Glottometrics 25
2013

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. 25 (2013), Lüdenscheid: RAM-Verlag, 2013. 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 25 (2013) **ISSN 2625-8226**

Contents

Report

Review

Glottometrics 25, 2013, 1-21

Hubiness, length, crossings and their relationships in dependency trees

Ramon Ferrer-i-Cancho[1]

Abstract. Here tree dependency structures are studied from three different perspectives: their degree variance (hubiness), the mean dependency length and the number of dependency crossings. Bounds that reveal pairwise dependencies among these three metrics are derived. Hubiness (the variance of degrees) plays a central role: the mean dependency length is bounded below by hubiness while the number of crossings is bounded above by hubiness. Our findings suggest that the online memory cost of a sentence might be determined not just by the ordering of words but also by the hubiness of the underlying structure. The 2nd moment of degree plays a crucial role that is reminiscent of its role in large complex networks.

Keywords: syntactic dependency, syntax, dependency length, crossings.

1. Introduction

According to dependency grammar (Mel'čuk 1988, Hudson 2007) the structure of a sentence can be defined by means of a tree in which arcs indicate syntactic dependencies between the occurrences of words (Fig. 1). In standard graph theory (Bollobás 1998), the black circles from which arcs arrive or depart in Fig. 1 (black circles) are called vertices. Vertices are usually labeled with words. Thus, each occurrence of a word of a sentence corresponds to a vertex. Arcs are also called edges or links. Here we focus on two aspects of dependency trees: the length of the dependencies (the distance between syntactically linked words) and the number of crossings of the dependency tree. The syntactic dependency structure of a sentence (Fig. 1) is perhaps the most inspiring and useful linguistic example of dependency tree. This article is motivated by those trees.

We assume that the words of a sentence are placed in a sequence in the same order as in the original sentence and define the concept of distance in this sequence. We adopt the convention that the position of the first word of the sentence (i.e. the 1st element of the sequence) is 1, the position of the second word of the sentence (i.e. the 2nd element of the sequence) is 2 and so on. $\pi(v)$ is defined as the position of a vertex v. In Fig. 1, π('she') = 1, π('loved') = 2 and so on. n is defined as the length of the sentence in words. n is also the

[1] Complexity and Quantitative Linguistics Lab. Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya (UPC). Campus Nord, Edifici Ω, Jordi Girona Salgado 1-3. 08034 Barcelona, Catalonia (Spain).Phone: +34 934137870, Fax: +34 934137787. E-mail: rferrericancho@lsi.upc.edu

number of vertices of the tree and the position of the last word of the sentence. d is defined as the distance between two vertices u and v as the absolute difference of their positions, i.e. $d = |\pi(u) - \pi(v)|$. If u and v are linked, then d is also the length of the edge formed by vertices u and v (Ferrer-i-Cancho 2004). Thus the distance or the length of the dependency between '*she*' and '*loved*' is $d = 1$ and the distance or the length of the dependency between '*loved*' and '*for*' is $d = 2$. d goes from 1 to $n - 1$.

Alternatively, dependency length has been defined so that consecutive words have distance zero (e.g. Hudson 1995, Hiranuma 1999). d_0 is used for referring to the length or distance defined using this alternative convention. This way, the length of the dependency between 'she' and '*loved*' is $d_0=0$ and that of the dependency between '*loved*' and '*for*' is $d_0=1$. d_0 goes from 0 to n-2.

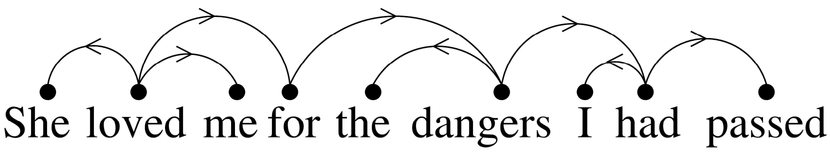

Figure 1. The syntactic structure of the sentence *'She loved me for the dangers I had passed'* following the conventions by Mel'čuk (1988). *'she'* and the verb *'loved'* are linked by a syntactic dependency. Arcs go from governors to dependents. Thus, *'she'* and *'me'* are dependents of the verbal form *'loved'*. Indeed, *'she'* and *'me'* are arguments of the verb form *'loved'* (the former as subject and the latter as object).

