If "C" and "E" are both continuous and differentiable in "p", "m", "s" and "v", that problem can be solved using a standard optimization technique known as the "Lagrange method". This method implies writing a Lagrangean function "L", which is defined as follows:

$$L = C(p, m, s, v) + \lambda \cdot [R(pg, gg, dg) - E(p, m, s, v)] \qquad (3)$$

and then finding the values of "p", "m", "s" and "v" for which the corresponding partial derivatives of "L" are equal to zero. These equalities are the "first-order conditions" of the problem, and can be written as:

$$\frac{\partial L}{\partial p} = \frac{\partial C}{\partial p} - \lambda \cdot \frac{\partial E}{\partial p} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C/\partial p)}{(\partial E/\partial p)} \qquad (4)$$

$$\frac{\partial L}{\partial m} = \frac{\partial C}{\partial m} - \lambda \cdot \frac{\partial E}{\partial m} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C/\partial m)}{(\partial E/\partial m)} \qquad (5)$$

$$\frac{\partial L}{\partial s} = \frac{\partial C}{\partial s} - \lambda \cdot \frac{\partial E}{\partial s} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C/\partial s)}{(\partial E/\partial s)} \qquad (6)$$

$$\frac{\partial L}{\partial v} = \frac{\partial C}{\partial v} - \lambda \cdot \frac{\partial E}{\partial v} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C/\partial v)}{(\partial E/\partial v)} \qquad (7).$$

In both the Lagrangean function and in its first-order conditions there is an additional "artificial variable" (λ), which is known as the "Lagrange multiplier" of the problem's constraint. This variable plays the role of converting the units in which the constraint is expressed (which in our case would be "expressivity units") into the units in which the objective function is expressed (i.e., complexity units). Due to that conversion, the first-order conditions can be stated as equations that relate infinitesimal changes in complexity with infinitesimal changes in expressivity, and establish optimal ratios between those changes.

Another role that the Lagrange multiplier plays is to include the fulfillment of the constraint as an additional first-order condition of the problem. This is due to the fact that, in order to minimize "C" subject to "R = E", we also need that:

$$\frac{\partial L}{\partial \lambda} = R(pg, gg, po) - E(p,m,s,v) = 0 \quad \rightarrow \quad R(pg,gg,dg) = E(p,m,s,v) \qquad (8)$$

and this last equation is included, together with equations (4) to (7), in a system whose solution is the one that determines the optimal values of "p", "m", "s" and "v" (and "λ").[1]

One relatively straightforward way to solve this system of equations is to use (4), (5), (6) and (7) to find the optimal relationships between each pair of partial complexity variables. By doing that, it is possible to express any complexity variable as a function of any other complexity variable (e.g., "m = m(p)", "s = s(p)", "s = s(m)", etc.). Making use of that possibility, we can replace those functions into (8), and write something like the following:

$$R(pg,gg,po) = E(p,m(p),s(p),v(p)) \quad \rightarrow \quad p = p(pg, gg, dg) \qquad (9)$$

$$R(pg,gg,po) = E(p(m),m,s(m),v(m)) \quad \rightarrow \quad m = m(pg, gg, dg) \qquad (10)$$

$$R(pg,gg,po) = E(p(s),m(s),s,v(s)) \quad \rightarrow \quad s = s(pg, gg, dg) \qquad (11)$$

$$R(pg,gg,po) = E(p(v),m(v),s(v),v) \quad \rightarrow \quad v = v(pg, gg, dg) \qquad (12) .$$

What equations (9) to (12) give us is actually the solution to our optimization problem. Each partial complexity variable is expressed as a function of the different phylogenetic, geographic and demographic factors that influence the language system under analysis, and can be inserted into equation (1) in order to get the minimum level of global complexity which is compatible with the fulfillment of the constraint stated in equation (2). By doing that, we obtain the following:

$$g = C(p(pg,gg,dg), m(pg,gg,dg), s(pg,gg,dg), v(pg,gg,dg)) \qquad (13)$$

which is an expression in which "g" is equated to a function that will ultimately depend on the actual levels of the non-linguistic variables. The whole process implied by this optimization model can therefore be represented by a graph like the one that appears in Figure 1.

[1] For a more complete explanation of this procedure, see Sundaram (1996), chapter 5.

As we can see, the idea behind this model is that the non-linguistic factors (i.e., phylogenetic, geographic and demographic), which influence the environment where language systems operate, have an effect on the way in which those systems choose the characteristics of their different domains (i.e., phonology, morphology, syntax and vocabulary). But as those characteristics are determined simultaneously, then their implied levels of partial complexity are related to each other, and they all have an impact on the global complexity of the system.

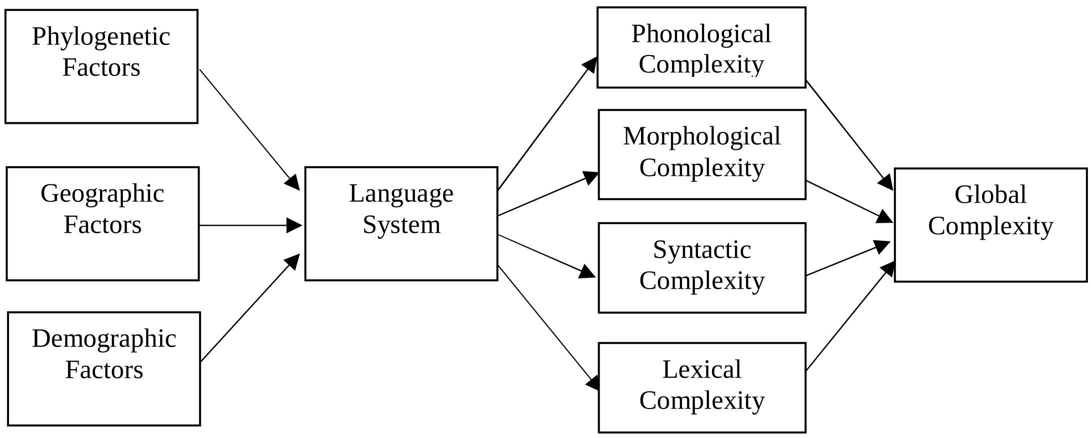

Figure 1. Language complexity model

3. Description of the data

In order to apply the theoretical model described in the previous section to an empirical example, we will use the same dataset that was previously employed in Coloma (2015, 2016), whose source is a series of articles published in IPA (1999) and in the *Journal of the International Phonetic Association*. It consists of a sample of 50 languages for which we have a version of the same text (the fable known as "The North Wind and the Sun"), on which we define different phonological, morphological, syntactic and lexical measures of complexity.[2] Those measures are the following:

Phonological inventory (*INV*): It is an index that consists of the sum of the number of consonant and vowel phonemes in each language, modified by the number of distinctive tones that such language possesses, and by the possible existence of distinctive levels of stress. This index is defined as:

$$INV = Consonants + Vowels*(Tones+Stress)$$

where *Consonants*, *Vowels* and *Tones* are numerical variables, and *Stress* is a binary variable that takes a value equal to one when stress is distinctive in a certain language (and zero otherwise).

Phoneme/word ratio (*PWR*): It is defined as the ratio between the total number of phonemes of "The North Wind and the Sun" text in each language, and the corresponding total number of words in that text.

[2] The list of languages and their complexity levels are reported in the appendix.

Word/clause ratio (*WCR*): It is defined as the ratio between the total number of words of “The North Wind and the Sun” text in each language, and the corresponding total number of clauses in that text.

Type/token ratio (*TTR*): It is defined as the ratio between the number of different words (types) of “The North Wind and the Sun” text in each language, and the total number of words (tokens) in that text.

To measure the global complexity of the texts under analysis we will use Kolmogorov complexity (*KC*). This will be defined as the ratio between the size of a compressed file (which contains the “The North Wind and the Sun” in a certain language) and the size of the original version of that file, both of them measured in bytes. Compression was made using the program “7zip”, version 4.32.

Another set of variables that we need for our empirical exercise is the one related to non-linguistic factors. This consists of one geographic variable, one phylogenetic variable and one demographic variable, all of which are categorical. The values of the geographic variable represent 10 different regions of the world, and each of them encompasses between 4 and 6 languages from our sample. The regions are North America, South America, Northern Europe, Southern Europe, Northern Africa, Southern Africa, West Asia, Central Asia, East Asia and the Pacific.

Additionally, the 50 languages in the sample belong to 24 different families, some of which are represented by more than one language. Those families are Indo-European (IE, 13 languages), Afro-Asiatic (AA, 5 languages), Niger-Congo (NC, 4 languages), Sino-Tibetan (ST, 3 languages), Altaic (Alt, 3 languages), Austronesian (Aus, 2 languages), Nilo-Saharan (NS, 2 languages), Oto-Manguean (OM, 2 languages), plus two languages that can generically be referred to as “Amazonian” (Amaz). The remaining 14 languages are grouped into another category named “Other families”.[3]

The demographic variable, finally, divides languages into three categories: “large languages”, “medium languages” and “small languages” (according to the number of speakers that the different languages possess). The group of large languages is constituted by the following 12 cases: Mandarin, English, Spanish, Hindi, Arabic, Portuguese, Russian, Japanese, Bengali, German, French and Malay. Correspondingly, the ones that belong to the category of “small languages” (less than 1 million speakers) are the following: Apache, Arrernte, Basque, Chickasaw, Dinka, Irish, Mapudungun, Nara, Sahaptin, Sandawe, Seri, Shiwilu, Tausug, Trique and Yine. The remaining 23 languages in the sample are considered to be “medium-sized”.

The main descriptive statistics of this database are summarized in Table 1, in which we find the average values of *INV*, *PWR*, *WCR*, *TTR* and *KC* for each group of languages. In that table we can see, for example, that East Asian and Oto-Manguean languages tend to be phonologically more complex, and that Amazonian languages tend to have higher phoneme/word ratios. The higher word/clause ratios, conversely, appear in Indo-European languages (especially in the Northern European ones), while the highest type/token ratios seem to occur in West Asia and in Sino-Tibetan languages. Finally, Kolmogorov complexity is higher in Southern Europe and East Asia, and in Sino-Tibetan and Altaic languages.

Another set of descriptive statistics that could be useful to analyze our complexity measures is the one formed by the correlation coefficients between the different variables. Those coefficients are reported in Table 2, in which we see that there are several partial complexity measures that display relatively significant negative correlation coefficients between themselves. The most important one is the coefficient

[3] To see which language belongs to each family, see the table included in the appendix.

between *PWR* and *WCR* (r = -0.7265), followed by the one between *WCR* and *TTR* (r = -0.5046), and by the one between *INV* and *PWR* (r = -0.3720). Conversely, Kolmogorov complexity exhibits relatively large positive correlation coefficients with *TTR* and *INV* ("r = 0.3639" and "r = 0.2543"), and small negative correlation coefficients with *PWR* and *WCR* ("r = -0.1395" and "r = -0.1108").

Table 1
Descriptive statistics for complexity variables

Category / Variable	INV	PWR	WCR	TTR	KC
Northern Europe	44.20	3.964	12.53	0.6336	0.7203
Southern Europe	33.60	4.086	11.93	0.6105	0.7490
Northern Africa	42.75	4.787	11.04	0.6417	0.6978
Southern Africa	49.00	4.469	10.98	0.6296	0.7002
North America	49.83	5.115	9.22	0.5938	0.6851
South America	25.00	7.076	7.58	0.6541	0.6357
West Asia	33.50	5.854	9.18	0.7634	0.7417
Central Asia	36.00	5.060	11.33	0.7067	0.6573
East Asia	68.50	4.709	10.32	0.6865	0.7911
Pacific	26.25	5.530	8.80	0.5838	0.6450
Indo-European	38.38	4.259	12.28	0.6555	0.7134
Afro-Asiatic	39.20	5.277	10.40	0.7164	0.6956
Niger-Congo	43.00	4.298	11.15	0.6231	0.7231
Sino-Tibetan	66.00	5.128	8.23	0.7469	0.8291
Altaic	31.00	6.009	8.52	0.7132	0.8103
Austronesian	22.00	5.592	9.63	0.5489	0.6252
Nilo-Saharan	46.50	4.157	11.76	0.5469	0.6792
Oto-Manguean	68.00	3.557	10.18	0.6173	0.7749
Amazonian	23.00	8.457	6.91	0.6904	0.5874
Other families	45.50	5.361	9.44	0.6302	0.6787
Large languages	37.67	4.439	11.15	0.6507	0.7153
Medium languages	44.48	5.048	10.30	0.6873	0.7315
Small languages	42.60	5.438	9.65	0.5992	0.6661
Average	**42.28**	**5.019**	**10.31**	**0.6521**	**0.7080**

Table 2
Correlation coefficients between complexity variables

Complexity variable	INV	PWR	WCR	TTR	KC
Phonological inventory	1.0000				
Phoneme/word ratio	-0.3720	1.0000			
Word/clause ratio	0.2136	-0.7265	1.0000		
Type/token ratio	-0.0428	0.5581	-0.5046	1.0000	
Kolmogorov complexity	0.2543	-0.1395	-0.1108	0.3639	1.0000

4. Empirical estimation

The dataset that we described in section 3 can be used to perform an estimation of the model developed in section 2. In order to do that, we first need to define which empirical variables will be used to approximate the theoretical variables of the model, and which functional forms can approximate the relationships that that model displays.

One obvious possibility is to use *INV*, *PWR*, *WCR* and *TTR* as proxies for "p", "m", "s" and "v", respectively. An easy way to use them to write expressions for the functions "C" and "E" is to suppose that the first of those functions is linear, and that the second one is log-linear. This implies that both functions will depend on four variables and four parameters each, and they can be written as:

$$C = c1 \cdot INV + c2 \cdot PWR + c3 \cdot WCR + c4 \cdot TTR \quad (14)$$

$$E = a1 \cdot ln(INV) + a2 \cdot ln(PWR) + a3 \cdot ln(WCR) + a4 \cdot ln(TTR) \quad (15)$$

where *c1*, *c2*, *c3* and *c4* are the parameters of the complexity function, and *a1*, *a2*, *a3* and *a4* are the parameters of the expressivity function.

