Glottometrics 24
2012

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 werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** 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 (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

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. 24 (2012), Lüdenscheid: RAM-Verlag, 2012. 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 24 (2012) **ISSN 2625-8226**

Contents

History of Quantitative Linguistics

Book Reviews

Glottometrics 24, 2012, 1-4

Diversifikation der starken Verben im Deutschen

Karl-Heinz Best

Abstract. In this paper the 1-displaced geometric distribution has been fitted to the ranked distribution of classes of the strong verbs in German. The classes are defined by the different vocalic alternations of the verbs. This way the paper brings a further corroboration of the hypothesis that diversification processes abide by laws.

Keywords: German, strong verbs, distribution, diversification

1. Starke Verben als Gegenstand der Quantitativen Linguistik

Die Quantitative Linguistik beruht zu wesentlichen Teilen auf der Grundannahme, das Prozesse und Zustände in den Sprachen ebenso wie in der Verwendung der Sprachen von Gesetzen gesteuert werden (Beispiele in: Altmann u.a. (2002); Köhler u.a. (Hrsg.) 2005). Eine ganze Reihe von Gesetzen sind bekannt, viele davon mehr oder weniger gut überprüft, manche bestehen bisher lediglich in Form von noch zu prüfenden Hypothesen.

Zu den bekannten und auch überprüften Sprachgesetzen gehören das Piotrowski-Gesetz ebenso wie das Diversifikationsgesetz. Beide Sprachgesetze lassen sich bei vielen verschiedenen Bereichen nachweisen. Eines der Phänomene, das sich dafür anbietet, ist die Entwicklung der starken Verben im Deutschen und ihre Diversifikation in der Gegenwartssprache. Dass der Verfall der starken Verben dem Piotrowski-Gesetz folgt, wurde bereits nachgewiesen (Best 2003, 12-14). In diesem Beitrag soll ihre Diversifikation untersucht werden.

2. Gesetzmäßigkeit der Ablautklassen der starken Verben

Generell gilt die Annahme, dass sprachliche Erscheinungen, die in unterschiedlicher Form auftreten, dem Gesetz der Diversifikation unterliegen (Altmann 1991, 2005). Dieses Gesetz kann je nach Phänomen unterschiedliche Formen annehmen, worüber Altmann (1991, 39-41 und 2005, 649-655) einen Überblick gibt. Die einfachste Version dieses Gesetzes ist die geometrische Verteilung, die in 1-verschobener Form

$$P_x = pq^{x-1}; \quad x = 1, 2, ...$$

lautet.

3. Starke Verben im Deutschen

Das Deutsche hat nur noch eine relativ kleine Anzahl starker Verben, insgesamt 173. Diese Verben unterscheiden sich danach, mit Hilfe welcher Ablaute sie ihre Flexionsformen bilden. Je nach dem, welche Ablautreihen sie enthalten, kann man sie zu Klassen zusammenfassen.

Die starken Verben verteilen sich derzeit auf 39 Ablautreihen. Ordnet man diese Verbklassen danach, wie viele Verben zu jeder von ihnen gehören, erhält man eine Rangordnung. Darauf angewendet ist die Hypothese zu überprüfen, dass diese nach Häufigkeitsrängen geordneten Verbklassen dem Diversifikationsgesetz unterliegen. Um dies zu erreichen, kann man z.B. die bereits genannte geometrische Verteilung an die entsprechende Datei anpassen und, wenn dies mit Erfolg durchgeführt ist, die Annahme der Gesetzmäßigkeit als bestärkt ansehen. Die Verteilung wird in 1-verschobener Form angepasst, da es keine Ablautklasse ohne wenigstens ein Verb als Element gibt.

Die Daten zur Verteilung der starken Verben auf die Ablautreihen sind der *Duden-Grammatik* (1998, 127) entnommen, die anders als die *Duden-Grammatik* (2009, 452-453) die Verben in der für unsere Zwecke erforderlichen Übersicht und vollständig aufführt.

4. Überprüfung der geometrischen Verteilung als Modell für die Diversifikation der starken Verben

Die Anpassung der 1-verschobenen geometrischen Verteilung mit Hilfe des *Altmann-Fitters* (1997) erbrachte das in Tabelle 1 dargestellte Ergebnis:

Tabelle 1
Anpassung der 1-verschobenen geometrischen Verteilung an die Ablautreihen der starken Verben im Deutschen (*Duden-Grammatik* 1998, 127)

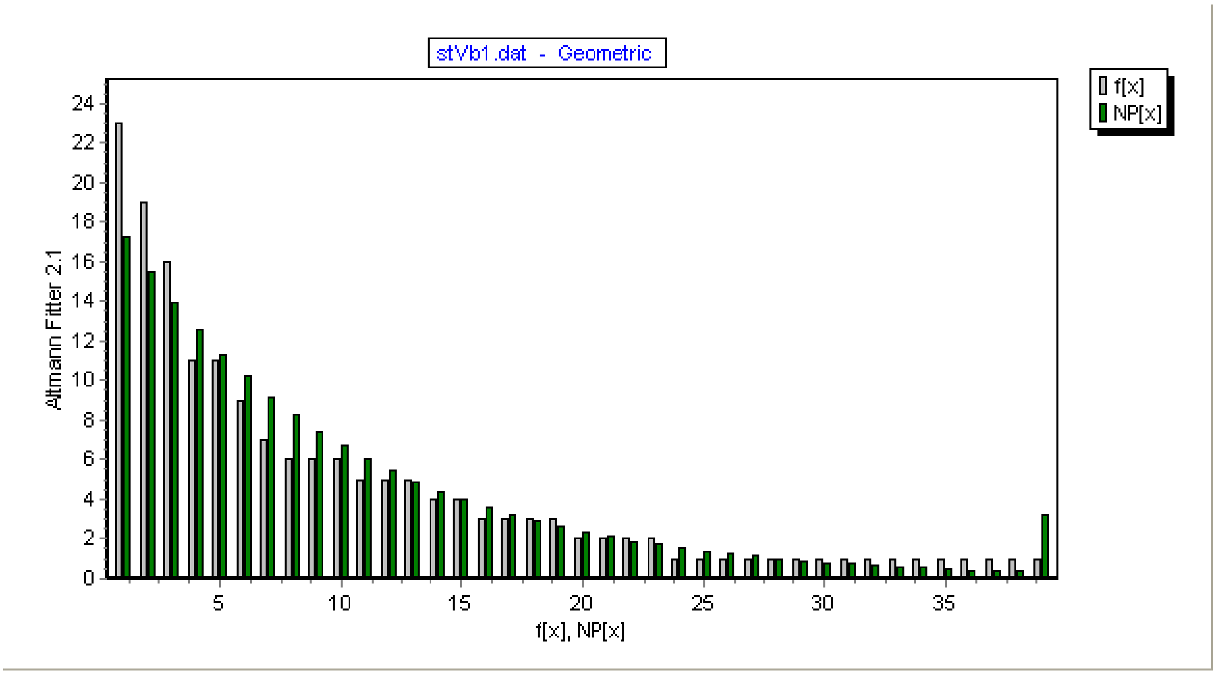
Rang	Ablautreihe	n_x	NP_x	Rang	Ablautreihe	n_x	NP_x
1	ei - i - i	23	17.26	21	au - i: - au	2	2.11
2	i - a - u	19	15.54	22	au - o: - o:	2	1.90
3	ei - i: - i:	16	13.99	23	a - i - a	2	1.71
4	i: - o - o	11	12.59	24	i - a: - e	1	1.54
5	i: - o: - o:	11	11.34	25	i - u - u	1	1.39
6	e - a - o	9	10.20	26	i - a: - e:	1	1.25
7	e - o - o	7	9.19	27	i: - a: - e:	1	1.12
8	i - a - o	6	8.27	28	a - o - o	1	1.01
9	a: - u: - a:	6	7.45	29	e: - u - o	1	0.91
10	e: - a: - e:	6	6.70	30	e: - a: -o	1	0.82
11	e - a: - o	5	6.03	31	o - a: - o	1	0.74
12	e - a: - e	5	5.43	32	o: - i: - o:	1	0.66
13	e: - o: - o:	5	4.89	33	u: - i: - u:	1	0.60
14	a – u: - a	4	4.40	34	ä - i - a	1	0.54
15	a: - i:- a:	4	3.96	35	ä: - a: - o:	1	0.48
16	a - i:- a	3	3.57	36	ö - o - o	1	0.44
17	e: - a: - o:	3	3.21	37	ö: - o: - o:	1	0.39
18	ä: - o: - o:	3	2.89	38	au - o - o	1	0.35
19	ü: - o: - o:	3	2.60	39	ei - i: - ei	1	3.19
20	i - o - o	2	2.34				
p = 0.0998 FG = 31 X^2 = 11.507 P = 0.99 (abgerundet)							

Legende zur Tabelle:
Der Doppelpunkt in den Ablautreihen zeigt die Länge der betreffenden Vokale an.
Rang: nach Zahl der Verben geordnete Rangfolge der Ablautreihen
p: Parameter der Verteilung

n_x: beobachtete Zahl der Verben der jeweiligen Ablautreihe
NP_x: durch Anpassung der 1-verschobenen geometrischen Verteilung berechnete Zahl der Verben der jeweiligen Ablautreihe
FG: Freiheitsgrade
X^2: Chiquadrat
P: Überschreitungswahrscheinlichkeit des Chiquadrats

Die Anpassung des gewählten Modells an die beobachteten Daten wird als erfolgreich angesehen, wenn $P \geq 0.05$; diese Bedingung ist erfüllt, so dass man feststellen kann, dass die 1-verschobene geometrische Verteilung sich als Modell für die nach Rängen geordneten Ablautreihen starker Verben im Deutschen bewährt.

Zur Veranschaulichung dient die folgende Graphik zu Tabelle 1; die hellen Balken zeigen die beobachteten, die dunklen die berechneten Werte.



Graphik zu Tabelle 1: Anpassung der 1-verschobenen geometrischen Verteilung an die Ablautklassen der starken Verben im Deutschen

5. Zusammenfassung

Als Ergebnis kann festgestellt werden, dass die Ablautreihen der starken Verben im Deutschen, in eine nach der Zahl der Verben geordnete Rangfolge gebracht, der 1-verschobenen geometrischen Verteilung unterliegt. Die Hypothese, dass sprachliche Phänomene entsprechend einem Sprachgesetz diversifizieren, konnte damit ein weiteres Mal bekräftigt werden.

Abschließend sei darauf hingewiesen, dass außer der geometrischen Verteilung, die hier als das einfachste Modell vorgezogen wurde, auch noch andere Verteilungen mit ähnlich guten Ergebnissen an die Daten angepasst werden können. Solche Möglichkeiten sind zu beachten, solange man nicht im Vorhinein begründen kann, welche Verteilung womöglich allein für ein bestimmtes Phänomen in Frage kommen kann.

Zu bemerken ist, dass die Klasse von starken Verben eine Menge darstellt, aus der von Zeit zu Zeit ein Verb in die Klasse der regelmäßigen Verben übergeht. Auch dieser Prozess verläuft gesetzmäßig und ist analog dem radioaktiven Zerfall in der Physik. Die Modellierung des Zerfalls ist zwar leicht, jedoch fehlen uns Daten, die jahrhundertelange Ereignisse darstellen. In der Regel fallen seltene Verben aus, ohne Rücksicht auf die Ablautreihe.

Literatur

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar* (S. 33-46). Hagen: Margit Rottmann Medienverlag.

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.) (2005), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 646-658.* Berlin-NewYork: de Gruyter.

Altmann, Gabriel; Bagheri, Dariusch; Goebl, Hans; Köhler, Reinhard; Prün, Claudia (2002). *Einführung in die quantitative Lexikologie*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Duden. Grammatik der deutschen Gegenwartssprache. 6., neu bearbeitete Auflage. Mannheim/Leipzig/Wien/Zürich: Dudenverlag 1998.

Duden. Die Grammatik. 8., überarbeitete Auflage. Mannheim/Wien/Zürich: Dudenverlag 2009.

Software

Altmann-Fitter. 1997. *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Glottometrics 24, 2012, 5-11

Non-traditional approach to the study of the rhythmics of Russian verse

Vadim S. Baevskij

Abstract. The paper suggests a more precise method of analysing the iambic and trochaic rhythms of Russian verse, as compared with the traditional one – from Andrey Bely to A.N. Kolmogorov and M.L. Gasparov inclusive. The method described is based on the linguistic ideas proposed by A.A. Potebnya.

Key words: *Andrey Bely, dynamically unstable words, syllable prominence, rhythm matrix.*

The present paper deals with the alternating rhythms only – the iamb and the trochee. The verse of alternating rhythm has been predominant over the whole history of the new Russian poetry that is nearly three centuries long. This is the verse rhythm of Pushkin, Pasternak and – with a little exception – of all the rest of Russian poets. Alternating rhythms, i.e. rhythms comprising the regularly alternating elements; here refers to syllabic positions that may be marked either by an unstressed syllable or an extremely prominent (stressed) one. These verse meters are characterized by a regular distribution of ictuses and non-ictuses. The ictus is a metrically strong syllabic position which is expected to be stressed and, in this way, to become extremely prominent. The non-ictus, on the contrary, is not expected to be marked by the stress. The expectance or non-expectance of the top prominence of syllables is fostered in us by our poetic culture.

A careful approach reveals, though, that prominence of both ictuses and non-ictuses may be of a different kind. However, when at the beginning of the twentieth century Andrey Bely began his studies of the Russian verse rhythm, he divided all syllables only into stressed and unstressed ones. After him, practically all the other scholars, including Gasparov and Kolmogorov, applied this procedure. Such approach suggests that the degree of prominence of the ictus and non-ictus should always be the same.

Some linguists (Potebnya, Koshutich, Shengeli, Bogoroditsky, Reformatsky, Panov, Shcherba, Gvozdev) treated the problem of syllable prominence according to its position in an utterance (with respect to the other syllables, the stress, the word-boundary) as well as according to the structure of this very syllable (open, closed, covert, overt). This problem has not been adequately dealt with, though.

■□

We shall call the methods used in this study as the *Potebnya effect* – after the scholar who was the first to formulate this. It plays an important role in the study of verse theory. The most detailed study of rhythm according to the Potebnya effect is given in my doctoral dissertation (Baevskij 1975: 141 – 231).The Russian language has a considerable layer of dynamically unstable words: pronouns, auxiliary verbs, monosyllabic adverbs, interjections, polysyllabic prepositions, conjunctions, particles. Some scholars of verse theory tend to consider dynamically unstable words as stressed ones, others as unstressed. Tomashevsky considered the dynamically unstable word to be stressed if it falls on the ictus and unstressed if it falls on the non-ictus.

However, it seems to be very confusing: following Andrey Bely's ideas, we try to be as accurate in the verse study as possible, but the divergence of opinions in this case destroys all the precision. Thus, studying Blok's trochaic pentameter, A.M. Astakhova records 45.5% of stressed syllables on the most significant first ictus (Astakhova 1926: 66), whereas R. Kemball records here 63.3% (Kemball 1965: 191). The results of these two papers can not be trusted as they diverge by nearly 20%.

The present paper suggests that one should consider the prominence of both ictuses and non-ictuses according to the Potebnya effect (Potebnya 1865: 62). Instead of dividing syllables into stressed and unstressed, it is recommended to supply them with a kind of scales, on the assumption of the abovementioned phonetic properties. In verse the gradation of syllable prominence becomes more complex because it is affected by the metrical structure (the position of the word either on the ictus or on the non-ictus).

■□

Potebnya did not have any special equipment for measuring the sounds in speech and the very study of speech sounds occupied a gray area. That is why we had to begin with conducting auditorial and instrumental experiments. The auditorial experiments were conducted in Smolensk State Teacher Training Institute where I am working (now it is the State University); the instrumental experiments were carried out in the laboratory of Experimental Phonetics in Minsk State Teacher Training Institute of Foreign Languages. I feel obliged to thank the head of the laboratory, Professor K.K. Baryshnikova, who contributed to the success of my work by leaving at my disposal all the equipment, and the laboratory stuff for their assistance. Apart from that, she took a deep interest in the subject matter of my research and gave me quite a number of extremely valuable pieces of advice. The procedures of the experiment carried out, as well as its results, were published in the following publications (Baevskij 1967: 50–55; Baevskij 1968: 16–22; Baevskij 1969: 244–250; Baevskij 1970: 157–168; Baevskij, Osipova 194: 174–195; Baevskij, Osipova 1974: 11–19; Baevskij 1975: 166–231; Baevskij 2001: 152–172, 307–309).

■□

In our study of the rhythm each syllable is indicated by a degree of prominence *B* where *B* each time possesses the value within the following numbers

$B = \{1.0; 1.5; 2.0; 2.5; 3.0\}$

according to the following rules:

the stressed syllable of the notional word on the ictus = 3.0;

the stressed syllable of the notional word on the non-ictus = 2.5;

the stressed syllable of a metrically dual word both on the ictus and on the non-ictus = 2.5;

the first pretonic, the second initial overt, the first and the second post-tonic, the final open = 2.0;

the third pretonic initial overt and the third post-tonic final closed = 1.5;

the prominence of the rest of the syllables = 1.0.

The experiment has shown that the differences in the syllable prominence which are smaller than 0.5 are not normally detected by human ear.

Let us illustrate it with an example.

Когда я думаю о Блоке,
когда тоскую по нему,
то вспоминаю я не строки,
а мост, пролетку и Неву.
И над ночными голосами
чеканный облик седока –
круги под страшными глазами
и черный очерк сюртука (Evtushenko 1962: 111).

Let us now supply it with a scheme showing the distribution of the ictuses and non-ictuses in each line of this text. The ictus is represented by «—», the non-ictus is represented by «υ»:

υ —υ —υ —υ — (υ)

The interpretation of the same verse by Andrey Bely would be rather unvaried. All the verses would be alike.

And now we shall present the rhythm matrix of this text; each syllable here is represented by the number which shows its prominence in speech according to the Potebnya effect.

2.0	2.5	2.5	3.0	1.0	2.0	2.0	3.0	2.0
2.0	2.5	2.0	3.0	2.0	1.0	2.0	3.0	
1.0	1.0	2.0	3.0	2.0	2.5	2.0	3.0	2.0
2.0	3.0	2.0	3.0	2.0	2.0	2.0	3.0	
1.5	1.0	2.0	3.0	2.0	1.0	2.0	3.0	2.0
2.0	3.0	1.0	3.0	1.0	1.0	2.0	3.0	
2.0	3.0	2.0	3.0	1.0	2.0	2.0	3.0	2.0
2.0	3.0	1.0	3.0	1.0	1.0	2.0	3.0	

In Andrey Bely’s interpretation the first, second, fourth, sixth, seventh and eighth lines are absolutely alike. In each of them the stress is on the second, fourth and eighth syllables. The other two lines, the third and the fifth, look different: the stress on the first ictus is omitted but it appears on the second and fourth ictuses. So, these eight lines demonstrate two varieties of rhythm: one characterizing the six lines, the other variety characterizing the other two.

However, if we carefully consider the rhythm matrix according Potebnya, we shall see that the abovementioned text does not contain any lines which are absolutely alike. Now we have the opportunity to detect even the slightest differences in the rhythm of this text.

■□

In order to make further description of Potebnya’s method of rhythm study easier, let us introduce the following symbols:

m – the quantity of the syllables under study
n – the number of the lines under study
M – the full quantity of the syllables in the line, the final ictus included
k – the quantity of the ictuses in the line
l – the quantity of the weak metrical positions in the line
N – the full quantity of the lines in the text

B – the value of the prominence of the syllable

B_{mn} – the prominence of syllable *m* in verse *n*.

So we get the following rhythm characteristics. The average prominence of each syllable in the line:

$$\overline{B}_m = \frac{1}{N} \sum_{n=1}^{N} B_{mn}$$

The average prominence of all ictuses:

$$\overline{B}_{(ict.)} = \frac{1}{k} \sum_{1}^{k} B_{(ict.)}$$

The average prominence of all non-ictuses:

$$\overline{B}_{(non-ict.)} = \frac{1}{l} \sum_{1}^{l} B_{(non-ict.)}$$

The next step in our research was suggested by Academician A.N. Kolmogorov. In his letter dated 24 June, 1971, A.N. Kolmogorov suggested to me introducing the index which became possible to calculate due to my study of the non-ictus prominence. This index represents the difference between the meanings of average prominence of all the ictuses and non-ictuses of the text:

$$P = \overline{B}_{(ict.)} - \overline{B}_{(non-ict.)}$$

This formula reflects not only the peculiarities of the distribution of syllable prominence regarding ictuses and non-ictuses, but the average word length as well. More than that, it shows the correlation of the average word length in different texts. The higher the value of P-index is, the more words the line may contain on the average. P-index, together with average prominence of a syllable in a text, is another relevant parameter. High average prominence of strong syllabic positions and low average prominence of weak syllabic positions create a sharp contrast and, therefore, emphasize the meter.

These regularities are especially noticeable in the texts which are quite big in volume. We systematically studied the rhythm of 442 poems (with a total of 11 282 lines) and a number of other poems (Baevskij, V.S. 2001. Ch.8 and other works in the reference). All these research studies allow to trace that from the middle of the 18th century to the middle of the 20th century; the differential prominence of ictuses and non-ictuses had been gradually becoming smaller. Though this index can not serve as the only criterion to make conclusions whether the author’s tendency is more of a traditional style or of the innovation, still this index is certainly worth considering when studying this subject.

Not only average characteristics of the verse rhythm can be studied with the help of the method shown in the paper, but the individual characteristics of each line as well. In order to obtain them we may calculate the average quadratic deviation (σ) of syllable prominence in this line regarding the average prominence of these very syllables in the whole text

$$\sigma_n = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(B_{mn} - \overline{B}_m)^{\,2}}$$

and set up an asymptotic normal test for testing differences.

The lines with the maximal quadratic deviation are the so-called rhythmic rarities. They differ from ideal average rhythm as much as possible; they are evidently characterized by the most remarkable rhythmic image. The purpose of careful philological research is to find out which role this rhythmic italics play in the semantics of the whole verse. Our observations show that the lines which are rhythmically marked are usually extremely significant from the point of view of the text message.

■□

The characteristic of syllable prominence according to the Potebnya effect has a number of advantages as compared with the traditional approach, which distinguishes only stressed and unstressed syllables.

1. Potebnya`s method allows to assess the degree of prominence of dynamically unstable words that do not entirely fit the binary opposition “stressed vs. unstressed” words.

2. The researcher has the opportunity to consider and characterize the rhythm of the text in a more careful and detailed way.

3. Quantitative analysis of all the syllables comprising the text allows assessing not only the prominence of ictuses but of non-ictuses as well, given that each one is contrasted with all the rest. So we are able to see and show the “living life” of non-ictuses in the alternating verse.

■□

The present paper has been written and published in hope that as I have found a friend in this generation and I shall find[1] a follower in the posterity.

Texts

[1] **Evtushenko, E.A**. *Vzmakh ruki.* [A Wave of the Hand]. (1962). Moscow: “Molodaia gvardiia”.

[2] **Pasternak, B.L.** *Sestra moia zhizn’.* [My Sister, Life]. (1922). Moscow: Izdatel’stvo Z.I. Grzhebika.

[1] Here two lines from Baratynsky’s verse are paraphrased:

И как нашел я друга в поколенье,
Читателя найду в потомстве я.

References

Astakhova, A.M. (1926). Iz istorii ritmiki khoreia. [From the History of the Trochaic Rhythm]. *Poetica. 1. Vremennik otdela slovesnykh iskusstv Gos. Instituta istorii iskusstv*. Leningrad: Academia.

Baevskij, V.S. (1966). O chislovoi otsenke sily slogov v stikhe al'terniruiushchego ritma. [On the Quantitative Assessment of Syllable Prominence in the Alternating Verse Rhythm]. *Voprosy iazykoznaniia, Iss. 2, 84 – 89.*

Baevskij, V.S. (1967). *Chislovye znacheniia sily slogov v stikhe al'terniruiushchego ritma.* [Numeric Values of Syllable Prominence in the Alternating Verse Rhythm]. Philologicheskiie nauki, Iss.3. P. 50 – 55.

Baevskij, V.S. (1968). *Ob eksperimental'nom issledovanii russkogo stikha al'terniruiushchego ritma // Metody eksperimental'nogo analiza rechi.* [On Experimental Study in Russian Alternating Verse Rhythm // Methods of Experimental Study of Speech]. Minsk, 16 – 22.

Baevskij, V.S. (1969). *Stikh al'terniruiushchego ritma v svete auditorskogo eksperimenta // Russkaia sovetskaya poeziia i stikhoved'eniie.* [Alternating Verse Rhythm in the Light of Auditorial Experiment // Russian Soviet Poetry and Verse Study]. Moscow, 244 – 250.

Baevskij, V.S. (1970). *K izucheniiu ritmiki (aktsentuatsii) russkogo stikha.* [On the Study of Rhythmics (Accentuation) of Russian Verse]. Uch'onyie zapiski, Vol.10. Smolensk State Pedagogical Institute of K. Marx. Novozybkovsk State Pedagogical Institute. Bryansk, 157 – 168.

Baevskij, V.S., Osipova L. Ya. (1974). *Issledovaniie stikhotvornogo ritma s ispol'zovaniiem EVM "Minsk-32"* [The Study of Verse Rhythm by Means of Computer]. Structural and Mathematical Linguistics, Kiev, Iss.2, 11 – 19.

Baevskij, V.S., Osipova L. Ya. (1974). *Algoritm i nekotoryie rezul'taty statisticheskogo issledovaniia al'terniruiushchego ritma na EVM "Minsk-32" // Mashinnyi perevod I prikladnaia linguistika.* [The Algorithm and the Results of Statistic Study of Alternating Rhythm by Means of Computer "Minsk-32" // Machine Translation and Applied Linguistics]. Moscow, Iss. 17. P. 174 – 195. (Republished: Glottometrica. 8. 1987 / Ed. L. Fickermann / Quantitative Linguistics. Vol. 32, 157 – 177).

Baevskij, V.S. (1975). *Tipologiia stikha russkoi liricheskoi poezii.* [The Typology of Verse in Russian Lyric Poetry]. Doctoral dissertation. Tartu, 166 – 231.

Baevskij, V.S. (1997). *Viacheslav Alexandrovich Sapogov.* Philologicheskiie nauki. Feb. P. 126 – 127. Republished in: Severo-Zapad, Iss.3. Vol. dedicated to V.A. Sapogov's memory. Cherepovets, 2000, 4 – 6.

Baevskij, V.S. (2001). Linguisticheskiie, matematicheskiie, semioticheskiie i komp'iuternye modeli v istorii i t'eorii literatury. [Linguistic, Mathematical, Semiotic and Computer Models in the History and Theory of Literature]. Ch.8. Moscow, *Yazyki slavianskoi kul'tury, 152 -172, 307 – 309.*

Baevskij, V. (2011). Academician Andrey Nikolaevich Kolmogorov as a Scholar of Verse Theory. *Glottometrics 22, 17 – 43.*

Baevskij, V.S. (2012). Shtrikhi k portretu: iz pisem Mikhaila Leonovicha Gasparova. [Some Strokes to the Portrait: from the Letters of Mikhail Leonovich Gasparov]. *Znamia. 2012. Iss. 2, 147 – 154.*

Kemball, R. (1965). *Alexander Blok. A Study in Rhythm and Metre.* The Hague.

Kolmogorov, A.N., Kondratov, A.M. (1962). Ritmika poem Maiakovskogo. [The Rhythmic System of Maiakovskii's Narrative Poems]. *Voprosy iazykoznaniia 3, 62 – 74.*

Meletinskii, E.M.[2] (1997). Piotr Alexandrovich Rudnev. *Philologicheskiie nauki. Iss. 3, 124 – 125.*

Potebnya, A.A. (1865). O zvukovykh osobennostiakh russkikh narechii. [On the Phonic Particularities of Russian Adverbs]. *Philologicheskiie zapiski, Iss.1.* Voronezh.

[2] The editorial staff of: “Philologicheskie nauki” explained to me that they could not publish two obituaries written by one and the same author. So I had to ask E.M. Meletinskii who kindly permitted to sign the second obituary by his name.

Glottometrics 24, 2012, 12-24

Vocabulary Growth of Content Words in ESP and General English.
A Contrastive Study Based on CMTE and SBNC

Zhao Xiaodong[1]
School of Foreign Languages
Dalian Maritime University

Abstract. This paper, based on Corpus of Maritime Transportation English (CMTE) and sampled British National Corpus (SBNC), employs FoxPro programs and SPSS analysis to study the dynamic growth patterns of words and content words of English for Specific Purposes (ESP) and general English at 4000-word intervals. Then it is tested in the paper whether Brunet's model can provide a good fit for the overall vocabulary growth of CMTE and SBNC, and whether this model is fit for describing the relationship between the vocabulary of content words and text length with the increase of tokens at 4000-word intervals. Lastly, the 95% confidence interval for content words in CMTE and general English is calculated.

Results of the study show that with the increase of cumulative tokens CMTE and SBNC exhibit a similar pattern of overall vocabulary increase, and the vocabulary increase curves of content words in the two corpora are also quite similar, with nouns increasing more rapidly than other content words. The difference is in SBNC overall number of words and content words increase more and more rapidly than those of CMTE, which means general English has greater vocabulary sizes of nouns, verbs, adjectives and adverbs. In addition, the vocabulary increase rate of SBNC tends to level with that of CMTE when the cumulative number of tokens reaches about 680000; the net increase of verbs in SBNC tends to slow down after the number of tokens reaches 350000. And in both general English and ESP, there is more inter-textual verb repetition, but less inter-textual adjective repetition. SPSS regression analyses show that Brunet's model can capture the vocabulary growth patterns of CMTE and the growth patterns of content words in CMTE and SBNC as well, with the determination coefficients (R^2) all close to 1.

Keywords: vocabulary growth, Corpus of Maritime Transportation English, content words, 95% confidence interval

1. Introduction

Many scholars (Altmann & Wagner, 1992; Baayen, 2001; Brunet, 1978; Fan, 2006, 2008a, 2010; Guiraud, 1954; Herdan, 1964; Köhler & Martináková, 1998; Somers, 1959; Tuldava, 1995) have studied the relationship between vocabulary and text length. They have either designed different quantitative models to describe vocabulary-text relationship, or they have tested these models by using different language data. Altenberg (1990), Francis and Kučera (1982), Johansson and Hofland (1989) have ever made static analyses of content words by calculating the proportions of content words in corpora LLC, Brown and LOB. Yet there are few studies on the dynamic vocabulary-text relationship of ESP (English for Specific Purposes) or vocabulary growth patterns of content words of general English and ESP. So this paper employs a quantitative method to make a dynamic study on the vocabulary growth of ESP and growth patterns of content words of ESP and general English at 4000-word intervals.

[1] Address correspondence to: dmuzhao@yahoo.com.cn

Since Fan's studies (2006, 2008a), which are based on a large scale of language data, reveal that Brunet's model can provide a very good fit for general English inter-textual vocabulary growth, this study also intends to analyse: (1) whether Brunet's model

$$V = \alpha(lnN)^{\beta} \quad \text{(Brunet, 1978)}$$

can capture the vocabulary growth features of ESP too; (2) whether this model can appropriately describe the relationship between the vocabulary of content words and text length for ESP as well as general English with the increase of tokens at 4000-word intervals.

Vocabulary growth is examined through the cumulative increase of word types against the cumulative increase of tokens. In this paper, vocabulary growth patterns of content words – nouns, verbs, adjectives and adverbs – are studied, whereas function words are excluded since there are far fewer function words(types), and they may not display an increasing pattern. Many linguists have noticed the sensitivity of type/token ratio (*TTR*) to the number of tokens (Guiraud, 1954; Orlov, 1982; Sichel, 1986; Holmes, 1994), and *TTR* is also calculated to measure lexical variation or lexical diversity (Malvern et al, 2004; Read, 2000). There are different ways to calculate *TTR*. Köhler and Galle (1993) employ the formula

$$TTR_x = \frac{t_x + T - \frac{xT}{N}}{N},$$

Laufer and Nation (1995) and Biber et al. (2000) employ the formula

$$TTR = 100 \times \frac{number\ \ of\ \ types}{number\ \ of\ \ tokens},$$

while Baayen (2001) uses

$$TTR = \frac{number\ of\ tokens}{number\ of\ types}.$$

Scott (1996) devises standardized *TTR*. This paper uses the formula

$$TTR = \frac{cumulative\ number\ of\ \ types}{cumulative\ number\ of\ \ tokens}$$

to work out the *TTR*s of general English and ESP. Standardized *TTR* is calculated on a 4,000-word basis, i.e., *TTR = number of types per 4,000 words/4,000.*

In this paper, English word tokens include all forms of English words, letters and abbreviations, etc. separated by spaces, but excluding punctuation marks. English word types refer to lemma types. That is, different word forms with the same sense, the same word class, but different inflections will be categorized into one same lemma type. Under this definition, *break*, *breaks*, *broke*, *breaking* and *broken* will be grouped into one lemma: *break*. In the process of lemmatization, Arabic numerals, punctuation marks and other non-alphabetic characters will all be excluded. In this paper, *type* and *vocabulary* have the same sense with *lemma*, and are used interchangeably.