The concept of link crossing (Hays 1964, Holan et al. 2000, Hudson 2000, Havelka 2007) will be defined next. Imagine that we have two pairs of linked vertices: (u,v) and (x,y), such that $\pi(u) < \pi(v)$ and $\pi(x) < \pi(y)$. The arcs (or edges) of (u,v) and (x,y) cross if and only if $\pi(u) < \pi(x) < \pi(v) < \pi(y)$ or $\pi(x) < \pi(u) < \pi(y) < \pi(v)$. We define C as the number of different pairs of edges that cross. For instance, $C = 0$ in the sentence in Fig. 1 and $C = 9$ in Fig. 2. When there are no vertex crossings ($C = 0$), the syntactic dependency tree of a sentence is said to be planar (Havelka 2007).

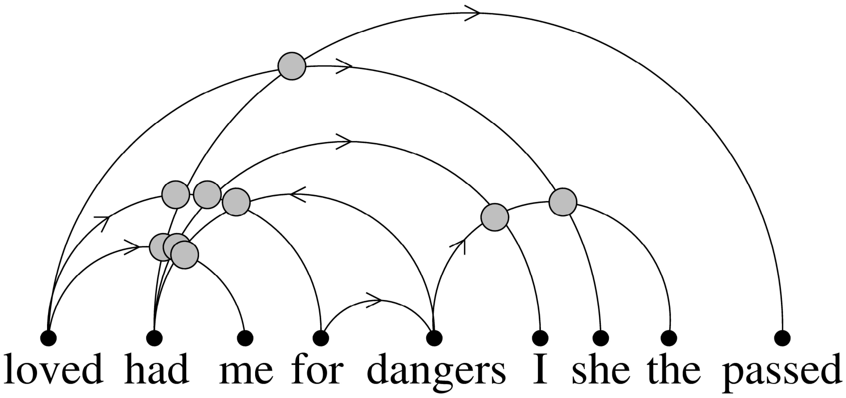

Figure 2. The structure of the sentence in Fig. 1 after having scrambled the words. Gray circles indicate edge crossings.

Although examples of real sentences with non-crossing dependencies are well-known (e.g., Mel'čuk 1988) the ungrammatical sentence in Fig. 2 has been chosen to introduce one of the problems that will be addressed in this article: what is a priori the maximum of number of crossings that can be reached? Crossings in syntactic dependency structures are rather rare (Havelka 2007) and it was hypothesized that this could be a side effect of minimizing the distance between syntactically linked words (Ferrer-i-Cancho 2006), which would be in turn a consequence of minimizing the online memory cost of the sentence (Morril 2000, Hawkins 2004, Grodner & Gibson 2005). Dependency lengths and crossings are no dissociated concepts as one may a priori believe (Hochberg & Stallmann 2003, Ferrer-i-Cancho 2006, Liu 2008).

This raises a very important research question for theoretical linguistics: is the minimization of crossings a principle by its own or is it a side-effect of a principle of dependency length minimization? Another related question is the origins of the low degree of vertices in syntactic dependency trees (in a sufficiently large sentence, vertices with a degree of the order of the length of the sentence are rare). In the sentence in Fig. 1, the maximum degree is three although it could be n - 1 = 8. Is it due to an autonomous principle of degree minimization or would it be again a side-effect of distance minimization? These questions are crucial for the development of a theory of language as simple as possible. A fundamental theoretical question is whether the low frequency of crossings or the low hubiness of syntactic dependency structures is due to an innate or biologically determined faculty for language that imposes universal constraints on world languages (e.g., the minimization of hubiness or the number of crossings) or these features could be simply due to the universal limitations of a complex brain for performing computations, being language production and processing particular cases of those computations (Christiansen et al 2012). Here it will be shown that the maximum number of crossings that can be achieved by a sentence (C_{max}) is bounded above by its mean dependency length ($\langle d \rangle$) and thus pressure for reducing crossings or hubiness could be a simple consequence of universal computational limitations of brains.