In order to perform a statistical estimation of "C", it is straightforward to assume that global complexity can be approximated by the value of *KC*. This implies that parameters *c1*, *c2*, *c3* and *c4* are going to be the results of a procedure in which *KC* is regressed as a linear function of *INV*, *PWR*, *WCR* and *TTR*. The estimation of "E", conversely, is considerably more cumbersome, since we do not have any empirical variable that can easily be associated to a measure of expressivity. What we can do, instead, is to work with the first-order conditions of the theoretical optimization problem described in section 2, and write them in the following way:

$$\frac{\partial C/\partial p}{\partial E/\partial p} = \frac{\partial C/\partial m}{\partial E/\partial m} = \frac{\partial C/\partial s}{\partial E/\partial s} = \frac{\partial C/\partial v}{\partial E/\partial v} \rightarrow \frac{c1}{a1/INV} = \frac{c2}{a2/PWR} = \frac{c3}{a3/WCR} = \frac{c4}{a4/TTR} \quad (16)$$

As those relationships imply equality signs, it is possible to write equations that relate complexity variables in pairs. Those pairs are the following:

$$INV = \frac{a1}{a2} \cdot \frac{c2}{c1} \cdot PWR \; ; \qquad INV = \frac{a1}{a3} \cdot \frac{c3}{c1} \cdot WCR \; ; \qquad INV = \frac{a1}{a4} \cdot \frac{c4}{c1} \cdot TTR \quad (17)$$

$$PWR = \frac{a2}{a1} \cdot \frac{c1}{c2} \cdot INV \; ; \qquad PWR = \frac{a2}{a3} \cdot \frac{c3}{c2} \cdot WCR \; ; \qquad PWR = \frac{a2}{a4} \cdot \frac{c4}{c2} \cdot TTR \quad (18)$$

$$WCR = \frac{a3}{a1} \cdot \frac{c1}{c3} \cdot INV \; ; \qquad WCR = \frac{a3}{a2} \cdot \frac{c2}{c3} \cdot PWR \; ; \qquad WCR = \frac{a3}{a4} \cdot \frac{c4}{c3} \cdot TTR \quad (19)$$

$$TTR = \frac{a4}{a1} \cdot \frac{c1}{c4} \cdot INV \; ; \qquad TTR = \frac{a4}{a2} \cdot \frac{c2}{c4} \cdot PWR \; ; \qquad TTR = \frac{a4}{a3} \cdot \frac{c3}{c4} \cdot WCR \quad (20) .$$

The equations that appear in (17), (18), (19) and (20) can also be added and reduced to four regression equations, so we end up with a system like this:

$$INV \cdot 3 = \frac{a1}{a2} \cdot \frac{c2}{c1} \cdot PWR + \frac{a1}{a3} \cdot \frac{c3}{c1} \cdot WCR + \frac{a1}{a4} \cdot \frac{c4}{c1} \cdot TTR \quad (21)$$

$$PWR \cdot 3 = \frac{a2}{a1} \cdot \frac{c1}{c2} \cdot INV + \frac{a2}{a3} \cdot \frac{c3}{c2} \cdot WCR + \frac{a2}{a4} \cdot \frac{c4}{c2} \cdot TTR \quad (22)$$

$$WCR \cdot 3 = \frac{a3}{a1} \cdot \frac{c1}{c3} \cdot INV + \frac{a3}{a2} \cdot \frac{c2}{c3} \cdot PWR + \frac{a3}{a4} \cdot \frac{c4}{c3} \cdot TTR \quad (23)$$

$$TTR \cdot 3 = \frac{a4}{a1} \cdot \frac{c1}{c4} \cdot INV + \frac{a4}{a2} \cdot \frac{c2}{c4} \cdot PWR + \frac{a4}{a3} \cdot \frac{c3}{c4} \cdot WCR \quad (24).$$

Another set of equations from the theoretical model that can be empirically estimated is the one that corresponds to the system formed by (9), (10), (11) and (12). One simple way to do it is working with the three non-linguistic variables described in section 3, and regressing each partial complexity variable (*INV*, *PWR*, *WCR* and *TTR*) against those categorical variables. What we obtain is something like this:

$$INV = b1r + b1f + b1p \quad ; \quad PWR = b2r + b2f + b2p \quad (25)$$

$$WCR = b3r + b3f + b3p \quad ; \quad TTR = b4r + b4f + b4p \quad (26)$$

where the different *bij* coefficients represent measures of the effect that each category (i.e., each region, family and population size group) has on our partial complexity variables.

If we estimate the system of equations represented in (25) and (26), using ordinary least squares,[4] we obtain a set of coefficients that can be used to build "instrumental variables". Those instrumental variables are created to replace the original partial complexity variables in a new set of regressions, and we will label them as $\widehat{INV}$, $\widehat{PWR}$, $\widehat{WCR}$ and $\widehat{TTR}$. They are formed by the fitted values of the regressions for equations (25)/(26), and their role is to represent the optimal values of *INV*, *PWR*, *WCR* and *TTR* in (21), (22), (23) and (24) (without including any endogenous elements that could make our estimation biased or inconsistent).[5]

The new set of regression equations can therefore be written in the following way:

$$KC = c1 \cdot \widehat{INV} + c2 \cdot \widehat{PWR} + c3 \cdot \widehat{WCR} + c4 \cdot \widehat{TTR} \quad (27)$$

$$INV \cdot 3 = c5 \cdot \widehat{PWR} + c6 \cdot \widehat{WCR} + c7 \cdot \widehat{TTR} \quad (28)$$

$$PWR \cdot 3 = (1/c5) \cdot \widehat{INV} + (c6/c5) \cdot \widehat{WCR} + (c7/c5) \cdot \widehat{TTR} \quad (29)$$

$$WCR \cdot 3 = (1/c6) \cdot \widehat{INV} + (c5/c6) \cdot \widehat{PWR} + (c7/c6) \cdot \widehat{TTR} \quad (30)$$

$$TTR \cdot 3 = (1/c7) \cdot \widehat{INV} + (c5/c7) \cdot \widehat{PWR} + (c6/c7) \cdot \widehat{WCR} \quad (31)$$

where "$c5 = (a1 \cdot c2)/(a2 \cdot c1)$", "$c6 = (a1 \cdot c3)/(a3 \cdot c1)$" and "$c7 = (a1 \cdot c4)/(a4 \cdot c1)$". Their results are reported in Table 3, and were obtained using three-stage least squares.

[4] This estimation was performed using the computing program EViews 3.1. The same software was used for the other regressions whose results are reported in this paper.
[5] For an explanation of the logic behind this procedure, see Kennedy (2008), chapter 10.

Table 3
Three-stage least square regression results

Parameter	Coefficient	Std. Error	t-statistic	Probability
c1	0.001300	0.000799	1.628061	0.1048
c2	0.026399	0.013069	2.019964	0.0445
c3	0.025719	0.005237	4.911356	0.0000
c4	0.392505	0.164358	2.388111	0.0177
c5	9.087457	0.588355	15.445540	0.0000
c6	4.342966	0.223163	19.461000	0.0000
c7	68.872710	3.391096	20.309870	0.0000

With the values that we have found, we can directly compute the parameters of the complexity function (*c1*, *c2*, *c3* and *c4*). Using an indirect calculation, we can also compute values for the parameters of the expressivity function (*a1*, *a2*, *a3* and *a4*). In particular, if we set those values so that they add up to one, we get the coefficients of equation (33), together with the complexity function written as equation (32).

$$C = 0.0013 \cdot INV + 0.026399 \cdot PWR + 0.025719 \cdot WCR + 0.392505 \cdot TTR \quad (32)$$

$$E = 0.0821 \cdot ln(INV) + 0.1836 \cdot ln(PWR) + 0.3742 \cdot ln(WCR) + 0.3601 \cdot ln(TTR) \quad (33).$$

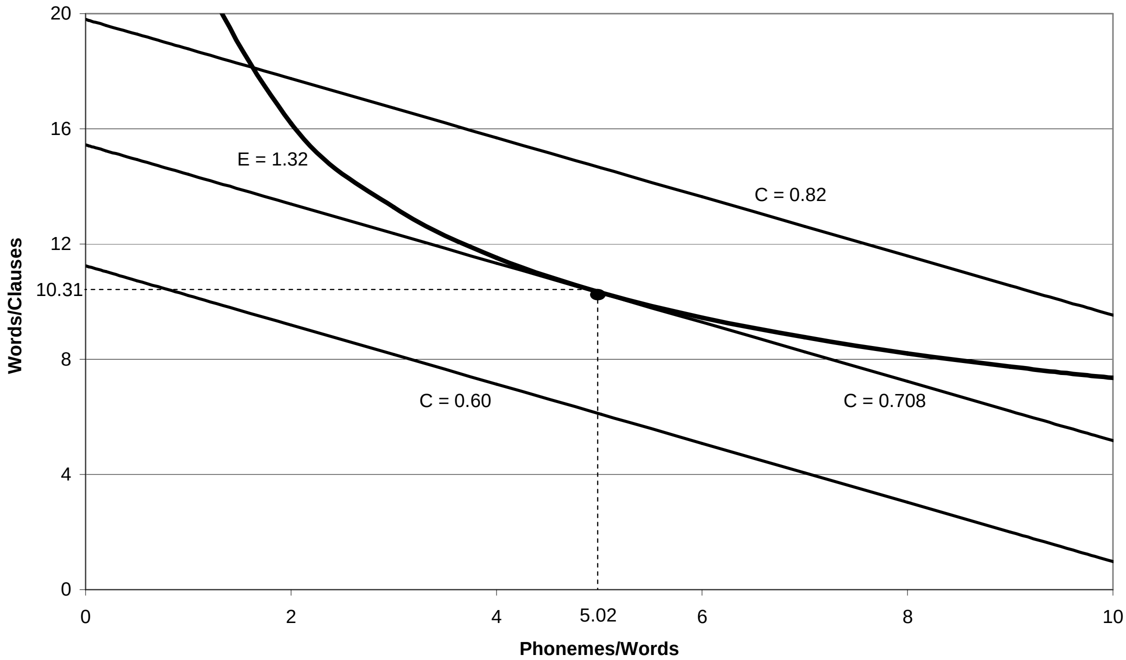

Figure 2. Iso-expressivity and iso-complexity curves

Equations (32) and (33) can be represented in a diagram like the one that appears in Figure 2, in which we have drawn one particular case of "E" (the one that corresponds to the expressivity levels implied by the average values of *INV*, *PWR*, *WCR* and *TTR*) and three particular cases of "C" (then one in which that function equates the average level of *KC*, plus two additional ones). As we can see, this diagram is depicted in the space of *PWR* vs. *WCR*, and in it we find that our iso-expressivity curve is tangent to the iso-complexity line for which $C = 0.708$ (which is the average level of *KC* in our sample). This means that such level of global complexity is the minimum one that could be obtained if we require that a language has the expressivity implied by the average levels of *INV*, *PWR*, *WCR* and *TTR*. Moreover, in that tangency point we see that the values for *PWR* and *WCR* are the ones that correspond to the average values of those variables (i.e., $PWR = 5.02$ and $WCR = 10.31$).

If we make a small variation in our model, we can also use it to estimate partial correlation coefficients between the complexity variables. In order to do that, we have to write equations (28)/(31) in the following way:

$$INV \cdot 3 = c5 \cdot \widehat{PWR} + c6 \cdot \widehat{WCR} + c7 \cdot \widehat{TTR} \tag{34}$$

$$PWR \cdot 3 = c8 \cdot \widehat{INV} + c6 \cdot c8 \cdot \widehat{WCR} + c7 \cdot c8 \cdot \widehat{TTR} \tag{35}$$

$$WCR \cdot 3 = c9 \cdot \widehat{INV} + c5 \cdot c9 \cdot \widehat{PWR} + c7 \cdot c9 \cdot \widehat{TTR} \tag{36}$$

$$TTR \cdot 3 = c10 \cdot \widehat{INV} + c5 \cdot c10 \cdot \widehat{PWR} + c6 \cdot c10 \cdot \widehat{WCR} \tag{37}$$

where we assume that *c8*, *c9* and *c10* are not necessarily equal to *1/c5*, *1/c6* and *1/c7*. Let us now define the correlation coefficients between our variables using the following formula:

$$r_{xy} = -\sqrt{\frac{c_{xy} \cdot c_{yx}}{9}} \tag{38}$$

where r_{xy} is the partial correlation coefficient between variables *x* and *y*, c_{xy} is the regression coefficient that corresponds to $\hat{y}$ in the equation where the dependent variable is $x \cdot 3$, and c_{yx} is the regression coefficient that corresponds to $\hat{x}$ in the equation where the dependent variable is $y \cdot 3$.[6]

Table 4
Partial correlation coefficients between complexity variables

Complexity variable	INV	PWR	WCR	TTR
Phonological inventory	1.0000			
Phoneme/word ratio	-0.2322	1.0000		
Word/clause ratio	-0.3720	-0.2592	1.0000	
Type/token ratio	-0.3947	-0.2750	-0.4406	1.0000

Using the results obtained in our new set of regressions, we used equation (38) to calculate the coefficients reported in Table 4. All of them turned out to be statistically significant at a 1% probability level, and the largest absolute value is the one that corresponds to the relationship between *WCR* and *TTR*. Note also that some variables that display positive product-moment correlation coefficients in Table 2 (*INV* vs. *WCR*, and *PWR* vs. *TTR*) are now negatively related. This is consistent with the idea that partial complexity measures are linked through the interaction between functions “C” and “E”, and must therefore be negatively correlated in all cases.

[6] For an explanation of the logic behind this formula, see Prokhorov (2002).

5. Concluding remarks

The two points that we believe are more important in this paper are related to the use of the proposed optimization model, and to its implementation through the concept of Kolmogorov complexity. On one hand, we think that our model is an elegant theoretical approach to the general problem of language as a self-regulated system, and that it is also good to include the interaction that the elements from that system may have with external forces such as phylogenetic, geographic and demographic factors.

On the other hand, we see the concept of Kolmogorov complexity (and its approximation through the ratio between the sizes of a compressed file and an original text file) as a promising empirical approach to global language complexity. Due to the fact that it is a measure that can be applied to different texts, it can also be correlated to other (partial) complexity measures for those texts, which can in turn be seen as their internal determinants.

The logic behind our results is that the relationship between the different complexity measures can be interpreted as the outcome of a process in which the language system defines certain levels of partial complexity in order to minimize a global complexity function, subject to an expressivity constraint. Using particular functional forms for those relationships, we were able to illustrate them through various parameters that are estimated in a simultaneous-equation regression procedure. In that procedure, we also used information from non-linguistic variables that define each observation in our sample (i.e., the region and family to which each language belongs, and its size in terms of number of speakers).

However, the empirical illustration included in this paper is not intended to test the accuracy of the proposed model to fit actual data. Its purpose is to show how the theoretical variables of the model can be interpreted as observable variables, and how those observable variables can be used to figure out plausible "shapes" for the functions postulated in the theoretical model. Of course, the model could also be used, in a different setting, to be contrasted with another theoretical alternative that provides a different explanation for language complexity phenomena.

Another possible use of the optimization model developed in this paper has to do with testing different definitions for the global complexity variables (apart from Kolmogorov complexity). It could also be possible to use the empirical version of the model to test different functional forms for the complexity and expressivity functions, since our linear and logarithmic versions of those functions are just one, relatively simple, alternative to write the relationships embedded in the theoretical model. That alternative can certainly be contrasted with other additional specifications.

Finally, the model could be applied in different contexts that were not necessarily cross-linguistic. An alternative sample to the one used could consist of texts written in the same language, but belonging to different authors, or genres, or styles, or time periods that cover different stages in the evolution of language.

Acknowledgments

This paper was financed by the Research Fund of CEMA University (Buenos Aires, Argentina), as part of the project entitled "The Use of Econometric Methods to Analyze Linguistic Problems". I thank Gabriel Altmann, Damián Blasi, Mariana Conte Grand and Ezequiel Koile for their useful comments.