2. Research design and methodologies

In this research Corpus of Maritime Transportation English (hereafter referred to as CMTE) is employed. CMTE is a corpus constructed by Dalian Maritime University in 2010. The running size of it is 1092258. The source texts of CMTE corpus are authoritative and representative enough, whose authors are all native English speakers, and most of these authentic texts were published in more than 30 maritime English journals or magazines between the year of 1995 and 2010. The sampled source texts are mostly between 2000 words and 5000 words, since according to Biber (1990:261) text samples of 2000 words to 5000 words can represent the linguistic characteristics of certain text categories. This corpus contains written English texts on marine transportation, covering various domains, such as port facilities, hazardous cargo, piracy, maritime logistics, maritime transportation, port transportation, shipping technology, nautical climate, marine incident investigation report, marine insurance, and various conventions, regulations or rules on ship collision, marine pollution, salvage, bills of lading, cargo handling, and marine search and rescue, etc. These texts cover various types of vessels, such as passenger ships, tankers, ro-ro ships, tugboats, ocean liners, freighters, ferries, fishing vessels, cruise ships, container ships, etc. In today's world economy, two thirds or more of the volume of trade is done by means of seaborne transportation. So it is meaningful to study the lexical features of CMTE as a kind of ESP.

The contrast corpus is a sampled corpus of written British National Corpus (BNC). First, a FoxPro program is used to remove all the tags in the tagged texts of BNC (written) corpus. Then another FoxPro program is used to draw a random sample of 28 texts from BNC corpus, with a total number of 1136347 words. Hereafter, the sampled BNC corpus is referred to as SBNC. Then the two corpora, CMTE and SBNC, are tagged with POS tags by using CLAWS4. A third FoxPro program is used to tokenize the two corpora and extract different classes of content words, i.e. nouns, verbs, adjectives and adverbs according to the POS tags, then all the tags of part of speech, punctuations, typographic signs and other non-alphabetic characters are removed.

Next, two FoxPro programs are employed to process the cleaned tokenized CMTE and SBNC respectively. These programs first randomly divide CMTE into 273 4000-word(token) chunks and SBNC into 284 4000-word chunks, then compute the number of word types for each 4000-word chunk, the cumulative number of word types, word tokens, as well as the TTR, standardized TTR, net increase of types, and the vocabulary growth data of four categories of content words at 4000-word intervals.

Finally, SPSS (16.0) is employed to test the fit of Brunet's model for the vocabulary growth of total number of words as well as that of content words in the two corpora.

3. Results and analysis

3.1 Vocabulary growth rates of SBNC and CMTE

Results show that the total number of vocabulary for the 284 SBNC chunks and 273 CMTE chunks is 37739 and 34566 respectively. The average number of vocabulary for each chunk is 1598 and 1567 respectively, which indicates on average a text of 4000 words in general English produces about 30 more word types than ESP. The mean standardized TTR for SBNC is 0.3995, and CMTE 0.3919, which indicates that SBNC has a greater lexical diversity than CMTE. This is testified by curves in Figure 1 below.

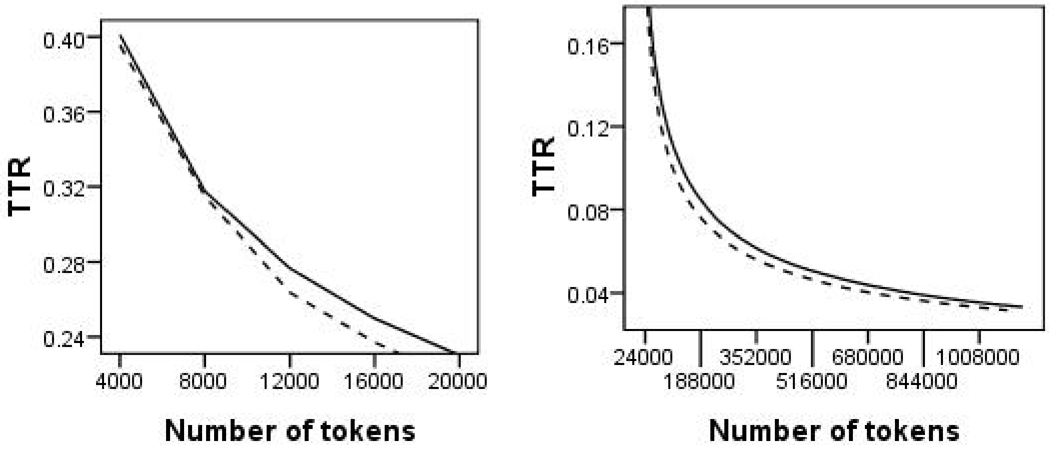


Figure 1. Left: *TTR* decrease curves before 20000 word tokens with the minimum scale of Y-axis set to 0.22, and maximum 0.40. The solid line is the decrease curve of SBNC, the dotted line CMTE.
Right: *TTR* decrease curves from 24000 word tokens downwards, with the minimum and maximum scales of Y-axis set to 0.03 and 0.17 respectively. The solid line is the decrease curve of SBNC, the dotted line CMTE.

Figure 1 shows that *TTR* curves of the two corpora both display a declining pattern. The decrease curve of SBNC has always been above that of CMTE and there has never been a touch point or cross point for the two curves. This suggests that the lexical diversity of SBNC has always been greater than that of CMTE.

Figure 2 is the growth curve of overall word types.

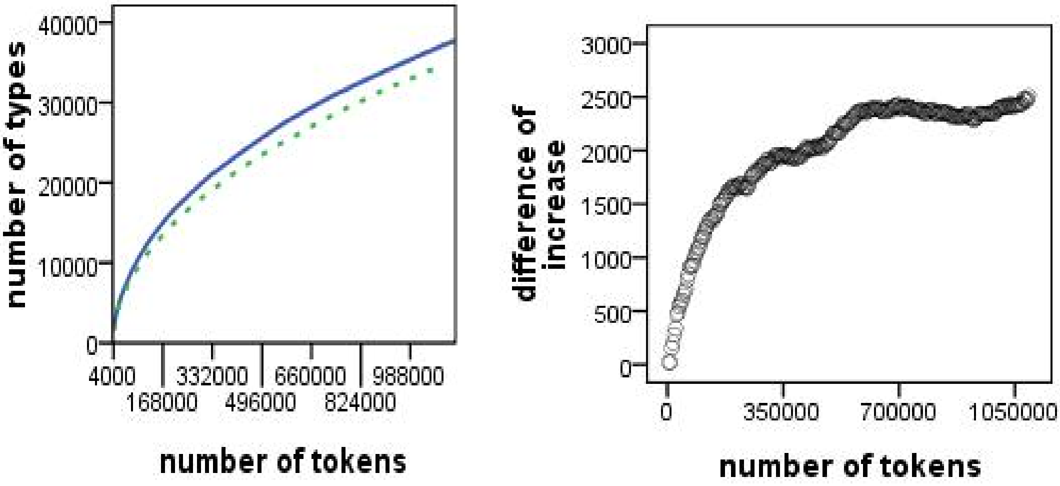


Figure 2 Left: growth curves of word types. The solid line is the growth curve of SBNC, the dotted line CMTE. Right: scatter plot of vocabulary increase differences between SBNC and CMTE.

Figure 2 (Left) shows that word types in SBNC and CMTE present a similar growth pattern: word types keep increasing until the end of the curves with the increase of cumulative tokens. Both of the curves rise sharply at the initial stage. Yet with the increase of tokens, the growth rate of types gradually slows down. This finding is similar to those research findings made by Baayen (2001) and Fan (2008a, 2008b, 2010). Despite the similarities, there are also obvious differences between the two curves: the growth curve of SBNC word types has always been above that of CMTE. This means for any text of similar size, general English has a larger vocabulary than ESP. In order to detect how larger the cumulative vocabulary of SBNC is than that of CMTE, we calculate the vocabulary differences for each pair of chunks (altogether 273 pairs of chunks) by subtracting the cumulative vocabulary number of CMTE from that of SBNC, the results of which are shown in Figure 2(Right). Figure 2(Right) gives a dynamic description of vocabulary differences between SBNC and CMTE with the increase of cumulative number of tokens. It shows that vocabulary differences between SBNC and CMTE fall between 0 to about 2500. With the increase of tokens, the value of differences becomes bigger and bigger, and it reaches a high point (about 2400) when the number of tokens reaches about 680000 and maintains at that level with minor fluctuations. That means, SBNC vocabulary increases more and more rapidly than does CMTE vocabulary, but from 680000 word tokens downwards, the growth rate of SBNC tends to level with that CMTE.

To take a closer look at the difference of vocabulary increase between SBNC and CMTE, we can examine the net increase of types for the 284 chunks of SBNC and 273 chunks of CMTE, as is shown in Figure 3 (Left). Figure 3 (Left) shows that with the increase of cumulative number of tokens, the net increases of types in those chunks of the two corpora both experience a sharp-slow decrease. In order to have a detailed account of the differences between the net increase of SBNC and CMTE, we work out the differences of net increase for each pair of chunks by subtracting the number of net increase in CMTE from that of SBNC, as is shown in the following scatter plot, Figure 3 (Right).

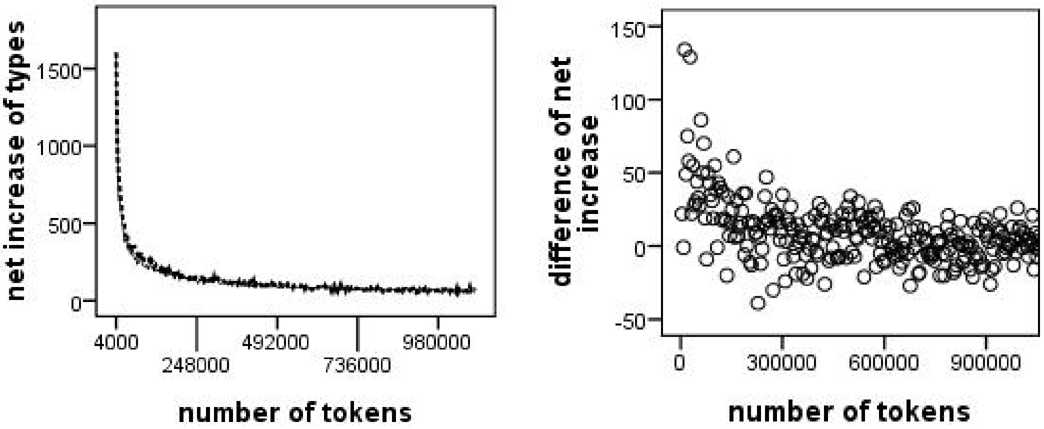


Figure 3 Left: decrease curves of net increase of types. The dotted line is the net increase of types in CMTE, the solid line SBNC. Right: differences of net increase between SBNC and CMTE for 273 pairs of chunks.

From Figure 3 (Right), we can find that the differences of net increase between SBNC and CMTE for 273 pairs of chunks mostly lie between -25 to 40, with SBNC taking the majority of larger values and the range of fluctuation being not so wide.

3.2. Vocabulary growth of content words

In both SBNC and CMTE, four types of content words exhibit similar growth patterns: nouns increase more sharply than adjectives, verbs and adverbs, indicating that in the two corpora the majority of vocabulary goes to nouns. The growth rate of adjectives ranks second, that of verbs ranks third, and adverbs the last. It also shows (see Figure 4) that there are far fewer verb types and adverb types 4. The growth curves of adverbs in the two corpora tend to level with the axis, which indicates that adverbs keep a steady increase in small amount.

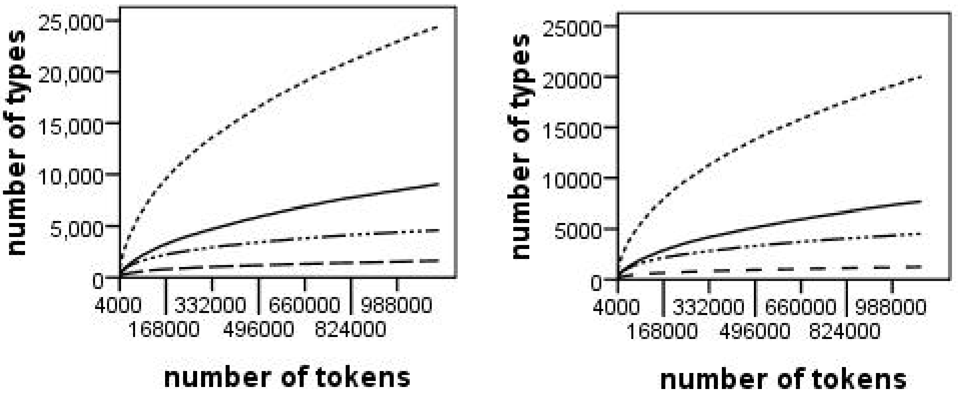


Figure 4. Left: growth curves of content words in SBNC. Curves from top to bottom are noun growth, adjective growth, verb growth and adverb growth. Right: growth curves of content words in CMTE. Curves from top to bottom are noun growth, adjective growth, verb growth and adverb growth.

Another difference is that content words in SBNC seem to increase more rapidly than those of CMTE. In order to verify this, we can compute the vocabulary differences of content words for each pair of chunks between SBNC and CMTE by subtracting the cumulative number of content words(types) in CMTE from that of SBNC with the increase of cumulative tokens, the result of which is shown in Figure 5.

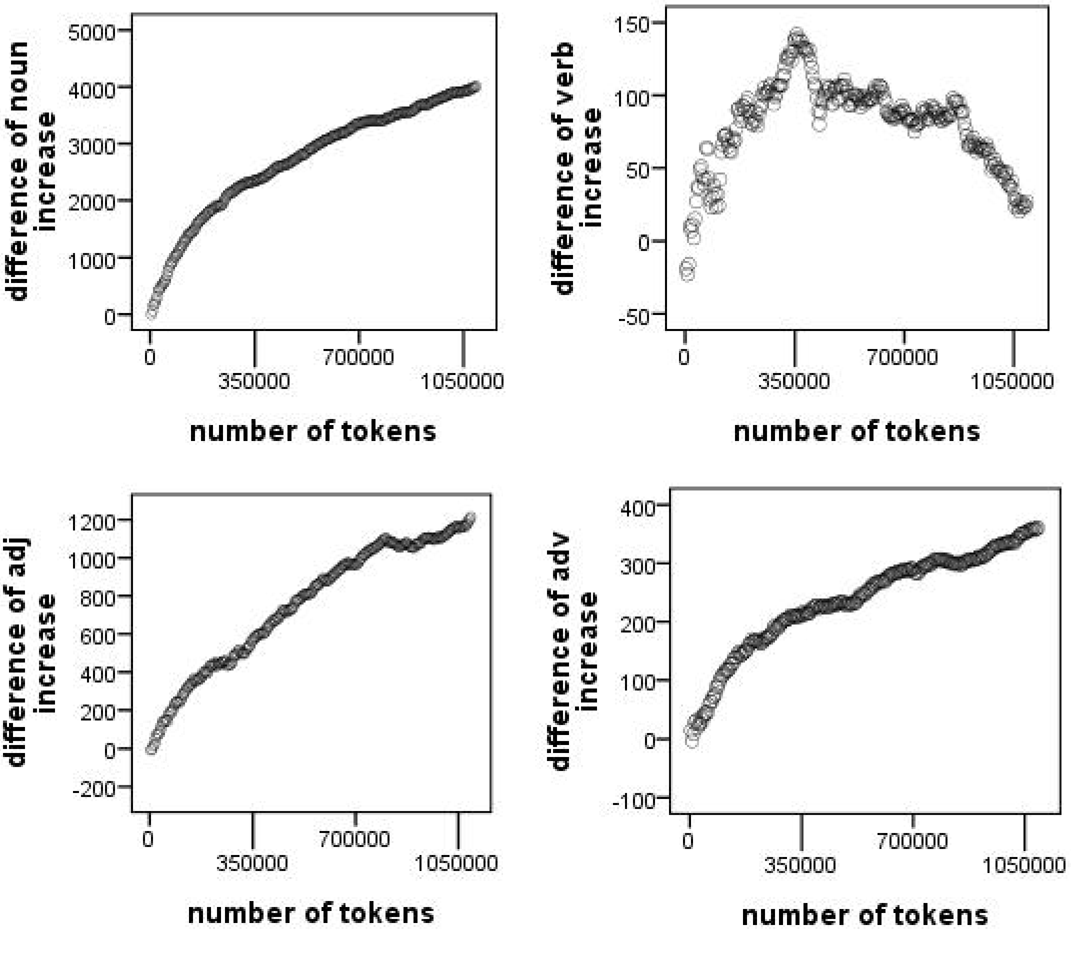


Figure 5. Vocabulary differences of content words for each pair of chunks between SBNC and CMTE with the increase of cumulative tokens.

Figure 5 shows that vocabulary differences of nouns, adjectives and adverbs between SBNC and CMTE all undergo a dynamic increasing pattern with the increase of cumulative number of tokens. And these values of differences are mostly positive and become bigger and bigger, with noun differences between 0-4000, verb differences between -25 to 150, adjectives about 0-1200, adverbs about 0-360. This suggests that in sampled BNC vocabularies of content words all go through a more and more rapid increase than those in CMTE. The exception is that the curve of verb differences between SBNC and CMTE undergoes an upside-down v-turn, with the apex stopping at 150 when the cumulative number of tokens reaches about 350000. After observing the data, we find this is mainly due to the slowdown of net increase of verbs in SBNC and greater net increase of verbs in CMTE. This means in SBNC the increase of verbs is limited after the cumulative number of tokens reaches 350000.

Features of the vocabulary growth of content words can also be clearly shown from the decrease curves of the net increase of content words in the two corpora (see Figure 6).

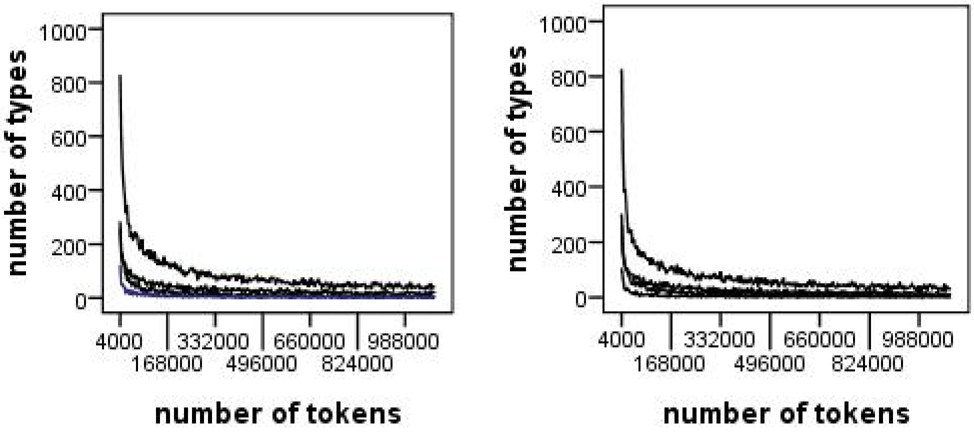


Figure 6. Left: decrease curves of the net increase of content word types in SBNC. Curves from top to bottom are net increases of nouns, adjectives, verbs and adverbs. Right: decrease curves of the net increase of content word types in CMTE. Curves from top to bottom are net increases of nouns, adjectives, verbs and adverbs.

Figure 6 shows that nouns have greater net increase from the initial stage to about 500000 cumulative word tokens. The decrease curves of the other 3 kinds of content words run very close. In other words, the vocabulary growth rates of adjectives, verbs and adverbs are much slower, with the decrease curves of adverbs roughly falling flat from almost the initial stage.

3.3. Fit of Brunet’s model to the vocabulary growth of CMTE and content words in the two corpora

Based on the observed data – the cumulative number of types in the two corpora, this paper employs SPSS regression analysis to work out Brunet’s model’s estimated values of SBNC and CMTE vocabulary growth at a 4000-word interval, and then the fitted curves are plotted, as shown in Figure 7. In addition, coefficients of determination (R^2) for the two corpora are also obtained through SPSS analysis.

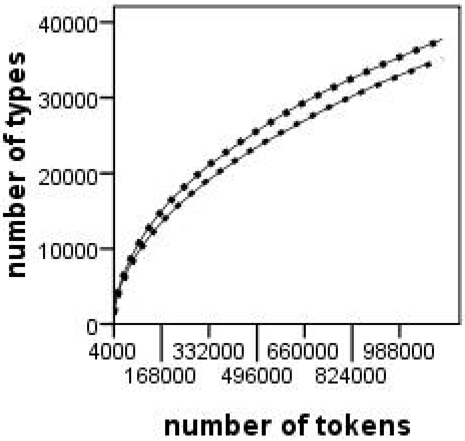


Figure 7. Fit of Brunet’s model to SBNC and CMTE. The upper solid line is the growth curve of types for observed SBNC, the upper dotted line Brunet’s model fit for SBNC. The lower solid line is the growth curve of types for observed CMTE, the lower dotted line Brunet’s model fit for CMTE.

Figure 7 shows that Brunet’s model provides a very good fit for the observed vocabulary growth of the two corpora. In regression analysis, R^2 is often used to test the fit of regression model to observed data, and the larger the coefficient of determination, the more adequate the model is. R^2 of Brunet’s model for observed SBNC and CMTE is respectively 0.999960 and 0.999737, which indicates that Brunet’s model can accurately reflect the vocabulary growth of SBNC and CMTE.

Next, the fit of Brunet’s model to the vocabulary growth of content words in the two corpora is analyzed. Figure 8 shows the fitted curves for nouns and verbs in SBNC and CMTE.

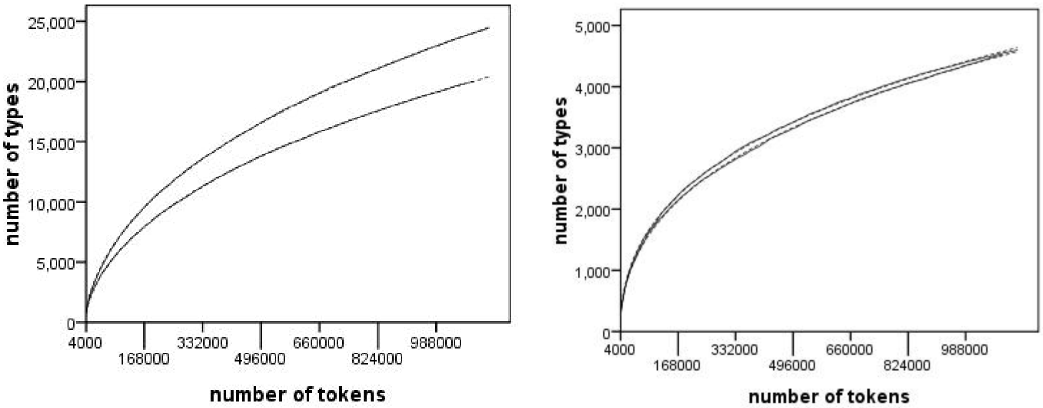


Figure 8. Left: fit of Brunet’s model to increase of nouns. The upper solid line is the growth curve of nouns for observed SBNC, the upper dotted line Brunet’s model fit for SBNC nouns. The lower solid line is the growth curve of nouns for observed CMTE, the lower dotted line Brunet’s model fit for CMTE nouns. R^2 for SBNC = 0.999945; R^2 for CMTE = 0.999955. Right: fit of Brunet’s model to increase of verbs. The upper solid line is the growth curve of verbs for observed SBNC, the upper dotted line Brunet’s model fit for SBNC verbs. The lower solid line is the growth curve of verbs for observed CMTE, the lower dotted line Brunet’s model fit for CMTE verbs. R^2 for SBNC = 0.999679; R^2 for CMTE = 0.999805.

Figure 8 shows that Brunet’s model provides a perfect fit for the vocabulary growth of nouns and verbs in both SBNC and CMTE. The coefficient of determination (R^2) of Brunet’s model for observed SBNC and CMTE nouns is 0.999945 and 0.999955 respectively, for observed SBNC and CMTE verbs, 0.999679 and 0.999805 respectively. The following figure (Figure 9) shows the fit curves for adjectives and adverbs in SBNC and CMTE.

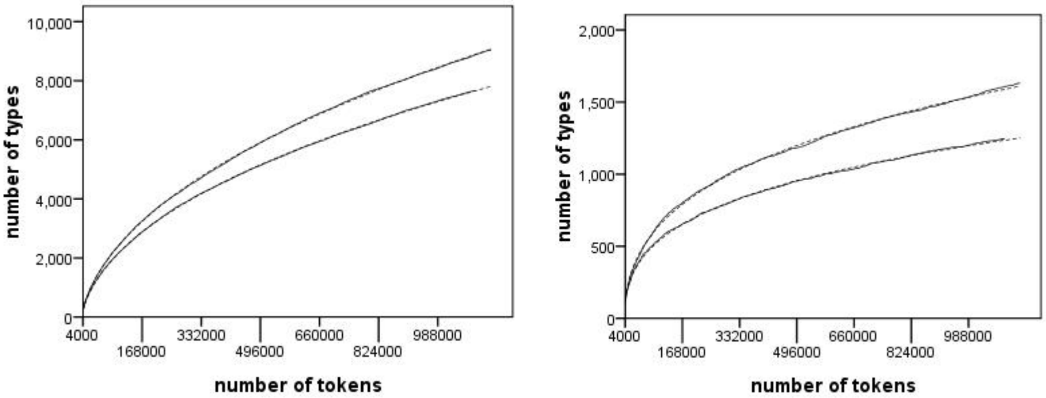


Figure 9. Left: fit of Brunet's model to increase of adjectives. The upper solid line is the growth curve of adjectives for observed SBNC, the upper dotted line Brunet's model fit for SBNC adjectives. The lower solid line is the growth curve of adjectives for observed CMTE, the lower dotted line Brunet's model fit for CMTE adjectives. R^2 for SBNC = 0.999889; R^2 for CMTE = 0.999900. Right: fit of Brunet's model to increase of adverbs. The upper solid line is the growth curve of adverbs for observed SBNC, the upper dotted line Brunet's model fit for SBNC adverbs. The lower solid line is the growth curve of adverbs for observed CMTE, the lower dotted line Brunet's model fit for CMTE adverbs. R^2 for SBNC = 0.999060; R^2 for CMTE=0.999143.

Figure 9 shows that predicted adjective and adverb values of Brunet's model match those observed values of adjectives and adverbs very well in both SBNC and CMTE. For the increase of adverbs, although coefficients of determination of Brunet's model for observed data of the two corpora are both close to 1, the fit curves are the least perfect of all the four types of content words.

3.4. Calculation of content words in a text at 95% confidence level

We employ the following formula for tolerance interval:

*Tolerance Interval = mean ± (tolerance critical value)*S* (Devore, 2000)

to capture 95% of all the possible values of the vocabulary sizes in texts of about 4000 words. Significance of doing this is that it enables us to predict the numbers of content words in both general English and ESP (CMTE).

In the formula, *mean* stands for the average number of types in those 4000-word chunks for each kind of content words; *tolerance critical value* is also a given value, which is determined by the number of observed values and the percentage of all such possible values the tolerance interval intends to include. *S* is the standard deviation of a set of normally distributed values.

By using SPSS, we can work out the means and standard deviations for nouns, verbs, adjectives and adverbs for those 4000-word chunks in the two corpora, the result of which is shown in Table 1. At the same time, the vocabulary distribution histograms of four types of content words can be drafted.

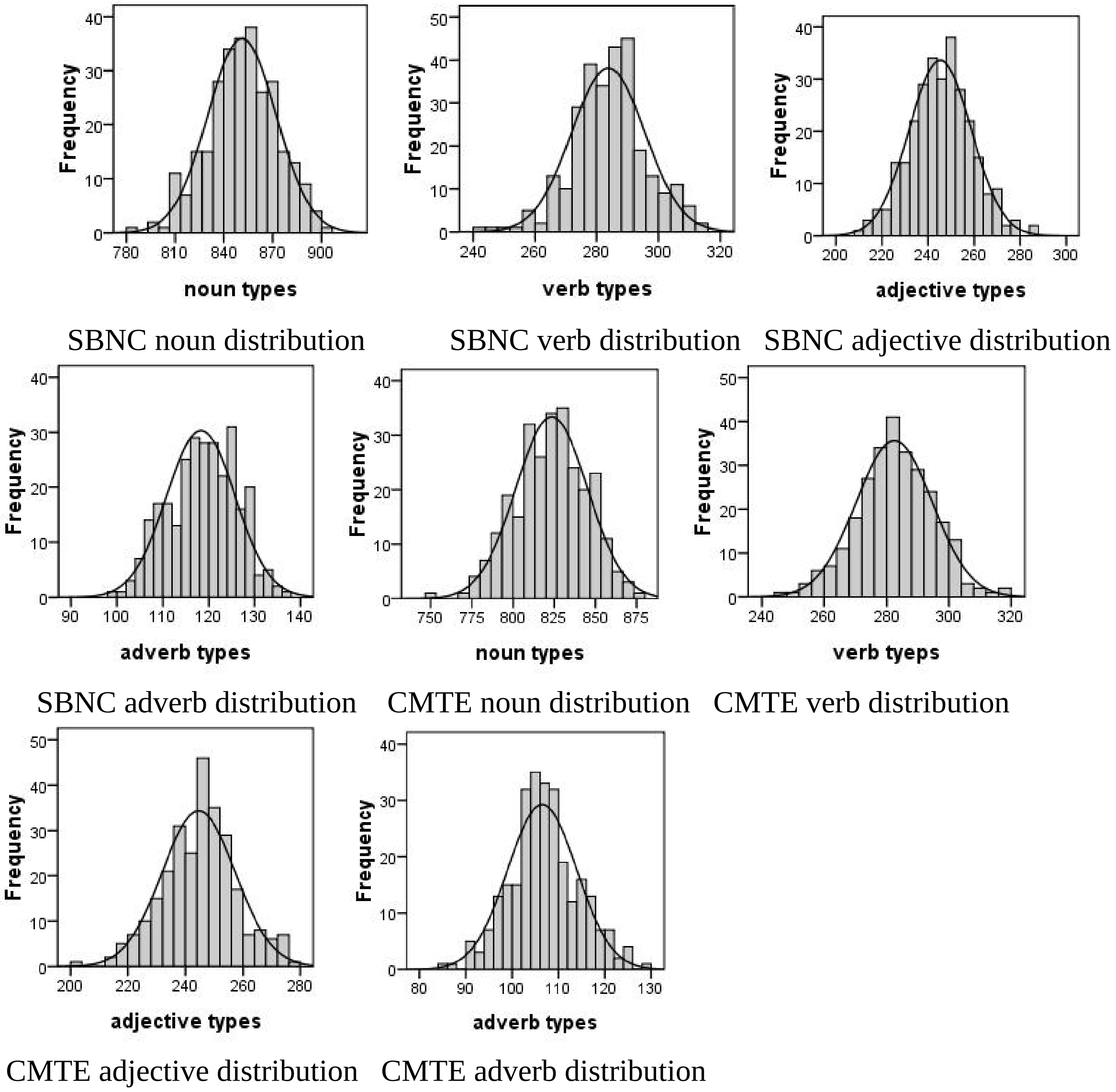
Table 1

Means and standard deviations for four types of content words

Corpus	Statistics	Nouns	Verbs	Adjectives	Adverbs
SBNC	Mean	851	284	245	118
	Std. Dev.	21.024	11.904	13.473	7.473
CMTE	Mean	823	282	245	107
	Std. Dev.	21.742	12.225	12.697	7.443

Observing the table, we may find that the means of verbs in the two corpora are higher than those of adjectives, or in other words, for each text of 4000 words, there are more verb types than adjective types. This is contradictory to the cumulative increase of content words, where the cumulative number of adjectives is always higher than that of verbs with the cumulative increase of tokens (*cf*. Figure 4). This suggests that in both general English and ESP, there is more inter-textual verb repetition, but less inter-textual adjective repetition.

The vocabulary distribution histograms of four types of content words are as follows:



SBNC noun distribution SBNC verb distribution SBNC adjective distribution

SBNC adverb distribution CMTE noun distribution CMTE verb distribution

CMTE adjective distribution CMTE adverb distribution

Figure 10. Vocabulary distribution histograms of content words in SBNC and CMTE

Figure 10 shows that nouns, verbs, adjectives and adverbs are all normally distributed in those 4000-word chunks in the two corpora. Then the one-sample Kolmogorov-Smirnov Test is used to test if the vocabulary distributions of content words in the two corpora are normally distributed. Results show that Kolmogorov-Smirnov Z values for nouns, verbs, adjectives and adverbs in CMTE and SBNC are 0.577, 0.661, 0.884, 1.163 and 0.673, 1.012, 0.630, 0.960. The asymptotic significance (2-tailed) values are 0.893, 0.775, 0.416, 0.134 and 0.755, 0.257, 0.822, 0.315 respectively, which are all greater than 0.05. So statistical analyses also show that nouns, verbs, adjectives and adverbs are all normally distributed in the two corpora.