Another important research question is whether the properties of dependency structures, when considered independently of how vertices are arranged sequentially, exhibit features that help to save computational costs. Here it will be shown that the variance of vertex degrees determines the minimum $\langle d \rangle$ the can be achieved ($\langle d \rangle_{\min}$), which in turn determines the minimum cognitive cost of sequences. This has a concrete consequence: the syntactic trees of long sentences cannot have hubs (hubs are vertices with a large number of links) due to the high online memory cost this would imply.

Those arguments are abstract enough to be valid not only for the communicative sequential behavior of other species but also for non-linguistic sequential behavior in general (human or not). In the present article, human language is the fuel to contribute to the development of a theory of natural sequential processing.

Besides illuminating the questions above, the present article aims at providing some mathematical results that are potentially useful for any research on (a) the mean dependency length (b) the number of crossings or (c) the relationship between mean dependency length and number of crossings in syntactic dependency trees. Lower and upper bounds for these quantities will be provided and the relationships between them will be unraveled.

The remainder of the article is organized as follows. Section 2 provides an introduction to graph theory that will help in the next sections. Sections 3 and 4 provide some results on dependency length and crossings, respectively. Sections 3 and 4 are essentially an enumeration of results aimed at facilitating their application. Readers interested in further details are referred to the appendices. The main article ends with a discussion in Section 5.

2. Graph theory

This section summarizes some results from standard graph theory and Appendix A. First we review elementary concepts of standard graph theory (Bollobás 1998). We neglect the direction of syntactic dependency arcs because our definition of dependency length and crossing is independent from it. A tree of *n* vertices has *n* - 1 edges. The degree of a vertex is the number of connections. For instance, 'she' in Fig. 1 has degree 1 while 'loved' has degree 3. Vertices with a large degree with regard to *n* are called hubs (Pastor-Satorras & Vespignani 2004) whereas vertices with degree one are called leaves (Bollobás 1998). It is convenient to label vertices not with the associated word (which is problematic if the same word appears more than once) but with natural numbers from 1 to *n*. Thus, k_i is the vertex degree of the *i*-th word of the sentence (e.g. $k_1 = 1$, $k_2 = 3$ in Fig. 1). The structure of the tree is defined by the adjacency matrix $A = \{a_{ij}\}$, where $a_{ij} = 1$ if the pair of vertices (i,j) is linked and otherwise $a_{ij} = 0$. The matrix is symmetric $a_{ij} = a_{ji}$ because we treat arcs as if they had no direction. Loops are not allowed ($a_{ii} = 0$). One has

$$k_i = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji} \ . \tag{1}$$

$\langle k \rangle$ and $\langle k^2 \rangle$ are the mean values of k_i and k_i^2 (the 1st and 2nd moments of k_i, respectively), i.e.

$$\langle k \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i \ . \tag{2}$$

$$\langle k^2 \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i^{\,2} \ . \tag{3}$$

For any tree, it is easy to see that (Noy 1998)

$$\langle k \rangle = 2 - \frac{2}{n} . \tag{4}$$

for $n \geq 1$, knowing Eq. 2 and

$$\sum_{i=1}^{n} k_i = 2(n-1) . \tag{5}$$

Since $\langle k \rangle$ is the same for any tree of a given length, $\langle k^2 \rangle$ determines $V[k]$, the variance of the vertex degrees as $V[k] = \langle k^2 \rangle - \langle k \rangle^2$.

Two kinds of extreme trees that will be very useful throughout this article, i.e. the linear tree and the star tree, will be introduced next. A linear tree (also called path tree) is a tree with no branching at all (Fig. 1 (a)). A star tree is a tree where all vertices except one (the hub) are connected to the hub (Fig 3 (b)). Star trees model the syntactic dependency structure of utterances with a single head (the head being the hub). $V[k]$ is maximized by star trees and thus $\langle k^2 \rangle$ alone can be regarded as a measure of "hubiness".