References

Atkinson, Quentin (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science 332, 346-349*.

Balasubrahmanyan, Viddhachalam & Sundaresan Naranan (2002). Algorithmic Information, Complexity and Zipf's Law. *Glottometrics 4, 1–26.*

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

Chiang, Alpha & Kevin Wainwright (2005). *Fundamental Methods of Mathematical Economics*, 4th edition. Boston: McGraw-Hill.

Coloma, Germán (2015). The Menzerath-Altmann Law in a Cross-Linguistic Context. *SKY Journal of Linguistics 28, 139-159*.

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

Culicover, Peter (2013). *Grammar and Complexity*. Oxford: Oxford University Press.

Ehret, Katharina & Benedikt Szmrecsányi (2015). An Information-Theoretic Approach to Assess Linguistic Complexity". In: R. Baechler & G. Seiler (eds.): *Complexity, Variation and Isolation*. Berlin: De Gruyter.

Fenk-Oczlon, Gertraud & August Fenk (2005). Crosslinguistic Correlations Between Size of Syllables, Number of Cases, and Adposition Order. In: G. Fenk-Oczlon & C. Winkler (eds.): *Sprache und Natürlichkeit: 75-86*. Tübingen: Narr.

Ferrer-i-Cancho, Ramón (2014). *Optimization Models of Natural Communication*, mimeo. Barcelona: Universitat Politècnica de Catalunya.

Futrell, Richard, Kyle Mahowald & Edward Gibson (2015). Large-Scale Evidence of Dependency Length Minimization in 37 Languages. *PNAS 112(33), 10336-10341*.

Hawkins, John (2004). *Efficiency and Complexity in Grammars*. New York: Oxford University Press.

Hay, Jennifer & Laurie Bauer (2007). Phoneme Inventory Size and Population Size. *Language 83, 388-400.*

IPA (1999). *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.

Juola, Patrick (2008). Assessing Linguistic Complexity. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 89-108.* Amsterdam: John Benjamins.

Ke, Jinyun, Mieko Ogura & William Wang (2003). Optimization Models of Sound Systems Using Genetic Algorithms. *Computational Linguistics 29, 1-18.*

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

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

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

Kolmogorov, Andrei (1963). On Tables of Random Numbers. *Sankhya 25, 369-376*.

Kusters, Wouter (2003). *Linguistic Complexity*. Utrecht: LOT.

Li, Ming & Paul Vitányi (1997). *An Introduction to Kolmogorov Complexity and its*

Applications, 2nd edition. New York: Springer.

McWhorter, John (2001). The World's Simplest Grammars Are Creole Grammars. *Linguistic Typology 5, 125-166*.

Miestamo, Matti (2008). Grammatical Complexity in a Cross-Linguistic Perspective. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 23-41*. Amsterdam: John Benjamins.

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

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

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

Sinnemäki, Kaius (2008). Complexity Trade-Offs in Core Argument Marking. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 67-88.* Amsterdam: John Benjamins.

Sundaram, Rangarajan (1996). *A First Course in Optimization Theory*. New York, Cambridge University Press.

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

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

Appendix
Dataset from "The North Wind and the Sun"

Language	Family	Region	Size	INV	PWR	WCR	TTR	KC
Amharic	AA	NAfrica	Medium	41	6.958	11.88	0.7263	0.6239
Apache	Other	NAmerica	Small	57	4.907	7.87	0.6017	0.6167
Arabic	AA	WAsia	Large	35	5.741	9.44	0.7647	0.6962
Arrernte	Other	Pacific	Small	35	5.892	6.17	0.6351	0.6150
Basque	Other	SEurope	Small	33	4.831	11.86	0.6506	0.7690
Bemba	NC	SAfrica	Medium	46	5.506	9.88	0.7468	0.7138
Bengali	IE	CAsia	Large	36	4.371	10.50	0.7143	0.7154
Berber	AA	NAfrica	Medium	37	3.873	8.78	0.7468	0.8087
Burmese	ST	EAsia	Medium	70	7.143	6.00	0.9048	0.9195
Cantonese	ST	EAsia	Medium	85	3.857	9.10	0.6484	0.7739
Chickasaw	Other	NAmerica	Small	34	8.316	5.70	0.6667	0.6376
Dinka	NS	NAfrica	Small	48	4.000	13.70	0.5474	0.7030
English	IE	NEurope	Large	35	3.389	12.56	0.5575	0.6945
French	IE	SEurope	Large	33	3.176	12.00	0.5926	0.7205
Georgian	Other	WAsia	Medium	38	6.058	7.67	0.8116	0.7731
German	IE	NEurope	Large	53	4.147	10.90	0.6514	0.6972
Greek	IE	SEurope	Medium	28	4.165	12.78	0.5478	0.7046
Hausa	AA	SAfrica	Medium	48	3.904	13.83	0.5241	0.6094
Hebrew	AA	WAsia	Medium	35	5.910	8.09	0.8202	0.7400
Hindi	IE	CAsia	Large	45	3.766	15.50	0.6290	0.6252
Hungarian	Other	NEurope	Medium	39	4.310	10.00	0.6300	0.7418
Igbo	NC	SAfrica	Medium	50	3.358	13.25	0.5094	0.8044
Irish	IE	NEurope	Small	46	3.147	18.43	0.5969	0.7421
Japanese	Alt	Pacific	Large	26	5.045	9.78	0.6023	0.7145
Kabiye	NC	SAfrica	Medium	39	4.758	10.11	0.6923	0.7441
Korean	Alt	EAsia	Medium	37	6.350	8.57	0.7833	0.8536
Malay	Aus	Pacific	Large	24	6.167	9.75	0.6154	0.6485
Mandarin	ST	EAsia	Large	43	4.385	9.60	0.6875	0.7940
Mapudungun	Other	SAmerica	Small	28	4.800	8.33	0.4800	0.7644
Nara	NS	NAfrica	Small	45	4.315	9.82	0.5463	0.6555
Nepali	IE	CAsia	Medium	38	5.340	10.44	0.8085	0.7198
Persian	IE	WAsia	Medium	35	5.308	10.11	0.7143	0.7220
Portuguese	IE	SEurope	Large	45	3.878	12.25	0.6429	0.7747
Quichua	Other	SAmerica	Medium	26	6.589	8.18	0.7556	0.6037
Russian	IE	NEurope	Large	48	4.825	10.78	0.7320	0.7261
Sahaptin	Other	NAmerica	Small	46	6.579	7.13	0.6140	0.7923
Sandawe	Other	SAfrica	Small	74	5.716	7.44	0.7612	0.6993
Seri	Other	NAmerica	Small	26	3.777	14.27	0.4459	0.5142
Shiwilu	Amaz	SAmerica	Small	25	7.750	7.71	0.6759	0.5322
Spanish	IE	SEurope	Large	29	4.381	10.78	0.6186	0.7761
Tajik	IE	WAsia	Medium	28	5.477	12.57	0.7159	0.6559
Tamil	Other	CAsia	Medium	25	6.763	8.89	0.6750	0.5686
Tausug	Aus	Pacific	Small	20	5.018	9.50	0.4825	0.6019
Temne	NC	SAfrica	Medium	37	3.568	11.36	0.5440	0.6302
Thai	Other	EAsia	Medium	66	3.664	11.91	0.5649	0.6437
Trique	OM	NAmerica	Small	101	3.355	10.70	0.5794	0.7059
Turkish	Alt	WAsia	Medium	30	6.631	7.22	0.7538	0.8629
Vietnamese	Other	EAsia	Medium	110	2.855	16.71	0.5299	0.7622
Yine	Amaz	SAmerica	Small	21	9.164	6.10	0.7049	0.6426
Zapotec	OM	NAmerica	Medium	35	3.759	9.67	0.6552	0.8440

Glottometrics 35, 2016, 64-84

On Russian Adnominals

Sergey Andreev[1], Ioan-Iovitz Popescu, Gabriel Altmann

Abstract. The aim of the article is to show that the class of adnominals in Russian behaves regularly and abides by a strict rank-frequency distribution, a fact giving them the status of "legal" linguistic units. It will be shown that there is a left-right asymmetry in placing adnominals.

Keyords: Russian, adnominals, rank-frequency distribution, asymmetry

Adnominals or adnominal modificators can be considered – cum grano salis – a part of the set of noun valencies. They are not compulsory and may be placed in front of or behind the noun. They can consist of parts of compounds, free words, phrases or clauses. They express some properties of "things" expressed by nouns. They are rather components of style, hence one can expect that in different text types we find a different set of adnominals. The writers use them consciously but choose them intuittively, without caring what form they choose. A good example is rhythmic poetry in which the author must care for rhythm and restrict his choice. In the course of human life the use of adnominals changes automatically, hence it should be studied also in the language of children.

Inspite of freedom and change of choice one can conjecture that adnominals set up a class whose members behave regularly and any of their properties abides by a regularity. If it is so, then adnominals can be considered quite usual linguistic units. There is no list of individual cases but one may set up a list of classes of adnominals. This will be different in different languages and – as one can suppose – different for every researcher because linguistic schools want to show that they are "right".

Without having this claim we consider a class of possible adnominals that were necessary for analyzing great works of the Russian literature. The list and some Russian examples are presented in Table 1. These adnominal types are proposed in Russian grammar sources (Russkaja grammatika, 1980; Valgina, 2003: 37-45). Each modificator is abbreviated and has as a last letter either L (left) or R (right) in order to be able to study the asymmetry of positioning the same type of adnominals. It must be remarked that not all of them occurred in all texts.

Analyzing a text we obtain a vector of abbreviations. In order to exemplify it we show the vector obtained from the long poem by A. Pushkin, *Vadim* (1822) containing 203 adnominals. Their meaning is given in Table 1.

[GR,AR,AL,AL,GR,AR,AL,AL,PtL,AL,GR,DETL,DETL,DETL,PtL,PtL,GR,PtL,AR,DETL,
AR,AR,AR,DETL,AL,GR,AR,AL,AR,AR,AL,GR,AR,AL,AL,AL,GR,GR,PtL,GR,ApR,
PtR, AL,PtL,AL,GR,PtL,ApR,AL,ApR,RCR,DETL,AL, DETL,AL,ConL,AL,AL,PtR,GR,
PrL,AL,GL,AR,DETR,AR,InstrR,PtR,PrR,AL,InstrR,PtL,InstrR,AL,PtL, AL,PtL,AL,AR,
AR,AL,DETR,AL,GL,GL,AL,AR,GR,AL,CR,AR,GL,AL,GR,AL,AL,GR,PtR,GR,GR, AR,
AR,AL,AR,RCR,AL,RCR,GR,AL,CR,AR,CR,RCR, AL,AR,GR, AL,AR,GR,AL,GR, DETL,
GR,AL,DETL,PtL,PtR,AL,GR,AR,PtL,AL,AR,AL,GL,AR,AL,AR,AR,AL,AL,GR, CNL,
DETL,AL,DETL,AL,AR,AR,AL,AL,DETR,DETL,PrL,AL,AL,GR,GR,AL,AR,AL,GR,AL,

[1] University of Smolensk. Email: smol.an@mail.ru

AL,AL,AL,AR,PrR,AL,AR,AL,AL,AL,AL,AL,AR,AR,DETL,AL,AR,PrL,AR,AL,AL, DETR,AL,GR,ApR,AR,ApR,AR,AR,AR, AL,PrR,GR,AL,AL,AL,PtR,AL,AL,AL]

The vectors of all the other texts are presented in the Appendix. Now, having data of this kind, a number of problems arise. (1) Is there some rank-frequency distributions of symbols or are they represented in equal proportions? (2) Is there a left-right symmetry of adnominals or are they placed equally? Some languages e.g. have the tendency to place adjectives on the left hand side but in poetry they stay frequently on the right hand side. (3) How is the distribution of distances between equal symbols? Is there some regularity or is it random? (4) Since we have to do with symbols, it is possible to construct Köhlerian (Köhler, Altmann 2014: 21-25; Köhler, Naumann 2016 Köhler, Naumann 2012) qualitative motifs and study their properties. (5) The symbols can be weighted, e.g. by their frequency; if one replaces the symbols by their frequency, one can obtain a quantitative vector which can be transformed in quantitative motifs.

Table 1
Adnominals in Russian

Type	**Head initial (modificator right)**	**Desig-nation**	**Head final (modificator left)**	**Desig-nation**
Componential noun (CN)	–	–	**khleb**zavod (**bread** factory)	**CNL**
Componential adjective (CA)	roza-**krasavitsa** (rose-**beauty**)	**CAR**	**krasavitsa**-roza (**beauty**-rose)	**CAL**
Adjective (A)	**krasnaja** roza (**red** rose)	**AL**	roza **krasnaja** (rose **red**)	**AR**
Participle (Pt)	**tantsujushchaja** devushka (**dancing** girl), **tsvetushchaja** roza (**blooming** rose), **pokrashennij** dom (**painted** house)	**PtL**	devushka **tantsujushchaja** (girl **dancing**), roza **tsvetushchaja** (rose **blooming**), dom **pokrashennij** (house **painted**)	**PtR**
Adverb (Av)	**nalevo** povorot (**to the left** turn), **sovsem** rebjonok (**absolutely** a child)	**AvL**	povorot **nalevo** (turn **to the left**), rebjonok **sovsem** (a child **absolutely**)	**AvR**
Infinitive (I)	**rabotat'** soglasije (**to work** agreement), **spat'** zhelanije (**to sleep** desire), **borotsja** gotovnost' (**to struggle** readiness)	**IL**	soglasije **rabotat'** (agreement **to work**), zhelanije **spat'** (desire **to sleep**), gotovnost' **borotsja** (readiness **to struggle**)	**IR**
Concretization (extension), nominative case	**Volga** reka **(Volga** the river)	**ConL**	reka **Volga** (the river **Volga**)	**ConR**
Noun, genitive case (no	**moloka** stakan ([of] milk glass), **raboti** nachalo ([of] work beginning)	**GL**	stakan **moloka** (glass [of] milk), nachalo **raboti** (beginning [of] work)	**GR**

preposition)				
Noun, dative case (*no preposition*)	**detjam** podarki ([to] **children** gifts), **druzjam** pomoshch ([to] **friends** assistance)	**DL**	podarki **detjam** (gifts [to] **children**), pomoshch **druzjam** (assistance [to] **friends**)	**DR**
Noun, instrumental case (*no preposition*)	**rukoj** tolchok ([with] a **hand** a push), **kartinoj** voskhishchenije ([of] **painting** admiration)	**Instr L**	tolchok **rukoj** (a push [with] a **hand**), voskhishchenije **kartinoj** (admiration of **painting**)	**Instr R**
Noun, *prepositional phrase*	**dlja devushki** podarok (**for the girl** a present), **k zhizni** ljubov (**for life** love), **o pomoshchi** pros'ba (**for help** plea),	**PrL**	podarok **dlja devushki** (a present **for the girl**), ljubov **k zhizni** (love **for life**), pros'ba **o pomoshchi** (plea **for help**),	**PrR**
Determiner	**moja** sestra (**my** sister), **takaja** kartina (**such** picture)	**DET L**	sestra **moja** (sister **my**), kartina **takaja** (picture **such**)	**DET R**
Conjunctional phrase	**Kak spetsialist** ja polnos-tju podderzhivaju etu ideju. (**As a specialist** I abso-lutely support this idea.)	**ConL**	Ja, **kak spetsialist**, polnostju podderzhivaju etu ideju. (I, **as a specialist**, absolutely support this idea.)	**ConR**
Apposition or Anteposition	**Znamenitaja telezvezda**, eta aktrisa nichego ne znajet o politike. (**A famous TV-star**, this actress does not know anything about politics.)	**ApL**	Eta aktrisa, **znamenitaja telezvezda**, nichego ne znajet o politike. (This actress, **a famous TV-star**, does not know anything about politics.)	**ApR**
Conjunctional clause	–	**–**	Znanije, **chto ona krasiva**, radovalo devushku. (Knowledge **that she is pretty** delighted the girl)	**CCR**
Relative clause	–	**–**	Film, **kotorij mi smotreli**, byl interesnym (The film **which we saw** was interesting)	**RCR**

Here we shall study the rank-frequency distribution of adnominals. Counting the numbers of individual adnominals in individual texts one may see that the Zipf-Mandelbrot distribution developed to this aim yields an excellent fit. There are merely two exceptions shown below.