To capture 95% of all the possible values of a normally distributed population at the 95% confidence level, the tolerance value is 1.96 for the number of observed values bigger than 120 (Butler, 172). Therefore, we can estimate the upper bounds and lower bounds of the vocabulary size of different content words in texts whose sizes are about 4000 word tokens. Based on the statistics in Table 1 and the given value of tolerance critical value, the 95% confidence interval for distribution of content words in SBNC and CMTE can be worked out (Table 2).

Table 2
95% confidence intervals for content words

Estimated Vocabulary Intervals		Nouns	Verbs	Adjectives	Adverbs
SBNC	Upper bounds	892	307	271	133
	Lower bounds	810	261	219	103
CMTE	Upper bounds	866	306	270	122
	Lower bounds	780	258	220	92

So for general English, there is a 95% probability that 95% of the nouns of a text of about 4000 words will lie between 810 and 892; for verbs, between 261 and 307; adjectives, between 219 and 271; adverbs, 103 and 133. For CMTE, there is a 95% probability that 95% of the nouns of a text of about 4000 words will lie between 780 and 866; for verbs, between 258 and 306; adjectives, between 220 and 270; adverbs, 92 and 122.

The significance of these findings is that by using the same method, we can predict the vocabulary of content words in texts of different sizes, such as texts of about 1000 words or 2000 words, etc.

Since the number of function words is limited, the complexity and diversity of a text mainly depends on the vocabulary or the number of content words, we can estimate the size of content words (types) of a text, predicting the lexical diversity of different types of content words of a certain text. This is very important in textbook compilation and choice of reading materials for students of different levels. Or in other words, in the choice of materials, the vocabulary sizes of content words for a text of certain size should fall within the lower bounds and upper bounds. It can also be used to assess language learners' English proficiency --- the acquisition of different types of content words in second language learners' written or spoken English, that is, to find how much difference there is between English language learners and native speakers with respect to the number of content words used. It is also applicable to text categorization or authorship attribution, according to which we can calculate numbers of nouns, verbs, adjectives and adverbs in chunks of different texts and pick out each kind of content words used by the author.

4. Conclusion

This study finds that the overall vocabulary increase and the vocabulary increase of content words in CMTE display a very similar pattern to those of general English, SBNC. The difference is that the overall number of words and content words in general English increase more and more rapidly than those of CMTE, which indicates general English has greater vocabulary sizes of nouns, verbs, adjectives and adverbs. In addition, the vocabulary increase rate of SBNC tends to level with that of CMTE when the cumulative number of tokens reaches about 680000; CMTE verbs experience greater net increase than do general English verbs after the number of tokens reaches 350000. And majority of vocabulary in the two corpora goes to nouns. In both general English and ESP, there is more inter-textual verb repetition, but less inter-textual adjective repetition. SPSS regression analysis shows that Brunet's model can capture the vocabulary growth of both general English and ESP very well, and this model also provides a perfect fit for the vocabulary growth of content words in both SBNC and CMTE. According to the analysis of content words of a text at 95% confidence level, we find that the estimated number of nouns in CMTE is smaller than that of SBNC, whereas the estimated numbers of verbs in the two corpora are similar, so are adjectives and adverbs.

References

Altenberg, B. (1990). Spoken English and the dictionary. In: Svartvik, J. (ed.), *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund Studies in English 82. Lund University Press.

Altmann, G., Wagner, R.K. (1992). Das Type-Token Verhältnis in der Kindersprache. In Wagner, K. B. (ed.), *Kindersprachstatistik: 47-56*. Essen: Blaue Eule.

Baayen, R.H. (2001). *Word Frequency Distribution*. Dordrecht: Kluwer Academic Publishers.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing 5, 257-269.*

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (2000). *Longman grammar of Spoken and Written English*. Beijing: Beijing Foreign Language Teaching and Research Press.

Brunet, E. (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution*. Genève: Slatkine.

Butler, C. (1985). *Statistics in Linguistics*. Oxford: Basil Blackwell Ltd.

Devore, J. (2000). *Probability and Statistics*. Pacific Grove: Brooks/Cole.

Fan, F. (2006). Models for Dynamic Inter-textual Type-token Relationship. *Glottometrics 12, 1-10*

Fan, F. (2008a). A corpus-based study on random textual vocabulary coverage. *Corpus Linguistics and Linguistic Theory 4(1), 1-17.*

Fan, F. (2008b). Hapax Legomena and Language Typology, a Case Study. In: E. Kelih, V. Levickij & G. Altmann (eds.), *Methods of Text Analysis: Omnibus Volume*. Chernivtsi: Chernivtsi National University.

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

Francis, W.N., Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

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

Holmes, D. (1994). Authorship attribution. *Computer and the Humanities 28(2), 87-106.*

Johansson, S., Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar* 2 vols. Oxford: Clarendon Press.

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

Köhler, R., Galle, M. (1993). Dynamic aspects of text characteristics. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 46-53*. Trier: WVT.

Köhler, R., Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W.A. (eds.), *Systems*: *New paradigms for the human sciences*: *514-546*. Berlin: de Gruyter.

Laufer, B., Nation, I. (1995). Vocabulary Size and Lexical Richness in L2 Written Production. *Applied Linguistics 16(4): 307-322.*

Malvern, D., Brian, R., Ngoni, C., Pilar, D. (2004). *Lexical diversity and language development: quantification and assessment*. New York: Palgrave Macmillan.

Orlov, J. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen*: *118-192*. Bochum: Brockmeyer.

Read, J. (2000). *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Scott, M. (1996). *Wordsmith Tools*. Oxford: Oxford University Press.

Sichel, H. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist 11, 45-72.*

Somers, H. (1959). *Analyse mathématique du langage: Lois générales et mesures statistiques*. Louvain: Nauwelaerts.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Trier: WVT

Glottometrics 24, 2012, 25-40

Stratification in musical texts based on rank-frequency distribution of tone pitches[1]

Ioan-Iovitz Popescu
Zuzana Martináková-Rendeková
Gabriel Altmann

Abstract. In this study we investigate the stratification in musical texts based on rank-frequency distribution of tone pitches. Preliminary investigations show that there are some similarities between music and language and that pitches play the same semiotic role as phonemes or graphemes, occupying approximately the same "corridor" in the ternary plot. In this case we try to apply the stratificational approach and to investigate possible strata composed of pitches having different functions.

Key-Words: stratification, self-regulation, musical texts, rank-frequency distribution, tone pitch

1. Introduction

Stratification is a common phenomenon in nature and culture. Homogeneity is always only partial or something perceived from a certain point of view. Real things consist of different elements and different subsystems. This holds for the most abstract human creation like language or arts down to the subatomic world. However, in many cases analysis is possible only if we suppose homogeneity or stipulate the ceteris paribus condition. Advance in research means in many cases the stepwise relaxing of the homogeneity condition and admitting further variables.

In language, we speak of a homogeneous text if it has been written by one author, does not contain different parts, in the best case it has been written in one go, without any pause, and has not been corrected subsequently. Evidently, texts of this kind are very seldom, the best examples are letters. But even if they exist, homogeneity has different faces. And even if in a certain sense they are homogeneous, we nevertheless discover strata of units which behave differently. The best example is the frequency of words which can be (fuzzily) partitioned in "meaning-words" and "auxiliaries" both having quite different frequencies. But even meaning-words themselves can be subdivided in "nominals" and predicates of different order. If we rank them all together according to frequency, we put up with the fact that we mix at least two different strata controlled by two different frequency regimes. This is why B. Mandelbrot corrected Zipf´s approach but was himself "corrected" many times in the history: some formulas improve the fitting at the beginning of the rank sequence but impair it at the tail of the sequence or vice versa. Still worse is the case when the empirical ranking sequence is not smooth as a whole – though monotonously decreasing – and the fitting gets weaker. Another problem is the fact that the traditional formulas can be derived also from the assumption of randomness of texts (ape typing) without involving any kind of economy.

[1] A short summary of the article appeared in: *Proceedings of the 10th WSEAS International Conference on Acoustics and Music: Theory and Applications, Prague, March 2009, pp. 116-119.*

In musical texts – except monophonic[2] compositions or compositions for one voice or one instrument – the text is clearly stratified in polyphonic[3] or homophonic[4] compositions. If we then rank the frequencies of pitches, we possibly obtain a good result using some special case of the Lerch function (encompassing the Zipf and Zipf-Mandelbrot formulas) but this need not always be the case. Besides, Zipf´s ranking approach does not yield an explanation of linguistic facts, and Mandelbrot´s approach is restricted to linguistics (where it can be obtained also for random texts), hence the transfer to music is made *per analogiam* in spite of very different empirical entities. In cases like this, one tries to find other ways which would yield at least as good fitting as the Zipf-Mandelbrot approach but at the same time bring a kind of explanation. We shall show two such possibilities.

2. Up-and-down self-regulation

The great majority of linguistic distributions is based on the assumption that if frequencies are ranked, the probability of class x is proportional to that of x-1. The proportionality need not be constant, it may be a function of x. Thus we obtain

(1) $$P_x = g(x)P_{x-1}\,.$$

Writing $P_{x\text{-}1}$ in the same way, we obtain at last

(2) $$P_x = P_0\prod_{i=1}^{x} g(i)$$

or for ranking from x = 1

(2a) $$P_x = P_1\prod_{i=2}^{x} g(i)$$

which is being sufficient in most cases. Hence the most frequent pitch in music (x = 0 or x = 1) controls the other ones. But in a musical composition the key may change and we obtain as many layers as there are key changes. By changing the key both the frequencies of pitches and their ranks get in disorder and the simple control by the smaller ranks (or the smallest rank) is not sufficient. In order to avoid this restriction, Wimmer proposed the use of partial-sums distributions in which the lower ranks are controlled by higher ranks (cf. Wimmer, Šidlík, Altmann 1999, Wimmer, Altmann 2000, 2001). This is a very extensive family of distributions and practically every distribution can be used for this purpose in different forms. A survey of forms can be found in Johnson, Kotz, Kemp (1992). But even if here a kind of different control becomes effective, the possible stratification is not expressed.

[2] Monophony in music is the simplest of textures consisting of melody (one note at a time or the same note duplicated or multiplicated at one or more octaves – unisono) without accompanying harmony.

[3] Polyphony in music is a texture consisting of two or more independent melodic voices.

[4] Homophony in music is a texture in which two or more parts move together in harmony and the relationship between them creates chords.

3. Simple stratification

Here we shall restrict ourselves to possible strata composed of pitches having different functions. However, ranking the pitches according to their frequency means a mixing of strata, an inserting one stratum into another. In some kinds of music a stratum can be fully dependent on another, whereby their number gets reduced; in other kinds of music the voices are so independent that there are several ranked strata.

All ranking of frequencies goes back to G.K. Zipf (1935) who conjectured that $f \times r = C$ (frequency times rank is constant). Later on an exponent has been added leading to a power curve which disseminated to many scientific disciplines and holds true for the great majority of linguistic rank-frequency data. Its corrections are cosmetic improvements for individual data (e.g. $C/(a+x)^b$, $C/(a+x)^{f(b,x)}$, $Cq^x/(a+x)^b$,…) but in spite of the fact that all of them can be derived from different assumptions, they do not have a linguistic foundation. But if Zipf´s approach is valid and ranking is something inherent to linguistic data, there is a possibility to start from a still simpler assumption. The power curve is given by the differential equation $y'/y = -a/x$, but if there are strata in texts, one can take very generally $y'/y = -a$, yielding $y = Ae^{-ax}$ and attain a look at the strata by adding analogous components, i.e. $y = Ae^{-ax} + Be^{-bx}$… Alternatively, the result follows from a differential equation of first order. This approach has been proposed as an alternative to Zipf´s approach (cf. Popescu, Altmann, Köhler 2009) and can be applied to the study of musical texts. The above sum is a sum of geometric sequences, since e^{-ax} can be written as q^x but we shall use the following exponential form of writing

$$y = 1 + A_1 e^{-x/r_1} + A_2 e^{-x/r_2} + A_3 e^{-v/r_3} \qquad (3)$$

Now, consider the rank-frequency of pitches in Palestrina´s work Pls05 *Missa Ascendo ad Patrem, 5. Movement Sanctus* in which there are 23 different pitches occurring 595 times (Table 1). Considering it a geometric series we apply the first exponential component of (1) and obtain the result given in the third column of Table 1.

Table 1
Rank-frequency of pitches in Palestrina's Pls05I
Missa Ascendo ad Patrem, 5. Sanctus

x	f_x	One component
1	70	77.29
2	69	67.85
3	64	59.57
4	56	52.32
5	45	45.96
6	42	40.40
7	34	35.52
8	33	31.24
9	30	27.50
10	26	24.22
11	22	21.34
12	20	18.83
13	18	16.62

14	16	14.68
15	11	12.99
16	10	11.51
17	9	10.20
18	9	9.07
19	4	8.07
20	2	7.19
21	2	6.42
22	2	5.75
23	1	5.16

Here, the determination coefficient is R^2 = 0,9806 and the parameters are the unique range r_1 = r_2 = r_3 = 7.57 and the cumulated amplitude A_1 + A_2 + A_3 = 3(29.02), as it results from the first position of Table 2. Now, taking two exponential components of (1) we shall state that the fitting does not change, the parameters of both components are identical and R^2 is the same. Even if we add a third exponential component, nothing changes. At this point we can conclude that the given composition is monostratal, as graphically illustrated in Figure 1.

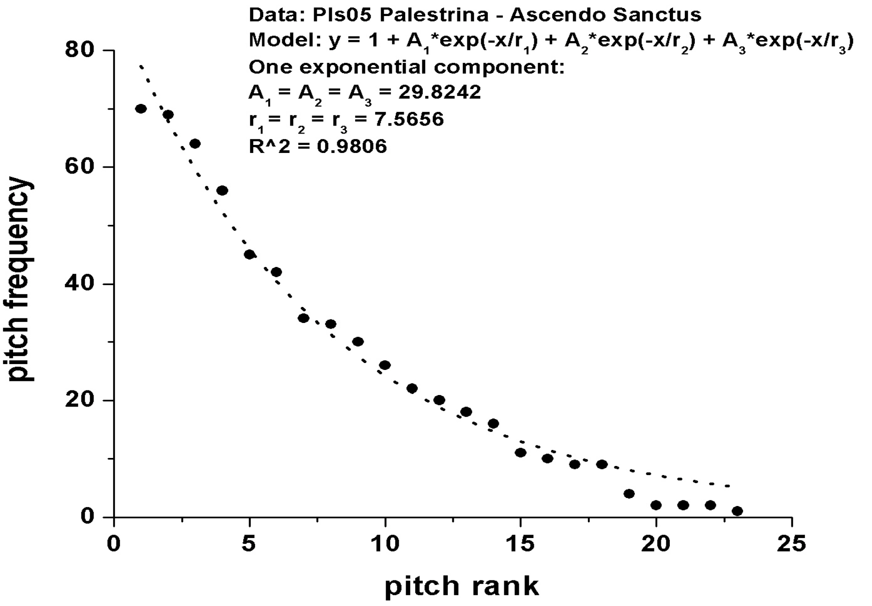


Figure 1. Fitting one exponential component to a work of Palestrina

Now, if we perform the same operations with 30 works of Palestrina that stay at our disposal, we shall state that all of them display this property: they are monostratal in the distribution of pitches as can be seen in Table 2. Here we computed the possibility of a composition having three strata, according to (1). As can be seen, the second and the third stratum are identical with the first. The procedure is as a matter of fact a test of stratification.

Table 2
The monostratal structure of 30 Palestrina´s works.
Ranking by R^2

ID	**Text**	A_1	r_1	A_2	r_2	A_3	r_3	R^2
Pls05	Missa Ascendo ad Patrem, Sanctus	29.02	7.57	29.02	7.57	29.02	7.57	0.9806
Pls20	Missa Papae Marchelli, 4. Sanctus	52.27	7.52	52.27	7.52	52.27	7.52	0.9787
Pls28	Missa Veni Sponsa Christi, 5. Benedictus	30.67	6.87	30.67	6.87	30.67	6.87	0.9782
Pls19	Missa Papae Marchelli 3. Credo	113.44	7.95	113.44	7.95	113.44	7.95	0.9777
Pls22	Missa Papae Marchelli 6. Agnus Dei I	34.58	7.56	34.58	7.56	34.58	7.56	0.9776
Pls18	Missa Papae Marchelli 2. Gloria	68.79	7.80	68.79	7.80	68.79	7.80	0.9731
Pls07	Missa Ascendo ad Patrem 7. Agnus Dei I	19.94	7.76	19.94	7.76	19.94	7.76	0.9711
Pls17	Missa Papae Marchelli 1. Kyrie	49.99	7.54	49.99	7.54	49.99	7.54	0.9698
Pls21	Missa Papae Marchelli 5. Benedictus	32.45	7.38	32.45	7.38	32.45	7.38	0.9665
Pls02	Missa Ascendo ad Patrem 2. Kyrie	40.46	8.34	40.46	8.34	40.46	8.34	0.9644
Pls03	Missa Ascendo ad Patrem 3. Gloria	58.61	8.69	58.61	8.69	58.61	8.69	0.9608
Pls01	Missa Ascendo ad Patrem 1. Motet	80.49	8.75	80.49	8.75	80.49	8.75	0.9605
Pls24	Missa Veni Sponsa Christi 1. Kyrie	29.96	8.29	29.96	8.29	29.96	8.29	0.9600
Pls08	Missa Ascendo ad Patrem 8. Agnus Dei II	21.18	8.55	21.18	8.55	21.18	8.55	0.9592
Pls10	Missa Ave Regina Coelorum 2. Kyrie	29.43	8.87	29.43	8.87	29.43	8.87	0.9578
Pls26	Missa Veni Sponsa Christi 3. Credo	66.88	9.20	66.88	9.20	66.88	9.21	0.9556
Pls12	Missa Ave Regina Coelorum 4. Credo	103.41	8.69	103.41	8.69	103.41	8.69	0.9550
Pls13	Missa Ave Regina Coelorum 5.Sanctus	18.50	8.96	18.50	8.96	18.50	8.96	0.9542
Pls25	Missa Veni Sponsa Christi 2. Gloria	44.21	8.67	44.21	8.67	44.21	8.67	0.9525
Pls29	Missa Veni Sponsa Christi 6. Agnus Dei I	15.65	7.99	15.65	7.99	15.65	7.99	0.9510
Pls09	Missa Ave Regina Coelorum 1. Chant	55.26	2.52	5.01	2.52	5.01	2.52	0.9487
Pls06	Missa Ascendo ad Patrem 6. Benedictus	26.55	7.91	26.55	7.91	26.55	7.91	0.9481

Pls27	Missa Veni Sponsa Christi 4. Sanctus	32.19	8.43	32.19	8.43	32.19	8.43	0.9472
Pls11	Missa Ave Regina Coelorum 3. Gloria	58.83	8.83	58.83	8.83	58.83	8.83	0.9445
Pls16	Missa Ave Regina Coelorum 8. Agnus Dei II	17.80	8.42	17.80	8.42	17.80	8.42	0.9423
Pls04	Missa Ascendo ad Patrem 4. Credo	90.66	8.92	90.66	8.92	90.66	8.92	0.9396
Pls30	Missa Veni Sponsa Christi 7. Agnus Dei II	16.48	9.39	16.48	9.39	16.48	9.39	0.9373
Pls14	Missa Ave Regina Coelorum 6. Benedictus	22.13	8.67	22.13	8.67	22.13	8.67	0.9304
Pls23	Missa Papae Marchelli 7. Agnus Dei II	32.77	9.22	32.77	9.22	32.77	9.22	0.9104
Pls15	Missa Ave Regina Coelorum 7. Agnus Dei I	13.79	11.52	13.79	11.52	13.79	11.52	0.9032

Let us consider now F. Liszt whose music differs from that of Palestrina not only because of the time gap between the composers. We compute always three components of the formula and comparing the parameters in the exponents we can draw conclusions about the stratification. The results are presented in Table 3. There are 7 monostratal compositions, 7 bistratal and 1 tristratal.

Table 3
Stratification of some compositions of F. Liszt

ID	Text	A_1	r_1	A_2	r_2	A_3	r_3	R^2
Liszt13	Liebesträume No. 3	69.47	2.29	43.87	20.64	43.54	20.64	0.9935
Liszt02	Paganini Capriccio No.3 La Campanella	424.66	1.67	95.11	21.90	86.62	21.90	0.9919
Liszt10	Hungarian Dance 5	44383.75	0.13	61.32	15.24	59.64	15.21	0.9909
Liszt12	Hungarian Rhapsody	68.84	2.49	31.23	13.86	27.33	13.86	0.9909
Liszt05	Venezia e Napoli: 1. Gondoliera	112546.76	0.13	96.80	6.05	148.89	15.94	0.9893
Liszt09	Hungarian Dance 1	55.00	17.62	55.00	17.62	55.00	17.62	0.9875
Liszt01	Concert Etude No.3: Un Sospiro	83.71	2.00	46.10	18.21	30.44	18.21	0.9874
Liszt14	Valse Oublieé No.1	113.16	1.00	48.16	19.17	48.14	19.17	0.9870
Liszt11	Hungarian Dance 6	102.26	1.98	89.20	18.04	83.81	18.04	0.9813
Liszt07	Venezia e Napoli: 3. Tarantella	104.81	26.47	104.81	26.47	104.81	26.47	0.9783
Liszt06	Venezia e Napoli: 2. Canzone	56.14	13.98	56.14	13.98	56.14	13.98	0.9740
Liszt15	Valse Oublieé No.2	77.10	19.19	77.10	19.19	77.10	19.19	0.9720
Liszt03	Transcendental Etudes: Eroica	34.42	31.81	34.42	31.81	34.42	31.81	0.9710
Liszt04	Transcendental Etudes: Feux Follets	74.23	20.90	74.23	20.90	74.23	20.90	0.9688
Liszt08	Sonata B minor	168.82	35.86	168.82	35.86	168.82	35.86	0.9607

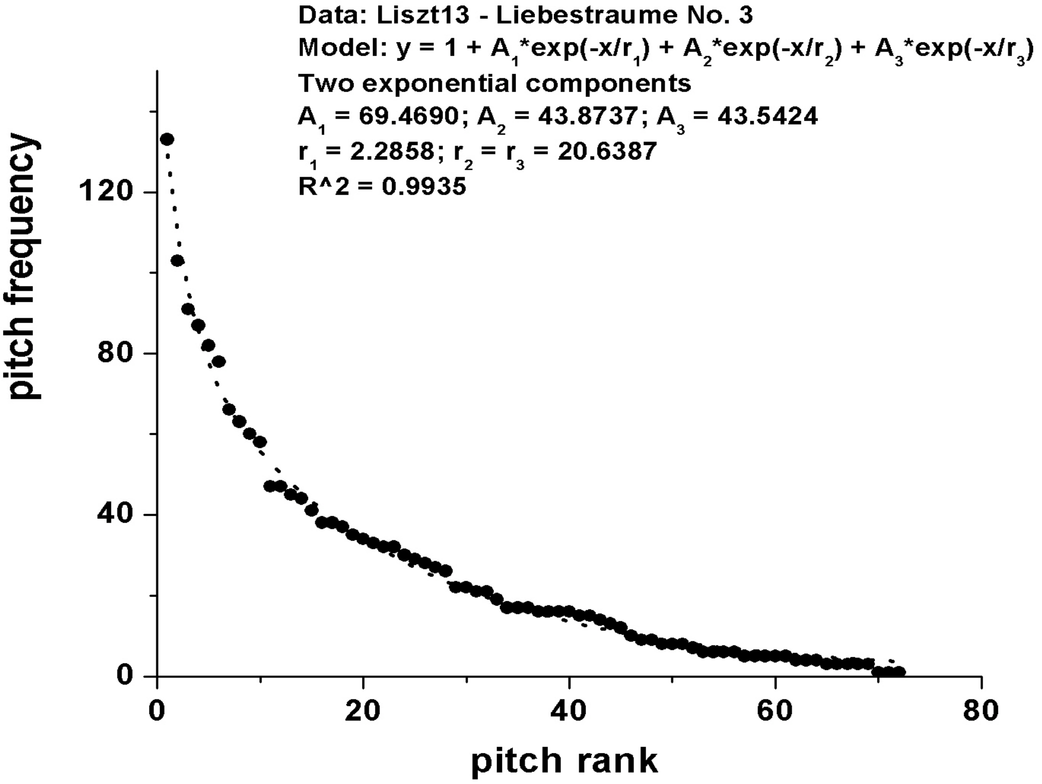


Figure 2. Fitting two exponential components to a work of Liszt

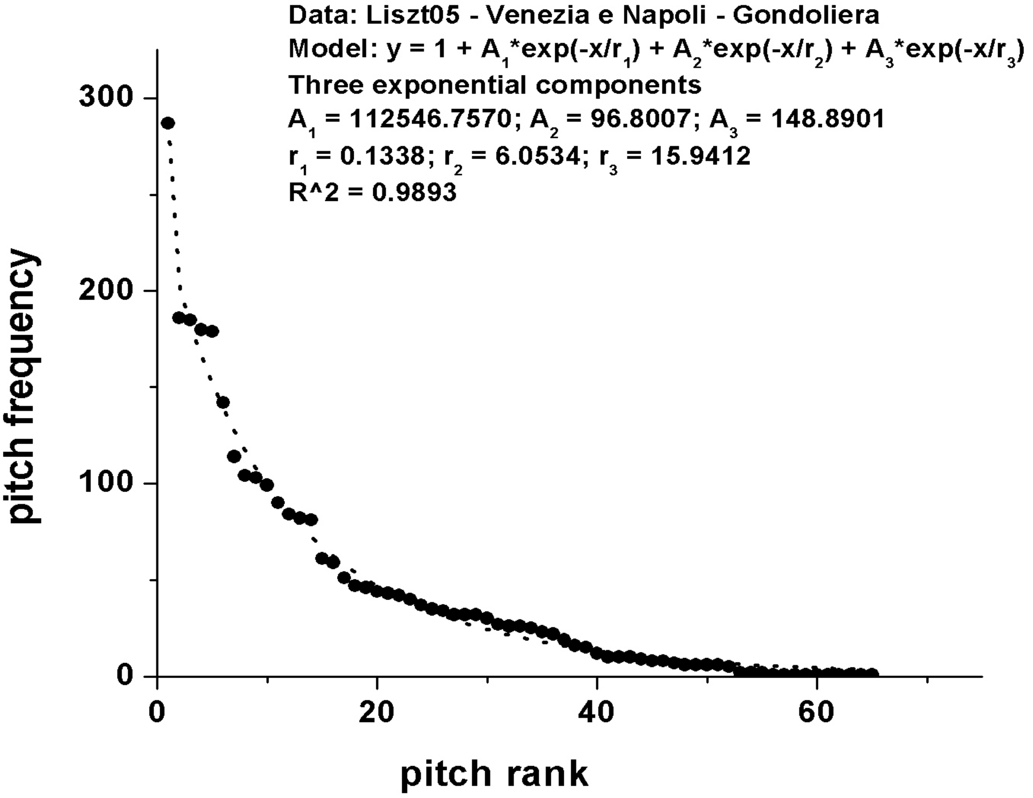


Figure 3. Fitting three exponential components to a work of Liszt

The analysis of different composers showed a rather variegated result given in Table 4.

Table 4
Strata with individual composers

Composer	Mean year	monostratal compositions	bistratal compositions	tristratal compositions
Palestrina (1525-1594)	1560	30	0	0
Gesualdo (1560/1-1613)	1587	5	2	0
Monteverdi (1567-1643)	1605	9	1	0
Bach (1685-1750)	1718	47	1	0
Mozart (1756-1791)	1774	8	0	1
Beethoven (1770-1827)	1799	17	15	0
Schumann (1810-1856)	1833	5	10	0
Wagner (1813-1883)	1848	2	1	0
Liszt (1811-1886)	1849	7	7	1
Skrjabin (1872-1915)	1894	14	11	1
Schoenberg (1874-1951)	1913	13	4	0
Stravinsky (1882-1971)	1927	12	13	0
Shostakovich(1906-1975)	1940	34	15	2

We considered only composers from whom we had at least 7 compositions As can be seen, out of 13 composers only 4 have tristratal works but the second stratum is present with 11 composers.

In order to compare the composers we set up a normalized vector [x, y, z] containing the proportion of compositions with 1, 2, and 3 strata. We obtain

Palestrina	= [1, 0, 0]
Gesualdo	= [0.71, 0.29, 0]
Monteverdi	= [0.90, 0.10, 0]
Bach	= [0.98, 0.02, 0]
Mozart	= [0.89, 0, 0.11]
Beethoven	= [0.53, 0.47, 0]
Schumann	= [0.33, 0.67, 0]
Wagner	= [0.67, 0.33, 0]
Liszt	= [0.47, 0.47, 0.07]
Skrjabin	= [0.54, 0.42, 0.04]
Schoenberg	= [0.76, 0.24, 0]
Stravinsky	= [0.48, 0.52, 0]
Shostakovich	= [0.67, 0.29, 0.04]

where $x + y + z = 1$. Due to this latter condition it is enough to investigate the (x, y) plot of the monostratal to bistratal proportion, as presented in Figure 4.

This result shows a peculiar picture: there are two classes of composers namely

{Palestrina, Monteverdi, Bach, Mozart} and

{Gesualdo, Beethoven, Schumann, Wagner, Liszt, Skrjabin, Schoenberg, Stravinsky, Shostakovich}.

It seems to be logical also from the musicological point of view: composers placed in the first class – Palestrina, Monteverdi, Bach and Mozart – wrote diatonic music, using the selection of the tone pitches.

On the other hand, composers placed in the second class – Gesualdo, Beethoven, Schumann, Wagner, Liszt, Skrjabin, Schoenberg, Stravinsky and Shostakovich – wrote more chromatic music using often chromatizations and all 12 tone pitches in their works.

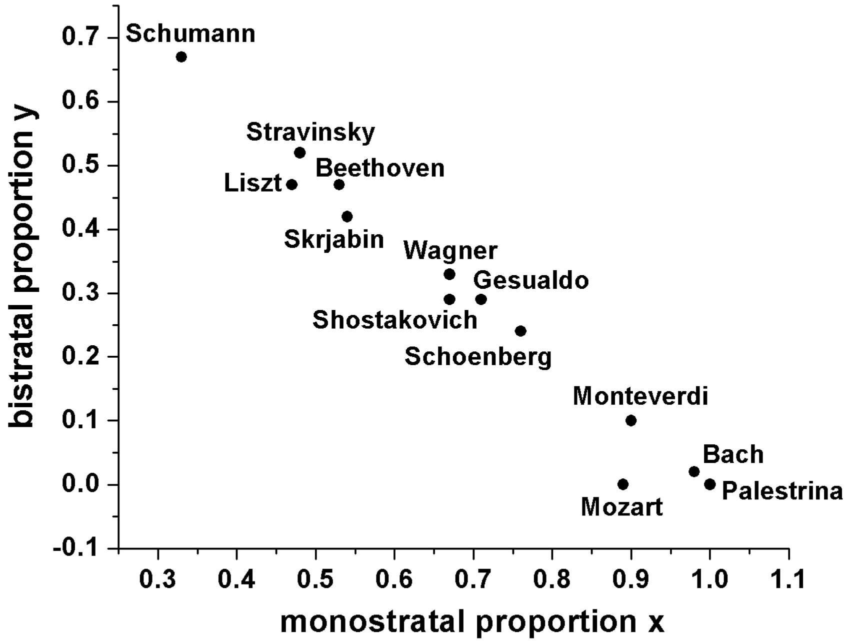


Figure 4. Monostratal to bistratal proportions for 13 composers

However, the results are not yet persuading because we have a small number of analysed works and the number of composers is not yet sufficient.

But in general, in the history of music, two main groups of compositions seem to alternate: the first is based on diatonic and the second on chromatic music.

In any case, this aspect can be examined easily by means of the above mentioned procedure. At the same time the result shows that in music and in language this formal aspect displays a kind of analogy. In language, stratification is a quite usual process because all classifications of entities are fuzzy and in text different classes are mixed. In music, stratification is caused not only by using several voices but also by stylistic preferences of individual composers. Looking at Figure 3 we see that there are pitches not lying in the direction of the exponential function, and there are also empty intervals of empirical values between two ranks. The thorough study of individual works could show us the stylistic means causing these “irregularities” Here we were interested only in discovering the general existence of stratification, individual analyses are left as future tasks.

4. The U-vector

Every rank-frequency distribution has several remarkable points which can serve both for characterization of individual works, genres or epochs and for testing a very general regularity which is known also from linguistics (cf. Popescu et al. 2010). In linguistics, these quantities are the inventory (vocabulary) of the entities used, *V*, the frequency of the most frequent unit, f_1, and the arc joining f_1 and *V* whose length is *L*. Though the rank-frequency sequence is usually very regular and the arc length *L* increases with increasing f_1 and *V*, it need not be constant for two texts with the same f_1 and *V*. Hence a text can be characterized by the vector

(4) $U(V, f_1, L)$

where *L* is the sum of the usual Euclidian distances between individual frequencies ordered in “one-step” distance., i.e.