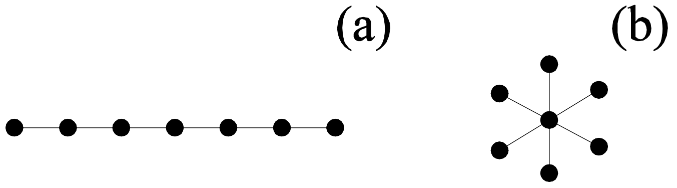

Figure 3. (a) a linear tree and (b) a star tree

Table 1 shows a summary of the second moment and the variance of linear and star trees (details of the calculation are given in Appendix A). It will be shown that $\langle k^2 \rangle$ is a key quantity for *d* and *C* that is maximized by star trees and minimized by linear trees. Table 2 shows some graph theoretic measurements on the dependency trees of Figs. 1 and 2.

Table 1
Summary of the properties of two extreme kinds of trees: star and linear trees. *n* is the number of vertices, $\langle k^2\rangle$ is the degree 2nd moment, $V[k]$ is the variance of the degree, $\langle d\rangle_{\min}$ is the actual minimum value of $\langle d \rangle$ that a linear arrangement of vertices can achieve and *C* is the number of link crossings

	Linear	Star
$\langle k^2 \rangle$	$4-\frac{6}{n}$	$n-1$
$V[k]$	$\frac{2}{\mathrm{n}}\left(1-\frac{2}{n}\right)$	$n-5+\frac{4}{n}\left(2-\frac{1}{n}\right)$
$\langle d \rangle_{\min}$	1	$\frac{n^2}{4(n-1)}$ if *n* is even $\frac{n+1}{4}$ if *n* is odd
C	$\leq \frac{n(n-5)}{2}+3$	0

3. Length theory

This section summarizes results from Appendix B. d_i is defined as the length of the *i*-th edge of dependency tree of *n* vertices. $d_1,\ldots,d_i,\ldots,d_{n-1}$ is the list of the lengths of the *n* - 1 edges of the tree. The mean dependency length of that tree is then

$$\langle d \rangle = \frac{1}{n-1}\sum_{i=1}^{n-1} d_i \qquad (6)$$

for $n \geq 1$. One has $\langle d \rangle$ = 11/8 ≈ 1.375 for the sentence in Fig. 1.

Table 2
Summary of the properties of the syntactic dependency trees of Fig. 1 and Fig. 2.

		Fig. 1	Fig. 2
Graph Theory	n	9	
	$\langle k^2 \rangle$	4	
Length Theory	$\langle d \rangle$	11/8 = 1.375	29/8 = 3.625
	$\langle d^2 \rangle$	17(8=2.125	133/8=16.625
	$\langle d \rangle_{\min}$	$\geq$ 19/16 = 1.1875	
	$E[d]$, $E[\langle d \rangle]$	= 10/3 ≈ 3.33	
	$\langle d^2 \rangle$	17/8=2.125	133/8=16.625
Crossing Theory	C	0	9
	C_{max} (by degree, Eq. 14)	$\leq$ 18	
	C_{max} (by length, Eq. 12)	$\leq$3	$\leq$21
	C_{max} (by length, Eq. 13)	$\leq$9	$\leq$32

We are interested in knowing the minimum and maximum values that $\langle d \rangle$ can take, $\langle d \rangle_{\min}$ and $\langle d \rangle_{\max}$, respectively. We would like to shed light on the extent to which actual sentences minimize or maximize $\langle d \rangle$. Since $1 \leq d_i \leq n - 1$, one has that $1 \leq \langle d \rangle \leq n - 1$. In general, 1 is the minimum value that $\langle d \rangle$ can take. This value is achieved by a linear tree whose vertices are arranged linearly. A linear tree is a tree where all vertices have degree 2 except two vertices that have degree 1. A linear arrangement of a linear tree consists of placing the vertices of degree 1 in both extremes the sequence (see Fig. 3 (a)) and placing the vertices of degree 2 immediately between its two linked vertices. Thus, $d_i = 1$ for all edges. While 1 is a reachable lower bound of $\langle d \rangle$ for linear trees, n - 1 is not a tight upper bound of $\langle d \rangle$ in general because there can only be a single edge of length n - 1. The number of edges that can be formed at distance d is $N(d) = n - d$, hence $N(n - 1) = 1$.

A non-crossing tree is defined as linear arrangement of a tree without link crossings. The tree in Fig. 1 is non-crossing (C=0) while the tree in Fig. 2 is not (C>0). It can be shown that the maximum value of $\langle d \rangle$ that a non-crossing tree of n vertices can achieve is

$$\langle d \rangle_{\max} = \frac{n}{2} \tag{7}$$

with $\langle d_0 \rangle_{\max} = \langle d \rangle_{\max} - 1$.