Table 2
Pushkin: Vadim, 1822

Rank	Frequency	Zipf-Mandelbrot
1	72	74.06
2	40	38.33
3	28	22.22

Table 3
Pushkin: Graf Nulin (Count Nulin), 1825

Rank	Frequency	Zipf-Mandelbrot
1	70	74.01
2	44	36.29
3	18	20.73

4	12	13.95
5	6	9.30
6	5	6.49
7	5	4.70
8	4	3.51
9	3	2.68
10	3	2.09
11	1	1.66
a = 3.1734, b = 3.3346, n = 11, DF = 7, X^2 = 4.2075, P = 0.76		

4	13	13.07
5	8	8.83
6	6	6.28
7	4	4.65
8	4	3.55
9	4	2.77
10	2	2.22
11	2	1.80
12	1	1.49
13	1	1.25
14	1	1.05
a = 2.6218, b = 2.2012, n = 14, DF = 10, X^2 = 3.25, P = 0.97		

Table 4
Pushkin: Mednij vsadnik (The Bronze Horseman), 1833

Rank	Frequency	Zipf-Mandelbrot
1	142	147.54
2	78	73.29
3	41	41.95
4	34	26.38
5	14	17.73
6	11	12.53
7	10	9.20
8	7	6.97
9	5	5.42
10	4	4.30
11	4	3.47
12	2	2.85
13	2	2.37
a = 2.7666, b = 2.4757, n = 13, DF = 9, X^2 = 4.22, P = 0.90		

Table 5
Derzhavin: Felitsa (Ode to Felica), 1782

Rank	Frequency	Zipf-Mandelbrot
1	27	27.00
2	26	31.88
3	23	22.40
4	21	16.03
5	15	11.77
6	9	8.87
7	6	6.83
8	5	5.36
9	5	4.27
10	4	3.45
11	3	2.82
12	1	2.34
13	1	1.95
14	1	1.64
15	1	1.39
a = 0.1091, b = 0.4247, n = 15, DF = 10 , X^2 = 5.47, P = 0.86		

Table 6
Derzhavin: Vodopad (Waterfall), 1791-1794

Rank	Frequency	Zipf-Mandelbrot
1	110	116.39
2	76	82.80
3	62	59.91
4	52	44.01
5	28	32.79
6	26	24.74
7	25	18.88
8	18	14.57

Table 7
Karamzin: Bednaja Liza (Poor Liza), 1792. Paragraphs 1-8

Rank	Frequency	Zipf-Mandelbrot
1	93	90.07
2	40	45.15
3	28	26.63
4	15	17.39
5	13	12.16
6	11	8.94
7	9	6.82
8	7	5.36

9	18	11.35
10	10	8.92
11	5	7.08
12	4	5.66
13	3	4.56
14	2	3.70
15	2	3.02
16	2	2.48
17	1	2.05
18	1	1.70
19	1	1.42
a = 6.5388, b = 17.7091, n = 19 , DF = 15 , X^2 = 13.8025, P = 0.5406		

9	4	4.31
10	3	3.54
11	3	2.95
12	2	2.49
13	2	2.13
14	1	1.84
15	1	1.60
16	1	1.41
17	1	1.25
a = 2,2286, b = 1,7481, n = 17, DF = 13, X^2 = 3,7913 , P = 0.9932		

Table 8
Pushkin: Vystrel (The Shot), 1830. Part 1

Rank	Frequency	Zipf-Mandelbrot
1	93	108.78
2	73	59.50
3	36	37.48
4	30	25.77
5	18	18.80
6	17	14.33
7	16	11.28
8	14	9.11
9	12	7.52
10	5	6.30
11	4	5.36
12	3	4.62
13	2	4.02
14	2	3.53
15	1	3.13
16	1	2.79
17	1	2.50
18	1	2.26
19	1	2.05
20	1	1.86
a = 1.9867, b = 1.8181, n = 20, DF = 16, X^2 = 21.9132 , P = 0.1469		

Table 9
Lermontov: Demon (The Demon), 1839. Part 1

Rank	Frequency	Zipf-Mandelbrot
1	55	53.98
2	30	30.56
3	23	19.76
4	10	13.88
5	9	10.31
6	7	7.98
7	7	6.36
8	7	5.20
9	7	4.34
10	4	3.67
11	4	3.16
12	3	2.74
13	2	2.40
14	1	2.13
15	1	1.90
16	1	1.70
17	1	1.54
18	1	1.39
a = 1.8768, b = 1.8247, n = 18, DF = 14, X^2 = 6.2019 , P = 0.9611		

Table 10
Lermontov: Fatalist (The Fatalist), 1837-1840

Rank	Frequency	Zipf-Mandelbrot
1	112	121.54
2	74	65.99
3	43	41.10
4	25	27.92

Table 11
Turgenev: Dvorjanskoe gnezdo (Home of the Gentry), 1856-1858. Part 1

Rank	Frequency	Zipf-Mandelbrot
1	84	80.02
2	38	41.48
3	24	25.25
4	18	16.94

5	22	20.13
6	18	15.17
7	17	11.81
8	15	9.45
9	8	7.72
10	5	6.42
11	4	5.42
12	4	4.63
13	2	4.00
14	1	3.49
15	1	3.07
16	1	2.72
17	1	2.42
a = 2.1193, b = 1.9942, n = 17, DF = 13, X^2 = 15.2228, P = 0.2937		

5	12	12.13
6	7	9.11
7	7	7.08
8	7	5.66
9	6	4.63
10	6	3.85
11	4	3.25
12	2	2.78
13	2	2.41
14	2	2.10
15	1	1.85
16	1	1.65
17	1	1.47
18	1	1.32
a = 2.0447, b = 1.6390, n = 18, DF = 14, X^2 = 4.3757 , P = 0.9927		

Table 12
Nekrasov: Zheleznaja doroga (The Railroad), 1864

Rank	Frequency	Zipf-Mandelbrot
1	40	42.39
2	29	24.24
3	16	15.70
4	12	11.01
5	6	8.16
6	4	6.29
7	4	5.00
8	4	4.07
9	4	3.38
10	3	2.85
11	3	2.44
12	3	2.11
13	2	1.84
14	2	1.63
15	1	1.44
16	1	1.29
17	1	1.16
a = 1.9543, b = 2.0209, n = 17, DF = 13, X^2 = 3.7158 , P = 0.9939		

Table 13
Tolstoy: Vojna i mir (War and Peace), 1863-1869. Chapter 2

Rank	Frequency	Zipf-Mandelbrot
1	81	71.92
2	34	45.49
3	29	30.23
4	21	20.90
5	15	14.93
6	13	10.96
7	11	8.24
8	11	6.32
9	4	4.93
10	2	3.90
11	2	3.13
12	2	2.55
13	2	2.09
14	1	1.73
15	1	1.45
16	1	1.22
a = 3.7988 , b = 6.8033, n = 16, DF = 12 , X^2 = 11.0012, P = 0.5288		

Table 14
Tolstoy: Anna Karenina, 1873-1877. Chapters 1-2

Rank	Frequency	Zipf-Mandelbrot
1	83	92.76
2	54	50.19
3	36	30.29
4	22	19.72

Table 15
Chekhov: Zhenshchina bez predrassudkov (The Woman without Prejudices), 1883

Rank	Frequency	Zipf-Mandelbrot
1	30	34.95
2	29	19.52
3	9	12.34
4	7	8.45

5	14	13.59
6	10	9.77
7	7	7.27
8	6	5.56
9	4	4.36
10	3	3.48
11	2	2.82
12	2	2.32
13	2	1.94
14	1	1.63
15	1	1.39
16	1	1.19
17	1	1.03
18	1	0. 90
19	1	0.79
a = 2.8502, b = 3.1579, n = 19, DF = 14 , X^2 = 3.5466 , P = 0.9976		

5	6	6.13
6	4	4.63
7	4	3.61
8	3	2.89
9	2	2.37
10	2	1.97
11	2	1.66
12	2	1.42
13	1	1.22
14	1	1.07
15	1	0.94
16	1	0.83
a = 2.1634, b = 2.2355, n = 16, DF = 11, X^2 = 7.0374 , P = 0.7960		

Table 16
Chekhov: Dama s sobachkoj (The Lady With The Dog), 1899. Chapers 1-2

Rank	Frequency	Zipf-Mandelbrot
1	107	94.85
2	39	56.12
3	30	35.55
4	27	23.74
5	23	16.52
6	20	11.90
7	6	8.81
8	5	6.68
9	5	5.17
10	5	4.07
11	3	3.25
12	2	2.63
13	1	2.16
14	1	1.79
15	1	1.50
16	1	1.26
a = 3.5188, b = 5.2189, n =16 , DF = 12, X^2 = 19.0603 , P = 0.0871		

Table 17
Kuprin: Chari (The Spell), 1897

Rank	Frequency	Zipf-Mandelbrot
1	92	84.84
2	35	46.23
3	31	28.33
4	20	18.79
5	15	13.20
6	9	9.68
7	9	7.34
8	8	5.72
9	6	4.56
10	1	3.70
11	1	3.05
12	1	2.55
a = 2.5430, b = 2.7085, n = 12, DF = 8, X^2 = 9.9898 , P = 0.2657		

Table 18
Kuprin: Junkera (The Junkers), 1928-1932. Chapter 2

Rank	Frequency	Zipf-Mandelbrot
1	125	111.73
2	39	49.71

Table 19
Bunin: Kavkaz (The Caucasus), 1937

Rank	Frequency	Zipf-Mandelbrot
1	100	92.46
2	31	40.10
3	18	21.86

3	22	27.77
4	13	17.62
5	13	12.13
6	12	8.83
7	9	6.71
8	7	5.26
9	6	4.23
10	4	3.47
11	3	2.90
12	2	2.46
13	2	2.11
14	1	1.83
15	1	1.60
16	1	1.41
17	1	1.25
a = 2.0919, b = 1.1152, n = 17 , DF = 13 , X^2 = 10.5270 , P = 0.6504		

4	16	13.57
5	13	9.16
6	6	6.56
7	5	4.90
8	5	3.79
9	3	3.01
10	2	2.44
11	2	2.02
12	1	1.69
13	1	1.44
a = 2.2369, b = 1.2088, n = 13, DF = 9 , X^2 = 6.3379 , P = 0.7057		

Table 20
Bunin: Stepa, 1938

Rank	Frequency	Zipf-Mandelbrot
1	103	91.96
2	33	50.01
3	29	29.41
4	26	18.38
5	18	12.05
6	4	8.21
7	4	5.78
8	4	4.19
9	3	3.10
10	2	2.35
11	1	1.81
12	1	1.41
13	1	1.12
14	1	0.90
15	1	0.73
16	1	0.60
a = 4.1363, b = 5.3031 , n = 16, DF = 19, X^2 = 16.8174 , P = 0.0785		

Table 21
Kuprin: Domik (Little House), 1929

Rank	Frequency	R-T.Zipf
1	125	113.81
2	21	37.90
3	19	19.92
4	16	12.62
5	10	8.86
6	6	6.63
7	5	5.20
8	5	4.20
9	4	3.49
10	4	2.95
11	3	2.54
12	3	2.21
13	2	1.95
14	1	1.73
a = 1.5863, R = 14, DF = 11, X^2 = 11.0760 , P = 0.4369		

In Table 21, the Zipf-Mandelbrot was not adequate, but the right-truncated Zipf distribution was satisfactory. Its formula is $P_x = x^{-a}/F(R)$, x = 1,2,…,R, where F(R) is simply the sum of x^{-a}. It is slightly simpler than Zipf-Mandelbrot, i.e. a simple power distribution.

Table 22
Bunin: Antonovskie jabloki (Apple Fragrance), 1900. Parts I-II

Rank	Frequency	Hyperpascal
1	192	185.47
2	47	51.18
3	36	37.36
4	23	28.35
5	21	21.82
6	20	16.91
7	13	13.16
8	12	10.27
9	12	8.04
10	12	6.30
11	4	4.94
12	3	3.88
13	2	3.05
14	1	2.39
15	1	1.88
16	1	1.48
17	1	1.17
18	1	4.35
k = 0.0477, m = 0.1369, q = 0.7921		
DF = 14, X^2 = 14.3679, P = 0.4227		

There are two esxceptions, namely *Domik* by Kuprin and *Antonovskie jabloki* by Bunin. Peculiar enough, in the former case a simpler distribution is adequate, in the latter, applying the continuous Zipf-Mandelbrot function the result of fitting is excellent, the discrete distribution is not. One must take another distribution . Here we apply the Hyperpascal distribution defined as

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0, \quad x = 0,1,2,...$$

which is, in our case, replaced one step to the right. This is, of course, only a preliminary proposal. We expect several other exceptions when other text types will be analyzed.

As can be seen, the ranking has an ordered character. In some places one of the classes acquires an extreme value but 19 out of 21 texts of Russian literature behave according to the same distribution. Needless to say, the hypothesis should be tested in other text types, and especially those creatred in 21th century. There are surely developments in the language of children, too.