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

Since both in music and in linguistics, the components of the vector *U* are variables, for further treatment it is advantageous to normalize them in some way. However, the upper boundaries of the components are unknown and one must restrict oneself to the given sample of texts and use the empirical minima and maxima of the components. The components will be normalized in the following way:

(6) $$X = \frac{V - V_{min}}{V_{\max} - V_{min}}, \quad Y = \frac{f_1 - f_{1,min}}{f_{1,max} - f_{1,min}}, \quad \mathrm{Z} = \frac{L - L_{min}}{L_{max} - L_{min}}.$$

and finally

(7) $$x = \frac{X}{X + Y + Z}, \quad y = \frac{Y}{X + Y + Z}, \quad z = \frac{Z}{X + Y + Z}.$$

In our data presented in Table 5, the minima and the maxima of the individual components are as follows:

$\mathrm{V_{min}}$ = 23 (Palestrina05);	$\mathrm{V_{max}}$ = 85 (Ligeti02)
$\mathrm{f_{1,min}}$ = 9 (Wagner02);	$\mathrm{f_{1,max}}$ = 1002 (Mozart05)
$\mathrm{L_{min}}$ = 34 (Wagner02;	$\mathrm{L_{max}}$ = 1012 (Mozart05),

from which the expressions (6) are computed as

$X = (V - 23)/(85 - 23)$
$Y = (f_1 - 9)/(1002 - 9)$
$Z = (L - 34)/(1012 - 34)$

and finally, formulas (7) whose values can be found in Table 5. It is to be noted that $x + y + z = 1$. If one adds to this sample a new text whose minima and maxima are more extreme that those above, the whole computation must be performed anew.

Table 5
Components of the vector U

Text ID	V	f_1	L	X	Y	Z	Sum	x	y	z
Bach01	44	106	124	0.3387	0.0977	0.0920	0.5284	0.6410	0.1849	0.1742
Bach02	45	147	163	0.3548	0.1390	0.1319	0.6257	0.5671	0.2221	0.2108
Bach03	45	155	169	0.3548	0.1470	0.1380	0.6399	0.5545	0.2298	0.2157
Bach04	47	140	158	0.3871	0.1319	0.1268	0.6458	0.5994	0.2043	0.1963
Bach05	44	113	129	0.3387	0.1047	0.0971	0.5406	0.6266	0.1937	0.1797
Beethoven01	59	537	550	0.5806	0.5317	0.5276	1.6400	0.3541	0.3242	0.3217
Beethoven02	62	626	644	0.6290	0.6213	0.6237	1.8741	0.3356	0.3315	0.3328
Beethoven03	63	625	636	0.6452	0.6203	0.6155	1.8810	0.3430	0.3298	0.3272
Beethoven04	63	868	879	0.6452	0.8651	0.8640	2.3742	0.2717	0.3644	0.3639
Beethoven05	63	473	492	0.6452	0.4673	0.4683	1.5807	0.4081	0.2956	0.2963
Gesualdo01	35	65	83	0.1935	0.0564	0.0501	0.3000	0.6451	0.1880	0.1670
Gesualdo02	34	51	68	0.1774	0.0423	0.0348	0.2545	0.6972	0.1662	0.1366
Gesualdo03	37	52	72	0.2258	0.0433	0.0389	0.3080	0.7332	0.1406	0.1262
Gesualdo04	36	67	82	0.2097	0.0584	0.0491	0.3172	0.6611	0.1842	0.1547
Gesualdo05	33	61	76	0.1613	0.0524	0.0429	0.2566	0.6286	0.2041	0.1674
Ligeti01	74	107	146	0.8226	0.0987	0.1145	1.0358	0.7942	0.0953	0.1106
Ligeti02	85	125	174	1.0000	0.1168	0.1431	1.2600	0.7937	0.0927	0.1136
Ligeti03	74	101	140	0.8226	0.0926	0.1084	1.0236	0.8036	0.0905	0.1059
Liszt01	65	128	166	0.6774	0.1198	0.1350	0.9322	0.7267	0.1286	0.1448
Liszt02	75	400	433	0.8387	0.3938	0.4080	1.6404	0.5113	0.2400	0.2487
Liszt03	78	113	152	0.8871	0.1047	0.1207	1.1125	0.7974	0.0941	0.1085
Liszt04	71	273	298	0.7742	0.2659	0.2699	1.3100	0.5910	0.2029	0.2061
Liszt05	65	287	317	0.6774	0.2800	0.2894	1.2467	0.5434	0.2246	0.2321
Monteverdi01	37	399	408	0.2258	0.3927	0.3824	1.0010	0.2256	0.3924	0.3820
Monteverdi02	30	242	248	0.1129	0.2346	0.2188	0.5664	0.1993	0.4143	0.3864
Monteverdi03	35	240	249	0.1935	0.2326	0.2198	0.6460	0.2996	0.3601	0.3403
Monteverdi04	32	341	347	0.1452	0.3343	0.3200	0.7995	0.1816	0.4182	0.4003
Monteverdi05	32	260	267	0.1452	0.2528	0.2382	0.6362	0.2282	0.3973	0.3745
Mozart01	58	692	707	0.5645	0.6878	0.6881	1.9405	0.2909	0.3545	0.3546
Mozart02	56	482	495	0.5323	0.4763	0.4714	1.4800	0.3596	0.3219	0.3185
Mozart03	59	499	510	0.5806	0.4935	0.4867	1.5608	0.3720	0.3162	0.3118
Mozart04	57	474	491	0.5484	0.4683	0.4673	1.4839	0.3695	0.3156	0.3149
Mozart05	55	1002	1012	0.5161	1.0000	1.0000	2.5161	0.2051	0.3974	0.3974

Palestrina01	27	209	215	0.0645	0.2014	0.1851	0.4510	0.1431	0.4466	0.4104
Palestrina02	24	101	108	0.0161	0.0926	0.0757	0.1844	0.0874	0.5023	0.4102
Palestrina03	26	157	164	0.0484	0.1490	0.1329	0.3304	0.1465	0.4512	0.4024
Palestrina04	27	243	248	0.0645	0.2356	0.2188	0.5190	0.1243	0.4541	0.4216
Palestrina05	23	70	76	0.0000	0.0614	0.0429	0.1044	0.0000	0.5886	0.4114
Schoenberg01	65	913	924	0.6774	0.9104	0.9100	2.4978	0.2712	0.3645	0.3643
Schoenberg02	67	68	112	0.7097	0.0594	0.0798	0.8488	0.8360	0.0700	0.0940
Schoenberg03	63	64	105	0.6452	0.0554	0.0726	0.7731	0.8345	0.0716	0.0939
Schoenberg04	70	51	97	0.7581	0.0423	0.0644	0.8648	0.8766	0.0489	0.0745
Schoenberg05	56	50	87	0.5323	0.0413	0.0542	0.6277	0.8479	0.0658	0.0863
Shostakovich01	30	73	87	0.1129	0.0645	0.0542	0.2315	0.4876	0.2784	0.2340
Shostakovich03	32	33	51	0.1452	0.0242	0.0174	0.1867	0.7775	0.1294	0.0931
Shostakovich04	34	23	45	0.1774	0.0141	0.0112	0.2028	0.8750	0.0695	0.0555
Shostakovich05	33	57	78	0.1613	0.0483	0.0450	0.2546	0.6335	0.1898	0.1767
Skrjabin01	48	19	55	0.4032	0.0101	0.0215	0.4348	0.9274	0.0232	0.0494
Skrjabin02	31	20	42	0.1290	0.0111	0.0082	0.1483	0.8701	0.0747	0.0552
Skrjabin03	51	33	67	0.4516	0.0242	0.0337	0.5095	0.8863	0.0474	0.0662
Skrjabin04	32	12	36	0.1452	0.0030	0.0020	0.1502	0.9663	0.0201	0.0136
Skrjabin05	38	23	52	0.2419	0.0141	0.0184	0.2744	0.8816	0.0514	0.0671
Stravinsky01	51	194	214	0.4516	0.1863	0.1840	0.8220	0.5494	0.2267	0.2239
Stravinsky02	64	407	430	0.6613	0.4008	0.4049	1.4670	0.4508	0.2732	0.2760
Stravinsky03	73	182	220	0.8065	0.1742	0.1902	1.1709	0.6888	0.1488	0.1624
Stravinsky04	65	396	428	0.6774	0.3897	0.4029	1.4700	0.4608	0.2651	0.2741
Stravinsky05	71	215	246	0.7742	0.2075	0.2168	1.1984	0.6460	0.1731	0.1809
Wagner01	29	33	50	0.0968	0.0242	0.0161	0.1371	0.7059	0.1763	0.1178
Wagner02	31	9	34	0.1290	0.0000	0.0000	0.1290	1.0000	0.0000	0.0000
Wagner03	81	484	501	0.9355	0.4783	0.4775	1.8913	0.4946	0.2529	0.2525
Wagner04	39	85	107	0.2581	0.0765	0.0742	0.4088	0.6313	0.1872	0.1815

Since we have three components, the graphical presentation would require a three- dimensional figure which is, of course, not quite lucid. Instead, we use the method of ternary plot which allows us to present the coordinates of *U* in a two-dimensional scheme as shown in Figure 5.

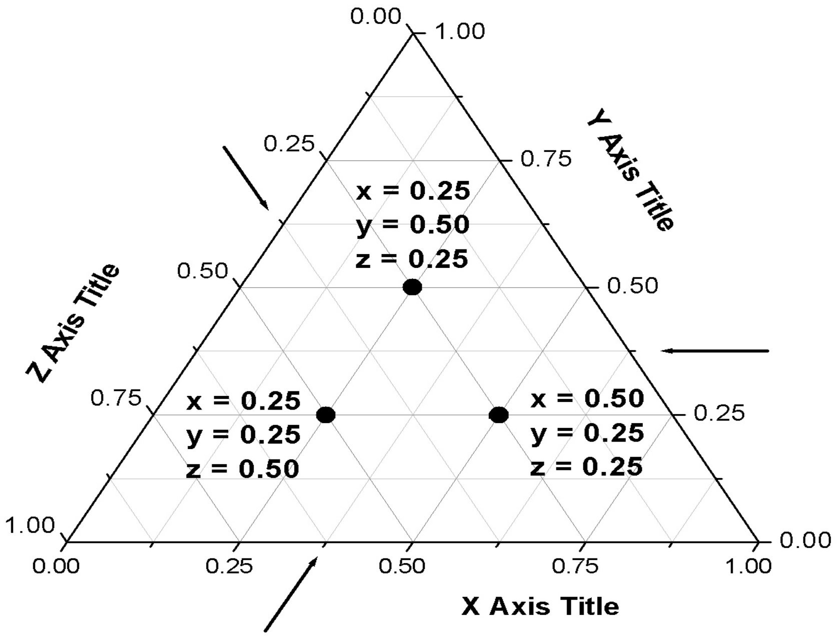


Figure 5. Ternary plot

The components of 61 musical compositions plotted in this scheme display a very regular behaviour as shown in Figure 6. Even if the composers have different styles, use different means and many times rework the composition, there is an unconscious mechanism behind their striving for originality leading to some stable interrelations.

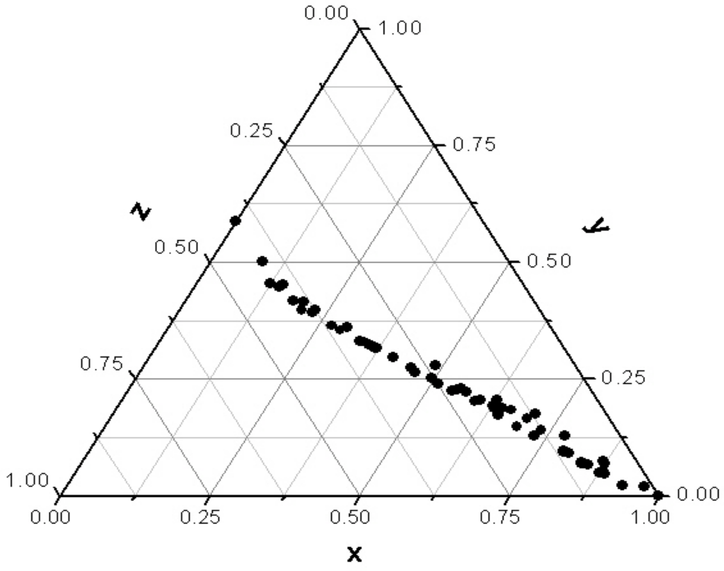


Figure 6. The components of the U-vector of 61 compositions

It may be noted that rank-frequency distributions of words in written texts abide by the same regularity, but the direction of the points in the ternary plot is quite different (cf. Popescu et al. 2010). Perhaps it is this direction that is characteristic for communication types using sound means. The research in this direction is not yet advanced, one would be obliged to analyse also the communication of some animals.

The relationship of individual components *x, y, z* to one another is quite straightforward. For each of them we obtain a linear relationship as shown in Figures 7 to 9. Though for small *x* and great *y* one can observe deviations form the linear relationships, we suppose that it is caused by the specificity of the given sample. Needless to say, further investigation will change some relationships but the deviation from linearity will not be essential.

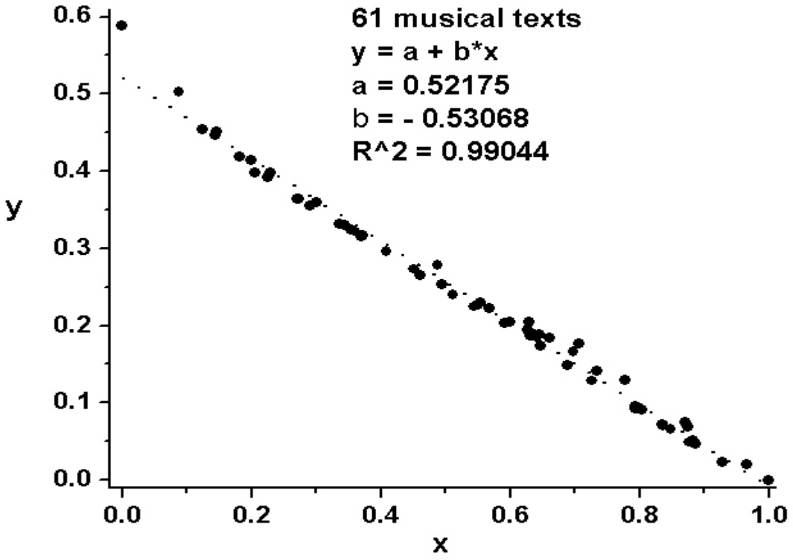


Figure 7. The relationship between *x* and *y*

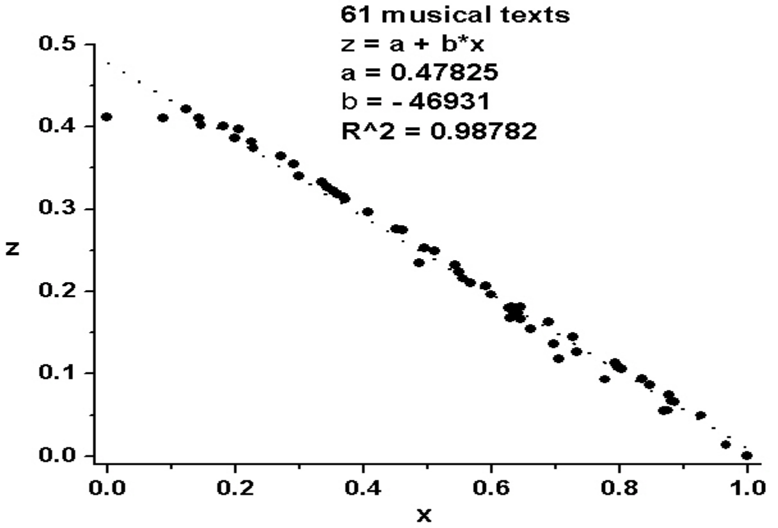


Figure 8. The relationship between *x* and *z*

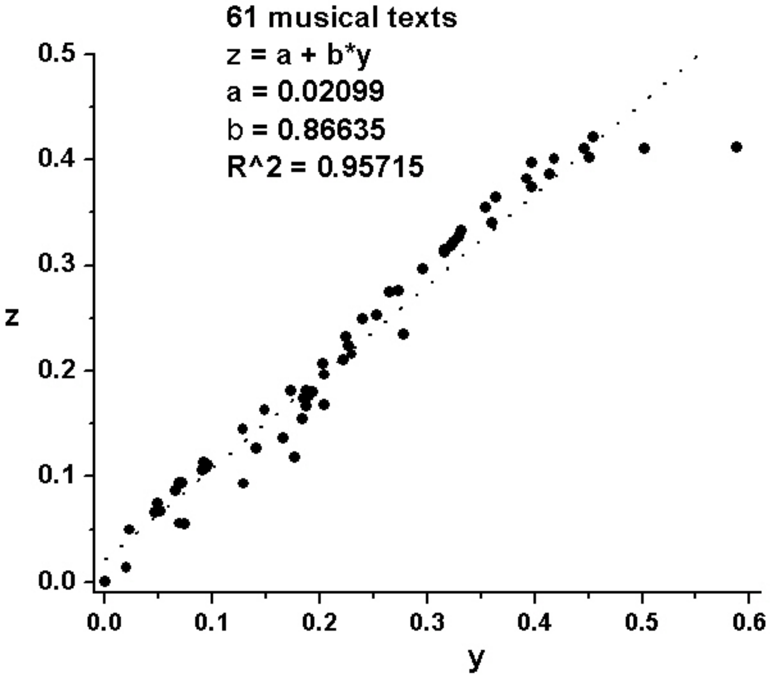


Figure 9. The relationship between *y* and *z*

Of course, the composers can also be averaged, and in that case we obtain a very unambiguous ordering. As shown in Table 6, only Schoenberg´s value of *z* slighty deviates. Increasing *x* is accompanied with decreasing *y* and *z*.

Table 6
Average *x*,*y*,*z* values of 13 composers

Composer	$\bar{x}$	$\bar{y}$	$\bar{z}$
Palestrina	0.1003	0.4885	0.4112
Monteverdi	0.2269	0.3965	0.3767
Mozart	0.3195	0.3411	0.3395
Beethoven	0.3425	0.3291	0.3284
Stravinsky	0.5592	0.2174	0.2235
Bach	0.5977	0.2069	0.1953
Liszt	0.6339	0.1780	0.1880
Gesualdo	0.6730	0.1766	0.1504
Shostakovich	0.6934	0.1668	0.1398
Wagner	0.7080	0.1541	0.1379
Schoenberg	0.7332	0.1242	0.1426
Ligeti	0.7971	0.0928	0.1100
Skrjabin	0.9064	0.0434	0.0503

Preliminary investigations show that in musical compositions pitches play the same semiotic role as the "lowest" linguistic entities like sounds, phonemes, graphemes occupying approximately the same "corridor" in the ternary plot. However, further investigations are necessary to show that the ternary plot is an appropriate means for scrutinizing the semiotic status of artistic or linguistic entities.

References

Johnson, N.L., Kotz, S., Kemp, A.W. (1992). *Univariate discrete distributions*. New York: Wiley.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf´s law – another view. *Quality and Quantity, 44(4), 713-731* .
Quality and Quantity (May 2009), DOI 10.1007/s11135-009-9234-y (2009);
http://www.springerlink.com/content/v268444738300073/?p=b28be84685b148ff8725b0cbc1671402&pi=0 (September 9, 2009);

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

Wimmer, G. Altmann, G. (2000). On the generalization of the STER distribution applied to generalized hypergeometric parents. *Acta Universitatis Palackianae Olomoucensis Facultas Rerum Naturalium, Mathematica 39, 215-247.*

Wimmer, G., Altmann, G. (2001). Models of rank-frequency distrubtions in language and music. In: Uhlířová L. et al. (eds.), *Text as a linguistic paradigm. Festschrift in honour of Luděk Hřebíček: 10-20.* Trier WVT

Wimmer, G., Šidlík, J. Altmann, G. (1999). A new model of rank-frequency distribution. *Journal of Quantitative Linguistics 6, 188-193.*

Zipf, G.K. (1935) *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton-Mifflin.

Glottometrics 24, 2012, 41-69

Aspects of the behaviour of parts-of-speech in Italian texts

Arjuna Tuzzi
Ioan-Iovitz Popescu
Peter Zörnig
Gabriel Altmann

Abstract. The present article is a continuation of the analysis performed in Tuzzi, Popescu, Altmann (2010). Here the parts-of-speech have been scrutinized. Their rank-frequency distributions have been characterized using the Repeat rate, the Entropy and Ord's criterion. The ranking of POS with individual Italian presidents has been used to characterize the homogeneity individually and as a whole of 63 texts using Kendall's concordance coefficient. There is high concordance. The last aspect is the computation of distances between identical parts-of-speech which yield a very unique picture represented by the Zipf-Alekseev function.

Keywords: parts-of-speech, Italian, rank-frequency distribution, Repeat rate, Entropy, Ord's criterion, concordance, distances

Introduction

Usually POS are studied from grammatical or semantic point of view. The results are classifications, one of which is known since antiquity and still used today. The number of classes lies between 8 and ca 100 according to the adopted criteria. It is to be remarked that no criteria warrant the "truth" and whatever kind of criteria we set up, they are merely our conceptual constructions. In some languages semantic criteria are sufficient, in other ones morphology may be helpful and if there is reduced morphology, syntactic criteria may be used. The choice of criteria depends also on the aim of investigation; it serves our elementary concept formation. The result of a classification procedure is never a theory but merely a taxonomic account (cf. Bunge 1983:17) of what is there. Since the ca 500 numerical classification methods yield different results, they have nothing to do with "truth" but merely with utility.

Nevertheless, if we are able to perform an elementary classification, we may compare both text and languages.

In a previous work (cf. Tuzzi, Popescu, Altmann 2010) the rank-frequency distribution of parts-of-speech in the end-of-year speeches of Italian presidents has been analysed. The ranking was performed according to frequency, i.e. each POS could attain different rank in different texts. But if we go back to the nominal scale and ascribe each POS its frequency rank in the given text, we obtain a different order. It would be possible to compare the texts using the absolute frequencies and the chi-square test but if we consider 60 texts and several thousands of words, the chi-square — which increases with increasing sample size — would simply signalize the heterogeneity of texts. If we ascribed the frequencies to the respective POS, we would obtain a still greater chi-square. Hence we simplify the procedure and ascribe the POS only their rank number. In Italian, the program works with the following classes: **noun (n), preposition (prep), verb (v), adjective (a), pronoun (pron), article (det), con-**

junction (cong), adverb (avv), numeral (num), interjection (esc), proper noun (nm), i.e. it follows the classical word class classification. (For other possibilities see Bergenholtz, Schaeder 1977; Best 2005; Kroeger 2005) Using this classification we ask the following questions:

(a) Which formal properties of the "speeches" change, and if so, in what way? That is, is there some development in the use of POS?

(b) Do all texts of an individual president display a concordant rank-order or are there non-homogeneities?

(c) Do the sequences of POS display different patterns? Here we ask what are the distances between equal POS in the sequence. Do they follow a certain distribution?

(d) One could, of course, ask whether there is some concordance between the digrams or trigrams of subsequent POS and show the entropy or the transinformation in the sequences, but this is a problem which can be omitted here and solved in another place.

1. Indicator development

The most common indicators of rank-frequency distributions are the (relative) entropy H_{rel}, the Repeat Rate *RR* and Ord's indicators *I* and *S* which are functions of moments of the distribution. In order to show them we present the distributions of POS in the 60 end-of-year speeches of Italian presidents as they were presented in Table 5.1 in Tuzzi et al. (2010: 117f.). Here the identity of individual POS is not taken into account, the frequencies are simply ranked.

Table 1
Ranked frequencies of parts-of-speech
in Italian end-of-year Addresses

Text	Parts-of-speech frequencies	*N*
1949Einaudi	41,37,33,30,17,15,14,6,1	194
1950Einaudi	42,36,20,15,15,9,8,4,1	150
1951Einaudi	50,41,40,34,21,18,15,11	230
1952Einaudi	46,35,28,27,13,12,11,7	179
1953Einaudi	47,42,34,24,15,12,9,6,1	190
1954Einaudi	57,54,43,36,20,18,17,14,1	260
1955Gronchi	83,78,64,51,31,30,29,21,1	388
1956Gronchi	180,121,88,87,71,58,34,26	665
1957Gronchi	267,241,170,126,93,84,79,59,6,5	1130
1958Gronchi	201,162,131,127,82,74,63,42,3,1	886
1959Gronchi	181,135,92,80,72,71,36,29,1	697
1960Gronchi	196,161,112,106,78,63,45,38,3,2	804
1961Gronchi	304,244,184,162,111,105,75,65,2	1252
1962Segni	196,147,120,83,73,54,36,29	738
1963Segni	257,219,170,131,93,68,52,45,14,8	1057
1964Saragat	102,85,84,64,42,40,28,17,3	465
1965Saragat	267,211,141,138,85,79,78,45,6,3	1053
1966Saragat	324,239,185,144,109,75,66,50,5,2	1199
1967Saragat	263,207,167,145,96,64,59,36,14,3,2	1056

1968Saragat	304,243,176,134,95,86,70,56,8,2	1174
1969Saragat	394,284,232,222,165,103,99,72,8,3,2	1584
1970Saragat	490,389,272,257,186,113,112,86,17,5,2	1929
1971Leone	70,51,37,35,30,17,11,6,3,2	262
1972Leone	182,149,134,111,69,45,45,24,5,3	767
1973Leone	298,232,205,174,103,97,76,63,1,1	1250
1974Leone	197,141,139,120,66,59,42,35,1,1	801
1975Leone	312,244,200,191,122,97,91,69,2	1328
1976Leone	321,239,211,196,113,112,97,73,3,1	1366
1977Leone	358,270,262,216,142,122,115,113,4,2	1604
1978Pertini	332,283,248,156,130,125,106,86,17,10	1493
1979Pertini	499,442,345,279,219,201,184,115,8,8,2	2302
1980Pertini	316,244,228,164,121,104,101,61,10,9,2	1360
1981Pertini	571,571,377,331,261,231,227,196,38,14,1	2818
1982Pertini	509,495,332,322,233,202,172,139,62,19,2	2487
1983Pertini	786,760,510,452,360,308,275,206,55,33,3	3748
1984Pertini	302,269,197,163,129,97,95,51,20,17	1340
1985Cossiga	612,427,404,289,207,192,120,93,10,3,2	2359
1986Cossiga	321,232,215,187,130,106,79,77,1,1	1349
1987Cossiga	501,414,349,248,184,163,107,103,11,11	2091
1988Cossiga	557,467,369,311,199,183,146,134,14,5	2385
1989Cossiga	441,399,302,231,154,145,102,101,31,6	1912
1990Cossiga	800,646,534,396,305,277,173,163,35,18	3347
1991Cossiga	95,71,64,57,48,29,26,22,4,2	418
1992Scalfaro	656,472,435,360,250,231,208,151,4,3,2	2772
1993Scalfaro	684,501,469,387,247,236,218,168,22,8,1	2941
1994Scalfaro	866,633,590,482,284,267,248,207,15,12,1	3605
1995Scalfaro	994,741,682,523,357,332,290,246,38,22,3	4228
1996Scalfaro	535,348,326,313,183,128,115,110,16,11	2085
1997Scalfaro	1113,1048,712,522,429,397,368,329,54,33,10	5015
1998Scalfaro	972,775,577,415,399,289,254,251,35,23,5	3995
1999Ciampi	504,347,291,278,206,110,89,82,24,9,1	1941
2000Ciampi	432,338,291,273,168,124,95,88,23,12	1844
2001Ciampi	549,395,338,262,224,109,96,89,18,15,2	2097
2002Ciampi	556,389,312,304,209,132,112,98,10,7	2129
2003Ciampi	408,297,231,214,142,112,79,75,4,2,1	1565
2004Ciampi	455,353,268,265,147,111,93,88,19,8	1807
2005Ciampi	290,235,181,166,114,89,55,40,12,10,1	1193
2006Napolitano	502,377,356,286,191,169,159,146,10,7,1	2204
2007Napolitano	419,352,274,242,144,123,115,104,13,5,3	1794
2008Napolitano	409,328,281,220,135,127,120,86,4,2,1	1713
2009Napolitano	528,410,374,268,175,173,166,166,24,8,1	2293
20010Napolitano	593,502,354,336,197,190,162,117,35,12	2498
20011Napolitano	551,458,341,330,193,167,153,134,27,8,3	2365

1.1. *Repeat Rate*

The Repeat Rate is defined as the sum of squares of relative frequencies $p_i^2 = (f_i/N)^2$, i.e.

$$(1) \qquad RR = \sum_{i=1}^{K} p_i^2 = \frac{1}{N^2}\sum_{i=1}^{K} f_i^2 .$$

Usually one uses it in a relative form given as

$$(2) \qquad RR_{rel} = \frac{1-RR}{1-1/K}$$

or in the McIntosh (1967) form as

$$(3) \qquad RR_{McInt} = \frac{1-\sqrt{RR}}{1-1/\sqrt{K}} .$$

For the extensive use of this indicator see e.g. Popescu et al. (2009), Popescu, Mačutek, Altmann (2009). The computed values are presented in Table 2. The inventory in each case is K = 11 even if in some texts not all POS are used.

Table 2
Repeat rates of Parts-of-speech in 60 end-of-year speeches of Italian presidents.

Text	***N***	***RR***	***RR_{rel}***	***RR_{McInt}***
1949Einaudi	194	0.1537	0.9309	0.8703
1950Einaudi	150	0.1810	0.9009	0.8226
1951Einaudi	230	0.1521	0.9327	0.8732
1952Einaudi	179	0.1666	0.9168	0.8474
1953Einaudi	190	0.1715	0.9113	0.8387
1954Einaudi	260	0.1556	0.9288	0.8669
1955Gronchi	388	0.1515	0.9333	0.8743
1956Gronchi	665	0.1641	0.9194	0.8516
1957Gronchi	1130	0.1563	0.9280	0.8656
1958Gronchi	886	0.1502	0.9348	0.8769
1959Gronchi	697	0.1610	0.9229	0.8572
1960Gronchi	804	0.1573	0.9270	0.8639
1961Gronchi	1252	0.1565	0.9279	0.8654
1962Segni	738	0.1684	0.9148	0.8442
1963Segni	1057	0.1596	0.9245	0.8598
1964Saragat	465	0.1537	0.9310	0.8704

1965Saragat	1053	0.1590	0.9250	0.8607
1966Saragat	1199	0.1680	0.9153	0.8449
1967Saragat	1056	0.1607	0.9232	0.8577
1968Saragat	1174	0.1632	0.9205	0.8533
1969Saragat	1584	0.1562	0.9282	0.8659
1970Saragat	1929	0.1610	0.9229	0.8572
1971Leone	262	0.1669	0.9165	0.8468
1972Leone	767	0.1615	0.9223	0.8563
1973Leone	1250	0.1566	0.9277	0.8651
1974Leone	801	0.1609	0.9230	0.8574
1975Leone	1328	0.1535	0.9312	0.8708
1976Leone	1366	0.1518	0.9331	0.8740
1977Leone	1604	0.1467	0.9386	0.8833
1978Pertini	1493	0.1470	0.9383	0.8827
1979Pertini	2302	0.1466	0.9388	0.8835
1980Pertini	1360	0.1502	0.9348	0.8768
1981Pertini	2818	0.1406	0.9453	0.8948
1982Pertini	2487	0.1400	0.9459	0.8959
1983Pertini	3748	0.1428	0.9429	0.8906
1984Pertini	1340	0.1489	0.9362	0.8793
1985Cossiga	2359	0.1629	0.9208	0.8538
1986Cossiga	1349	0.1530	0.9317	0.8717
1987Cossiga	2091	0.1575	0.9268	0.8636
1988Cossiga	2385	0.1536	0.9310	0.8705
1989Cossiga	1912	0.1544	0.9301	0.8690
1990Cossiga	3347	0.1542	0.9304	0.8695
1991Cossiga	418	0.1473	0.9380	0.8822
1992Scalfaro	2772	0.1502	0.9348	0.8769
1993Scalfaro	2941	0.1482	0.9370	0.8806
1994Scalfaro	3605	0.1529	0.9318	0.8718
1995Scalfaro	4228	0.1488	0.9363	0.8794
1996Scalfaro	2085	0.1581	0.9261	0.8625
1997Scalfaro	5015	0.1473	0.9379	0.8821
1998Scalfaro	3995	0.1518	0.9330	0.8739
1999Ciampi	1941	0.1609	0.9230	0.8574

2000Ciampi	1844	0.1533	0.9314	0.8712
2001Ciampi	2097	0.1637	0.9199	0.8523
2002Ciampi	2129	0.1619	0.9220	0.8557
2003Ciampi	1565	0.1627	0.9211	0.8542
2004Ciampi	1807	0.1606	0.9233	0.8579
2005Ciampi	1193	0.1584	0.9258	0.8619
2006Napolitano	2204	0.1471	0.9382	0.8826
2007Napolitano	1794	0.1532	0.9314	0.8712
2008Napolitano	1713	0.1562	0.9282	0.8658
2009Napolitano	2293	0.1474	0.9379	0.8820
20010Napolitano	2498	0.1535	0.9311	0.8707
20011Napolitano	2365	0.1512	0.9336	0.8749

Taking the values individually, we obtain rather an ellipse, though a slight decrease — which is probably its main axis — can be seen, as shown in Figure 1

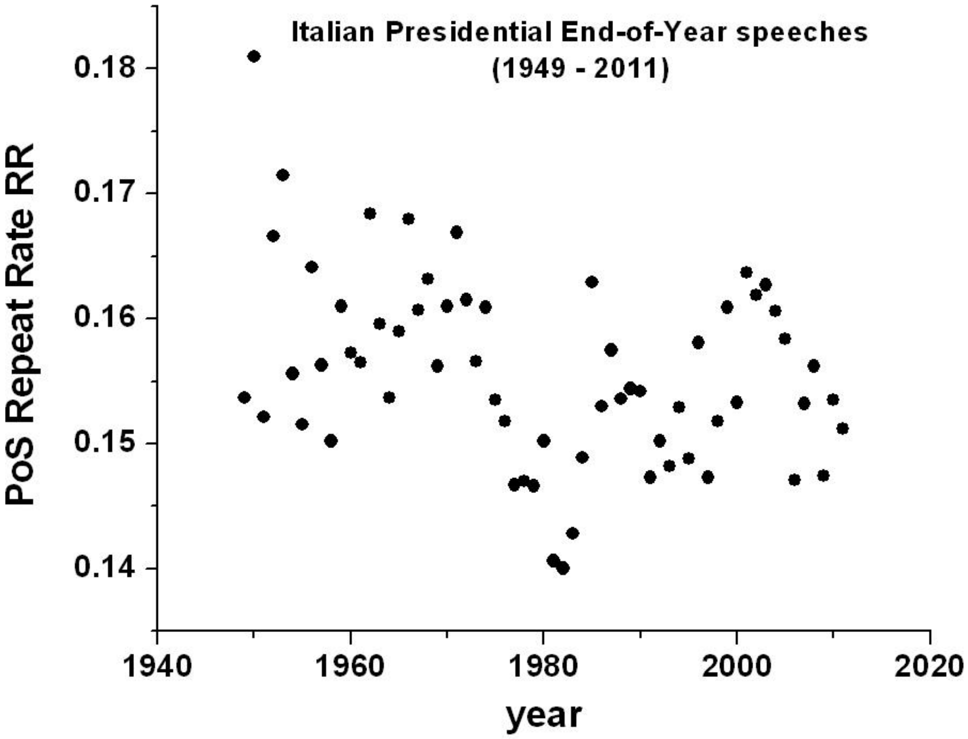


Figure 1. Repeat rates in the course of years

The decrease, though not very smooth, can be better seen, if we take the means of individual presidents and obtain the results presented in Table 3 and Figure 2.