As a star tree cannot have crossings because all vertices except the hub are connected to the hub, Eq. 7 gives the maximum value of $\langle d \rangle$ that a star tree can achieve. This maximum is achieved when the hub is placed first or last in the sequence of vertices. In contrast, the minimum value of $\langle d \rangle$ that a star tree can achieve is obtained when the hub is placed at the center and half of the leaves to its left and half of the leaves to its right (at position $(n + 1)/2$ if n is odd and either at positions $n/2$ or $n/2 + 1$ if n is even).

If the vertices of an edge are placed at random positions of a sentence (being a priori all the n sentence positions equally likely), it can be can also be shown that the expected length of a single edge and its variance for $n \geq 2$ are

$$E[d] = \frac{n+1}{3} \tag{8}$$

and

$$V[d] = \frac{(n+1)(n-2)}{18}, \tag{9}$$

respectively. One has $E[d_0] = E[d] - 1$ and $V[d_0] = V[d]$. E[$\langle d \rangle$], the expected mean length of the edges of a tree in which vertices have been placed at random, satisfies $E[\langle d \rangle] = E[d]$.

The minimum value that $\langle d \rangle$ can achieve is 1, which is only achieved by a linear tree. However, notice that $\langle d \rangle$ = 1 is impossible to achieve in a tree with at least one vertex of degree three or greater. Hence, what about non-linear trees?

Table 1 shows the value of $\langle d \rangle_{\min}$ for star trees. A lower bound for $\langle d \rangle_{\min}$ can be derived from $\langle d \rangle_{\min}$ for star trees. $\langle d \rangle_{\min}$, the minimum value of $\langle d \rangle$, obeys

$$\langle d \rangle_{\min} \geq \frac{1}{2(n-1)} \sum_{i=1}^{n} \left(\left\lfloor \frac{k_i}{2} \right\rfloor \left(\left\lfloor \frac{k_i}{2} \right\rfloor + 1 \right) + \frac{k_i + 1}{2} (k_i \bmod 2) \right), \tag{10}$$

where x mod y is the modulus of the division of x by y. Eq. 10 is obtained by looking at the whole tree as an ensemble of star trees formed by each vertex and its neighbours (the star tree of the i-th vertex has k_i+1 vertices) and considering that every star tree is arranged sequentially in the best possible way, independently from other star trees. A much simpler lower bound for $\langle d \rangle_{\min}$ with regard to Eq. 10 is

$$\langle d\rangle_{\min} \geq \frac{n\langle k^2\rangle}{8(n-1)} + \frac{1}{2}. \tag{11}$$

Eq. 11 shows that the minimum dependency length is bounded below by the variance of the degrees. Table 2 shows some dependency length measurements for the dependency trees of Figs. 1 and 2.

4. Crossing theory

This section summarizes results from Appendix C. Crossings are impossible ($C = 0$) for $n \leq 3$. When $n > 3$, simple upper bounds for C_{max}, the maximum number of crossings, are offered by the linear arrangement of vertices and by the structure of the tree. As for the former, one has

$$C_{\max} \leq \binom{n-1-M}{2}, \tag{12}$$

where M is the number of uncrossable edges (edges of length 1 or n - 1 are not crossable). Incorporating information from all dependency lengths, one also has

$$C_{\max} \leq \frac{n-1}{2}\left(n\langle d\rangle - \langle d^2\rangle - n + 1\right), \tag{13}$$

where $\langle d^2\rangle$ is the 2nd moment of dependency length. It is easy to see from the previous inequality that crossings are impossible ($C = 0$) when $\langle d\rangle$ takes its absolute minimum value ($\langle d\rangle = 1$). Notice that Eq. 10 indicates that not all trees can reach $\langle d\rangle = 1$. As for an upper bound derived from the structure of the tree, one has

$$C_{\max} \leq C_{pairs} = \frac{n}{2}\left(n-1-\langle k^2\rangle\right), \tag{14}$$

where C_{pairs} is the number of edge pairs that can cross (edges departing from the same vertex cannot cross).