Now, since the trend is evident, one may conjecture that the parameters of the Zipf-Mandelbrot distribution are also linked in some way. Since n is merely the number of classes, we may study the relation <a,b> as shown in Table 23

Table 23
Relation between the parameters *a* and *b* of the Zipf-Mandelbrot distribution

Author	**Title**	**year**	**a**	**b**
Derzhavin	Felitsa	1782	0.1091	0.4247
Lermontov	Demon	1839	1.8768	1.8247
Nekrasov	Zheleznaja doroga	1864	1.9543	2.0209
Pushkin	Vystrel	1830	1.9867	1.8181
Turgenev	Dvoryanskoe gnezdo	1856-1858	2.0447	1.639
Kuprin	Junkera	1928-1932	2.0919	1.1152
Lermontov	Fatalist	1837-1840	2.1193	1.9942
Chekhov	Zhenshchina bez predrassudkov	1883	2.1634	2.2355
Karamzin	Bednaja Liza	1792	2.2286	1.7481
Bunin	Kavkaz	1937	2.2369	1.2088
Kuprin	Chari	1897	2.5430	2.7085
Pushkin	Graf Nulin	1825	2.6218	2.2012
Pushkin	Mednij vsadnik	1833	2.7666	2.4757
Tolstoy	Anna Karenina	1873-1877	2.8502	3.1579
Pushkin	Vadim	1822	3.1734	3.3346
Chekhov	Dama s sobachkoy	1899	3.5188	6.2189
Tolstoy	Vojna i mir	1863-1869	3.7988	6.8033
Bunin	Stepa	1938	4.1363	5.3031
Derzhavin	Vodopad	1791-1794	6.5388	17.7091

As can be seen in Figure 1, *b* increases according to the increase of *a*. There are only slight deviations. The course of the function can be captured by a simple exponential or a power function. Ordering the table according to increasing *a*, we obtain the results in Table 24.

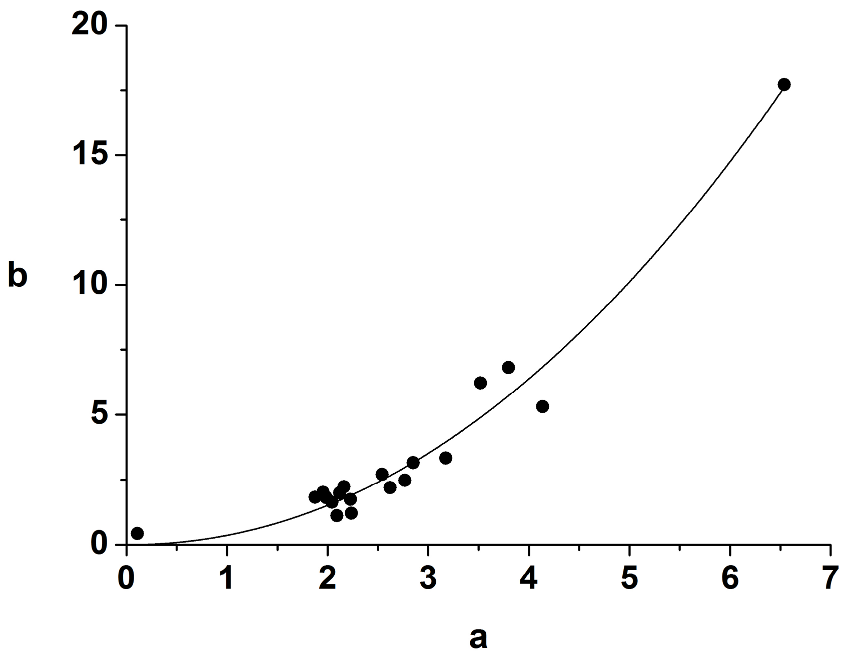

Figure 1. The link between parameter *a* and *b*, <a,b>

We adhere here to the simple power function and obtain $b = Ka^M$, here $b = 0.3626a^M$ with R^2 = 0.9695. Only two texts were omitted in which one should search for boundary conditions. Nevertheless, there is a very regular link. If the relationship can be shown also for other text types and languages, we are on the trace of background law.

Table 24
Relation between parameters *a* and *b*

Parameter a	**Parameter b**	**Power function**
0.1091	0.4247	0.0037
1.8768	1.8247	1.3331
1.9543	2.0209	1.4494
1.9867	1.8181	1.4995
2.0447	1.6390	1.5915
2.0919	1.1152	1.6684
2.1193	1.9942	1.7139
2.1634	2.2355	1.7885
2.2286	1.7481	1.9017
2.2369	1.2088	1.9164
2.5430	2.7085	2.4985
2.6218	2.2012	2.6612
2.7666	2.4757	2.9741
2.8502	3.1579	3.1630
3.1734	3.3346	3.9497
3.5188	6.2189	4.8905
3.7988	6.8033	5.7295
4.1363	5.3031	6.8322
6.5388	17.7091	17.6137
K = 0.3626, M = 2.0680, R^2 = 0.9695		

Left-right asymmetry

In order to see how one placed the adnominals in Russian, it is possible to perfom a test for the equality of both positions (L and R). Adding all adnominals ...L and those of ...R, we obtain the full sample. Now to test whether the two proportions are equal, one can apply the binomial test; but if the number of cases is too large, one rather uses an asymptotic test yielding almost the same result. Let L be the number of adnominals ending with L, and R those ending with R, we set up the chi-square criterion and test

$$X^2 = \frac{(L-R)^2}{L+R}$$

Representing the chi-square with 1 degree of freedom whose root is identical with the quantile of the normal distribution. The critical value is 3.84. In this way we obtain the following classes:

SL = significantly left (L > R, X^2 > 3.84, u > 1.96)
L = left (L > R, X^2 < 3.84, u < 1.96)
NE = neutral (X^2 < 0.5, – 0.71 < u < 0.71)
R = right (L < R, X^2 < 3.84, u < −1.96)
SR = significantly right (L < R, X^2 > 3.84, o < − 1.96)

For the sake of simplicity we show the test using the first text, namely Vadim by A. Pushkin where one finds L = 107, R = 96, hence

$$X^2 = (107 – 96)^2/(107+96) = 0.5961$$

which is not significant, and since L > R, the class obtained is L (non-significant left trend). For the other texts we obtain the results presented in Table 25.

Table 25
Asymmetry in positioning the adnominals

	Text	Left (L)	Right (R)	X^2	Class
Pushkin	Vadim	107	96	0.5961	L
Pushkin	Graf Nulin	119	114	0.1073	NE
Puskhin	Mednij Vsadnik	263	178	16.3832	SL
Pushkin	Vystrel	185	146	4.5952	SL
Derzhavin	Felitsa	64	86	3.2267	R
Derzhavin	Vodopad	203	242	3.4180	R
Karamzin	Bednaja Liza	137	97	6.8376	SL
Turgenev	Dvorjanskoe gnezdo	132	91	7.5381	SL
Lermontov	Demon	82	91	0.4682	NE
Lermontov	Fatalist	205	148	9,2040	SL
Nekrasov	Zheleznaja doroga	71	64	0.3630	NE
Tolstoj	Vojna i mir	134	96	6.2792	SL
Chekhov	Zhenshchina bez predrassudkov	65	39	6.5000	SL
Chekhov	Dama s sobachkoj	152	124	2.8406	L
Kuprin	Chari	148	80	20.2807	SL
Kuprin	Domik	156	68	34.5714	SL
Kuprin	Junkera	155	106	9.1992	SL
Bunin	Antonovskie jabloki	240	162	15.1343	SL
Bunin	Kavkaz	150	53	46.3498	SL
Bunin	Stepa	164	68	39.7241	SL

As can be seen, 14 out of 20 cases have a left tendency concerning adnominals, a phenomenon that is in accordance with Russian grammar. Only the works by Derzhavin display a right hand tendency. This trend is caused either by stylistic or by text type boundary conditions. The number of analyzed texts is not sufficient to make conjectures but it would be helpful for literary scientists.

References

Shvedova, N.U. (chief ed.) (1980). *Russkaja grammatika.* [Russian Grammar] Toma [Volumes] 1-2. Moskva: Nauka.

Köhler, R., Altmann, G. (2014). *Problems in quantitative linguistics*. Vol. 4. Lüdenscheid: RAM-Verlag.

Köhler, R., Naumann, S. (2012). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: *Proceedings of COLING 2012* (Mumbai, December 2012). Technical Papers. Mumbai, 2012. P. 263–278.

Köhler, R., Naumann, S. (2016) Syntactic text characterisation using linguistic S-motifs. *Glottometrics 34, 1-8.*

Valgina, N.S. (2003). *Sovremennij russkij yazik: Syntaksis* [Modern Russian: Syntax]. Moskva: Vysshaja shkola.

Appendix

A.S. Pushkin. *Vadim* (long poem), 1822, words 846, attributes 203. From: A.S. Pushkin. Polnoje sobranije sochinenij v desyati tomakh (Complete collection of works in ten volumes). Tom 4 [Volume 4]. Izdatelstvo Akademii nauk SSSR (Publishing house of the Academy of Sciences of the USSR). Moskva (Moscow), 1963.

[GR,AR,AL,AL,GR,AR,AL,AL,PtL,AL,GR,DETL,DETL,DETL,PtL,PtL,GR,PtL,AR,DETL, AR,AR,AR,DETL,AL,GR,AR,AL,AR,AR,AL,GR,AR,AL,AL,AL,GR,GR,PtL,GR,ApR,PtR, AL,PtL,AL,GR,PtL,ApR,AL,ApR,RCR,DETL,AL,DETL,AL,ConL,AL,AL,PtR,GR,PrL,AL, GL,AR,DETR,AR,InstrR,PtR,PrR,AL,InstrR, PtL, InstrR,AL,PtL,AL,PtL,AL,AR,AR,AL, DETR,AL,GL,GL,AL,AR,GR,AL,CR,AR,GL,AL,GR,AL,AL,GR,PtR,GR,GR,AR,AR,AL, AR,RCR,AL,RCR,GR,AL,CR,AR,CR,RCR, AL,AR,GR, AL,AR,GR,AL,GR,DETL,GR, AL,DETL,PtL,PtR,AL,GR,AR,PtL,AL,AR, AL,GL,AR,AL,AR,AR,AL,AL,GR,CNL,DETL, AL,DETL,AL,AR,AR,AL,AL,DETR,DETL,PrL,AL,AL,GR,GR,AL,AR,AL,GR,AL,AL,AL, AL,AR,PrR,AL,AR,AL,AL,AL,AL,AL,AR,AR,DETL,AL,AR,PrL,AR,AL,AL,DETR,AL, GR,ApR,AR,ApR,AR,AR,AR,AL,PrR,GR,AL,AL,AL, PtR, AL,AL,AL];

Pushkin: *Graf Nulin* (*Count Nulin*) (long poem), 1825, words 1567, attributes 231. From: A.S. Pushkin. Polnoje sobranije sochinenij v desyati tomakh (Complete collection of works in ten volumes). Tom 4 [Volume 4]. Izdatelstvo Akademii nauk SSSR (Publishing house of the Academy of Sciences of the USSR). Moskva (Moscow), 1963.

[AL,PrR,DETL,AL,AL,GR,PtR,PrR,AL,PrR,PrL,PrL,AL,PrR,AL,PrL,AL,PrL,AR, AL,AL, GR,AL,AL,AL,GR,DR,AL,DETL,AL,AR,PrR,GR,AvR,GL,AR,DETR, DETL,AL,AL,AL, CR,AL,GR,GR,GR,GR,AR,AR,AvR,AR,AR,AR,AR, AR,PrR,AL,PrL,PtL,GR, AL,AL, AL,AR,AR,DETR,AR,AR,ApR,GR,AR,PtR,CR,AL,AL, DETL,PtL,AR,AvL,PtL,AL,AR, CNL,AR,CR,AL,RCR,GR,DETL,PtL, ConR,AL,AL,AL,ConR,AL,GR,AL,GR,AL,GR,AL, GR,GR,GR,GR, DETL,AL,DETL,AR,PrR,AR,DETL,PrL,DETL,DETL,AL,AR,AR,DETR, GL,PtR,AR,AL,AL, PtL,AL,DETR,DETL,PtL,DETR, CR,AL,AL,PrR, PtL,AL,AR,AR,AL, AL,AL,DETL,AR,AL,AL,AL,AR,GR,AR, AR, DETL, GR,GR,AR,AR,ApR,AL, PtL,AL, AL,AL,GR,AR,AR,AL,AL,GR,AL,GR, PtL,AR,PtR,AvR,AL,AR,GR,DETL,DETL, AL,GR, DETL,AR,DETL,AR,PtL,AR,AR,DETR,CR,DETR,AR,AR,AL,DETL,DETL,AL, AL,AL, DETR,AL,GR,AL,DETR,PtL,AL,AL,AR,AL,AR,PtL,PtL,AvR, DETL,DETR,CR, AL,AL, DETL,AL,ApR,AR, DETL,AL,DETR,DETL,DETL,ApR,GR,DETL,DL,AL,DETR]

Pushkin: ***Mednij vsadnik*** (*The Bronze Horseman)* (long poem), 1833, words 2017, attributes 411. From: A.S. Pushkin. Polnoje sobranije sochinenij v desyati tomakh (Complete collection of works in ten volumes). Tom 4 [Volume 4]. Izdatelstvo Akademii nauk SSSR (Publishing house of the Academy of Sciences of the USSR). Moskva (Moscow), 1963.

[AL,GR,GR,AR,AR,AL,AL,AL,AL,GR,PtR,DR,PrR,PtL,GR,AL,AR,AL,DETL,AL,ApR,AL ,GL,GR,GR,AL,AL,GR,AL,AL,AL,DETL,AL,AvL,AL,AR,GR,GR,GR, DETL, PrR,AL, CAL,AL,DETL,AL,AL,AL,AL,GL,ApR,DETL,AL,AL,GL,AL,AL, DETL, DETL,GL,AR, DETL,AL,GL,AL,AR,DETR,AL,PtL, AL,GR,AL,AL,AR,AL, AL,DETR,GL,DETR,AR, AL,GR,AvR,AR,AL,AR,GR,GR,AR,AL,GR,GR,AR,AL,AL,AL, GR,AL,GL,GL,AL, DETL, AvL,PtL,DETL,GR,AR,GR,DETR,AR,AvL,PtL, PrR,AL,DETL,GL,AL,AL,PrR, DETL,AL, AL,AR,CR, AL,AR,DETR,AR,AL,AL,GR,AL,AL,DETR,DETL,AL,DETL, PtL,AL,AL, DETL,GR,AR,DETL,AR,DETL, AR,DETL,DETL,PrL,DETL,AR,DETR,PtL, AL,DETL, PtL, PtL,AL,AR, DETL,AL,GR,AR,RCR,AR,AR,AR,AR,IR,AR,AR,AR,DETR, InstrR, DETL,AL,AL,GR,AL,AL,DETL,AL,DETL,InstrL,GL,GL,GL,PtL,GR, PrR,PtL, AR,AR, ConR,AL,ConR,AL, ConR, AL,GR,AL,GR,AL,GR,InstrL,PtL,PrR,PtL,AL,DETL, AL,AL, AL,AL,AL, AL, AL,AL,AR,AR,AR,AL,AL,DETL,PtL,PtL,AR,CL,AR,PtL,PtL,AL, AR,AR, AvR, AL, AvL,AL,AR,AL,PtL,AR, AR,ConL,PtL,AvL,PrL,DETL,PtR,AL,DETL, DETL, DETL,DETL,ConR,AR,ConR,ConR,AL, PtL,PtL,AL,PrR,AL,DETL, DETL,PrR,AL, DETR, PtL,PtL,DETR,PtL,GR,AL,DETR,PrL,PtL,ConR,AR,AR, AL,AL,DETL,AL,AL,AR,AR,AR, AR,AR,PtR,AR,AL,ConR,AL,AL,RCR,DETL,AL,AL,AL,AR,GL,AL,AL,AL,GR, AR,AR, AR,DETL,AR,AL,DETL,AL,AR,PtL,DETL,AR, PtR,InstrR,AL,AL,GR,AL,AL,DETL, DETL,AL,AL,AL,GR,GR,DETL,AL,AL,DETL,DETL,AL,AL,DETL,AR,DETL,PrL,PtL, AR,AL,AL,PtR,AL,GR,DETL,AL,GR,AR,AL,GR,AL,AL,AL,ConR, DL,PtL,GR,AR, AR, PrR,GR,AL,PtL,ConL,AR,AL,PtL, PtL,AL,AR, RCR,AL,RCR, AR,AL,AL,DETL,AL, AL, GR,DETL,PrR,AR,GR,AR,AR,GR,GR, DETR,AL,AR,AL,AR,ConL,AR,AL,AR,GL, CAL, AL,PtL,AR,AR,CR,CAvL,PtL,DETL, AR,CR,AL,DETL,DETL,DETL, DETR,DETL, PtR, PtL,AL,InstrL,PrR,AR,AL, DETR,AL,AL,AL, ConR,PtL,AR,PtR,DETR,AL,DETR]

Pushkin: ***Vystrel*** (*The Shot*) (Novella) in *Povesti Belkina (The Belkin Tales*), 1830, Part 1, words 2044, attributes 331. From: A.S. Pushkin. Polnoje sobranije sochinenij v desyati tomakh (Complete collection of works in ten volumes). Tom 4 [Volume 4]. Izdatelstvo Akademii nauk SSSR (Publishing house of the Academy of Sciences of the USSR). Moskva (Moscow), 1963.