Table 3
Mean Repeat Rates of individual presidents

President	Years	Mean RR
Einaudi	1949-54	0.1634
Gronchi	1955-61	0.1567
Segni	1962-63	0.1640
Saragat	1964-70	0.1603
Leone	1971-77	0.1568
Pertini	1978-84	0.1452
Cossiga	1985-91	0.1547
Scalfaro	1992-98	0.1510
Ciampi	1999-2005	0.1602
Napolitano	2006-11	0.1514

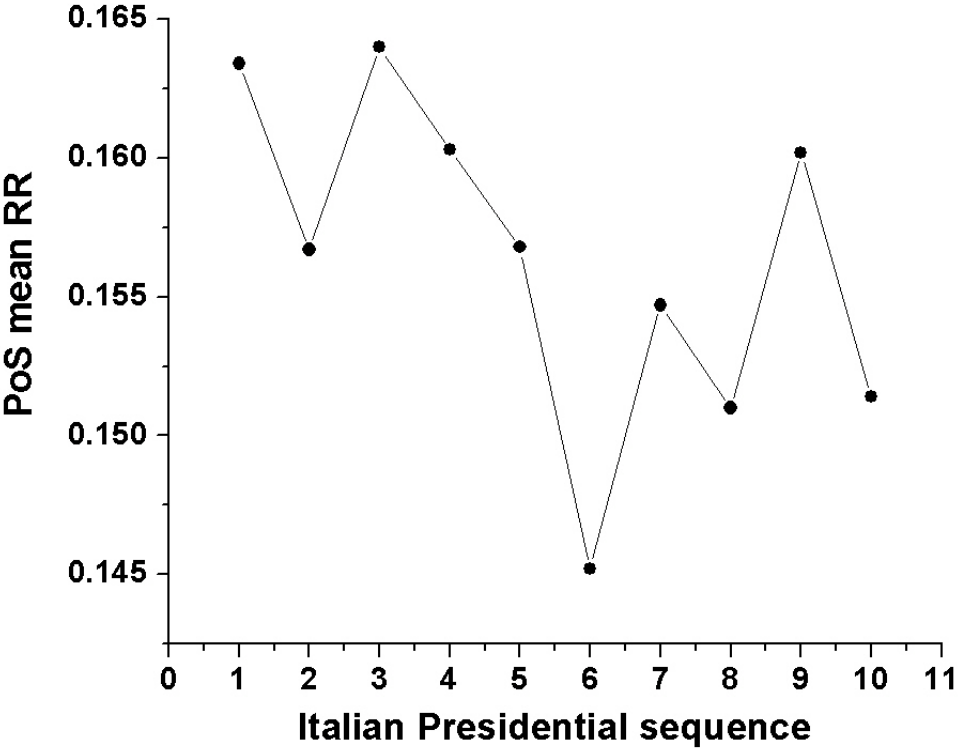


Figure 2. Mean Repeat Rates of individual presidents

In Figure 2 the irregular decrease is evident. But if it is existent, it means that the newer presidents use a more vivid language, more complex sentences, and more different facts. Though the nouns are the most frequent POS, their attributes and predicates get a more complex structure. Using only these texts it cannot be said whether this is a special property of these texts, or whether Italian as a whole has changed in this sense. In any case, one sees a slight change within the last 60 years. Of course, the mean *RR* could be computed also using the addition of frequencies of the same POS in all texts of the given president but we omit this possibility because it would mean text mixing.

1.2. Entropy

Considering the second frequently used indicator, the Entropy, defined as

$$H = -\sum_{i=1}^{K} p_i \log_2 p_i = \log_2 N - \frac{1}{N}\sum_{i=1}^{K} f_i \log_2 f_i \qquad (4)$$

which can be transformed in Repeat Rate, we obtain the results in Table 4 and Figure 3.

Table 4
The development of Entropy in 60 years

Text	H	Text	H
1949Einaudi	2.8423	1981Pertini	2.9843
1950Einaudi	2.7170	1982Pertini	3.0097
1951Einaudi	2.8383	1983Pertini	2.9842
1952Einaudi	2.7608	1984Pertini	2.9458
1953Einaudi	2.7477	1985Cossiga	2.8178
1954Einaudi	2.8417	1986Cossiga	2.8517
1955Gronchi	2.8613	1987Cossiga	2.8610
1956Gronchi	2.7813	1988Cossiga	2.8759
1957Gronchi	2.8720	1989Cossiga	2.8959
1958Gronchi	2.8787	1990Cossiga	2.8934
1959Gronchi	2.8071	1991Cossiga	2.9262
1960Gronchi	2.8495	1992Scalfaro	2.8813
1961Gronchi	2.8334	1993Scalfaro	2.9180
1962Segni	2.7546	1994Scalfaro	2.8800
1963Segni	2.8174	1995Scalfaro	2.9227
1964Saragat	2.8513	1996Scalfaro	2.8675
1965Saragat	2.8518	1997Scalfaro	2.9391
1966Saragat	2.7916	1998Scalfaro	2.9100
1967Saragat	2.8566	1999Ciampi	2.8504
1968Saragat	2.8294	2000Ciampi	2.9004
1969Saragat	2.8639	2001Ciampi	2.8326
1970Saragat	2.8503	2002Ciampi	2.8298

1971Leone	2.8135	2003Ciampi	2.8088
1972Leone	2.8226	2004Ciampi	2.8538
1973Leone	2.8301	2005Ciampi	2.8629
1974Leone	2.8060	2006Napolitano	2.9136
1975Leone	2.8467	2007Napolitano	2.8795
1976Leone	2.8662	2008Napolitano	2.8424
1977Leone	2.9003	2009Napolitano	2.9358
1978Pertini	2.9440	2010Napolitano	2.9043
1979Pertini	2.9176	2011Napolitano	2.9165
1980Pertini	2.9214		

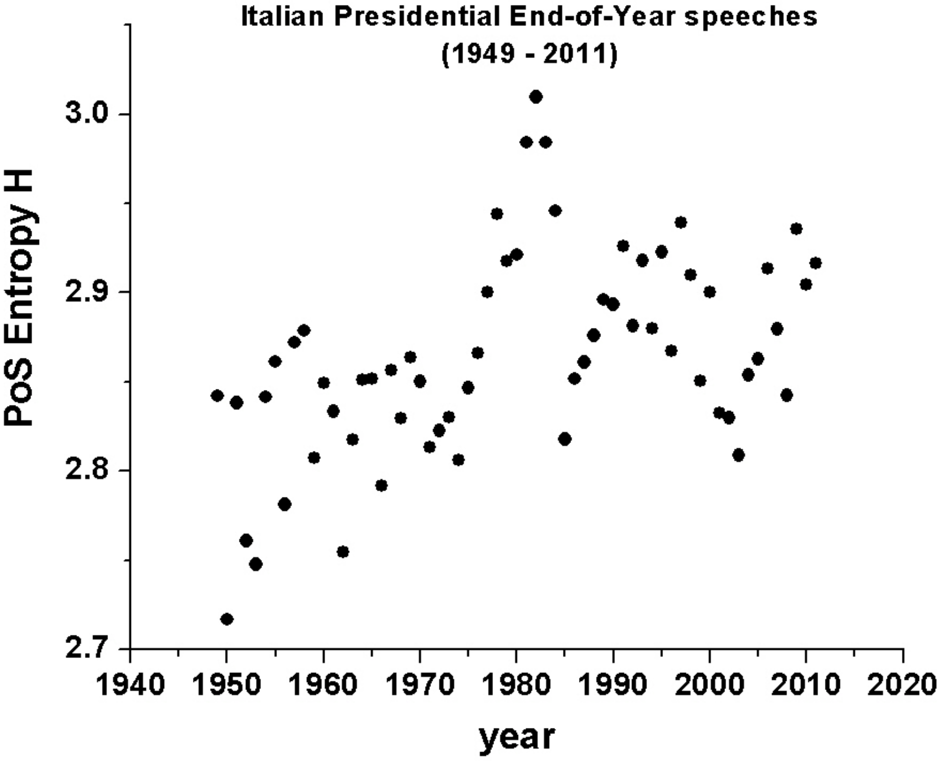


Figure 3. Entropies in the 63 speeches

The motion, except for Pertini, is quite regular and displays a slightly increasing entropy, i.e. more homogeneous use of POS, in other words, a motion away from stereotypy.

The computing of relative Entropy is here irrelevant because for each text we must take into account the same number of classes (11).

Again, the means of Entropy could be computed. It is not recommended to add the frequencies in texts of individual presidents because it would mean the creation of mixed samples; further, the ranks in individual texts do not represent always the same POS, hence it would be a very heterogeneous sample. For this reason we take the means of individual presidents considering *H* a variable. We obtain the results presented in Table 5 and Figure 4 in which the slight increase of Entropy in the course of years can be observed. This is, of course the same story as told by the Repeat Rate. Pertini is again an outlier. But only historians or literary scientists could explain why.

Table 5
Mean Entropies in texts of individual presidents

President	Years	Mean H
Einaudi	1949-54	2.7913
Gronchi	1955-61	2.8405
Segni	1962-63	2.7860
Saragat	1964-70	2.8421
Leone	1971-77	2.8408
Pertini	1978-84	2.9581
Cossiga	1985-91	2.8745
Scalfaro	1992-98	2.9027
Ciampi	1999-2005	2.8484
Napolitano	2006-11	2.8987

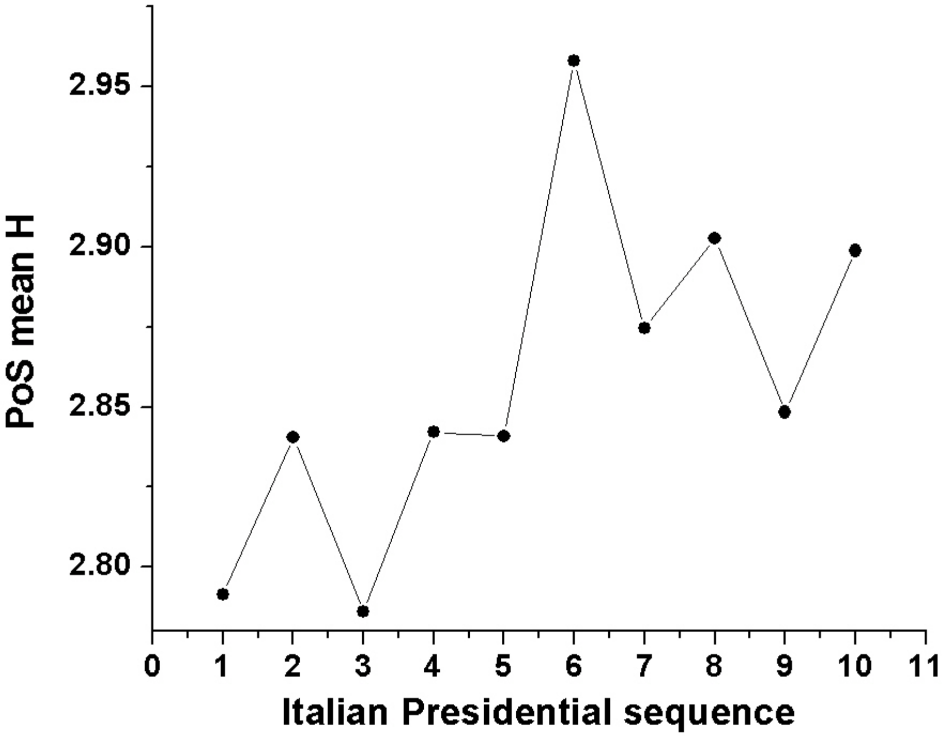


Figure 4. Mean Entropies in texts of individual presidents

1.3. Ord’s criterion

Ord (1972) defined indicators based on moments of the distribution. The first one, *I,* is the usual variance divided by the mean, i.e.

(5) $$I = \frac{m_2}{m_1'}.$$

The second, *S*, takes into account the asymmetry in its relation to the variance, i.e.

(6) $$S = \frac{m_3}{m_2}.$$

Both have been used frequently in linguistic literature. Here m_r is the central moment of *r*-th order, *m’*$_1$ is the mean of the rank-frequency distribution. If one computes these indicators using Table 1, one obtains the results presented in Table 6.

Table 6
Ord’s criterion for the rank-frequency distributions of POS

Text	***N***	***I***	***S***
1949Einaudi	194	1.2189	1.2690
1950Einaudi	150	1.3551	1.7796
1951Einaudi	230	1.2456	1.2328
1952Einaudi	179	1.2934	1.4916
1953Einaudi	190	1.2558	1.7159
1954Einaudi	260	1.3259	1.5080
1955Gronchi	388	1.3265	1.3657
1956Gronchi	665	1.3169	1.2636
1957Gronchi	1130	1.4418	1.6456
1958Gronchi	886	1.3211	1.2585
1959Gronchi	697	1.3430	1.2641
1960Gronchi	804	1.3621	1.4859
1961Gronchi	1252	1.3493	1.3737
1962Segni	738	1.3019	1.4926
1963Segni	1057	1.4416	1.9439
1964Saragat	465	1.2493	1.3227
1965Saragat	1053	1.4232	1.5455
1966Saragat	1199	1.3796	1.6931
1967Saragat	1056	1.3945	1.8499

1968Saragat	1174	1.4314	1.6946
1969Saragat	1584	1.3685	1.5092
1970Saragat	1929	1.4163	1.7434
1971Leone	262	1.3658	1.8151
1972Leone	767	1.2888	1.6668
1973Leone	1250	1.3204	1.4109
1974Leone	801	1.2808	1.4538
1975Leone	1328	1.3247	1.3087
1976Leone	1366	1.3463	1.2948
1977Leone	1604	1.3791	1.2881
1978Pertini	1493	1.4681	1.5592
1979Pertini	2302	1.3731	1.3268
1980Pertini	1360	1.4281	1.6122
1981Pertini	2818	1.4712	1.3681
1982Pertini	2487	1.4968	1.5657
1983Pertini	3748	1.4769	1.5485
1984Pertini	1340	1.4785	1.7668
1985Cossiga	2359	1.3437	1.5755
1986Cossiga	1349	1.3284	1.3096
1987Cossiga	2091	1.3910	1.6895
1988Cossiga	2385	1.3920	1.5291
1989Cossiga	1912	1.4487	1.7592
1990Cossiga	3347	1.4227	1.6663
1991Cossiga	418	1.3746	1.4217
1992Scalfaro	2772	1.3698	1.2874
1993Scalfaro	2941	1.4133	1.4010
1994Scalfaro	3605	1.4114	1.5144
1995Scalfaro	4228	1.4571	1.5545
1996Scalfaro	2085	1.4205	1.6725
1997Scalfaro	5015	1.5348	1.6463
1998Scalfaro	3995	1.5167	1.6602
1999Ciampi	1941	1.4041	1.7731
2000Ciampi	1844	1.4005	1.7182
2001Ciampi	2097	1.4260	1.9352
2002Ciampi	2129	1.3727	1.6078

2003Ciampi	1565	1.3642	1.5782
2004Ciampi	1807	1.4228	1.8024
2005Ciampi	1193	1.3865	1.8298
2006Napolitano	2204	1.4168	1.3981
2007Napolitano	1794	1.4365	1.6750
2008Napolitano	1713	1.3698	1.5059
2009Napolitano	2293	1.4967	1.5254
2010Napolitano	2498	1.4541	1.6773
2011Napolitano	2365	1.4538	1.6501

The values of <I, S> can be seen in Figure 5

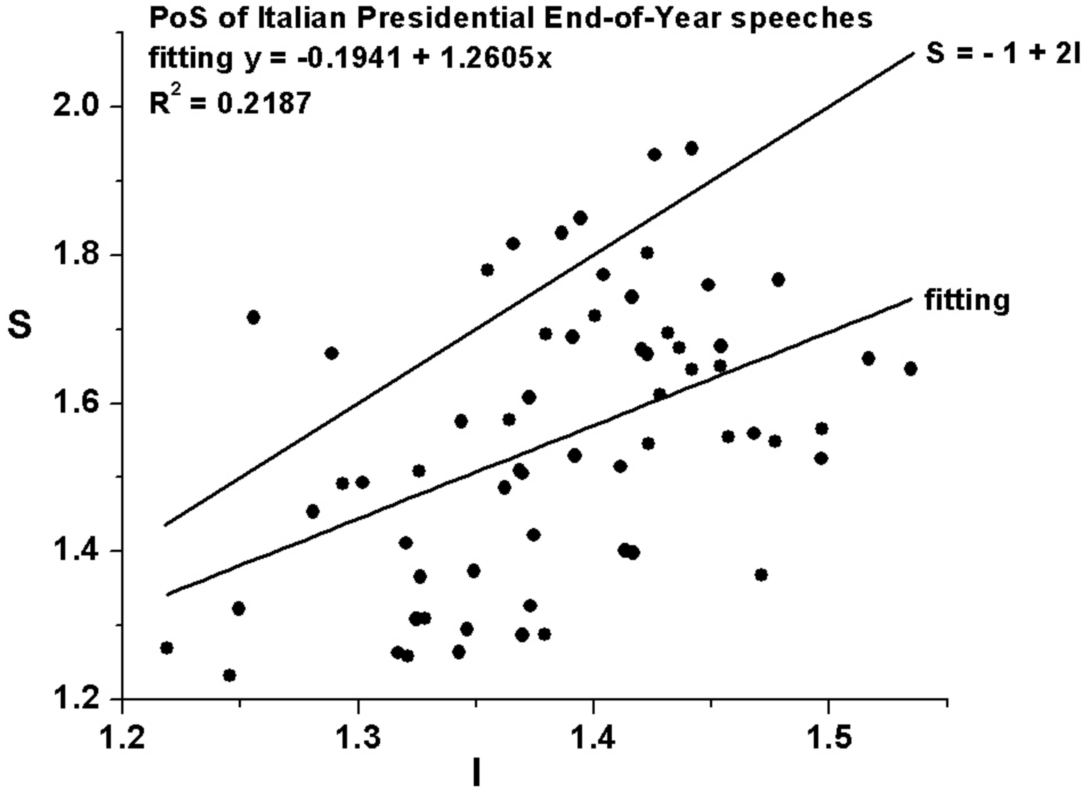


Figure 5. The <I, S> relation for the rank-frequency distributions.

Again, the increasing trend is visible but the dispersion is very great. In any case, the <I,S> points are placed in an ellipse whose main axis is the given straight line. The values lie in the domain of the negative hypergeometric (beta-binomial) distribution with eight exceptions (Einaudi 1950, 1953, Segni 1963, Saragat 1967, Leone 1971, 1972, Ciampi 2001, 2005) represented by the eight highest points in Figure 5. In general, the *S*-points that are smaller than 2*I* - 1 lie in the beta-binomial domain, that means, almost all of the texts can be modelled by the beta-binomial distribution. It cannot be said whether all Italian texts abide by this background mechanism.

The <I,S>-scheme of individual presidents looks more concentrated as can be seen in Figure 6 for Gronchi, Ciampi, Cossiga, and Scalfaro. Yet, the remainig data cannot be satisfactorily captured by a linear fitting ($y = a + bx$), as can be seen in the associated Table 7 below. Actually, it yields good R^2 only with two presidents, Gronchi and Ciampi. In general, perhaps the points spread out within an ellipse, depending on the actual writer(s) and style.

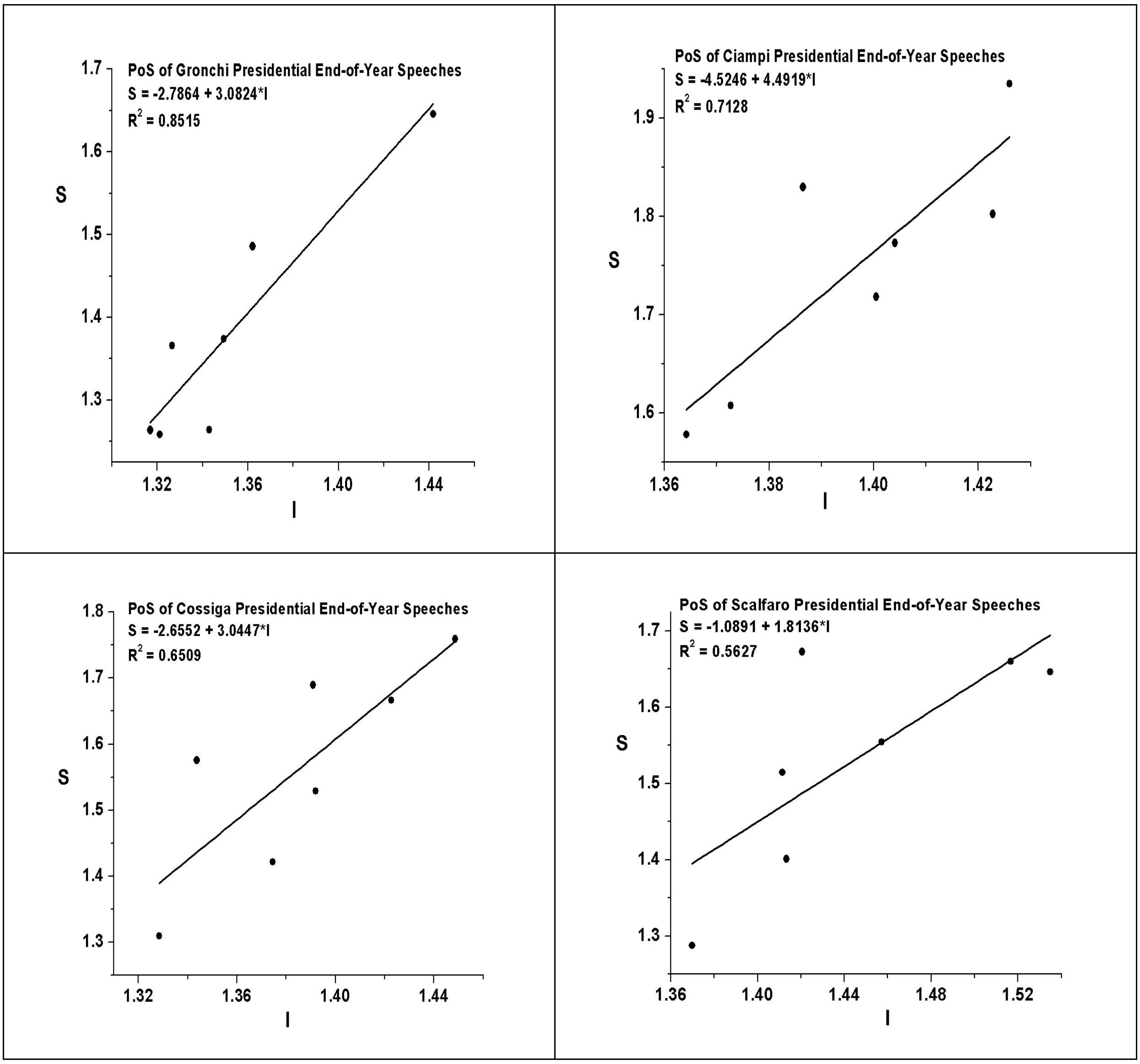


Figure 6. Some <I,S>-relation of individual presidents

Table 7
The slope b of the linear relation btween I and S

President	Speeches	Slope *b*	R^2
Einaudi	6	2.9375	0.4629
Gronchi	7	3.0824	0.8515
Segni	2	3.2305	1.0000

Saragat	7	2.0871	0.5470
Leone	7	-0.6531	0.0139
Pertini	7	1.8439	0.2727
Cossiga	7	3.0447	0.6509
Scalfaro	7	1.8136	0.5627
Ciampi	7	4.4919	0.7128
Napolitano	6	0.9091	0.1159

Ordering, e.g. the *I*-values according to years we obtain a relatively stable image. The slope of the straight line is positive but not significant. That means, the presidential speeches try to maintain the same distributional strategy.

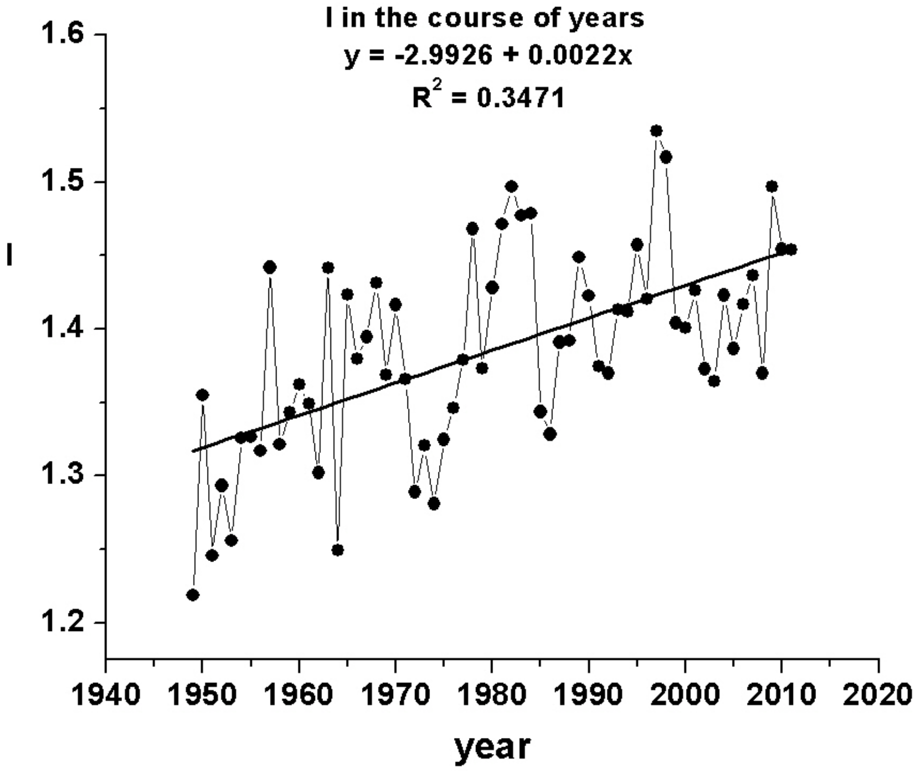


Figure 7. The values of *I* in the course of years

2. Comparisons

For comparing the individual speeches or presidents, the frequencies must be ascribed to the same POS, not to the ranks. Afterwards the frequencies could be compared and tests for homogeneity performed. However, we want to avoid the usual chi-square test for homogeneity because of some bad properties of this statistics. Instead, we fix the classes and each class obtains a rank number on the basis of its frequency. The ties will be taken into account. For a group of texts we compute Kendall's concordance coefficient $W \in <0;1>$ yielding zero for no concordance and 1 for full concordance. An example for rewriting the texts by Einaudi are shown in Table 8.

Table 8
Ranking the representation of individual POS in texts by Einaudi
(E49 = Einaudi 1949,…)

		E49	E50	E51	E52	E53	E54	T_i	T_i^2
1	**noun**	1	1	1	1	1	1	6	36
2	**preposition**	2	2	4	2	2	2	14	196
3	**verb**	3	3	2	3	4	4	19	361
4	**adjective**	4	4	3	4	3	3	21	441
5	**pronoun**	5	6	5	5	5	5	31	961
6	**article**	7	7	6	6	7	7	40	1600
7	**conjunction**	6	5	8	7	6	6	38	1444
8	**adverb**	8	8	7	8	8	8	47	2209
9	**numeral**	9	9	10	10	9	11	58	3364
10	**interjection**	10,5	10,5	10	10	10,5	9,5	61	3721
11	**proper noun**	10,5	10,5	10	10	10,5	9,5	61	3721
	Sums							396	18054

Here T_i is the sum of the row, T_i^2 is the square of the sum. K = 11 is the number of different POS and m the number of texts, here m = 6. As can be seen, in row 9 to 11 there are 4 ties of two cells and in 2 ties of three cells. The weight of the ties will be computed as

$$V = \sum_{k=1}^{m}(v_k^3 - v_k)\,, \tag{7}$$

yielding in our case $V = 4(2^3 - 2) + 2(3^3 - 3) = 24 + 48 = 72$. All these values are inserted in Kendall's formula (SS = sum of squares)

$$W = \frac{12SS}{m^2(N^3-N)-mV} = \frac{12\left(\sum_{i=1}^{m}T_i^2 - \frac{(\sum_{i=1}^{m}T_i)^2}{N}\right)}{m^2(N^3-N)-mV}. \tag{8}$$

Inserting the computed values in (8) we obtain

$$W = \frac{12\left(18054 - \frac{396^2}{11}\right)}{6^2(11^3-11)-6(72)} = 0.9679\,.$$

This value is so high that no test of significance is necessary. One usually transforms it to a chi-square or normal variable (cf. e.g. Bortz, Lienert, Boehnke 1990:469), but for us it is sufficient to know that concerning the distribution of POS, all texts by Einaudi display a high concordance.

If we perform this test for all presidents separately, we obtain the results presented in Table 9.

Taking all texts together we obtain for the 63 speeches W = 0.9450, that means, as to the distribution of POS all texts are concordant; no change can be observed, even if the common value is slightly smaller than the individual ones.

In general, the use of parts of speech is relatively constant in Italian. It agrees with our expectation, but at the same time it shows that some tests, e.g. the chi-square, must be used with caution. If one uses the chi-square test for homogeneity, one obtains in many cases signs of great non-homogeneities which are due to the weakness of the chi-square.

Table 9
Concordance of POS in texts of individual presidents

President	**Years**	**W**
Einaudi	1949-54	0.9679
Gronchi	1955-61	0.9469
Segni	1962-63	0.9954
Saragat	1964-70	0.9712
Leone	1971-77	0.9621
Pertini	1978-84	0.9775
Cossiga	1985-91	0.9466
Scalfaro	1992-98	0.9523
Ciampi	1999-2005	0.9711
Napolitano	2006-11	0.9584

3. Distances

While the preceding sections are related to the frequency distribution of POS, we turn now to another aspect, namely the question which are the regularities underlying the repetition of POS. To be concrete, we inerprete any president’s speech as a formal sequence $S = (s_1,\ldots,s_n)$ of length n, the elements of which are chosen from the set of word classes

$$W = \{\text{a, avv, conj, det, esc, n, nm, num, prep, pron, v}\}.$$

The elements of W represent adjective, adverb etc. (see above). For the sake of illustration we consider the following hypothetical short “speech”

$$S = (\text{a, v, n, v, det, a, a, v, v, n, det, a}).$$

For any $r \in W$ we define the *distance* between two consecutive elements of type *r* as the number of elements ≠ *r*, lying between them. For example, the distance between the second and the third *v* in the foregoing sequence is 3. For a given text we are interested in the distribution of the distance *x*. For this small example we obtain the following observed frequencies:

x	frequency
0	2
1	1
2	0
3	1
4	2
5	1
6	1

In particular, the distance 4 appears two times: between first and second *a*, and between the third and fourths *a*. Further details on the distribution of the distances between identical elements of a sequence can be found in Zörnig (2010, 2013).