Knowing that $\langle k^2\rangle = n - 1$ in a star tree (Table 1), Eq. 14 gives that a star tree cannot have crossings ($C_{max}= 0$) regardless of how its vertices are arranged linearly. Since $C \geq 0$ it follows from Eq. 14 that a tree with $\langle k^2\rangle > n - 1$ cannot exist because it would have a negative number of crossings. Therefore, a star tree has maximum $\langle k^2\rangle$.

5. DISCUSSION

It has been shown that $\langle d\rangle_{\min}$ is bounded below by $\langle k^2\rangle$, i.e. the larger the value of $\langle k^2\rangle$ (Eq. 11) the larger the value of $\langle d\rangle_{\min}$. It has also been shown that C_{max} is bounded above by both $\langle d\rangle$ (Eq. 13) and $\langle k^2\rangle$ (Eq. 14), i.e. the smaller the value of $\langle d\rangle$ the smaller the value of C_{max} while the larger the value of $\langle k^2\rangle$ the smaller the value of C_{max}. This suggests that the low frequency of crossings in languages could be due to pressure for high degree variance but also to pressure for short dependency lengths. However, a high degree variance increases the minimum arc length that can be achieved and therefore raises the minimum cognitive cost of the sentence and thus the true reason for the low frequency of crossings in language might not hubiness but online memory limitations of the human brain.

Temperley (2008) has suggested that the structural properties of syntactic dependency trees (leaving aside the linear arrangement of vertices) might reflect pressure for dependency length minimization. With this regard, our results have implications for the presence of hubs in sentences. Eq. 14 implies that the more skewed the degree distribution of vertices (the higher the value of $\langle k^2\rangle$), the higher the minimum value of $\langle d\rangle$ that can be achieved. Reading this result in terms of the cognitive cost implied by $\langle d\rangle$ (Ferrer-i-Cancho 2006), long sentences with large $\langle k^2\rangle$ would be cognitively too expensive in practice. If actual sentences minimize $\langle d\rangle$, then a necessary condition is that $\langle d\rangle_{\min}$ is not too high. Thus, $\langle k^2\rangle$ must be reduced and hubs must be avoided. This is in contrast with the large-scale organization of syntactic dependency networks (Ferrer-i-Cancho et al. 2004), where vertices with high degree do exist. The absence of hubs at the sentence scale is likely to be caused by the constraints of short term memory (Morrill 2000, Hawkins 2004, Grodner and Gibson 2005) while the existence of hubs at the large-scale could be due to the fact that dependencies at this scale are kept by long-term memory. In sum, the limited resources of our brains lead to the principle of dependency length minimization (Morrill 2000, Hawkins 2004, Grodner and Gibson 2005, Ferrer-i-Cancho 2006), which in turn make hubs expensive in syntactic dependency trees.

Our theoretical framework suggests new questions for empirical research. If there is actually cognitive pressure for reducing hubiness (V[k]) or mean arc lengths ($\langle d\rangle$), an important research question is: how do these quantities scale with n, the length of the sentence? As the maximum number of crossings depends on V[k] or $\langle d\rangle$ (Section 3), how does C scales as a function of V[k] or $\langle d\rangle$? As the minimum value of $\langle d\rangle$ depends on V[k] (Section 2), how does $\langle d\rangle$ scale as a function of V[k]? The growing availability of dependency treebanks (e.g. Civit *et al.* 2006, Böhmová *et al.* 2003, Bosco *et al.* 2000) suggests that the questions above could be answered for syntactic dependency trees in a near future.

Our results have also implications for the parallel research on complex network physics. It has been shown that $\langle k^2\rangle$ is a crucial quantity for $\langle d\rangle_{\min}$ (Eq. 11), C_{max} (Eq. 14) in dependency trees. This result is reminiscent of the key role played by $\langle k^2\rangle/\langle k\rangle$ in large complex networks (Pastor-Satorras & Vespignani 2004), for instance, concerning the diffusion of epidemics in Internet (if $\langle k^2\rangle/\langle k\rangle$ diverges then the pandemics cannot be stopped). In syntactic dependency trees, one has that $\langle k^2 \rangle / \langle k \rangle$ = $\langle k^2 \rangle / (2 - 2/n)$). Our findings support the idea that $\langle k^2 \rangle / \langle k \rangle$ is a general fundamental property of the network of many real systems.