[CR,AL,GR,AR,AvL,AL,PrR,AL,PrR,AvL,AL,AL,AL,PrR,DETL,DETL,AR,DETL,AL,AL, AL,AL,AL,DETL,DETL,DETL,AL,PtR,AL,RCR,PtL,AL,AL,DETL,PrR,DETL,GR,DETL, PtR,AL,DETL,DETL,AR,GL,PtR,AL,DETR,PrR,DETL,GR,PrR,ConR,AR,AL,GR,AL,AL, GR,RCR,RCR,PrR,DETL,DETL,PrR,RCR,DETL,PrL,DETL,DETL,AL,AL,GR,AR,RCR, AL,DELT,PtL,DETL,GR,GR,IR,AL,PtR,PrR,AL,DETL,CCR,PtR,GR,GR,GR,GR,PrR,AL, RCR,PtL,AL,PrR,AL,AR,CR,AL,AL,PtL,AL,DETL,DETL,DETR,PtR,AL,PrR,AL,GR,GR, AL,RCR,AL,GR,GR,AR,GR,DETL,AL,RCR,RCR,AL,DETL,GR,AR,AL,DETL,AL,AL,AL, AL,CCR,DETL,DETL,AL,DETL,IR,DETL,DETL,AL,AL,DETR,AL,GR,PrR,AL,DR,GR, GR,PrR,AL,GR,AL,DETL,InstrR,DETR,AR,PtR,DETL,PrR,RCR,AL,GR,DETR,PtR,DETL, AL,DETL,AL,DETL,DETL,PtL,DETL,DETL,AL,AL,PtL,DETR,AL,AL,AL,AvL,GR,DETL ,DETL,RCR,DETL,AL,GR,AL,PtL,AL,PtR,AL,GR,AL,DETL,AL,DETL,DETL,DETL,AL, PrR,CR,IR,DETR,DETL,DETL,DETL,DETL,DETL,IR,DETR,DETR,DETL,GR,AL,AL,PrR ,PrR,PrR,CL,AL,DETR,DETL,AL,PrR,AL,PtR,DETL,PtL,AL,AvL,PtR,AL,ConR,DETL,AL ,AL,AL,GR,AR,AR,AR,AL,RCR,RCR,PtL,DETL,DETL,DETR,PrR,PrR,GR,AL,DETR, RCR,AR,AR,DETR,RCR,AR,AL,PrR,GR,DETL,GR,DETL,GR,PtR,PrR,PrR,DETL,AL, DETL,PtL,DETL,PtL,DETL,AL,PrR,PrR,PtR,PtR,AR,GR,AR,GR,IR,AL,DETL,AL,AL,GR,

DETR,DETL,AL,GR,AL,RCR,DETL,AL,DETR,DETL,DETL,DETL,AL,RCR,DETR,PtL, DETL,ApR,PrR,AL,AL,AL,AL,AL,DETL,DETL,DETL,DETL,AL,AL,RCR,PrR,DETL, PrR]

Derzhavin: ***Felitsa*** *(Ode to Felica)* (long poem), 1782, words 1151, attributes 150. From: G.R. Derzhavin. Stikhotvorenija. Biblioteka poeta. Bol'shaja serija. (Poems. Poet's library. Big series). Leningrad, Sovetskij pisatel', 1957.
[AL,AL,GR,RCR,AR,AL,AL,CR,DETL,AL,RCR,RCR,DETL,DETL,RCR,RCR,GL,DETL, DETL,AL,DL,DETL,AL,DETL,DETL,DETL,AR,PrR,GR,PtL,AL,DETR,PrR,PrR,AR,RCR, RCR,RCR,AL,AL,AR,GR,AR,AR,RCR,AR,RCR,RCR,PrL,AL,AL,AL,AR,AR,DETR,AL, DETL,GR,AL,AL,GR,DETL,GR,GR,AR,AL,ApR,RCR,RCR,AL,ConL,PrR,DETL,PrL,PrL, AR,PtR,PtR,DETR,GR,PtR,RCR,DETL,GL,GL,AR,AL,ConR,AL,DETL,AL,AR,DETL,GL, GR,AR,GR,DETR,DETR,AR,AR,PtR,PrR,AR,GR,DETL,AR,AL,DETL,PrL,AR,AR,AR, DETR,GR,GR,AL,PrL,ConR,RCR,RCR,RCR,AR,AR,DETL,DETL,RCR,AL,DETL,RCR, RCR,AL,DETL,AR,DETL,DR,GR,RCR,AR,GR,DETR,GL,DETR,AR,AL,DETL,AL,DETR, PrL,ConR]

Derzhavin: ***Vodopad*** (*Waterfall)* (long poem), 1791-1794, words 2042, attributes 446. From: G.R. Derzhavin. Stikhotvorenija. Biblioteka poeta. Bol'shaja serija. (Poems. Poet's library. Big series). Leningrad, Sovetskij pisatel', 1957.

[AL,GR,GL,GR,AL,PrR,AL,AL,GR,ConL,PtL,AL,AL,PrR,AR,GR,GR,GR,AR,ConR,AR, DETL,AL,DETR,PtR,GR,GL,AL,PrR,AL,AR,AL,AL,DETR,PtL,AL,DETL,GR,AL,RCR, GR,AL,DETL,AR,AR,GR,DETR,PtL,PrR,DETR,AR,AR,ConR,AL,GR,DETL,GR,DETR, GR,DETL,RCR,DETL,PtL,GR,GR,GR,GL,AR,DETL,AR,AR,DETL,ConR,AL,PtL,GR,AR, AL,GR,PtL,GR,PtL,GR,PtL,DETL,GL,DETR,AL,DETL,AL,AL,APR,AL,GR,ConR,PtL, PtR,AL,PtR,DETR,AL,GR,RCR,AL,AL,DETL,DETL,RCR,RCR,AL,PtL,DETL,PtL,AL, PtR,AR,DETR,PtL,AR,AL,DETL,PrR,AR,DETL,DETL,PtR,GR,PrR,AL,AL,AR,AL,AL, AR,AL,GR,GR,AR,AL,GR,DETL,AL,DETL,DETR,AL,AL,GR,AL,PrR,AL,DETL,PrR, DETL,AL,AL,RCR,DETL,AR,GR,GR,GR,AR,GR,GR,AL,AvL,AL,PrR,GR,AL,AL,PrR, PtL,AR,AL,DETL,PtR,PtR,ConR,,AL,PrR,PtR,PtR,PtR,GR,DETL,GR,AL,AR,DETL,GR, AL,AL,DETR,GR,AL,AR,DETR,DETR,DETR,AR,PrR,AR,DETL,DR,AR,DR,PtL,GR, DETR,DETR,GR,AR,AL,Av,PtL,InstrL,DETL,GR,AL,GR,AL,GR,AR,DETL,DETL,GR, DETR,AR,AR,AL,AR,AL,AL,AL,GR,AR,DETL,PrR,DETL,PrL,AL,GR,AL,GR,PtL,AL, PtR,AR,AR,GL,AL,CR,GR,AL,GR,GR,GR,GR,PrR,PrR,AR,PrR,RCR,GR,GR,DETL,DETL, RCR,AL,DETL,RCR,AR,GR,AR,AL,AR,CR,AR,GR,RCR,AL,AL,DETR,AL,DETR,GR, DETR,ApR,PrR,PtL,AR,AL,CR,RCR,AR,AR,GR,RCR,AL,DETL,AL,DETL,AL,AL,AL, GR,PtR,PrR,DETR,RCR,AR,AR,AL,DETL,PtL,AL,DETR,AR,AR,GR,GR,GL,AL,AL,RCR, DETL,PrR,AL,AL,GR,AL,AL,AL,GR,DETL,DETL,AR,AL,DETL,AL,PtR,RCR,GL,GL, GR,PtR,AR,AL,AL,GR,DETL,DETL,PtL,CR,DETR,GL,DETR,GR,PtR,PrR,PtL,GR,GR, GR,GR,AL,AL,GL,AL,PtL,PtR,GR,DETL,DETR,PrR,AL,DETL,AL,AL,AL,DETL,PtL, ApR,AR,AR,DETL,AL,DETR,AL,AL,AR,PtL,PtL,DETL,AR,DETL,RCR,RCR,AL,GR,GR, InstrL,AL,DETL,AL,AL,AR,DETL,AR,AR,AL,GR,DETR,DETR,PtL,AL,GR,PrR,PtR,PrR, PtR,AL,AL,AL,DETL,RCR,AR,AR,AR,DETR,AR,AR,PrR,PrR,AR,AR,PrR,AL,GR,AL, PrL,AL,DETL,AL,GR, DR]

Karamzin: *Bednaja Liza* (*Poor Liza)* (novella), 1792, Paragraphs 1-8, words 1117, attributes 234. From: N.M. Karamzin. Izbrannije sochinenija v dvukh tomakh (Selected works in two volumes). Khudozhestvennaja literatura, Moskva-Leningrad, 1964. Tom 1.

[GR,PrR,GR,DETR,DETL,AL,AL,AL,DETL,RCR,AL,AL,CL,DETL,AL,DETL,DETL,AL, GR,GR,RCR,AL,ApR,AL,AL,DETR,AL,AL,AL,PrR,PrR,AL,AL,PtL,AL,AL,PtR,AL,AL, GR,PtR,AL,GR,RCR,AL,GR,AL,DETL,GR,AL,RCR,AL,GR,AL,AL,AL,RCR,AR,AL,AL, GR,AL,AL,GR,CL,AL,AL,AL,InstrL,PtR,CR,AL,PrR,DETR,DETL,AL,GR,PtL,ApR,PtL, AL,AL,PrL,GR,AL,GR,AL,GR,GR,PtR,RCR,DETL,RCR,AL,ApR,AL,PtL,PtL,AL,AL,GR, DETR,PrR,PrR,DETR,CR,GR,GR,AL,AL,PrR,AL,PrR,GR,AL,PtR,GR,AL,PrR,DETR,AL, GR,AL,DETR,GR,GR,DETL,PrR,PtR,GR,PtR,AL,GR,DETL,DETL,GR,AL,DETL,RCR, AL,AL,GR,RCR,AL,PtL,ConR,AL,AL,DETL,CL,GR,AL,GR,AL,GR,RCR,DETL,AL,GR, PrL,AL,PrR,AL,PrR,AL,AL,PrR,PrR,PrR,AL,AL,ApR,DETR,AL,AL,AL,AL,GR,DETL,AL, AL,PrL,GR,DETL,RCR,DETL,AL,AL,DETR,AL,AL,GR,PtL,AL,GR,DETR,CCR,DETL, DETL,IR,DETL,AL,AL,DETR,GR,GR,DETL,DETL,AL,AL,DETL,AL,AL,DETR,AL,GR, AR,AL,PtL,AL,AL,PtL,AL,GR,DETL,AL,AL,DETL,DETL,AL,AL,DETL,DETL,DETR]

Lermontov: *Fatalist* (*The Fatalist*) (novella) in *Geroy nashego vremeni* (*A Hero of Our Time*), 1837-1840, words 2625, attributes 355. From: M.Yu. Lermontov. Sobranije sochinenij v chetirekh tomakh. (M.Yu. Lermontov. Collection of works in four volumes). Moskva. Khudozhestvennaja literatura. 1976. Tom 4.

[AL,AL,PrR,GR,CR,PrR,RCR,GR,CR,DETL,AL,AL,AL,PrR,DETL,AL,RCR,DETL,AL, DETL,AL,PtL,RCR,DETL,GR,DETL,DETL,PrR,DETL,PtR,GR,AL,DETL,GR,CR,DETL, AL,AL,GR,AL,AL,AL,AL,AL,DETL,AL,AL,PtR,PrR,DETR,GR,AR,AR,RCR,DETL,AL, AL,AL,RCR,GR,DETL,AL,RCR,AL,DETL,ApR,AL,PrR,AL,AR,AL,GR,RCR,DETL,PrR, AL,GR,CR,DETL,AL,AR,AR,DETL,GR,AL,AL,IR,AL,GR,GR,DETL,DETL,DETL,AL,AL, AL,DETL,PtL,AL,DETL,DETL,GR,AL,PrR,DETL,AL,DETL,RCR,RCR,DETL,AL,AL,GR, AL,IR,AL,AL,DETR,AL,AL,AL,DETL,AL,AL,RCR,DETL,RCR,PtR,PrR,DETL,DETL, RCR,AL,AR,AL,PrR,PrR,AL,PrR,RCR,DELT,GR,AL,RCR,PrL,AL,AL,GR,AR,AR,CR,GR, AL,GR,AL,AR,PtR,AR,DETL,AL,PrR,GR,DETL,PtL,,PrR,DETL,PtR,DETL,PrR,DETL,AP, DETL,CR,PtR,GR,AL,GR,AL,DETL,AL,PrR,AR,PtR,AL,ApR,PtR,DETL,AL,PrR,PtR,AL, PrR,AL,PrR,GR,PtL,PrR,DETL,DETL,PtL,AL,RCR,DETL,PrR,PrR,AL,GR,DETR,DETL, AL,AL,DETR,AL,AL,RCR,AL,AL,RCR,AL,PrR,AL,PtL,GR,AL,GR,GR,PtR,AL,DETL, ConR,RCR,AL,AL,DETL,GR,AL,DETL,DETL,DETL,DETL,AL,DETL,DETL,AL,IR,IR, AR,AR,AR,PtR,AL,RCR,AL,DETL,AL,GR,AL,AL,DETL,AL,RCR,AL,DETL,AL,CR,PtR, DETL,AL,PtL,AL,DETL,DETL,DETL,GR,PrL,GR,PtR,CCR,AL,GR,AL,GR,RCR,AL,PrR, AL,AL,PtR,PtR,PtR,AL,AL,DETL,DETL,DETL,DETL,PtL,AL,GR,AL,PrR,GR,GR,DETL, PtL,RCR,AL,GR,AL,AL,DETL,InstrR,DETL,GR,AL,AL,PtL,AL,DETL,AL,AL,CR,AR,AL, DETR,DETL,PtL,AL,PtL,AL,PtL,AR,PtL,DETL,PrL,AvR,PtL,DETL,PtR,PrL,GR,GR, DETL,GR,GR,DETL,GR,DETL,DETL,AR,DETL,AR,DETL,AL,AL]

Lermontov: *Demon* (*The Demon*) (long poem), 1839, Part 1, words 1640, attributes 173. From: M.Yu. Lermontov. Sobranije sochinenij v chetirekh tomakh (M.Yu. Lermontov. Collection of works in four volumes). Moskva. Khudozhestvennaja Literatura, 1976. Tom 2.