For the 63 studied speeches, the observed distance frequencies are given in the upper part of Table 10. For example, in speech 3, the distance 5 occurs 13 times. The line "Rest" contains the sum $\sum_{x>20} f_x$, where f_x denotes the frequency of the distance *x*. We took into consideration only the distances 0 to 20 because the "rest" is mostly filled with zeroes. We tried to fit the extended positive negative binomial distribution (EPNB) to the observed distance frequencies in Table 10. By means of the software Altmann-Fitter (1997), the EPNB has been proved to be one of the best suited discrete probablity models for the given data. This model, which may be justified linguistically as follows, is given in Wimmer, Altmann (1999: 49) as:

$$(9) \qquad P_x = \begin{cases} 1-\alpha, & x=0 \\ \dfrac{\alpha \binom{k+x-1}{x} p^k (1-p)^k}{1-p^k}, & x=1,2,3,... \end{cases}$$

$$k > 0,\ 0 < p, \alpha < 1$$

The EPNB can be considered as a zero-modified negative binomial distribution, i.e. it is obtained by setting the theoretical frequency of the zero class equal to the observed one and by adjusting the negative binomial distribution to the remaining observed frequencies.

In the six lines below the line "Rest" of Table 9, we indicate the result in fitting model (9): *k*, α , *p* denote the optimal parameter values, determined iteratively, X^2 is the observed chi-square value and *P* the probability to exceed this value. Finally, *DF* means the number of degrees of freedom. This number is always 18 except for the second speech where *DF* = 17.

Modelling frequency distributions in linguistics, we always start from the assumption that there is an attractor value prescribed by the given language (say *a*) which is steadily changed by the speaker/writer on the basis of some boundary circumstances concerning style, aim, text sort, etc. This "force" of the speaker is usually symbolized as *g*(*x*), being mostly a simple function. However, language must be held in equilibrium, hence the hearer controls the changes made by the speaker and does not allow him to wander in another attractor, otherwise the text could get incomprehensible. This “force” can be called *h*(*x*). Hence, if we approach the data using a discrete model, we use either a stochastic process, or, in order to hold the mathematics at lower level, we model the phenomenon by the resulting difference equation. As a matter of fact, if there is a distribution of data in linguistics, then the neighbouring classes are always linked by some kind of proportionality. In most cases we have the basic model

(10) $$P_x = f(x)P_{x-1} = \frac{g(x)}{h(x)}P_{x-1}.$$

Considering *g*(*x*) as the expression of the state of the phenomenon in general plus contribution of the speaker, we obtain in simple cases $g(x) = a + bx$. The hearer controls this influence by $h(x) = cx$. Inserting these suppositions in (10), we obtain the negative binomial distribution, and modifying the zero-class yields model (9). The zero-class plays a special role: in some languages the sequence of equal POS is not allowed by grammar, in other ones it is quite usual. Hence both for Italian or other languages this class seems to play a special role in any language.

However, the EPNB was not as satisfactory as it seemed to be after fitting it to the first four presidential speeches. In fact, it could only be well fitted in 15 of 63 cases, namely to the texts No. 1–4, 7, 9, 10, 17, 19, 23, 26–29, 41. Probably a greater family of distributions linked with the negative binomial would be necessary in order to obtain better results (e.g. text No. 5 can be better captured by the mixed negative binomial, etc.). An explanation for the difficulties encounterd in fitting the EPNB might be the fact that a presidential speech is corrected by several persons, formulations are exchanged, it is made “smooth”, etc. Other reasons for the problems in adjusting the model could be the smallness of samples or possible irregularities which cannot be captured by the given model.

Hence, in order to obtain an adequate probability distribution we should know all boundary conditions responsible for the structuring of each individual text. This is simply impossible. Anyway, we have discused the EPNB here for comparison purposes, since it appears to be one of the best adequate discrete models.

Since no model is that good that it captures the “truth”, we approach it merely step by step. But since “discrete” and “continuous”, “finite” and “infinite” etc. are not properties of reality but merely properties of our models – our concepts – we may approach our problem using also a continuous model or a continuous summing (integration) and consider the result as discrete. We venture a change in *g*(*x*) and set $g(x) = c + k\ ln\ x$ and considering *g*(*x*)/*h*(*x*) as the relative rate of change of frequencies which we call simply *y*, we obtain the differential equation

(11) $$\frac{dy}{y} = \frac{c + k\ \ln\ x}{mx}dx.$$

After simplification of the parameters its solutions yields

(12) $$y = Cx^{a\,+\,b\,ln\,x},$$

representing the Zipf-Alekseev function. Here *y* is the frequency and *x* is the distance. In order to use this function, we consider it discrete and count the distances in the following way: distance is the number of steps which must be made in order to come from an entity to the next identical entity. Practically it means that we shift the frequencies in Table 9 one place to the right, i.e. distance zero gets distance 1, etc., otherwise we would obtain no result for $x = 0$.

Since now we have a discrete sequence of values, we fit (12) to the data and consider the determination coefficient R^2 as decisive for acceptation. The results can be found in the lower parts of Table 10. Of course, one can change (12) in a discrete distribution if one considers C a normalizing constant and restricts the support to $x = 1,2,\ldots,n$. For our purposes it is not necessary.

In this model a is again the status quo in the language, b is the "force" of the writer and C is the initial value. In this way we avoid the special marking of the smallest distance because it is estimated by C. It is to be noted that (11) is merely a "lenification" of the force of the writer. In the discrete model, we have the difference equation $P_x = \frac{a+bx}{cx} P_{x-1}$, yielding the negative binomial distribution, in (11) the writer influences only the logarithm of the distance.

Table 10a
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	1	2	3	4	5	6	7	8
0	15	4	24	16	19	16	26	40
1	31	30	31	34	18	35	54	101
2	29	19	26	20	29	29	46	101
3	18	16	24	13	26	35	39	68
4	12	12	12	16	17	27	34	62
5	12	10	13	13	15	11	33	33
6	7	11	18	9	9	19	24	31
7	9	4	10	8	6	15	23	37
8	8	3	12	3	4	7	11	24
9	6	5	9	3	2	12	11	15
10	4	1	7	6	3	4	10	15
11	5	3	2	4	3	4	13	17
12	2	1	6	2	1	4	5	18
13	2	3	0	1	1	1	6	4
14	1	1	2	4	0	8	3	14
15	3	1	4	4	2	3	3	7
16	3	4	1	0	4	0	3	5
17	3	1	1	1	2	0	1	4
18	5	1	3	1	1	2	1	6
19	0	0	0	1	4	2	2	3
20	0	0	1	0	3	1	2	6
Rest	10	10	15	12	12	16	28	45

k	0.5671	0.3093	0.6259	0.2124	0.9519	0.7748	0.7115	0.6809
p	0.0965	0.077	0.1052	0.0653	0.109	0.1225	0.1105	0.1039
α	0.9189	0.9714	0.8914	0.9064	0.895	0.9363	0.9312	0.939
X^2	16.05	14.95	21.14	16.52	44.57	30.83	21.85	13.12
P	0.5892	0.5989	0.2723	0.5561	0.0005	0.0302	0.2388	0.0161
DF	18	17	18	18	18	18	18	18
a	1.2603	2.1369	0.6423	0.9981	1.4156	1.5566	1.3109	1.5914
b	-0.8099	-1.1789	-0.5123	-0.7252	-0.8213	-0.8153	-0.7257	-0.8941
C	17.4036	9.1407	24.3861	18.992	14.5721	16.5245	27.9738	47.8433
R^2	0.93	0.87	0.92	0.89	0.89	0.91	0.97	0.95

Table 10b
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
0	70	49	34	39	70	46	66	36
1	178	140	105	121	209	101	176	56
2	162	102	112	113	161	125	140	66
3	118	90	86	93	135	74	122	55
4	69	85	60	62	109	51	91	46
5	76	63	38	49	89	54	55	27
6	62	50	31	50	53	50	59	23
7	59	39	27	41	52	30	44	23
8	32	30	11	32	37	27	36	14
9	37	30	30	31	42	22	39	4
10	27	31	19	28	42	13	19	10
11	26	23	16	13	39	12	8	14
12	22	17	18	10	26	15	25	11
13	20	9	16	8	20	10	8	6
14	15	13	11	10	15	14	12	7
15	9	7	4	1	20	8	15	6
16	14	11	7	9	9	5	12	8
17	16	10	3	8	7	10	8	3
18	8	6	5	3	4	7	11	2
19	5	8	7	3	11	2	10	4
20	11	9	5	7	7	4	4	4
Rest	84	54	42	63	85	49	87	30

k	0.5759	0.5732	0.7797	0.666	0.5408	0.779	0.5001	0.914
p	0.092	0.1024	0.1133	0.1003	0.096	0.1112	0.0842	0.1182
a	0.9375	0.9441	0.9505	0.9505	0.9436	0.9369	0.937	0.9209
X^2	28.2	18.14	48.59	43.09	30.45	46.8	44.27	37.57
P	0.0591	0.4463	0.0001	0.0008	0.0333	0.0002	0.0005	0.0044
DF	18	18	18	18	18	18	18	18
a	1.4238	1.3838	2.0137	1.6926	1.4929	1.6153	1.5507	1.4084
b	-0.8297	-0.7641	-1.0733	-0.8967	-0.8456	-0.8868	-0.8984	-0.7919
C	87.2605	63.2462	41.8791	51.0835	92.6848	51.1299	80.6542	34.0047
R^2	0.95	0.95	0.94	0.96	0.95	0.94	0.96	0.96

Table 10c
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
0	67	80	78	60	104	112	16	51
1	201	217	174	217	254	344	45	117
2	134	157	139	157	198	241	43	113
3	100	136	98	121	161	188	23	75
4	81	94	79	97	139	149	18	54
5	55	86	75	63	99	121	17	58
6	43	54	48	57	105	116	16	54
7	53	61	49	58	55	89	11	33
8	41	40	40	36	46	72	10	40
9	31	25	38	40	55	59	7	16
10	21	22	22	27	52	59	3	18
11	27	18	19	21	41	38	2	15
12	20	24	22	29	25	29	9	14
13	21	18	13	14	28	37	3	10
14	13	19	16	17	27	17	2	4
15	7	8	8	14	16	16	1	5
16	10	7	15	11	13	31	1	11
17	11	11	11	13	14	11	3	7
18	12	10	8	4	18	23	3	4
19	4	6	7	5	9	10	3	5
20	6	6	6	4	9	8	2	3
Rest	84	89	80	98	105	148	14	50

k	0.3013	0.4434	0.4811	0.3827	0.5610	0.3897	0.6015	0.6951
p	0.0685	0.0854	0.0854	0.076	0.0981	0.0790	0.1050	0.1105
a	0.9357	0.9327	0.9254	0.9488	0.9339	0.9416	0.9365	0.9326
X^2	19.15	35.28	17.13	30.98	30.64	41.82	18.32	37.01
P	0.3824	0.0007	0.514	0.0289	0.0317	0.0012	0.4345	0.0052
DF	18	18	18	18	18	18	18	18
a	1.4127	1.5061	1.1934	1.6529	1.2897	1.3774	1.6672	1.389
b	-0.8861	-0.9027	-0.7493	-0.9599	-0.757	-0.8147	-0.9844	-0.7971
C	91.4858	99.4463	93.1266	88.229	126.7531	154.7161	20.3653	59.5986
R^2	0.91	0.96	0.96	0.92	0.96	0.92	0.92	0.95

Table 10d
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
0	79	58	96	98	91	143	208	103
1	196	139	208	218	211	204	257	186
2	167	100	162	161	153	156	229	137
3	130	80	142	123	145	144	243	142
4	122	64	97	113	118	109	234	141
5	82	43	75	95	90	87	152	100
6	79	54	84	68	71	94	163	81
7	50	33	64	74	70	62	123	66
8	39	23	41	56	61	70	86	41
9	33	32	37	51	43	54	85	45
10	25	18	40	35	35	44	62	39
11	22	18	35	36	30	38	60	27
12	17	17	29	27	31	37	37	23
13	20	10	25	20	23	25	37	21
14	15	5	14	13	18	22	24	25
15	22	9	16	16	24	16	31	14
16	14	4	14	21	15	11	17	13
17	10	6	12	9	11	18	18	13
18	14	6	10	11	9	12	29	11
19	8	7	8	12	8	5	23	11
20	5	8	8	5	13	15	10	10
Rest	91	57	101	94	86	117	163	100

k	0.6134	0.4303	0.5190	0.4929	0.5643	0.5489	0.9115	0.6433
p	0.0980	0.0845	0.0891	0.0917	0.1003	0.0902	0.1229	0.1026
α	0.9363	0.9267	0.9272	0.9277	0.9329	0.9036	0.9092	0.9236
X^2	39.88	23.44	21.70	27.62	20.40	35.39	84.71	42.21
P	0.0022	0.1742	0.2457	0.0681	0.3111	0.0084	0.0000	0.0010
DF	18	18	18	18	18	18	18	18
a	1.4964	1.2396	1.1551	1.0677	1.1827	0.6564	0.881	1.1044
b	-0.8452	-0.7755	-0.7098	-0.6605	-0.6921	-0.4978	-0.5408	-0.6443
C	92.722	70.8384	112.8847	117.6646	108.5249	149.513	191.2796	106.7775
R^2	0.97	0.94	0.96	0.95	0.96	0.98	0.96	0.97

Table 10e
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
0	283	253	335	114	163	111	123	147
1	331	282	421	174	429	228	384	391
2	295	269	443	139	341	168	281	301
3	232	248	379	130	248	145	215	248
4	236	195	338	124	161	121	177	204
5	212	166	282	98	144	65	127	158
6	206	162	195	73	121	67	96	114
7	131	117	165	50	90	54	81	106
8	134	101	142	67	84	61	74	95
9	77	70	111	49	87	36	59	70
10	76	65	89	29	65	31	47	65
11	47	55	97	23	57	34	48	52
12	55	54	82	34	25	22	28	37
13	55	49	71	24	25	24	31	40
14	44	48	54	17	30	26	32	48
15	42	35	57	14	24	13	31	30
16	32	27	36	12	25	8	22	27
17	29	28	33	10	19	11	22	26
18	30	22	34	13	21	10	21	16
19	25	15	33	4	14	5	12	11
20	29	13	25	12	9	4	6	15
Rest	206	201	315	120	166	96	164	173

k	0.7709	0.789	0.823	0.6093	0.4727	0.4717	0.4036	0.5194
p	0.1093	0.1052	0.1046	0.0913	0.088	0.0885	0.0782	0.091
α	0.8992	0.8978	0.9104	0.9143	0.9306	0.9172	0.9409	0.9381
X^2	69.09	44.53	101.86	50.83	47.01	35.9	27.43	28.77
P	0	0.0005	0	0.0001	0.0002	0.0073	0.0712	0.0513
DF	18	18	18	18	18	18	18	18
a	0.5967	0.623	0.9076	0.904	1.4698	1.1495	1.5822	1.3702
b	-0.4532	-0.4698	-0.5711	-0.5766	-0.9058	-0.7454	-0.9393	-0.7960
C	274.9587	243.4574	318.7885	115.5113	204.8422	127.1658	166.0769	185.9161
R^2	0.97	0.99	0.99	0.98	0.95	0.96	0.94	0.96

Table 10f
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
0	112	203	26	198	230	272	297	143
1	357	542	73	405	484	599	623	388
2	262	511	70	352	377	483	562	278
3	196	341	42	327	287	354	435	231
4	130	262	31	247	233	293	344	150
5	105	227	22	166	189	244	250	137
6	89	159	17	137	151	177	258	84
7	71	142	12	106	114	137	198	95
8	60	100	17	106	91	127	139	69
9	63	90	11	79	85	109	115	52
10	63	69	14	69	77	89	113	44
11	39	75	4	67	52	66	96	44
12	29	61	5	46	74	70	80	35
13	27	54	6	54	48	61	72	16
14	27	48	7	32	35	55	70	25
15	31	47	3	29	38	45	51	25
16	15	27	3	29	31	33	39	21
17	16	31	8	33	22	26	33	20
18	21	26	1	25	32	15	44	22
19	20	27	7	23	22	27	38	13
20	15	26	1	13	15	16	27	16
Rest	154	268	27	218	243	296	333	167

k	0.365	0.5612	0.5401	0.6342	0.4522	0.4775	0.6146	0.3816
p	0.072	0.0872	0.0899	0.0949	0.0792	0.0823	0.0932	0.0753
α	0.9411	0.9391	0.9361	0.9283	0.9215	0.9243	0.9296	0.9311
X^2	22.23	88.86	31.35	71.97	37.01	60.91	69.13	38.77
P	0.222	0	0.0262	0	0.0052	0	0	0.0031
DF	18	18	18	18	18	18	18	18
a	1.602	1.6304	1.8809	1.3311	1.1643	1.2603	1.2581	1.5122
b	-0.9742	-0.9373	-1.1182	-0.7754	-0.7467	-0.7866	-0.7432	-0.9345
C	154.5795	249.1718	31.599	217.242	267.2656	318.2097	338.4312	179.4397
R^2	0.92	0.96	0.94	0.98	0.97	0.97	0.98	0.95

Table 10g
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**
0	452	318	118	135	125	128	84	118
1	677	588	310	266	363	332	273	311
2	577	473	292	251	264	280	200	248
3	549	417	184	201	204	255	163	150
4	421	364	145	153	186	160	132	144
5	309	242	137	108	116	143	109	121
6	285	210	122	97	128	106	88	87
7	224	184	78	77	93	96	62	86
8	191	132	76	71	74	77	57	61
9	148	109	55	43	53	79	37	38
10	95	103	35	53	57	48	32	52
11	106	86	42	48	41	42	48	39
12	94	69	31	30	34	38	34	36
13	88	66	25	20	24	33	28	37
14	76	56	24	17	34	21	11	24
15	50	51	21	25	19	26	18	24
16	55	45	13	23	28	23	15	18
17	43	38	12	22	20	11	8	18
18	38	35	19	21	12	16	10	8
19	37	33	11	9	19	18	10	13
20	36	25	8	8	15	13	8	15
Rest	451	340	170	156	174	170	124	148

k	0.6022	0.5469	0.5384	0.6165	0.4084	0.5802	0.4466	0.4182
p	0.0878	0.0861	0.0848	0.0897	0.0773	0.0919	0.0835	0.0772
α	0.9096	0.9202	0.9388	0.9264	0.94	0.9395	0.9458	0.9343
X^2	121.21	66.58	74.41	59.81	39.15	57.03	39.57	28.19
P	0	0.0005	0	0	0.0027	0	0.0024	0.0593
DF	18	18	18	18	18	18	18	18
a	0.9398	1.1009	1.5334	1.3026	1.4224	1.5107	1.5494	1.3291
b	-0.6186	-0.686	-0.8792	-0.7717	-0.8366	-0.8503	-0.8814	-0.8142
C	459.7704	345.5195	145.878	147.836	164.9233	155.1139	115.8548	151.8823
R^2	0.99	0.98	0.96	0.98	0.94	0.97	0.94	0.93

Table 10h
Distances between equal parts-of-speech
(First row: order number of the text. First column: distance)

d/T	**57**	**58**	**59**	**60**	**61**	**62**	**63**
0	103	123	118	112	168	160	150
1	208	330	284	261	369	407	395
2	153	277	266	243	241	354	314
3	114	228	177	161	213	286	238
4	99	178	132	139	208	173	191
5	75	150	111	118	157	185	172
6	53	109	93	91	121	109	145
7	54	103	72	75	104	91	69
8	42	90	64	70	87	91	88
9	24	78	45	50	66	85	64
10	22	70	38	45	61	55	60
11	17	50	36	45	52	44	54
12	19	35	28	28	51	40	51
13	14	48	35	26	45	32	36
14	16	27	28	26	33	41	21
15	14	20	25	17	31	27	29
16	7	31	21	14	31	26	21
17	11	18	20	11	28	25	20
18	5	19	17	10	25	14	22
19	9	17	17	12	10	17	13
20	4	13	8	18	9	14	16
Rest	117	178	147	131	172	212	185

k	0.3498	0.582	0.5398	0.6059	0.3962	0.5259	0.5003
p	0.0655	0.0918	0.0839	0.0943	0.0798	0.0842	0.087
a	0.9127	0.9439	0.9338	0.9342	0.9264	0.9357	0.9363
X^2	37.14	33.45	37.21	37.25	54.08	79.59	51.75
P	0.005	0.0147	0.0049	0.0049	0	0	0
DF	18	18	18	18	18	18	18
a	1.1842	1.3794	1.4606	1.3587	1.0215	1.5485	1.4210
b	-0.7953	-0.766	-0.8659	-0.789	-0.6428	-0.9015	-0.8329
C	116.5316	156.4256	141.7161	133.2378	199.8833	192.342	187.1029
R^2	0.97	0.96	0.96	0.97	0.94	0.96	0.96

While the *C*-values in the Zipf-Alekseev function represent the first frequency, the parameters *a* and *b* (i.e. the state of the language and the influence of the writer) are strongly associated. As can be seen in Table 11, there is a highly significant linear relationship between them. The relationship is computed separately for each president.

Table 11
Linear relationhips between the parameters *a* and *b* in Zipf-Alekseev function concerning the distances ($x = a, y = b$)

President	**a,b linear fitting**	**R^2**
Einaudi	y = - 0.2624 - 0.4105x	0.9424
Gronchi	y = - 0.1375 - 0.4645x	0.9622
Segni	y = - 1.1769 + 0.1796x	1.0000
Saragat	y = - 0.1297 - 0.5034x	0.8605
Leone	y = - 0.1156 - 0.5061x	0.9520
Pertini	y = - 0.2510 - 0.3520x	0.9708
Cossiga	y = - 0.1256 - 0.5182x	0.9517
Scalfaro	y = - 0.1180 - 0.5212x	0.9332
Ciampi	y = - 0.4521 - 0.2710x	0.8097
Napolitano	y = - 0.1317 - 0.4894x	0.9425

4. Conclusion

The class of POS considered here is very small, hence the results are quite uniform. The speeches do not diverge, showing that Italian is very stable in this respect. The Repeat rates display a very slow decrease with great dispersion, the opposite trend can be found with Entropy.

As can be seen, sometimes an approximation of the distribution of frequencies using a continuous function based on the same principles as its discrete counterpart but a slightly changed “speaker force” may yield much more satisfactory results. All distributions of distances have a short concave part at the beginning which cannot be adequately captured by a discrete distribution even if one modifies P_0. This is at the same time a good occasion to show that modelling is no capturing of “truth” or an intrinsic property of phenomena but merely our conceptual approach. It is the better the more acceptable is our deduction of the model based on linguistic circumstances and the easier its systematization in a theory.

References

Bergenholtz, H. Schaeder, B. (1977). *Die Wortarten des Deutschen.* Stuttgart: Klett

Best, K.-H. (2005^3). *Linguistik in Kürze*. Göttingen: Script.

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

Bunge, M. (1983). *Treatise on Basic Philosophy 6. Understanding the World*. Dordrecht: Reidel.

Kroeger, P. (2005). *Analyzing Grammar: An Introduction*. Cambridge: Cambridge University Press.

Ord, J.K. (1972). *Families of discrete distributions.* London: Griffin.

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

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM.

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

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

Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational Statistics & Data Analysis 54, 2317-2327.*

Zörnig, P. (2013). Distances between words of equal length in a text. To appear in: Köhler, R. Altmann, G. (eds.), *Issues in Quantitative Linguistics 3. To honour Karl-Heinz Best on the occasion of his 70th birthday*. Lüdenscheid: RAM-Verlag.

Glottometrics 24, 2012, 70-78

Models of morph lengths: Discrete and continuous approaches

Emmerich Kelih[1]
Peter Zörnig[2]

Abstract. We discuss a discrete and a continuous approach for modelling the distribution of morph length. The proposed models (1-displaced extended binomial distribution, beta function) have been successfully fitted to Spanish, Russian and Slovenian data records. It can be shown that the models are suitable for modelling word form types as well as word form tokens.

Keywords: Morph length, Spanish, Russian, Slovenian

0. Introduction

Unlike sentence, syllable and word lengths, the distribution of the morph length has so far received little attention in the literature (see Best 2001, 2005a). In the studies of Best it is assumed that the morph length basically follows the same laws as the word length, which can be justified linguistically by a variable proportional relationship or a generalized relation (see for example Eq. (1) and (3) in Best (2005 b)). It is now reasonable to try to fit these probabilistic models to morph length distributions that have proved to be adequate for the word length. In particular, Best (2001) has successfully fitted the 1-shifted Hyperpoisson distribution to the distribution of morph lengths in journalistic texts of a German newspaper and Rottmann (2003) has fitted the extended positive binomial distribution (EPB) to word length distributions (measured in the number of syllables) of Latvian and Lithuanian texts. Generally, it is an inductive attempt to find a suitable model based on the proportionality approach of Wimmer and Altmann (2005).

In the present paper, we show that the EPB is suitable to model the morph length distribution. We start with a data set of Saporta (1966), which is mentioned by Köhler and Altmann (2009: 78–79) as an open problem of quantitative linguistics. In addition, we provide new data sets from Russian and Slovenian, for which the EPB has also been proved appropriate. The morph length is measured in the number of phonemes and its determination is performed on the level of word form types and word form tokens. Since the morph length is a discrete random variable it is natural to model it by means of a discrete probability model. However, in quantitative linguistics, continuous models have also been considered to describe discrete observations, since “discrete” and “continuous” are considered to be mere conceptual properties see for example Tuzzi et al. (2012, section 3) and in particular Mačutek and Altmann (2007). Therefore we also present as an alternative to the discrete approach a continuous model based on the beta function to describe the morph length.

[1] Institut für Slawistik, Universität Wien, emmerich.kelih@univie.ac.at.
[2] Departamento de Estatística, Universidade de Brasília, peter@unb.br.

1. Spanish data record (Saporta, 1966)

Before re-analyzing the data from Saporta (1966: 69) it should be noted that the author does not provide further information about the performed morph segmentation. From the data record it is only evident that a class of zero-morphs is considered. To ensure the comparability of our study we do not consider zero morphemes[3] as a separate class either in the re-analysis of the Spanish data or in our analysis of the Slovenian and Russian data. In Saporta (1966) the morph length has been measured in the number of phonemes. In Table 1, the observed morph frequency f_x is listed as a function of the phoneme number x, wherein, as mentioned above, the class of the zero-morphs was omitted and the last five classes have been pooled.

Table 1
Spanish morph lengths (Saporta 1966)

x	$\mathbf{f_x}$	**discrete, NP$_x$**	**continuous, y(x)**
1	59	59.00	25.40
2	97	114.95	139.04
3	307	265.89	284.86
4	387	378.45	370.30
5	327	370.34	353.59
6	261	263.57	257.81
7	143	140.68	140.41
8	64	57.21	52.06
9	19	17.81	10.26
10	9	5.08	0.47
EPB d.	n = 14, p = 0.2625, α = 0.964, C = 0.0097		
Beta f.	m = 0, M = 11, C = 0.000295, a = 3.2027, b = 4.9355, R^2 = 0.9736		

Based on the unified theory (cf. Wimmer, Altmann 2005) we suppose that the occurrence probabilities of magnitudes of properties develop in such a way that neighbouring classes stay in proportional relation to each other. Practically,

$$(1) \qquad P_x = f(x)P_{x-1}$$

where $f(x)$ is a proportionality function changing with x. In many cases it is simply the ratio $f(x) = g(x)/h(x)$, where $g(x)$ contains a language constant and the function of change performed by the speaker, and $h(x)$ is the controlling function of the hearer. In our case we consider the possibility that the language constant is a, and the effect of the speaker is $-bx$,

[3] The question of whether so-called zero-morphemes must be considered in the analysis of morphs must be decided prior to the examination. In our analysis no zero-morphemes are considered. However, one should bear in mind that considering a "zero" class causes a number of serious methodological and theoretical problems. The situation is quite similar to that of zero-syllabic words in Slavic languages – a problem which has been discussed at great length in Antić, Kelih and Grzybek (2006). When considering zero-classes in statistical modelling, in particular the following must be considered: the number of zero-morphemes is probably very low in any language and this low frequency presumably causes a "high jump" from this class to the class of morphemes of length 1, which in turn can lead to difficulties in modelling.

while the control function of the hearer is $h(x) = cx$. Inserting these assumptions into (1) we obtain

(2) $$P_x = f(x)P_{x-1} = \frac{g(x)}{h(x)}P_{x-1} = \frac{a-bx}{cx}P_{x-1} = \frac{\frac{a}{b}-x}{x}\frac{b}{c}\mathrm{P}_{x\text{-}1}$$

By substituting $a/b = n + 1$ and $b/c = p/q$, we obtain the well-known recurrence formula

(3) $$P_x = \frac{n-x+1}{x}\frac{p}{q}P_{x-1},$$

whose solution is the binomial distribution

(4) $$P_x = \binom{n}{x}p^x q^{n-x}, \quad x = 0,1,...,n.$$

Now, since morphs of zero lengths are excluded, we truncate (4) at $x = 1$ and obtain the truncated binomial distribution

(5) $$P_x = \binom{n}{x}\frac{p^x q^{n-x}}{1-q^n}, \quad x = 1,2,...,n.$$

Since the first frequency, i.e. P_1, yields in each case a special value, we displace the distribution one step to the right and define P_1 separately by setting $P_1 = 1 - \alpha$. This yields a modified distribution that can be called 1-displaced extended binomial distribution
(1-displaced EPB), defined as

(6) $$P_x = \begin{cases} 1-\alpha, & x = 1 \\ \dfrac{\alpha\dbinom{n}{x-1}p^{x-1}q^{n-x+1}}{1-q^n}, & x = 2,3,...,n+1 \end{cases}$$

where $0 < \alpha$,$p < 1$, $q = 1- p$.

By means of the software Altmann-Fitter (1997), the EPB has been proved to be suitable for the Spanish data.

The probability $P_1 = 1- \alpha$ is set equal to the corresponding relative frequency, i.e. the estimator $\hat{\alpha}$ satisfies

(7) $$1- \hat{\alpha} = \frac{f_1}{N},$$

where $N = \sum_x f_x$ denotes the number of all morphs (sample size). The parameters p and n are iteratively determined.

By fitting the 1-shifted EPB to the data in Table 1, we obtain the theoretical frequencies in the third column. The penultimate line of Table 1 lists the optimal parameter values n, p, α and the contingency coefficient $C = \chi^2/N$.

Since the sample size is very large (here as well as in all following data sets) the chi-square value is also high and, as is often done in quantitative linguistics in this case (see for example Rottmann (2003: 53)), one can use the coefficient C as a criterion to decide on the goodness of fit. A fit is considered good if $C \le 0.01$ and satisfactory if $C \le 0.02$. Therefore the fit of the EPB in Table 1 can be considered satisfactory.

Our continuous approach is based on the assumption that the relative rate of change of the morph frequency y is proportional to the rate of change of the number x of phonemes, i.e.

(8) $$\frac{dy}{y} \sim dx.$$

As in the discrete case we assume that the proportionality is not given by a constant but by a function g(x), involving impacts of speaker and hearer. This leads to

(9) $$\frac{dy}{y} = g(x)dx.$$

Now the function $g(x)$ is composed of a difference of speaker and hearer portions. The former can be expressed as

(10) $$\frac{a}{x-m},$$

where a is the speaker-force and m the minimum value of x. The farther away x is from m, the less the impact of the speaker is. Similarly, the hearer portion is expressed as

(11) $$\frac{b}{M-x},$$

where b is the permanent force of the hearer and M the maximum value of x. The farther away x is from M, the stronger the impact of the hearer is. The two forces are considered to be in equilibrium.

Expressing the function $g(x)$ in (9) by (10) and (11), we get the relation

(12) $$\frac{dy}{y} = (\frac{a}{x-m} - \frac{b}{M-x})dx$$

which has the simple solution

(13) $$y = C(x - m)^a (M - x)^b \quad \text{for} \quad m \le x \le M.$$

This is a function with five parameters C, m, M, a and b ($C > 0$; $m < M$; $a, b > -1$) which can have many different shapes; for details see Altmann and Grotjahn (1988) and Köhler and Altmann (1986). The parameters m and M can be directly estimated from the observed phoneme numbers. To ensure that the range [*m, M*] of the model contains the observed *x*-values, it must hold $m \leq x_{min}$ and $M \geq x_{max}$, where x_{min} and x_{max} are the minimum and maximum of the observed phoneme numbers, respectively. In the present article we have always chosen

(14) $m = x_{min} - 1$ and $M = x_{max} + 1$.