ACKNOWLEDGEMENTS
This manuscript derives from a longer manuscript finished in 2007. We are grateful to G. Altmann, O. Jiménez and H. Liu for helpful comments.. This work was supported by the grant “Iniciació i reincorporació a la recerca” from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

References

Böhmová, A., Hajič, J., Hajičová, E. & Hladká, B. (2003). The Prague dependency tree bank: three-level annotation scenario. In: Abeille, A. (ed.), *Treebanks: building and using syntactically annotated corpora: 103-127.* Dordrecht: Kluwer.

Bollobás, B. (1998). *Modern graph theory*. New York: Springer-Verlag.

Christiansen, M.H., Conway, C.M & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from eventrelated brain potentials. *Language and Cognitive Processes 27, 231-256.*

Civit, M., Martí, M.A., Bufí, N. (2006). *Cat3LB and Cast3LB: from Constituents to dependencies*: 141-153. Berlin: Springer Verlag,

DeGroot, M. H. (1989). *Probability and statistics*. 2nd edition. Reading, MA: Addisson-Wesley.

Ferrer-i-Cancho, R., Solé, R.V. & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E 69, 051915.*

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E 70, 056135.*

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters 76, 1228-1235.*

Grodner, D. & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science 29, 261-291.*

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science 9, 159-191.*

Havelka, J. (2007). Beyond projectivity: multilingual evaluation of constraints and measures on non-projective structures. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07): 608-615.* Prague, Czech Republic: Association for Computational Linguistics.

Hawkins, J.A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hays, David G. (1964). Dependency theory: a formalism and some observations. *Language 40, 511-525.*

Hiranuma, S. (1999). Syntactic difficulty in English and Japanese: a textual study. *UCL Working Papers in Linguistics, 11, 309-322.*

Hochberg. R. A. & Stallmann. M. F. (2003). Optimal one-page tree embeddings in linear time. *Information processing letters 87, 59-66.*

Holan, T., Kubon, V., Plátek, M. & Oliva, K. (2000). On complexity of word order. *Traitement automatique des langues 41 (1), 273-300.*

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper. http://www.phon.ucl.ac.uk/home/dick/difficulty.htm

Hudson, R. (2000). Discontinuity. *Traitement automatique des langues 41 (1), 15-56.*

Hudson, R. (2007). *Language networks. The new word grammar.* Oxford University Press.

Mel'čuk, I. (1988). *Dependency syntax: theory and practice*. Albany, N.Y.: SUNY Press.

Morrill, G. (2000). Incremental processing and acceptability. *Computational Linguistics, 26, 3, 319-338.*

Noy, M. (1998). Enumeration of noncrossing trees on a circle. *Discrete Mathematics 180, 301-313.*

Pastor-Satorras, R. & Vespignani, A. (2004). *Evolution and Structure of the Internet: A Statistical Physics Approach.* Cambridge: Cambridge University Press.

Spiegel, M. & Liu, J. (1999). *Mathematical handbook of formulas and tables*. 2nd edition. New York: McGraw-Hill.

Temperley, D. (2008). Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics 15, 256-282.*

Yuret, D. (2006). *Lexical attraction models of language.* Unpublished.

Zörnig, P. (1984). The distribution of the distance between like elements in a sequence. *Glottometrika 6, 1-15.*

APPENDIX A: GRAPH THEORY

A.1. 2nd moment and variance of degree in linear and star trees

Knowing Eq. 3, it is easy to see that a linear graph (i.e. two vertices of degree 1 and the remainder of degree 2) has

$$\left\langle k^2 \right\rangle = \frac{1}{n}(2+4(n-2)) = 4-\frac{6}{n} \tag{A1}$$

whereas a star graph has

$$\left\langle k^2 \right\rangle = \frac{1}{n}(n-1+(n-1)^2) = n-1 \tag{A2}$$

for $n \geq 2$. While $\left\langle k^2 \right\rangle$ never exceeds 4 in a linear graph it grows linearly with n in a star graph. Knowing that the degree variance is $V[k]$ = $\left\langle k^2 \right\rangle - \left\langle k \right\rangle^2$ and Eqs. 4, A1 and A2, it is easy to show that a linear graph has

$$V[k] = \frac{2}{\mathrm{n}}\left(1-\frac{2}{n}\right). \tag{A3}$$

and a star graph has

$$V[k] = n-5+\frac{4}{n}\left(2-\frac{1}{n}\right). \tag{A4}$$

See Noy (1998) for $\left\langle k^2 \right\rangle$ and $V[k]$ in random trees and random trees without crossings.