[AL,GR,ApR,AL,AL,GL,DETL,RCR,GR,AL,ApR,RCR,PtL,AL,GR,RCR,AL,AL,PtL,PtR, GR,RCR,AL,GR,DETR,AR,GR,AR,PtR,GR,AL,AL,DETR,GR,GR,ConR,GR,AR,ConR, ApR,GR,AL,ConR,AL,PrR,PrR,AL,AL,GR,DETR,AL,AL,AvR,AL,InstrR,AL,GR,AR,PtL, GR,PrR,GR,PrL,AL,CR,DETL,AL,AL,AL,GR,DETR,DETR,AR,DETL,GL,AR,AL,GL,AL, AL,GR,AL,AL,PrR,AR,GR,RCR,AR,AL,DETL,GR,GL,AL,AL,PtL,RCR,PtL,AL,GR,AL, GR,GR,AL,GL,AL,AR,ConR,ConR,GR,AR,AR,GL,GR,AR,AL,AL,RCR,AL,AL,AL,AR, AL,AL,GR,DETR,AR,PtL,AL,CR,AR,AL,AR,DETR,DETL,PtR,AL,GR,DETR,AR,AL, AvL,AL,DETR,AL,AL,DETL,AL,AR,AR,GR,AR,PtR,DETL,ConR,ConR,AL,GR,GR,GR, AR,AL,AR,DETL,GL,PtL,AL,DETL,AL,InstrL,AL,DETL,AR,AL]

Turgenev: *Dvorjanskoje gnezdo* (*Home of the Gentry*) (novel), 1856-1858, Part 1, words 1640, attrbutes 224. From: I.S. Turgenev. Polnoje sobranije sochinenij i pisem v dvadtsati vos'mi tomakh (Complete collection of works and letters in twenty eight volumes). Moscow-Leningrad, Nauka, 1964. Tom 7.
[AL,AL,AL,AL,AL,DETL,GR,PtL,AL,GR,AL,AL,PrR,AL,GR,CR,AL,PrL,DETL,AL,AL, ApR,AL,DETL,ApR,AR,AR,AR,AR,AL,PtL,AR,IR,IR,PrR,ApR,CR,AL,DETL,CR,AL, DETR,PrR,AL,AL,DETL,DR,AL,DETL,PrR,RCR,DETR,DETL,AR,AR,PrR,PrR,CR,RCR, AL,AL,AL,CR,AL,CR,AL,PrR,DETL,AL,AL,PrR,AL,PrR,AL,PrR,DETL,GR,DETL,DETL, DETLAL,GR,DETL,AL,AL,DETL,DETL,DETL,DETL,AL,PrR,AR,AR,PtR,InstrR,AL,AL, GR,PrR,PtR,DETL,DETL,DETL,ApR,RCR,PtL,PrR,PtR,AL,AL,DETL,DETL,RCR,AL,PrR, AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,AR,AL,DETR,AL,AL,AL,AL,DETL,AL, AR,ConR,DETL,AL,DETR,AL,GR,AL,AL,GR,AL,PrR,AL,PrR,AL,AL,PrR,PrR,AR,AvR, AR,AvR,AL,PtL,PrR,PrR,GR,GR,DETL,GR,AR,AR,DETR,RCR,AR,PrR,DETL,PrR,PtL, AL,DETL,AL,RCR,DETL,DETL,AL,DETL,DETL,DETL,DETR,DETL,AL,AL,GL,ApR, DETR,DETL,AR,AL,DETL,AL,PtL,DETL,AL,AL,GL,AL,AL,RCR,DETL,PrR,ApR,PtR, AL,AL,DETL,DETL,AR,AL,AL,AL,AL,GR,AL,AL,AL,PrR,PtL]

Nekrasov: *Zheleznaja doroga* (*The Railroad*) (long poem), 1864, words 1640, attributes 224. From: N.A. Nekrasov. Polnoje sobranije sochinenij i pisem v 15 tomakh (Complete collection of works and letters in 15 volumes). Nauka, Moskva, 1981. Tom 2.

[AL,AL,AL,AL,PtR,PrR,AR,PtL,ConR,AL,ConR,ConR,AL,AL,AL,AL,PrR,AL,AR,AL,AR, DETR,AL,AL,AL,DETR,AR,AL,DETL,AR,AR,AR,AR,GR,AR,GR,AR,DETR,AR,DETL, PtL,CNL,AL,ApR,AL,GR,DETL,ApR,DETL,AR,PrL,CR,PrL,AL,AL,GR,AR,DETR,ApR, PtL,AL,AL,AR,AR,AL,PrP,PrL,PrL,PtL,PrR,InstrL,RCR,DETL,DETL,DETL,AR,AL,AL, PrR,AR,AR,AR,AL,DETL,AR,RCR,AL,AL,DETL,AR,DETL,AR,GR,AR,GR,AL,GL,GL, DETL,GL,GR,AL,CR,DETL,DETL,DETR,AR,DETR,CR,AR,AL,DETL,DETL,GR,IL,IL, AL,GR,GR,GR,AL,AL,DETL,AR,AL,AL,AL,AL,AL,AR,AR,AR,CR,DELT,AL,GR,AL]

Tolstoy: *Vojna i mir* (*War and Peace*) (novel), 1863-1869, Chapter 2, words 1062, adnominals 233. From: L.N. Tolstoy. Vojna i mir [War and Peace]. Moskva, Sovremennik, 1978.

[GR,AL,GR,AL,PrR,PrR,AL,PrR,RCR,GR,CR,ApR,CR,PtR,GR,AL,AL,ConL,AL,AL,CR, PtR,PtR,PtR,AL,DETL,PtR,AL,ApR,GL,CR,RCR,CR,PtL,AL,AL,PrR,PtR,DETL,GR,AL, AL,AL,GR,AL,AL,DETL,DETL,,PrR,DETL,PrR,PrR,DETL,GR,RCR,GR,PtL,AL,GR, DETL,AL,CR,PtL,AL,AL,DETL,AL,PtL,PrL,AL,PrR,AL,DETR,ApR,GR,PtL,DETL,AL, DETL,DETL,AL,GR,GR,AL,AL,PTL,DETL,PTL,AL,AL,PtR,DETL,DETR,AL,PtL,AL, RCR,AL,AL,AL,AL,PrR,AL,DETL,DETL,DETL,PrR,AL,AL,GR,AL,DETL,CR,GR,CR,AL, CR,AL,AL,AL,AL,PtL,PrR,AL,PrR,AL,PrR,AL,PrR,AL,PrR,AL,AL,AL,AL,AL,GR,ApR, GR,CR,PtR,RCR,AL,PtR,PrR,AP,GR,DETL,PrR,DETL,AL,DETL,PrL,PtL,,GR,RCR,AR, AR,AL,PtL,DETL,DETL,AL,AL,AL,AL,PtR,PtL,RCR,AL,ConR,AL,GR,PrR,DETL,CR,AL, DETL,AL,GR,DETL,GR,GR,AL,GR,DETL,RCR,AL,GR,AL,DETL,GR,GR,DETL,RCR, AL,AL,AL,AL,GR,AL,DETL,PtL,PtL,AL,AL,AL,DETL,AL,PrR,DETL,RCR,PtL,RCR,PtL, DETL,AL,RCR,DETL,GR,ConR,PrR,AL,PrR,AL,RCR,PtL,AL,GR,PtR,IR,DETL,AL]

Tolstoy: *Anna Karenina* (novel), 1873-1877, Chapters 1-2, words 1062, attributes 233. From: L.N. Tolstoy. Anna Karenina. Moskva, Khudozhestvennaja literatura, 1985.

[AL,AL,AL,GR,CR,AL,DETL,CNR,AL,AL,DETL,DETL,GR,DETL,GR,DETL,PrR,DETL, AL,PtL,ApR,GR,GR,DETL,AL,DETL,ConR,PtR,AL,DETL,GR,AL,AL,PrR,CR,AL,GR,GR, DETL,AL,DETL,AL,PtL,GR,DETL,AR,AvL,AL,DETL,AL,GR,AR,GR,PtL,AL,AL,GR,PtL, PrR,GR,AL,PrR,DETL,AL,PrL,AL,AL,DETL,RCR,PrR,GR,DETL,CCR,DETL,GR,DETL, DETL,AL,DETR,CCR,InstrR,InstrL,DETL,AL,PrR,DETL,PrR,AL,DETL,AL,RCR,AL,PrR, PrR,AL,PtL,PrR,DETL,AL,AL,AL,RCR,CR,PrR,GR,GR,GR,DETL,DETL,DETL,GR, DETL,DETL,RCR,DETL,GR,DETL,AL,GR,RCR,AL,AL,AL,DETL,AL,DETL,AL,AL,AL, GR,DETL,DETL,AL,AR,PrR,,DETR,DETL,RCR,,AL,AL,AL,PtL,ApR,ApR,PtL,InstrL,AL, DETL,DETL,DETL,DETL,DETL,PtL,PtL,AL,ApR,AL,AL,AL,ApR,GR,GR,AL,DETL,PrR, PrR,DETL,PrR,AL,AL,CR,DETL,PrR,RCR,AL,RCR,DETL,AL,PtL,GR,DETL,RCR,CNL, GR,AL,AL,AL,DETL,AL,AL,AL,AL,PtR,GR,PtR,DETL,AL,AL,ApR,CR,PrR,PrR,AL,GR, RCR,GR,DETL,GR,DETL,DETL,PtL,PtL,DETL,CR,AL,AL,GR,PtR,AL,AL,AL,PrR,DETL, DETL,AL,GR,PrR,AL,PtL,AL,PrR,PrR,AL,AL,AL,AL,GR,AL,AL,AL,AL,AL,GR,AL,ApR, GR,DETL,InsrL,PtL,AL,AL,GR]

Chekhov: *Zhenschina bez predrassudkov* (*The Woman without Prejudices*) (story), 1883, words 1063, attributes 114. From: A.P. Chekhov. Polnoje sobranije sochinenij i pisem v tridtsati tomakh (Complete collection of works and letters in thirty volumes). Nauka, Moskva, 1975. Tom 2.

[DETR,AR,DETR,PrL,AL,PrL,RCR,DETL,AL,CNL,DETL,DETL,AL,AL,AL,PtL,DETL, AL,AL,AL,AL,GR,DETR,AL,AL,DETL,AL,GR,GR,AL,AL,ApR,DETL,DETL,RCR,DETL, AL,DETL,DETR,PrR,AL,DETL,DETL,DETL,DETL,DETL,DETL,DETL,DETL,IR,DETL, DETL,AL,AL,AR,AR,AR,AR,AL,DETL,GR,DETR,PrR,AL,ConR,AL,AL,DETL,RCR, RCR,DETL,ConR,DETL,DETL,PrR,AL,DETR,AL,CCR,AL,GR,DETL,AL,ApR,PtL,AR, PtR,PtR,GR,PtR,DETL,DETL,AL,DETL,AL,AL,AL,PtR,AR,GR,AvL,GR,DETL,GR,]

Chekhov: *Dama s sobachkoj* (*The Lady With The Dog*) (novella), 1899, Chapers 1-2, words 2456, Attributes 277. From: A.P. Chekhov. Polnoje sobranije sochinenij i pisem v tridtsati tomakh (Complete collection of works and letters in thirty volumes). Nauka, Moskva, 1975. Tom 10.

[AL,ApR,PrR,PtR,PtR,AL,PrR,AL,AL,GR,ApR,PrR,AL,AL,AL,DETL,PrR,AL,PrR,PrR,GR, CNR,AL,AR,PrR,AL,AR,AR,AR,PtR,AR,AR,AR,AL,AL,AL,DETL,AR,AR,RCR,DETL, AR,AL,RCR,AL,AL,AL,ApR,AR,AR,AR,AR,AL,PrR,AL,DETL,AL,AL,AR,AL,GR,DETL, RCR,AL,DETL,AL,PrR,PrR,PrR,AL,AL,AL,PrR,PrR,AL,PrR,RCR,AL,DETL,CR,AL,AL, AR,AR,RCR,AL,AR,AR,AL,AL,RCR,DETL,RCR,AL,AL,AL,DETL,CR,DETL,PrR,AL, PrR,DETL,RCR,AL,AL,RCR,DETL,AL,AL,AL,AL,PrR,AL,AL,AL,GR,PrR,AL,GR,RCR, AL,PrL,AL,AL,AR,AR,AR,DETR,ConR,DETL,RCR,AL,DETL,RCR,AR,AR,RCR,AL,AL, IT,IR,AL,AL,PtL,AL,AL,DETR,DETL,PrR,AL,GR,AL,GR,PrR,CCR,DETL,GR,AL,AL, PrR,AL,PrR,AR,AR,PtL,GR,AL,PtR,DETL,AL,AL,DETL,DETL,AR,AR,DETL,AL,AL, RCR,AL,DETL,DETL,AR,AR,AL,AL,PtL,PtL,DETL,AL,AL,DETL,CR,DETL,PrR,AL,GR, AL,AL,AL,GR,PtR,AL,RCR,AL,DETL,AL,GR,AL,GR,GR,AL,GR,AL,RCR,AL,DETL, CCR,AL,GR,DETL,AL,DETL,ApR,DETL,PtL,AL,PrR,DETL,PtR,CCR,AL,DETL,AL,CCR, GR,AL,PrR,AL,AL,AL,GR,AL,ApR,PrR,PrR,RCR,DETL,AL,AL,GR,AL,DETL,DETL,AL, DETL,AL,AL,GR,GR,AL,GR,DETL,RCR,AL,AL,DETL,AL,RCR,DETL,AL,AL,AL,GR, RCR,CCR]

Kuprin: ***Chari*** (*The Spell*) (Story), 1897, words 935, attributes 229. From: A.I. Kuprin. Sobranije sochinenij v shesti tomakh (Collection of works in six volumes), Gosudarstvennoje izdatel'stvo khudozhestvennoj literaturi, Moskva, 1957. Tom 2.