We fitted (13) to the data in Table 1, where *m* and *M* have been chosen as in (14) and *C, a, b* are considered as three freely selectable parameters ($C > 0$; $a, b > -1$) which have been optimized iteratively. The optimal values of *C, a* and *b* and the coefficient of multiple determination R^2, used as a measure for the "goodness of fit", are listed in the last line of Table 1. The coefficient R^2, also known as "proportion of variance explained", is defined by

(15) $$R^2 = 1 - \frac{\sum_x (f_x - y(x))^2}{\sum_x (f_x - \bar{f})^2}$$

where $\bar{f} = \frac{1}{x_{\max} - x_{\min} + 1} \sum_x f_x = \frac{N}{x_{\max} - x_{\min} + 1}$ is the mean of the observed frequencies.

The frequencies predicted by the continuous model, i.e. the values *y*(*x*) obtained from (13) for the optimal parameter values, are shown in the last column of Table 1. A fit is considered very good if $R^2 > 0.9$ (Altmann 1997), so the fit of the continuous model can be considered as very good.

As an alternative to the continuous model above where *C* is assumed to be a free parameter, we could consider *C* as the normalizing constant. In this case the curve (13) becomes the density of the beta distribution on the interval [*n, M*], in which *C* is expressed in terms of the other four parameters as

(16) $$C = \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)(M-m)^{a+b+1}},$$

where Γ denotes the gamma function. This density with the four parameters *m, M, a* and *b* could be fitted to the observed data (where the optimal values of *a* and *b* in the last line of Table 1 could be used as starting values for the iterative fitting). However, we will not pursue this idea in the present article.

2. New data record: Russian

The focus of the analysis to be carried out is the morph length in Russian. For this purpose we study the Russian novel "Kak zakaljalas' stal' (KZS)" by N. Ostrovskij; see Kelih (2009a, 2009b). For comparison purposes we also examine the Slovenian translation of this text, i.e. we study the morph length in a South Slavic and in an East Slavic language. In each case only the first chapter is considered. The investigation steps are:

1. These texts are subjected to a tagging procedure, in which chapter headings, abbreviations, digits etc. are processed by a common principle (cf. Antić, Kelih and Grzybek (2006)), i.e. omitted.

2. The word form is determined according to orthographic criteria (Kelih (2007)), i.e. each sequence which is separated by a space is regarded as one word form. The hyphen is considered to have a delimiting function.
3. In the created lists of word form types the number of morphs per word form is determined manually. Further details of the segmentation are given in the language-specific analyses below.
4. Zero-morphemes are not considered in the segmentation. The length of a morpheme is measured in the number of phonemes.
5. Only word forms with a frequency greater than one are analyzed[4].
6. The morph length is determined at the level of word form types and word form tokens. Thus the behaviour of the morph length can be modelled and compared both on the paradigmatic and syntactic level.

The morphological segmentation causes a series of theoretical problems, since – as with other linguistic units – a whole range of different definitions of morph or morpheme is provided. Slovenian as well as Russian are both, typologically, highly inflectional languages, so the morphological segmentation can be done by means of the same analytical procedure.

For the present analysis a pragmatic approach has been chosen. The segmentation is based on Russian morphological dictionaries, which provide in-depth morphological information about Russian word forms. For Russian the (exemplary) morphological dictionary by Kuznecova and Efremova (1986), which provides a detailed morphological segmentation of more than 52,000 Russian lexemes (for details see Kempgen (1999)), was used. As another reference source for Russian we used Tichonov (2002). To give at least one example of the performed segmentation: the Russian verb form *vyneslas'* (3.P. f. Sg. Past Tense, reflexive) is segmented into {*vy*} (prefix) – {*nes*} (root) – {*l*} (suffix, marking past tense) – {*a*} (suffix marking feminine) –{*s'*} (postfix, marking of reflexivity), resulting in one morph with three phonemes, three morphs with one phoneme and one morph with two phonemes.

The fitting of both models to the Russian data is illustrated in Table 2 for the data at the level of word form types and in Table 3 for word form tokens. One empirical fact of the analyzed texts is worth mentioning: in both cases the most frequent morph has the length of one phoneme, which can presumably be explained by the fact that Russian is a highly inflectional language, where just one phoneme can be the carrier of different grammatical information. A rather similar picture can be seen in the Slovenian data.

Table 2
Russian morph lengths (Types)

x	f_x	discrete, NP_x	continuous, y(x)
1	623	623.00	614.73
2	223	239.92	281.42
3	242	218.26	160.04
4	95	115.82	96.73
5	48	39.51	58.32
6	11	8.99	33.24
7	2	1.36	16.43
8	3	0.14	5.51
EPB 1-d.	$n = 9$, $p = 0.1853$, $\alpha = 0.5004$, $C = 0.0033$		
Beta f.	M = 0, M = 9, C = 32.6042, a = -0.8552, b = 1.4123, $R^2 = 0.9649$		

[4] This approach results from the fact that elsewhere a separate analysis of the Hapax Legomena has been made (Kelih 2011) and the same word-form lists are used here.

Table 3
Russian morph lengths (Tokens)

x	f_x	discrete, NP_x	continuous, y(x)
1	2518	2518.00	2498.78
2	1200	1190.86	1331.57
3	986	907.98	743.09
4	290	423.07	399.66
5	168	134.41	196.05
6	38	30.74	81.32
7	4	5.21	24.42
8	7	0.73	3.28
EPB 1-d.	$n = 13$, $p = 0.1127$, $\alpha = 0.5168$, $C = 0.0117$		
Beta f.	$m = 0$, $M = 9$, $C = 6.9700$, $a = -0.3632$, $b = 2.8286$, $R^2 = 0.9832$		

All fittings of this section can be considered good or satisfactory. A first indication is given that the EPB distribution and the used beta function are suitable to model Russian morph lengths.

3. New data record: Slovenian

Unlike the Russian, for the segmentation of Slovenian word forms there are not so many reference books (morphological and word formation dictionaries) available. Nevertheless the morphological segmentation was performed in analogy to the Russian analysis.

The prefixes, stems and suffixes have been identified step by step (as described in Toporišič (2000: 149) and SSKJ (1970ff), which were used as the main resources for the determination of the morphs). Other word formation issues were resolved with the help of Stramljič-Breznik's (2004) analysis.The "depths" of the morpheme identification are identical to the segmentation of the Russian data. Tables 4 and 5 show the corresponding fittings for the two Slovenian data records.

Table 4
Slovenian morph lengths: word form types

x	f_x	discrete, NP_x	continuous, y(x)
1	453	453.00	440.86
2	258	276.86	313.96
3	284	251.10	216.51
4	136	147.08	139.46
5	64	62.53	79.95
6	22	20.56	36.85
7	1	6.89	9.91
EPB 1-d.	$n = 33$, $p = 0.0536$, $\alpha = 0.6281$, $C = 0.0054$		
Beta f.	$m = 0$, $M = 8$, $C = 11.4039$, $a = -0.0720$, $b = 1.8782$, $R^2 = 0.9486$		

Table 5
Slovenian morph lengths: word form tokens

x	f_x	discrete, NP_x	continuous, y(x)
1	1963	1963.00	1970.92
2	2170	2050.24	2145.73
3	1178	1284.23	1213.25
4	424	509.47	430.79
5	188	143.60	89.26
6	62	30.58	7.84
7	2	5.88	0.10
EPB 1-d.	$n = 21$, $p = 0.0589$, $\alpha = 0.6721$, $C = 0.0098$		
Beta f.	$m = 0$, $M = 8$, $C = 0.0041$, $a = 1.6186$, $b = 6.7270$, $R^2 = 0.9971$		

Both considered models have proved to be suitable also for Slovenian data.

4. Concluding remarks

In the present study we consider a discrete model for the morph length distribution for Spanish, Slovenian and Russian data, which requires three parameters. It should be mentioned at this point that, considering a language individually, more suitable discrete models than the EPB come into play. However, the EPB is the only distribution (out of the stock of about 200 discrete distributions available with the Altmann-Fitter (1997)) which could be fitted to all five data records. Since the data sets in Tables 1–5 have very small classes for large lengths, we have pooled the classes so that the minimum class size is 50.

Choosing a continuous function, we restricted ourselves to the beta function. The use of this continuous function is satisfactory for all analyzed languages. In any case, it could be shown that for the newly analyzed data (Slovenian, Russian) the same probabilistic model can be employed, both for the type as well as the token level. It appears that in addition to the 1-shifted Hyperpoisson distribution (cf. Best (2005a: 258)) also the EPB distribution and the Bbeta function can be considered as good models for the distribution of the morph length. Since only a small number of languages and texts have been investigated to date, further studies of other languages are indispensable in future. Furthermore, the impact of the used (different) morph segmentation should be analyzed in detail.

References

Altmann, Gabriel (1997): The art of quantitative linguistics. In: *Journal of Quantitative Linguistics 4, 13–22.*

Altmann, Gabriel; Grotjahn Rüdiger (1988): Linguistische Meßverfahren. In: Ulrich Ammon, N. Dittmar und K.J Matheier (Eds.): *Sociolinguistics. Soziolinguistik. Band 2.* Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 3), p. 1026–1039.

Antić, Gordana; Kelih, Emmerich; Grzybek, Peter (2006): Zero-syllable Words in Determining Word Length. In: Peter Grzybek (Ed.): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht, NL: Springer (Text, Speech and Language Technology, 31), p. 117–156.

Best, Karl-Heinz (2001): Zur Länge von Morphen in deutschen Texten. In: Karl-Heinz Best (Eds.) *Häufigkeitsverteilungen in Texten*. Göttingen: Pest & Gutschmidt (Göttinger linguistische Abhandlungen, 4), p. 1–14.

Best, Karl-Heinz (2005a): Morphlängen. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), p. 255–260.

Best, Karl-Heinz (2005b): Wortlänge. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), p. 260–273.

Kelih, Emmerich (2007): Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen. In: Kaliuščenko, Volodymir; Köhler, Reinhard; Levickij Viktor (Eds.): *Problems of Typological and Quantitative Lexicology*. Chernivtsi: Ruta, p. 91–105.

Kelih, Emmerich (2009): Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: Kelih, Emmerich; Levickij, Viktor V.; Altmann, Gabriel (Eds.): *Methods of Text Analysis. Metody analizu tekstu*. Černivci: ČNU, p. 106–124.

Köhler, Reinhard; Altmann, Gabriel (2009): *Problems in quantitative linguistics 2*. Lüdenscheid: RAM-Verlag (Studies in Quantitative Linguistics, 4).

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

Kuznecova, A.I.; Efremova, T.F. (1986): *Slovar' morfem russkogo jazyka*. Moskva: Russkij jazyk.

Mačutek, Ján; Altmann, Gabriel (2007): Discrete and Continuous Modeling in Quantitative Linguistics. *Journal of Quantitative Linguistics 14* (*1*), *81–94.*

Rottmann, Otto A. (2003): Word lengths in the Baltic languages – are they of the same type as the word lengths in the Slavic languages? *Glottometrics 6, 52–60.*

Saporta, Sol (1966): Phoneme distribution and language universals. In: Greenberg, Joseph H. (Hg.): *Universals of language*. 2nd edition. Cambridge, MA: M.I.T. Press, p. 61–72.

Stramljič-Breznik, Irena (2004): *Besednodružinski slovar slovenskega jezika: poskusni zvezek za iztočnice na B.* Maribor: Slavistično društvo.

Toporišič, Jože (2000): *Slovenska slovnica*. Maribor: Obzorja.

Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Zörnig Peter; Altmann Gabriel (2012): Aspects of the behaviour of parts-of-speech in Italian texts (*In this volume*).

Wimmer, Gejza; Altmann, Gabriel (2005): Unified derivation of some linguistic laws. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), p. 791–801.

Software

Altmann-Fitter (1997): *Iterative fitting of probability distributions,* Lüdenscheid: RAM-Verlag.

Glottometrics 24, 2012,79-88

Length and complexity of NPs in Written English

Wang Hua[1]

School of Foreign Languages
Dalian Maritime University

Abstract. This paper investigates the length distribution and the complexity of NPs of written English using the written section of the ICE-GB corpus as the data source. The results show that NPs have very complex patterns. The distribution of NPs and their patterns is affected by NP length. Such relationships can be exactly described with mathematic models.

***Key words**: NP, pattern, length, frequency, distribution, model*

1. Introduction

NPs are essential components of sentences and have received attention from generations of linguists. Different grammars, whatever their theoretical frames, deal with NPs as to its components and syntactic functions. The methods used by these linguists are mainly qualitative. The present study approaches NPs from a different angle using the quantitative methods, focusing on the length of NPs and its relationship with the structure and complexity of NPs.

Length of linguistic units plays a very important role in quantitative linguistics and is an essential measurement in synergetic linguistics (Köhler, 2005). There have been a large number of publications on length of linguistic constructs, such as word and sentences length and their interrelations with other linguistic components (Menzerath 1928; Altmann, 1980, 1988; Wimmer, Köhler, Grotjahn & Altmann, 1994; Wimmer & Altmann, 1996; Grzybek, Kelih &, Stadlober**,** 2008; Fan, Grzybek & Altmann, 2010; Levitsky & Melnyk, 2011). For example, Altmann (1988) studied sentence length and concluded that it may depend on many different factors, such as sentence complexity, sentence structure, text length and so on. Wimmer and Altmann (1996) examined word length distributions and discovered that the compound Poisson and Ord family model can best capture word length distributions in language. However, literature concerning the quantitative study on the major sublevels of the sentence, the phrase, is few and far between. This paper intends to tackle the quantitative aspects of the noun phrase (NP) such as its length, structures, its structural complexity and the distribution of its structures with different length. NP structures are described in terms of their patterns. For example, a grammatical description of the NP *a big red apple* has the structure *determiner adjective adjective noun*. NP length is based on the number of syntactic components of the phrase rather than the number of words. NP complexity is described in terms of the number of

[1] Address correspondence to: *whdeworld@yahoo.cn*

different NP patterns NPs of certain length may have. For example, If NPs all have the pattern *determiner noun*, then there is no complexity in NP structures. However, in actuality NP structures must be much more varied than that.

NPs have premodifiers and postmodifiers. The latter is extremely messy to deal with since postmodifiers can be very long and complex, containing in turn other NPs with their own premodifiers or postmodifiers. In the present study, the NP postmodifier, irrespective of the number of their syntactic components, is regarded as one single syntactic component of the preceding NP, but the NPs in the postmodifiers are all counted as NPs. For example, in the sentence *this is the dog that chased the cat that killed the mice that ate the rice* there are 4 NPs, i.e., *the dog that…*, *the cat that…*, *the mice that…* and *the rice*. Their NP structures are respectively *article noun postmodifier-clause*, *article noun postmodifier-clause*, *article noun postmodifier-clause* and *article noun*, and the length of these NPs are respectively 3, 3, 3, and 2. Another example is the structure of the NP *the man with a large green umbrella*, and its NP structure is *article noun postmodifier-prepositional phrase*, and in the prepositional phrase, there is another NP. The length of the main NP is 3, and that of the nested NP is 3 as well: *article adjective adjective noun*, i.e., *a large green umbrella*. The main NP contains 7 words but its length is three.

2. The Data

The ICE-GB corpus was used as the data source. It is a 1,000,000-word corpus of contemporary British English, which contains both ICE-GBS (corpus of spoken English) and ICE-GBW (corpus of written English). The former contains 300 2,000-texts totalling about 600,000 words, while the latter 200 texts totalling about 400,000. These texts are grammatically tagged. Since the study focuses on NPs of written English, only ICE-GBW was used. The syntactic tags of NP components were extracted for the study of NP length and structures. The frequencies of NPs of different patterns were also calculated. Table 1 displays the major syntactic tags used in ICE-GBW and the NP components they represent.

Table 1

Major syntactic tags and the NP components they represent

A	adverbial	AVP	adverb phrase
ADJ	adjective	AVPO	adverb phrase postmodifier
ADV	adverb	AVPR	adverb phrase premodifier
AJHD	adjective phrase head	CF	focus complement
AJP	adjective phrase	CJ	conjoin
AJPO	adjective phrase postmodifier	CL	clause
AJPR	adjective phrase premodifier	CLEFTIT	cleft it
ART	article	CLOP	cleft operator
AUX	auxiliary	CO	object complement
AVB	auxiliary verb	COAP	appositive connector
AVHD	adverb phrase head	CONJUNC	conjunction

CONNEC	connective
COOR	coordinator
CS	subject complement
CT	transitive complement
DEFUNC	detached function
DISMK	discourse marker
DISP	disparate
DT	determiner
DTCE	central determiner
DTP	determiner phrase
DTPE	predeterminer
DTPO	determiner postmodifier
DTPR	determiner premodifier
DTPS	postdeterminer
ELE	element
EMPTY	empty
EXOP	existential operator
EXTHERE	existential there
FNPPO	floating NP postmodifier
FOC	focus
FRM	formulaic expression
GENF	genitive function
GENM	genitive marker
IMPOP	imperative operator
INDET	indeterminate
INTERJEC	interjection
INTOP	interrogative operator
INVOP	inverted operator
MVB	main verb
N	noun
NADJ	nominal adjective
NONCL	nonclause
NOOD	notional direct object
NOSU	notional subject
NP	noun phrase
NPHD	noun phrase head
NPPO	noun phrase postmodifier
NPPR	noun phrase premodifier
NUM	numeral
OD	direct object
OI	indirect object
OP	operator
P	prepositional
PARA	parataxis
PAUSE	pause
PAUSE	pause
PC	prepositional complement
PMOD	prepositional modifier
PP	prepositional phrase
PREDEL	predicate element
PREDGP	predicate group
PREP	preposition
PROD	provisional direct object
PROFM	proform
PRON	pronoun
PRSU	provisional subject
PRTCL	particle
PS	stranded preposition
PU	parsing unit
PUNC	punctuation
PUNC	punctuation
REACT	reaction signal
SBHD	subordinator phrase head
SBMO	subordinator phrase modifier
SU	subject
SUB	subordinator
SUBP	subordinator phrase
TAGQ	tag question
TO	particle to
TO	'to' infinitive marker
UNTAG	missing/unidentifiable items
UNTAG	untag
V	verb
VB	verbal
VP	verb phrase

In ICE-GB, all the NPs have an NP head symbolized by the tag NPHD and a postmodifier symbolized by the tag NPPO; these two tags are kept in the extracted data but do not count as an independent NP component. For example, the NP structure ART ADJ NPHD-N NPPO-CL has 4 components, instead of 6. NPHD-N means the noun phrase head is a noun, while NPPO-CL means the noun phrase postmodifier is a clause.

3. Length and NP frequency

The 430,416-word ICE-GBW contains 27,647 sentences, averaging 15.57 words per sentence. The total number of NPs is 115,809. As the length of NPs increases, the number of NPs with such length decreases. NPs with only one syntactic component have the largest number, 44,629, and those with 15 components have only two occurrences. Table 2 shows NP length and its corresponding NP frequency.

Table 2
NP length and its corresponding frequency

Length	**Frequency**
1	44629
2	33844
3	24681
4	8477
5	2629
6	960
7	352
8	138
9	54
10	21
11	10
12	8
13	4
15	2

The exponential regression model can best capture the relationship between NP length and NP occurrences. The exponential regression model is in the following form:

(1) $$y = ae^{bx}$$

The frequency of NPs and their corresponding length can be described with the following exponential model (N_{freq} = NP frequency, N_{len} = NP length; *a*, *b*: model parameters):

(2) $$N_{freq} = ae^{bN_{len}}$$

The fit is good, with $R^2 = 0.980$, $a = 136677.074$, $b = -08135$. Figure 1 displays the model fit.

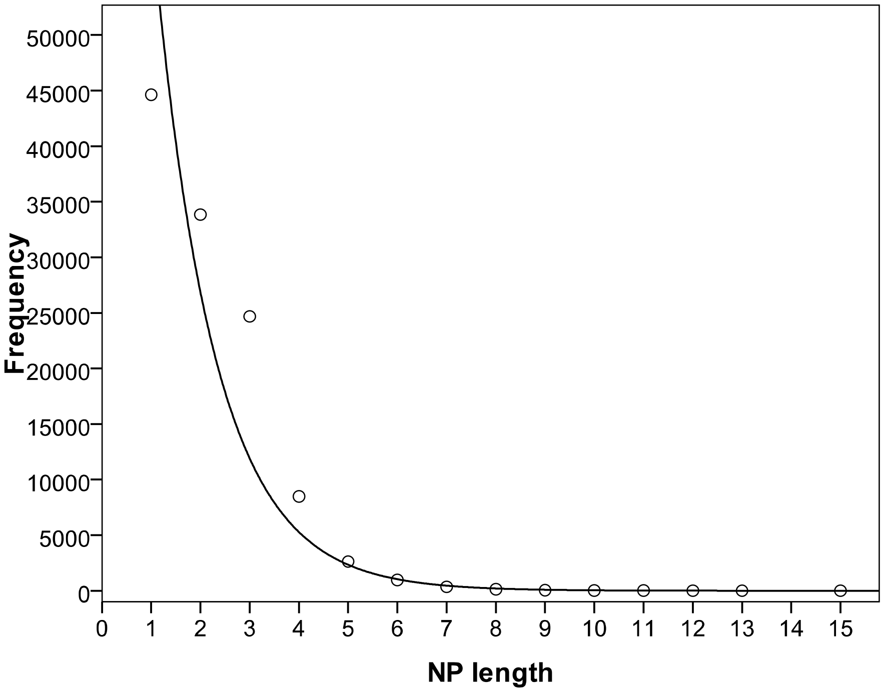


Figure 1. The exponential regression model fit to the NP frequency curve. The solid line: the model fit, the small circles, the observed value

4. Distribution of NP types and patterns

Of all the 115,809 NPs, there are 1,894 different patterns. This shows the complexity of NP structures. These NPs can be classified into the following types: bare NPs, i.e., NPs with a bare head; determiner + NPHD; Premodifier + NPHD; Postmodifirer + NPHD and Premodifier + NPHD + Postmodifier. The NPs of the type bare NP/Det+NPHD are the most frequently used, totalling 70,790, accounting for 61.13/% of all the NPs. Detailed information on the distribution of NP types is in Table 3.

Table 3
Distribution of NP types

Type	**Frequency**	**Percentage**
Bare NP/Det+NPHD	70790	61.13/%
Pre+NPHD	13744	11.87%
Post+NPHD	24027	20.75%
Pre+NPHD+Post	7248	6.25%
Total	115809	100%

The complexity of NPs of different length is shown in Table 4. NPs with length 5 have 439 different patterns, and those with length 15 have only 2.

Table 4
NP length and the corresponding patterns

Length	**Number of Patterns**	**Examples**
1	4	NPHD-N
2	53	NPHD-N NPPO-PP
3	183	ADV ADJ NPHD-N
4	341	ADV ART NUM NPHD-N
5	439	ART ADJ CONJUNC NUM NPHD-N
6	405	ART ADJ ADJ NPHD-N PUNC NPPO-AVP
7	244	ADV ART ADV ADJ ADJ NPHD-N NPPO-PP
8	128	ADJ NPHD-N CONJUNC ART ADJ ADJ NPHD-N NPPO-NP
9	52	ART ADV ADJ NPHD-N COMJUNC ART ADV ADJ NPHD-N
10	21	ART ADJ NPHD-N COMJUNC ADV ART ADV ADJ CONJUNC ADJ NPHD-PRON
11	10	ART ADJ ADJ NPHD-N CONJUNC ART ADV ADJ CNOJUNC ADJ NPHD-N
12	8	PRON ADJ CONJUNC ADJ NPHD-N CONJUNC ADV PRON NUM ADJ ADJ NPHD-N
13	4	ADJ CONJUNC ADJ ADJ ADJ NPHD-N CONJUNC AUX ASV V PERP ADJ NPHD-N
15	2	ART ADJ ADJ ADJ COMJUMC ADJ NPHD-N CONJUNC ART ADJ ADJ ADJ COMJUNC ADJ NPHD-N

The following are the 183 different NP patterns for NPs with length 3:

1. ADJ ADJ NPHD-N
2. ADJ ADJ NPHD-PRON
3. ADJ ADV NPHD-N
4. ADJ CONJUNC NPHD-N
5. ADJ NPHD-N NPHD-N
6. ADJ NPHD-N NPHD-PRON
7. ADJ NPHD-N NPPO-AJP
8. ADJ NPHD-N NPPO-AVP
9. ADJ NPHD-N NPPO-CL
10. ADJ NPHD-N NPPO-NP
11. ADJ NPHD-N NPPO-PP
12. ADJ NPHD-NUM NPHD-N
13. ADJ NPHD-PRON NPPO-CL
14. ADJ NPHD-PRON NPPO-PP
15. ADJ NUM NPHD-N
16. ADJ PREP NPHD-PRON
17. ADJ PUNC NPPO-AJP
18. ADJ PUNC NPPO-AVP
19. ADJ PUNC NPPO-CL
20. ADJ PUNC NPPO-DISP
21. ADJ PUNC NPPO-PP
22. ADJ UNTAG NPHD-N
23. ADV ADJ NPHD-N
24. ADV ADJ NPHD-PRON
25. ADV ADV NPHD-N
26. ADV ADV NPHD-NUM
27. ADV ADV NPHD-PRON
28. ADV ART NPHD-N

29. ADV ART NPHD-NADJ
30. ADV ART NPHD-NUM
31. ADV NPHD-N NPHD-N
32. ADV NPHD-N NPPO-AVP
33. ADV NPHD-N NPPO-CL
34. ADV NPHD-N NPPO-DISP
35. ADV NPHD-N NPPO-NP
36. ADV NPHD-N NPPO-PP
37. ADV NPHD-NUM NPPO-AVP
38. ADV NPHD-NUM NPPO-CL
39. ADV NPHD-NUM NPPO-NP
40. ADV NPHD-NUM NPPO-PP
41. ADV NPHD-PRON NPPO-AJP
42. ADV NPHD-PRON NPPO-CL
43. ADV NPHD-PRON NPPO-PP
44. ADV NUM NPHD-N
45. ADV PRON NPHD-N
46. ADV PRON NPHD-PRON
47. ADV PUNC NPPO-CL
48. ART ADJ NPHD-N
49. ART ADJ NPHD-NADJ
50. ART ADJ NPHD-NUM
51. ART ADJ NPHD-PRON
52. ART ADV NPHD-N
53. ART ADV NPHD-NADJ
54. ART ADV NPHD-PRON
55. ART ART NPHD-N
56. ART NPHD-N NPHD-N
57. ART NPHD-N NPPO-AJP
58. ART NPHD-N NPPO-AVP
59. ART NPHD-N NPPO-CL
60. ART NPHD-N NPPO-DISP
61. ART NPHD-N NPPO-NP
62. ART NPHD-N NPPO-PP
63. ART NPHD-NADJ NPHD-N
64. ART NPHD-NADJ NPPO-AVP
65. ART NPHD-NADJ NPPO-CL
66. ART NPHD-NADJ NPPO-PP
67. ART NPHD-NUM NPHD-N
68. ART NPHD-NUM NPPO-AVP
69. ART NPHD-NUM NPPO-CL
70. ART NPHD-NUM NPPO-NP
71. ART NPHD-NUM NPPO-PP
72. ART NPHD-PRON NPHD-PRON
73. ART NPHD-PRON NPPO-CL
74. ART NPHD-PRON NPPO-PP
75. ART NUM NPHD-N
76. ART NUM NPHD-NADJ
77. ART NUM NPHD-NUM
78. ART NUM NPHD-PRON
79. ART PREP NPHD-N
80. ART PRON NPHD-N
81. ART PUNC NPPO-AJP
82. ART PUNC NPPO-AVP
83. ART PUNC NPPO-CL
84. ART PUNC NPPO-DISP
85. ART PUNC NPPO-NP
86. ART PUNC NPPO-PP
87. ART UNTAG NPHD-N
88. CONJUNC ADV NPHD-PRON
89. CONNEC PUNC NPPO-PP
90. NPHD-N CONJUNC NPHD-N
91. NPHD-N CONJUNC NPHD-NADJ
92. NPHD-N CONJUNC NPHD-NUM
93. NPHD-N CONJUNC NPHD-PRON
94. NPHD-N NPHD-N NPHD-N
95. NPHD-N NPHD-N NPPO-AVP
96. NPHD-N NPHD-N NPPO-CL
97. NPHD-N NPHD-N NPPO-NP
98. NPHD-N NPHD-N NPPO-PP
99. NPHD-N NPHD-NUM NPHD-N
100. NPHD-N NPHD-NUM NPPO-CL
101. NPHD-N NPHD-NUM NPPO-PP
102. NPHD-N PUNC NPPO-AJP
103. NPHD-N PUNC NPPO-AVP
104. NPHD-N PUNC NPPO-CL
105. NPHD-N PUNC NPPO-DISP
106. NPHD-N PUNC NPPO-NP
107. NPHD-N PUNC NPPO-PP
108. NPHD-NADJ CONJUNC NPHD-N
109. NPHD-NADJ CONJUNC NPHD-NADJ
110. NPHD-NUM CONJUNC NPHD-N
111. NPHD-NUM CONJUNC NPHD-NUM
112. NPHD-NUM CONJUNC NPHD-PRON
113. NPHD-NUM NPHD-N NPPO-NP
114. NPHD-NUM NPHD-N NPPO-PP
115. NPHD-NUM NPHD-NUM NPPO-PP
116. NPHD-NUM PUNC NPPO-AJP
117. NPHD-NUM PUNC NPPO-AVP
118. NPHD-NUM PUNC NPPO-CL
119. NPHD-NUM PUNC NPPO-NP
120. NPHD-NUM PUNC NPPO-PP
121. NPHD-PRON CONJUNC NPHD-N
122. NPHD-PRON CONJUNC NPHD-NADJ
123. NPHD-PRON CONJUNC NPHD-PRON
124. NPHD-PRON NPHD-N NPPO-CL
125. NPHD-PRON NPHD-N NPPO-PP
126. NPHD-PRON NPHD-PRON NPPO-PP
127. NPHD-PRON PUNC NPPO-AJP
128. NPHD-PRON PUNC NPPO-CL
129. NPHD-PRON PUNC NPPO-PP
130. NUM ADJ NPHD-N
131. NUM ADJ NPHD-PRON
132. NUM ADV NPPO-PP
133. NUM ART NPHD-N
134. NUM NPHD-N NPHD-N
135. NUM NPHD-N NPPO-AJP
136. NUM NPHD-N NPPO-AVP
137. NUM NPHD-N NPPO-CL
138. NUM NPHD-N NPPO-DISP
139. NUM NPHD-N NPPO-NP
140. NUM NPHD-N NPPO-PP
141. NUM NPHD-NUM NPHD-NUM
142. NUM NPHD-NUM NPPO-PP

143. NUM NUM NPHD-N
144. NUM PRON NPHD-N
145. NUM PUNC NPPO-AVP
146. NUM PUNC NPPO-CL
147. NUM PUNC NPPO-PP
148. NUM UNTAG NPHD-N
149. PREP ART NPHD-N
150. PREP NPHD-N NPPO-CL
151. PREP PRON NPHD-N
152. PRON ADJ NPHD-N
153. PRON ADJ NPHD-NADJ
154. PRON ADJ NPHD-NUM
155. PRON ADJ NPHD-PRON
156. PRON ADV NPHD-N
157. PRON ADV NPHD-NADJ
158. PRON ART NPHD-N
159. PRON ART NPHD-NADJ
160. PRON ART NPHD-NUM
161. PRON NPHD-N NPHD-N
162. PRON NPHD-N NPHD-NUM
163. PRON NPHD-N NPPO-AJP
164. PRON NPHD-N NPPO-AVP
165. PRON NPHD-N NPPO-CL
166. PRON NPHD-N NPPO-DISP
167. PRON NPHD-N NPPO-NP
168. PRON NPHD-N NPPO-PP
169. PRON NPHD-NADJ NPPO-CL
170. PRON NPHD-NADJ NPPO-PP
171. PRON NPHD-NUM NPHD-N
172. PRON NPHD-NUM NPPO-AVP
173. PRON NPHD-NUM NPPO-CL
174. PRON NPHD-NUM NPPO-PP
175. PRON NPHD-PRON NPPO-CL
176. PRON NPHD-PRON NPPO-PP
177. PRON NUM NPHD-N
178. PRON NUM NPHD-PRON
179. PRON PRON NPHD-N
180. PRON PRON NPHD-PRON
181. PRON PUNC NPPO-PP
182. PRON UNTAG NPHD-N
183. UNTAG NPHD-N NPPO-NP

Nemcová and Serdelová (2005) use the following to describe the relationship between the number of synonyms (*y*) of a word and the length of the word in syllables *x*:

(3) $$y = ax^{b}e^{cx} + 1$$

(3) is a special case of Wimmer & Altmann (2005). This relationship also holds for NP length (N_{len}) and the number of NP patterns ($N_{pattern}$) NPs have with the corresponding length:

(4) $$N_{pattern} = aN_{len}^{\ b}e^{cN_{len}} + 1$$

The fit is very good. R^2 = 0.991, a = 2.0369, b = 8.8827 and c = -1.7831. Figure 2 is the model fit.