[DETL,DETL,AL,AL,DETL,RCR,AL,DETL,PrR,AL,PrR,PrR,AL,DETL,AL,AL,PtR,GR, PtL,GR,PtL,PrR,AL,PrR,GR,PtR,AL,GR,AL,AL,AL,AL,AL,AR,PtR,AR,ApR,GR,GR, DETL,AL,GR,DETL,AL,AL,AL,PtL,GR,AL,AR,PtR,AL,PtL,AL,PtL,CNR,PrR,PtL,AL,GR, GR,DETL,AL,RCR,DETL,AL,PtL,AL,GR,AL,GR,DETL,AR,PtL,DETL,PtL,AL,AL,DETL, AL,RCR,DETL,AL,AL,AL,AL,AL,AL,AL,PtL,PR,PrR,IR,AL,PtL,RCR,AL,GR,DETL,AL, GR,AL,GR,AL,AL,PtR,AL,AL,AL,PrR,AL,AL,DETL,AL,GR,AL,DETL,AL,AL,AL,PrR, ALR,AR,ApR,DETL,AL,AL,PrL,RCR,AL,AL,PrR,AL,AL,DETL,RCR,AL,AL,DETL,AL, AL,DETL,RCR,GR,GR,GR,AL,PtL,AL,GR,GR,DETL,PtL,DETL,PtR,GR,DETL,ApR, DETL,AL,ApR,DETL,AL,ApR,DETL,ApR,AL,AL,GR,AL,PtL,PtR,DETL,AL,PrR,AL,PrR, PtL,AL,ApR,PtL,PtL,AR,PtR,AL,AL,ApR,GR,PtL,AL,ApR,AL,PrR,DETL,RCR,GR,PtL, DETL,AL,AL,PrR,AL,AL,GR,AL,GR,GR,PtR,PrR,AL,DETL,AL,AL,DETL,AL,AL,DETL, AL,DETL,AL,DETL,AL,DETL,AL,PtL,GR,GR,GR,AL,AL]

Kuprin: ***Domik*** (*Little House*) (story), 1929, words 1064, attributes 224. From: A.I. Kuprin. Sobranije sochinenij v shesti tomakh (Collection of works in six volumes). Moskva, Gosudarstvennoje izdatel'stvo khudozhestvennoj literaturi, 1957. Tom 6.

[AL,CL,AL,CR,AL,ApR,AL,AL,AL,AL,AL,AL,AL,PrR,AL,AL,InstrR,AL,AL,AL,,InstrR, AL,PrR,DETL,AL,AL,AL,AL,DETL,AL,GR,DETL,AL,AL,PtL,AL,PrR,AL,AL,AL,AL, RCR,AL,AL,AL,CNR,AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,AL,DETL,PtL,PrR,AL,GR, AL,AL,ApR,DETL,AL,PtR,AL,AL,DETL,AL,RCR,AL,PtR,PtL,ApR,CR,AR,PtL,AL,AL, AL,AL,AL,AL,GR,RCR,AL,AL,AL,CR,AL,AL,AL,AL,AL,PrR,DETL,AL,AL,AL,DETL, RCR,ConR,AL,AL,GR,RCR,DETL,AL,AL,AL,AL,PrR,AL,PrR,AL,CL,AL,GR,DETL,AL, AL,GR,AL,AL,DETL,AL,AL,DETL,AL,AL,AL,AL,AR,RCR,ConR,PrR,CL,AL,GR,CL,AL, AL,DETL,CL,AL,DETL,CR,GR,AL,AL,PrR,DETL,DETL,AL,AL,ConR,AL,DETL,DETL, AL,DETL,AL,CR,GR,AL,PrR,PrR,AL,DETL,AL,GR,CR,GR,CR,AL,AL,AR,AL,AL,AL, AL,PrR,PrR,PrR,AL,AL,AL,AL,GR,AL,ConR,AL,AL,AL,AL,GR,GR,CR,PrR,GR,GR, InstrR,CR,AL,GR,AL,AL,GR,GR,AL,AL,AL,AL,PrR,AL,CR,AL,AR,PtL,AL,DETL]

Kuprin: ***Junkera*** (*The Junkers*), 1928-1932, Chapter 2, words 1182, attributes 261. (Novel). From: A.I. Kuprin. Sobranije sochinenij v shesti tomakh (Collection of works in six volumes), Moskva, Gosudarstvennoje izdatel'stvo khudozhestvennoj literaturi, 1957. Tom 6.

[AL,CR,AL,AR,AR,AR,AL,AL,DETL,AL,AL,CNR,AR,AL,AR,AL,DETL,DETL,DETL, GR,CR,AL,AL,AL,GR,AL,AL,DETL,AL,IR,DETL,IL,GR,IL,GR,PtR,CR,GR,AL,AL,ConR, AL,AL,AL,AL,AL,AL,AL,AL,CR,AL,PtR,AL,AL,PrR,CR,AL,AR,AR,CL,AR,AR,PtR, DETL,PtL,AL,AL,AL,AL,AL,AL,DETL,GR,AL,GR,AL,DETL,DETR,AL,DETL,AL,PrR, GR,AL,GR,CR,ApR,CR,AR,AL,CR,AL,GR,AL,GR,AL,ApR,DETR,PtR,DETL,AL,CR,AL, AL,AL,DETR,AL,AL,AL,AL,AL,AL,AL,AL,AL,PtR,GR,RCR,AL,AL,AL,PrR,AL,PrR,AL, AL,AL,AL,AL,AL,GR,RCR,DETL,DETL,AL,AL,GR,AL,DETL,AL,GR,AL,GR,ApR,ApR, AL,AL,GR,GR,PtR,AL,PtR,AL,AL,AL,GR,AL,AL,AL,AL,AL,PtL,PrR,RCR,GR,GR,DETL, AL,DETL,DETL,PrR,CR,AL,GR,ApR,AL,GR,RCR,AL,AL,AL,AR,AL,GR,AL,AL,GR,AL, AL,AL,AL,AL,AL,GR,CR,DETL,CR,AL,ApR,PtR,AL,GR,AR,AR,AL,AL,GR,RCR,AL,AL, GR,AL,PrR,AL,AL,GR,CR,AL,PtL,AL,GR,AL,AL,GR,GR,GR,AL,GR,AL,GR,GR,PtR,AL, AL,IR,AL,PrR,PtR,PtR,AL,AL,PtR,RCR,PtL,AL,AL,GR,AL,AL,DETL,PrR,DETL,AL,GR,AvL,DETL,DETL,AL,AL,RCR,AL]

Bunin: ***Antonovskie jabloki*** (*Apple Fragrance*) (story), 1900, Parts I-II, words 1999, attributes 405. From: Ivan Bunin. Sobranije sochinenij v shesti tomakh (Ivan Bunin. Collection of works in six volumes). Moskva, Santax, 1994. Tom 1.

[AL,AL,AL,PtR,PrR,AL,PrR,GR,PrR,CL,GR,AL,AL,AL,AR,AL,AL,AL,AL,AL,PtL,PtL,AL,AL,PtR,AL,GR,GR,AL,GR,DETL,GR,CNR,RCR,AL,GR,AL,GR,AL,AL,PtL,AL,AL,GR,AL,PrR,PrR,GR,PtL,GR,PtL,AL,PrR,PtL,DETL,RCR,AL,PTL,DETL,PtL,AL,AL,PrR,CR,AL,AL,AL,AL,AL,AL,CNR,PrR,PtL,DETL,AL,AL,AL,PrR,AL,AR,AL,AL,PrR,AL,AL,CR,GR,PrR,PrR,PrR,AL,GR,PtL,AL,AL,AL,AL,AL,PrR,AL,Pr,AL,PtL,AL,AL,PtR,PrR,PrR,AL,AL,PrR,AL,PrR,AL,AL,ApR,RCR,AL,AvL,GR,AL,AL,GR,GR,AL,PrR,GR,AL,AL,AL,GR,ApR,GR,AL,GR,AL,PtR,AL,PtL,AL,PrL,AL,PrR,DETL,AL,AL,ApR,DETL,ApR,DETL,AL,RCR,RCR,AL,GR,CR,AL,ConR,AL,AR,DETL,AL,AL,ConL,AL,AL,PrR,AL,AL,AL,DETL,PrR,AL,AL,PtL,AL,PtR,AL,AL,AL,AL,AL,RCR,AL,PtL,AL,RCR,AL,AL,DETR,AL,AL,PrR,AL,AL,AL,AL,AL,AL,GR,AL,GR,DETL,AR,AL,DETL,GR,AL,AL,GR,RCR,DETR,ApR,PrR,PtL,AL,AR,AR,AL,PrR,AL,AL,PrR,PrR,PrR,PtR,PrR,AR,PtR,ApR,AL,CR,CR,CR,DETL,CNR,AL,GR,AL,AL,PtR,AL,RCR,AL,AL,PtR,AL,AL,PrR,AL,AL,PrR,AL,AL,DETL,AL,PrR,AL,AL,AL,PrR,AL,PrR,AL,PrR,AL,AR,PrR,AL,PrR,PrR,AL,AL,GR,DETR,AL,AL,GR,DETL,AL,AL,DETL,GR,CR,PtL,DETL,ApR,AL,AL,AL,AR,AR,AR,AR,PtL,AL,IstrL,AL,AL,AL,AL,AL,AL,PtL,AL,DETR,ConR,AL,ApR,AL,AL,AL,CR,AR,AR,AR,PtL,AL,AR,AR,AL,AL,AL,PtL,RCR,AL,AL,GR,AL,AL,PrR,AR,AL,PtR,AL,PtL,AL,DETL,InstrR,RCR,PrR,DETL,GR,DETL,DETL,AL,AL,AL,PtR,PtR,AL,AL,AL,ApR,AL,DETL,GR,AL,AL,PrR,DETL,AL,GR,DETL,AL,AL,GR,AL,AL,GR,AL,AL,RCR,ApR,AL,GR,ApR,ApR,PrR,AL,AL,AL,PrR,AL,AL,AL,PrR,PrR,ApR,AL,CR,CR,CR,CR,AL,AL,AL,PrR,PtL,AL,AR,AR,PrR,AL,AL]

Bunin: ***Kavkaz*** (*The Caucasus*) (story), 1937, words 1259, attributes 205.From: Ivan Bunin. Sobranije sochinenij v shesti tomakh [Ivan Bunin. Collection of works in six volumes]. Moskva, Santax, 1994. Tom 6.

[AL,PrR,PrR,DELT,DETL,AL,PtL,PtL,PrR,DETL,DETL,DETL,PrR,DETL,AL,AL,DETL,ApR,DETL,DETL,DETL,AL,AL,DETL,AL,DETL,AR,AR,AL,AL,AL,AL,AL,AL,DETL,AL,AL,AL,PtL,GR,PtL,PtL,PrR,AR,GR,AL,AL,AL,AL,RCR,AL,AL,DETL,AL,AL,PtR,AL,AL,GR,IR,AL,DETL,AL,AL,AL,AL,RCR,DETL,GR,GR,PtL,RCR,DETL,AL,AL,DETL,AL,DETL,AL,AL,PtL,AL,AL,PtL,AL,AL,PrR,GR,AL,PrR,AL,AL,PrR,PrR,PtL,AL,AL,GR,PrR,AR,PtL,AL,PtL,AL,RCR,AL,RCR,AL,AL,AL,AL,AL,AL,AL,GR,AL,PtL,PrL,PtL,DETL,AL,AL,GR,AL,AL,PrL,AL,PtL,AL,AL,PtR,DETL,AL,AR,PtL,PrL,AL,AL,DETL,AL,AL,AL,AL,GR,GR,PtR,PtR,PrR,GR,DETL,DR,DETL,DR,AL,AL,AL,AL,AL,AL,GR,AL,GR,RCR,DETL,ApR,AL,PrR,AL,AL,AL,AL,AL,PtR,AL,AL,AL,DETL,DETL,AL,DETL,AL,AL,AL,AL,AL,AL,GR,AL,AL,AL,AL,GR,PtL,AL,DETR,DETL,PTL,DETL,AL,AL,DETL,GR,GR,PrR,DETL]

Bunin: ***Stepa*** (story), 1938, words 1520, attributes 233. From: Ivan Bunin. Sobranije sochinenij v shesti tomakh [Ivan Bunin. Collection of works in six volumes]. Moskva, Santax, 1994. Tom 6.

[PrR,AL,CR,IstrR,PrR,PtL,AR,PtL,PrR,RCR,AL,AL,AL,AL,PtL,AL,AL,AL,AL,AL,PtR,AL,GR,AL,AL,AL,AL,PrR,DETL,AL,AL,DETL,PtR,AL,GR,GR,PtL,AL,AL,GR,AL,AL,PtL,

PtL,PtL,DETL,PtL,DETL,DETL,AL,AR,AL,CNR,PtR,PrR,AL,PtL,DETL,AL,AL,AL,GR, AL,PrR,AL,PrR,DETL,AL,AL,PtL,AL,PrL,PtL,AL,PrR,PrR,CL,AL,GR,PtR,PrR,CR,PrR, PrR,PrR,PrR,PrR,PrR,GR,AL,AL,DETL,AL,DETL,PtL,AL,DETL,AL,PtL,GR,AL,AL,CNR, CR,DETR,DETL,PrR,AL,PrR,AL,AL,PrR,AL,AL,PrR,PtL,AL,AL,AL,PtL,PrL,DETL,GR, AL,AL,AL,PrR,AL,AL,PrL,PrR,AL,PtL,PrL,AL,GR,AL,AL,AL,PrR,AL,PrR,AL,AL,AL, AL,PtL,GR,ApR,DETL,AL,AL,DETL,PtL,PtL,PrR,AL,AL,AL,AvL,AL,GR,PrR,AL,GR, DETL,AL,Ap,DETL,PtL,PtL,DETL,DETL,AL,PtL,PrR,PtL,AL,PtL,AL,AL,GR,AL,AL,AL, CCR,DETL,AL,AL,DETL,AL,AL,DETL,PtL,DETL,AL,AL,AL,PrR,AR,AL,AR,GR,GR,AL, GR,AL,PrR,DETL,PtL,PtL,AL,AL,AL,AL,AL,AL,PtL,PrR,AL,AL,PrR,PrR,PrR,PtL,DETL, GR,DETL,AL,DETL,AL,AL,AL]

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language.* 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis.* 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1.* 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences.* IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5.* 2015, III+146 pp.
22. P. Zörnig et al. (2016). *Positional occurrences in texts: Weighted Consensus Strings.* 2016. II+179.pp