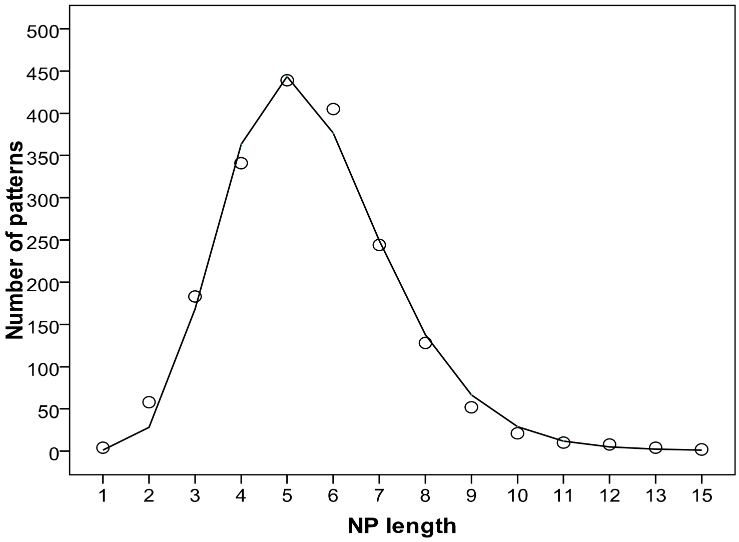


Figure 2. The relationship between NP length and the number of patterns NPs with corresponding length have. Solid line: model fit, small circles: the observed value.

5. Conclusion

This study reveals the complexity of NPs and the relationship among NP length, NP frequency and NP patterns. A 400,000-word ICE-GBW has 1,894 different types of NPs. Shorter NPs generally have higher occurrences and NPs with 5 components have the highest complexity. Such relationships can be exactly described with mathematic models.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. In: *Glottometrika 2, 1-10*. Bochum: Brockmeyer.

Altmann, G. (1988). *Verteilungen der Satzlängen.* In: Schulz, K.-P. (ed.), *Glottometrika 9: 147-169*. Bochum: Brockmeyer.

Fan, F., Grzybek, P., Altmann, G. (2010). Dynamics of word length in sentence. *Glottometrics 20, 70-109.*

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

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

Levitsky, V. & Melnyk, Y. (2011). Sentence length and sentence structure in English prose. *Glottometrics 21, 14-24.*

Menzerath, P. (1928). Über einige phonetische Probleme. In: *Actes du premier Congres international de linguistes.* Leiden: Sijthoff.

Nemcová, E, & Serdelová, K. (2005). On synonymy in Slovak. In: Altmann, G, Levickij, V & Perebyinis, V. (eds.), *Problems of Quantitative Linguistics*: *194-209*. Chernivtsi: Ruta.

Wimmer, G. & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Hrsg.), *Glottometrika 15: 112-133*. Trier: Wissenschaftlicher Verlag Trier.

Wimmer, G. & Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.). *Contributions to the science of language: Word length and related issues*: *93-117*. Boston: Kluver.

Wimmer, G, Köhler, R, Grotjahn, R, & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Glottometrics 24, 2012, 88-94

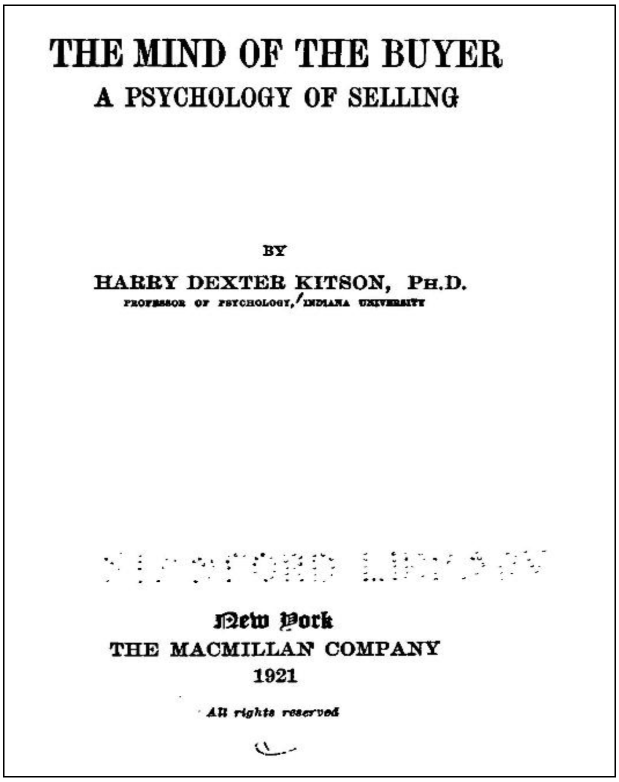
History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

Harry Dexter Kitson (1886-1959)

Peter Grzybek

Harry Dexter Kitson, born in 1886 in Mishawaka, Indiana, taught applied psychology at Teachers' College, Columbia University. He was a charter member of the American Psychological Association and a pioneer in the field of vocational guidance. His main field of professional interest throughout his life (see references below), and it would definitely be incorrect to rank him among the precursors of quantitative linguistics. Yet, some ideas and analyses represented in his 1921 book *The Mind of the Buyer. A Psychology of Selling* illustrate the need of his time for a solid basis in linguostatistics and quantitative linguistics, and therefore deserve mention in an historical flashback.



THE MIND OF THE BUYER
A PSYCHOLOGY OF SELLING

BY

HARRY DEXTER KITSON, PH.D.
PROFESSOR OF PSYCHOLOGY, INDIANA UNIVERSITY

New York
THE MACMILLAN COMPANY
1921

All rights reserved

Kitson's booklet was meant to be a guide in advertisement strategies for salesmen, in his words "every one who is engaged in influencing men to buy" (p. v). For Kitson, such a work must necessarily be based on theoretical psychology and deal with profound psychological questions, particularly mental processes such as attention, interest, desire and confidence (ibd., v).

With this orientation and phrasing, Kitson's booklet was a typical child of its time. After all, the booklet was published in the very same year when the famous *AIDA* formula was first used as an acronym by C.P. Russell (1921) to refer to the relevant components (or steps, as which they were considered at that time) of successful advertising: "attract **A**ttention, maintain **I**nterest, create **D**esire and get **A**ction". Such components were usually traced back to psychological theories of that time (usually related to some kind of association psychology).

Yet, although Russell is considered to have casted this concept into a concise verbal form – i.e., the *AIDA* formula –, he is not responsible for having developed the general idea and concept behind it. It is commonly held that it is American advertising and sales pioneer Elias St. Elmo Lewis who should be credited for having established the term and approach in 1903: postulating at least three principles to which a successful advertisement should conform, for Lewis, the "mission of an advertisement" was "to attract a reader […]; then to interest him, […]; then to convince him […]." The first published instance of the general concept seems to be a 1904 article by Frank Hutchinson Dukesmith, according to whom the four most important steps were attention, interest, desire, and conviction. Later important references are Ralph Starr Butler's (1911) *Advertising, Selling, and Credits. Part II: Selling and Buying,* with a whole chapter on "Principles of Salesmanship" (p. 410ff.) focusing on attention, interest, desire, action. Butler, in turn, refers to Arthur F. Sheldon, founder of the Sheldon School of Scientific Salesmanship, and his 1911 book *The art of selling, for business colleges*, containing similar ideas.

In this respect, Kitson's approach is not genuinely innovative. What makes him differ from preceding approaches, however, is his definition and treatment of what he termed the "collective buyer". According to him, persons who are served by a given selling medium constitute a collectivity. For Kitson, such a public is not a simple arithmetic summation of individual minds, nor is a some kind of super-mind transcending its components (ibd., 54). For him, newspapers and magazines offer good evidence of the existence of the collective mind: "psychologically speaking, the readers of a sales medium constitute an entity, a public, which is not a loose aggregation of isolated and individual minds but an organic union, coalesced into one collective mind" (ibd., 55). Furthermore, each public is unique, and readers of different newspapers differ from each other, what does of course not exclude the possibility that a given individual may belong to more than one public.

In trying to develop measurements of such publics which, in Kitson's terms, are "buying publics" (ibd., 56), Kitson suggests to study a number of relationships, mainly geographical, economic, sociological, and psychological. In his effort to establish some kind of "yard sticks" for the psychological, or mental, dimension (including ideas, feelings, motives), Kitson suggested, among others, the analysis of linguistic criteria of different journals and newspapers. Admitting that the kind of measures he suggested are still very fragmentary (ibd., 63), he suggested to concentrate on word length and sentence length, which he considered to be indicators of psychological differences between periodicals.

With regard to word length, Kitson first chose the *Chicago Evening Post* and the *Chicago American* for his analyses. From the editorial, news and feature columns of six parallel issues of each of these two papers, ca. 5000 words were taken in consecutive order and tabulated according to the number of syllables they contained. Likewise, two magazines were analyzed, the *Century* and the *American* magazines. Unfortunately, Kitson did not give the complete results, concentrating on words with more than two syllables only. The data are represented in Table 1.

Table 1

	Word Length			
	> 2	> 3	> 4	> 5
Chicago Evening Post	13,20	4,60	1,20	0,00
Chicago American	7,70	2,70	0,70	0,00
Century	13,50	4,30	1,00	0,20
American	9,90	2,70	0,60	0,10

In Kitson's interpretation, the results show that the number of words with more than two syllables in the *Post* is greater than that in the *American* by ca. 71%, a ratio approximately holding for all the polysyllabic words. The results of the magazine analyses are quite similar to the ones from the newspaper analysis; here, the number of words with more than two syllables in the *Century* is greater than the corresponding number in the *American* by ca. 36%. Kitson therefore concludes that the two journals and the two magazines clearly differ in their profiles.

A re-analysis of his data shows that his interpretation, albeit correct, is not unproblematic: first and foremost, because the dominating amount of one- and two-syllable words have been totally omitted from the analyses – after all, their percentage is > 90% in all cases, ranging from 90.8% to 95% across the four samples. But even concentrating on the word length frequencies of words with more than two syllables shows that Kitson's conclusions are far from being self-evident. Table 2 offers Kitson's data in re-ordered form, presenting them in non-cumulative form.

Table 2

	Word Length			
	3	4	5	> 5
Chicago Evening Post	8,60	3,40	1,20	0,00
Chicago American	5,00	2,00	0,70	0,00
Century	9,20	3,30	0,80	0,20
American	7,20	2,10	0,50	0,10

Comparing word length of both the two journals (*Chicago Evening Post, Chicago America*n) and the two magazines (*Century, American*) by way of the non-parametric Mann-Whitney *U*-test yields non-significant differences in both cases *(p* = 0.73 and $p \approx 1$, respectively), after weighting word length by the percentages given; the same holds true for the two journals' data and the two magazines's data taken together in combination (*p* = 0.84). Also a Kruskal-Wallis test for differences between all four groups shows the differences to be non-significant (*p* = 0.98); quite logically, post-hoc comparisons of averages yield no homogeneous subgroups. It seems reasonable, therefore, to conclude that, in contrast to Kitson's observations, there are no significant differences across the four journalistic samples with regard to word length.

In case of his sentence length analyses, based on parallel issues and columns of the same four journals and magazines, Kitson provides a better data basis which, as a consequence, allows for more reliable re-analyses. A total of 8000 sentences were analyzed for sentence length, measured in the number of words per sentence. Here, all data are presented, pooled in intervals per 10, in a similar fashion as the word length data in Table 1; the data are accordingly reproduced in Table 3.

Table 3

	Sentence Length (in intervals per 10)										
	1-10	>10	>20	>30	>40	>50	>60	>70	>80	>90	>100
Chicago Evening Post	16,9	83,1	49	22,3	8,5	2,7	0,8	0,2			
Chicago American	23,1	76,9	43,4	20,6	10,3	2,3	1,8	0,6	0,3	0,2	0,2
Century	22,8	77,2	45,4	24,4	10,6	5,5	2,4	0,9	0,4	0,2	
American	30,5	69,5	33,5	14,5	5,2	1,8	0,7	0,3	0,1	0,1	0,1

Comparing sentence length for the two journals, the *Chicago Evening Post* and the *Chicago American*, Kitson states that the results show a greater number of “long” sentences (considering sentences with > 20 words as long) in the *Post*; he likewise finds the *Century* to favor long sentences as compared to the *American*.

Attempting to re-analyze the data, it seems reasonable to re-order them without cumulation, in analogy to the word length data presented in Table 2. The corresponding sentence length data are presented in Table 4.

Table 4

	Sentence Length (in intervals per 10)										
	1-10	>10	>20	>30	>40	>50	>60	>70	>80	>90	>100
Chicago Evening Post	16,9	34,1	26,7	13,8	5,8	1,9	0,6	0,2			
Chicago American	23,1	33,5	22,8	10,3	8	0,5	1,2	0,3	0,1		0,2
Century	22,8	31,8	21	13,8	5,1	3,1	1,5	0,5	0,2	0,2	
American	30,5	36	19	9,3	3,4	1,1	0,4	0,2			0,1

Again, a re-analysis is not unproblematic, since not the raw data are given, but the pooled data in intervals per 10. Nevertheless, after weighting the sentence length categories with the percentages given allows a test for differences, in analogy to the above word length analyses, first between the two journals and the two magazines, then between all four samples. Whereas a Mann-Whitney *U*-test yields no significant differences between the two journals (p = 0.31), it shows the differences between the two magazines to be significant (p = 0.03). As to a comparison between all four groups, a Kruskal-Wallis test shows the differences to be significant (p = 0.03), but a post-hoc comparison of means yields no homogeneous subgroups.

This seemingly contradictory result might well be due to the fact that all four samples follow a common frequency distribution model, though with different weights for the individual length classes, what should result in different parameter values for the given model. In order to test this assumption, it would be necessary to have the original data at hand, what is not the case. Nevertheless, by way of some approximation, one might try to reconstruct original sample sizes given the fact that on the whole, 8000 sentences were analyzed, based on four approximately equal sample sizes. The results of reconstructing the corresponding absolute frequencies are represented in Table 5.

Table 5

	Sentence Length (in intervals per 10)										
	1	2	3	4	5	6	7	8	9	10	11
	1-10	>10	>20	>30	>40	>50	>60	>70	>80	>90	>100
Chicago Evening Post	338	682	534	276	116	38	12	4	0	0	0
Chicago American	462	670	456	206	160	10	24	6	2	0	4
Century	456	636	420	276	102	62	30	10	4	4	0
American	610	720	380	186	68	22	8	4	0	0	2

In trying to find a theoretical frequency distribution as an adequate model for these data, it turns out that the negative binomial distribution defined as

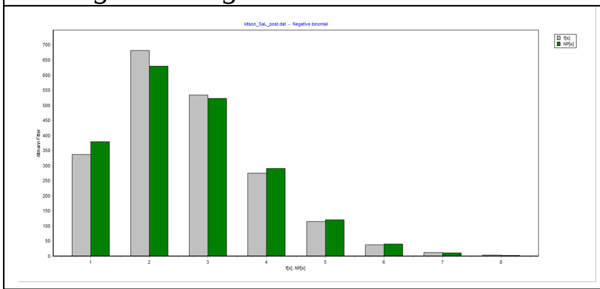
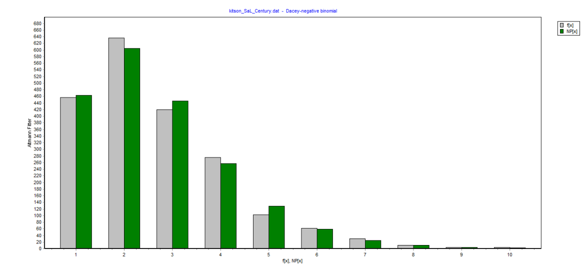
$$(1) \qquad P_x = \binom{k+x-1}{x} p^k q^x$$

is an excellent model for the three of the data sets (*Chicago Evening Post*, *Century*, *American*), whereas the *Chicago American* data can be fitted by the mixed negative binomial distribution defined as

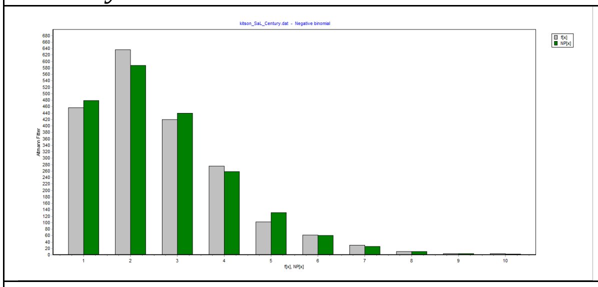
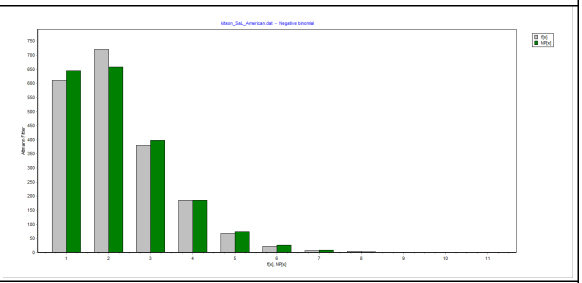
$$(2) \qquad P_x = \alpha \binom{k_1+x-1}{x} p_1^{k_1} q_1^{\ x} + (1-\alpha) \binom{k_2+x-1}{x} p_2^{k_2} q_2^{\ x},$$

both of course in one-displaced form. Taking into account that we are concerned with mixed data, in all cases, the need for a mixed model seems to be reasonable – quite evidently, with α = 0 or α = 1, the mixed model (2) has the ordinary model (1) as a special case..

Figures 1-4 show the fitting results, with parameter values for k and p given below the graphs, as well as the discrepancy coefficients $C = X^2/N$, with $C < 0.02$ indicating a good, $C < 0.01$ a very good fitting result.[1]

Chicago Evening Post	*Chicago American*
	
$k =$ 826.59, $p = 0.998$, $C = 0.0051$	$k = 3.36$, $p = 0.70$, $\alpha = 0.23$, $C = 0.0060$

[1] In case of the *Chicago Evening Post* data, with parameter $k \to \infty$ and $1\text{-}p = q \to 0$, the negative binomial distribution converges against the Poisson distribution, yielding an equally good fit with a = 1.66 and C = 0.0051. – The Century data can also be modeled by the Mixed Poisson distribution: with a = 3.31, b = 1.30, and α = 0.19 the result is almost identical, with C = 0.0054.

<table>
<tr><th><em>Century</em></th><th><em>American</em></th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td><em>k</em> = 4.59, <em>p</em> = 0.73, <em>C</em> = 0.0077</td><td><em>k</em> = 5.41, <em>p</em> = 0.81, <em>C</em> = 0.0056</td></tr>
</table>

Figures 1-4. Fitting the distribution of sentence length by the negative binomial distribution

The sentence length data thus indeed follow one and the same model, albeit with some “local” modifications.

As a result, one can say that Kitson has indeed raised interesting and important questions which, in one way or another, would today be treated between the fields of applied and quantitative linguistics. Whereas earlier word length and sentence length studies had mainly treated them in terms of individual author characteristics, on the basis of literary texts – in order to determine authorship, for example, or literary development –, Kitson, not referring to the works of Sherman, Mendenhall and others, extended the field of interest to “everyday” journalistic texts, asking for recipient specific and, in this sense, pragmatic differences. Almost simultaneously with and subsequent to his work, the influential discipline of text difficulty and readability research would become increasingly important: after the first readability formula suggested by Lively and Pressey in 1923, this line of research faced a first highlight in Rudolf Flesch’s works (e.g., Flesch 1948), very much later leading to, among others, systematic analyses of different journalistic sources (e.g., Amstad 1978). And although Kitson did not create a readability formula, he is considered to have shown how the principles work (cf. DuBay 2004: 13), since in almost all readability studies, word and sentence length have always played a crucial role, though not as separate, but specifically related factors, more often combined with further linguistic levels and units.

In this context, it may be important to emphasize that Kitson explicitly stated that from these findings he would not reason that superiority in long words or sentences proves conclusively a corresponding intellectual superiority. Admitting that “long words and long sentences are not an absolute criterion of erudition or short of ignorance” (ibd., 63), he nevertheless admits “that in the long run, the chances favor a greater number of long words being associated with more enlightened people” (ibd., 63). Interestingly enough, Kitson (ibd., 63) refers to a relation between vocabulary richness (in terms of the size of an individual’s subjective lexicon) on the one hand, and word length on the other: “Measurements made by various vocabulary tests have shown that there are more words in the vocabulary of the more enlightened; hence we might expect a greater number of long words there.”

References

Amstad, Toni (1978). *Wie verständlich sind unsere Zeitungen?* Diss., Univ. Zürich.

Butler, Ralph Starr (1911). *Advertising, Selling, and Credits. Part II: Selling and Buying.* New York: Alexander Hamilton Institute.

DuBay, William H. (2004). *The Principles of Readability*. Costa Mesa, CA: Impact Information. [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.4042]

Dukesmith, Frank Hutchinson (1904). Three Natural Fields of Salesmanship. *Salesmanship 2(1), January 1904, p. 14.*

Flesch, Rudolf (1948). A New Readability Yardstick. In: *Journal of Applied Psychology* 32(3), 1948; 221-233.

Kitson, Harry Dexter (1916). *The scientific Study of the College student***.** Phil. Diss, Chicago. [1st reprint: Princeton, NJ, 1917]

Kitson, Harry Dexter (1916). *How to use your mind***.** *A psychology of study; being a manual for the use of students and teachers in the administration of supervised study***.** Philadelphia**:** Lippincott Co.

Kitson, Harry Dexter (1920). *Manual for the study of the psychology of advertising and selling***.** Philadelphia, London: J.B. Lippincott Co.

Kitson, Harry Dexter (1921). *The Mind of the Buyer. A Psychology of Selling*. New York: Macmillan. [http://archive.org/stream/mindbuyerapsych00kitsgoog#page/n76/mode/2up]

Kitson, Harry Dexter (1929). *How to find the right vocation.* **N**ew York: Harper & Brothers Publishers**.**

Kitson, Harry Dexter (1933). *Finding a job during the depression*. New York City: The Robert C. Cook Co.

Kitson, Harry Dexter (1947). How to find the right vocation**.** New York: Harper & Brothers Publishers.

Kitson, Harry Dexter; Newton, Juna Barnes (1950). *Helping people find jobs: how to operate a placement office.* New York: Harper & Brothers Publishers.

Kitson, Harry Dexter (1954). *I find my vocation.* New York: McGraw-Hill.

Lewis, Elias St. Elmo (1903). Catch-Line and Argument. In: *The Book-Keeper, Vol. 15, February*; 124.

Lively, Bertha A.; Pressey, Sidney L. (1923). A method for measuring the ‘vocabulary burden’ of textbooks. *Educational administration and supervision*, 9; 389–398.

Russell, C.P. (1921). How to Write a Sales-Making Letter. *Printers' Ink, June 2.*

Sheldon, Arthur Frederick (1911). *The art of selling, for business colleges.* Published/ Created: Libertyville, Ill., The Sheldon University Press.

Book Reviews

Gordana Đuraš: *Generalized Poisson models for word length frequencies in texts of Slavic languages.* Diss. University of Graz, Austria. Reviewed by E. Nemcová.

Usually, one does not review dissertations but the above work has been written at the Department of Statistics, hence it plays a special role. The fact that statisticians admit linguistic problems as worth of being studied signalizes a kind of paradigm change and inclusion of linguistics in the circle of at least "more exact" sciences. This trend can be observed in different official places. At the University of Lancaster, Rosie Knight has been awarded the inaugural Anna Siewierska Memorial Prize for her dissertation "Laws Governing Rank Frequency and Stratification in English Texts", and in Bucharest, the physicist Ioan-Iovitz.Popescu obtained the Prize for Exact Sciences "Grigore Moisil" of the National Grand Lodge of Romania in partnership with the Romanian Academy for the book "The Lambda Structure of Texts". It is not easy to break a paradigm which is frequently similar to a religion but in the present time the breaks come in shorter time intervals.

The second positive feature of this dissertation is the fact that it has been written in English. That means, English has been admitted in Austria as a dissertation language. For international communication in science it is a step forward, and for Englishmen it means to accept deviations not as errors but as inherent parts of the "intelligent Middle European English" which is one of the dozens of English dialects.

The author concentrates on some Slavic languages, discusses the problem of zero-syllabic words which may be considered proclitics, and computes word length in terms of syllable numbers. Any other definition of word length evokes insurmountable difficulties.

The book can be used also as a text-book on the Poisson distribution. Of course, no book can encompass all possible derivations, interrelations and uses of the Poisson distribution but for Slavic languages the author prepared a very thorough introduction. One finds here chapters on the Fucks-Poisson mixture, Singh-Poisson, Cohen-Poisson, Hyper-Poisson and some generalized Poisson distributions based on Lagrange development. Each chapter contains displacements, size-bias and different estimation methods for moments. In each chapter there are many details concerning the distributions (expectation, variance, probability generating function), there are figures showing the course of the distribution. In a separate chapter one finds the fitting to data, some other related distributions, theoretical chapters about generating functions and estimators. .

The book is a contribution to two disciplines at the same time. Of course, the problem of word length distribution is not finished. The more languages one analyzes, the more new distributions will appear. This is not caused by the availability of software but by a number of boundary conditions which are different in all languages. The idea to find a unique law working without any boundary conditions must be given up. This holds not only for physics but especially for the social sciences where the abilities, inheritance, social status, education, age, gender, aim, etc. of the individual play a decisive role. Since language is the most complex product of Man, whose highest principle is undisturbed communication, the means for reaching it are manifold. Whatever background model we use, e.g. urn model, Poisson birth process, waiting time, etc., every language uses different technique based on its grammar. Strongly analytic languages use short words, strongly synthetic languages use a mixture of long and short words (synsemantics); hence in every language the technique of placing words of different length is different. This may lead not only to differences in parameters but also to differences in models.

But even in one language text sorts may be different. A model adequate for poems need not be adequate for a stage play, etc. Texts of a language cannot be considered equal to a big basin full of homogeneous water, but rather to a garden in which every flower is different. Just as there is no mean height or colour of the flowers, there is no mean word length in a language, there is no mean word length even for Shakespeare (only for his known written works) and there are no populations in language having some mean properties. This boils down to the fact that for the study of word length corpuses as a whole cannot be used, only single texts are admissible. This fact has been very consequently taken into account in the reviewed book: the texts are studied individually but for the same text several models have been proposed. Usually the procedure of fitting is as follows: One fits several distributions to all texts and chooses the simultaneously best for all. Unfortunately, there are usually several "best" fittings. Decisions could be made if we knew the boundary conditions, but this aim will — perhaps — be attained in the distant future when hundreds of languages and millions of texts have been analyzed. For the time being we proceed rather inductively: that which is empirically better wins.

The problem of boundary conditions intervenes even with individual texts. Are they homogeneous? Were they written in one go or did the author make coffee pauses? Did (s)he or some editor make many corrections in the ready text? Frequently, we must modify a "good" model and create modified distributions, or we try with mixed distributions of which the Fucks-Poisson model is the best known. A possible "theory" in this domain runs through our fingers and we at least hope that we approximate some fictive truth.

Mathematically, the book is an excellent and very systematic introduction to different problems of probability distributions. It can be used not only for word length problems but mutatis mutandis for any linguistic problem that needs probabilistic modelling. The book can be recommended both to linguists and mathematicians.

Emmerich Kelih, *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation.* München-Berlin-Washington D.C.: Sagner 2012, 188 pp. (Specimina Philologiae Slavicae 168)
Reviewed by **Emília Nemcová**.

The book under review is both an excellent historical survey of the syllable problem and a methodological analysis of different approaches. While the majority of linguists believe that a linguistic unit is a strictly defined real entity having crisp boundaries - a belief which is a remnant of older "essentialistic" times - the author adheres rather to the conjecture of L. Hřebíček saying that "Let there be some (hypothetical) text constructs composed of some (hypothetical) components. If the size of the components is a function of the size of the constructs, according to Menzerath's law, then both the constructs and the components are textual units and they lie on two different levels" (cf. Altmann 2001: 11). Since the best criteria for establishing a linguistic unit are laws, and syllable length depends on word length, the "existence" of syllables can be considered as given, though some philosophies and linguists deny its theoretical relevance on different grounds. Of course, there are different models of syllable, but these are merely means of description.

In the first chapter the author succeeds in integrating the syllable in Köhler's (1986, 2005) self-regulation cycle representing the best way to a theory of language and text. This is done *passim*, hence the book is at the same time both a survey of some results and an introduction to theoretical thinking and to a deductive approach in linguistics. Syllable has the same fate as all other linguistic units: one discusses their definitions (there are dozens of definitions of word and sentence, too), rules of forming, criteria for separation based on some rules and describes the inventories or types. This is, of course, necessary, but it is only a necessary, but not a sufficient condition for theoretical work. Theoretical work begins with hypotheses and their testing. Rules are not part of a theory. In the best case they are criteria, and as such they are conventions.

In the second chapter different conceptions and definition of syllable are presented: for structuralism, it is Pulgram (1970), for generative phonology it is Hooper (1972, 1976) who begins to deviate from the conception of crisp boundaries and begins to measure. The measurement is the basis of Vennemann's (1988) approach who presents a dynamic syllable and whose approach can further be developed, and Lehfeldt's (1971) method who applies a probability approach for syllable segmentation. His algorithm has been programmed and applied in several languages.

Vennemann (1982, 1983, 1986, 1988) elaborated a number of preference rules - which are erroneously called laws and criticized on this reason by Kelih - representing tendencies of syllable formation. Many of these tendencies could be statistically tested. Kelih criticizes Vennemann's conception of a "better" syllable which should perhaps be replaced by "more frequent" or "easier to pronounce" or something else that is measurable.

Optimality "theory" is a classical case of a search for "essences": what is syllable and how to separate it from the next one. But even if we know it, nobody says what is the aim of this procedure? What are the results useful for? Usually, one sets up a hypothesis and samples data which are relevant for its testing. If the data force us to reject the hypothesis, we first check the data, then the criteria and at last, the hypothesis. In inductive/exploratory research we produce data and search for a "regularity" under which it could be subsumed and we generalize stepwise. But in optimality theory one cannot recognize either an existing or a future theory. Kelih is frugal with his criticism but evidently this is a case of proto-scientific approach asking merely "what is there?"

The third chapter revives the one hundred years old discussion concerning sonority. Is it an intuitive concept or can it be objectively measured? Can it be used for syllable definition and separation? Or is it a conglomerate of several properties? Some authors treat it as a crisp property, other ones speak about approximations and probability. Perhaps the theory of fuzzy sets would bring an acceptable solution to several problems.

The chapter on phonotactics - a concept appearing in the whole book - is very critical. Are there mirror-effects, asymmetries, positional preferences, correlation with phonetic distances, links to other properties, how to take a sample, etc.? Kelih is keen on showing the main deficiencies in statistical processing (or better: non-processing) of data. But even if one of the hypotheses has been positively tested in one language, is the trend a general phenomenon? How many languages must be included in the sample in order to render both types of errors small? Or would it not be better to strive for a theoretical approach from which individual hypotheses could be derived? In that case the hypothesis would be at least theoretically corroborated. The whole discipline, strongly supported by universalists, has a methodological shortage: it searches for universals, not for laws. All laws are universals but not vice versa. There is a long way from a universal to a law but in qualitative linguistics even rules or frequent events are called laws as mentioned above.

Chapter 5 brings a survey of trials to use statistics, in most cases for classification. The author shows the results and their weaknesses. Deviations are “explained” by the authors using some ad hoc criteria. One does not try to derive a function containing parameters for boundary conditions, one relies on absolute numbers. Nevertheless, these first steps in Slavic languages were the initiators of more sophisticated approaches in this century. The only reasonable hypothesis is the statement that the more syllable types there are in a language, the longer are they on the average. This relationship is clearly linear as shown on p. 137.

The last chapter is, from the point of Quantitative Linguistics, of course, the most progressive. After is has been shown that the syllable is a “legal” linguistic unit having all properties required by Altmann (1996), its placing in theory is performed by stating its relationship to different other language properties. It fits into the Köhlerian control cycle which contains additional links and interrelations to other linguistic properties. From the synergetic point of view there is no other and better way to “theorify” a linguistic entity. The author specifies a number of hypotheses linking the syllable and its properties with other entities. Unfortunately, functions, data and tests are not given - this is a task for generations of linguists - but if a theory of syllable is to be set up, the book under review furnishes us all necessary requirements. As a matter of fact, any future quantitative or synergetic approach can rely on it.

References

Altmann, G. (1996).The nature of linguistic units. *Journal of Quantitative Linguistics 3(1),1-8.*

Altmann, G. (2001). Theory building in text science. In: L. Uhlířová, G. Wimmer, G. Altmann, R. Köhler, (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 10-20*. Trier: WVT.

Hooper, J.B. (1972). The syllable in phonological theory. *Language 48(3), 525-540.*

Hooper, J.B. (1976). *An introduction to natural generative phonology***.** New York: Academic Press.

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

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

Lehfeldt. W. (1971). Ein Algorithmus zur automatischen Silbetrennung. *Phonetica 24, 212-237.*

Pulgram, E. (1970). *Syllable, word, nexus, cursus*. The Hague-Paris: Mouton.

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change.* Berlin: de Gruyter.