A.2. Linear trees have minimum degree variance.

Next it will be proven that a linear tree has minimum $\left\langle k^2 \right\rangle$ by induction on n. Consider the sum of the squares of degrees of a tree of n vertices is

$$K_2(n) = \sum_{i=1}^{n} k_i^{\ 2} \tag{A5}$$

and thus $\left\langle k^2 \right\rangle$ = $K_2(n)/n$. In a linear, tree Eq. A1 gives $K_2(n) = 4n$ - 6. We want to prove that

$$K_2(n) \geq 4n - 6 \tag{A6}$$

for any tree (with $n \geq 2$). When $n = 2$, Eq. 6 holds trivially as only a linear tree is possible. We hypothesize that A6 holds for n and wonder it holds for $n + 1$, too. Imagine that the degree sequence of a tree of $n + 1$ vertices is k_1, k_2, k_3,…, k_n, k_{n+1}. A leaf is defined as a vertex of degree 1. It is well-known that any tree has at least two leaves (Bollobás 1998, pp. 11). Without any loss of generality, consider that the (n+1)-th vertex is a leaf and that the vertex that must be attached to that leaf is the n-th vertex (a leaf, by definition, has one connection). As $k_{n+1} = 1$, the tree of n+1 vertices has

$$K_2(n+1) = \sum_{i=1}^{n+1} k_i^{\ 2} = \sum_{i=1}^{n} k_i^{\ 2} + 1\,. \tag{A7}$$

The degree sequence k_1, k_2, k_3,…, (k_n – 1) defines a tree of n vertices as we only have substracted a leaf. As $k_n^2 = (k_n - 1)^2 + 2k_n - 1$, Eq. A7 can be rewritten as

$$K_2(n+1) = \sum_{i=1}^{n-1} k_i^{\ 2} + (k_n - 1)^2 + 2k_n = K_2'(n) + 2k_n\,, \tag{A8}$$

where $K_2'(n)$ is the value of $K_2(n)$ for the degree sequence of length n above.
By the hypothesis of induction, $K_2'(n) \geq 4n - 6$ and thus

$$K_2(n+1) \geq 4n - 6 + 2k_n\,. \tag{A9}$$

Notice that $k_n \geq 1$ as the n-th vertex is connected to the (n+1)-th vertex. Furthermore, notice also that $k_n \geq 2$ when $n > 2$ because the n-th vertex must be connected to vertices other than the (n + 1)-th to keep the graph connected (connectedness of the graph of n+1 nodes requires $k_n > 1$ except when $n = 1$, but we are considering the case $n > 2$). Applying $k_n \geq 2$ to Eq. A9 yields

$$K_2(n+1) \geq 4n - 2 = 4(n+1) - 6 \tag{A10}$$

as we wanted to prove.

APPENDIX B: LENGTH THEORY

B.1. The distribution of dependency lengths in random linear arrangements.

First we study the distribution of dependency lengths in trees where vertices are placed at random in a sequence. The probability that two randomly placed vertices in a sequence of length *n* are at distance *d* is (Ferrer-i-Cancho 2004)

$$p(d) = \frac{N(d)}{\sum_{i=1}^{n-1} N(i)}, \tag{B1}$$

where *N*(*d*) = *n* - *d* is the number of vertex pairs at distance *d* (*N*(*d*) = 0 if *d* < 1 or *d* > *n* - 1). Knowing Table 3 and *N*(*d*) = *n* – *d,* Eq. B1 is transformed into

$$p(d) = \frac{2(n-d)}{n(n-1)}. \tag{B2}$$

for $n \geq 2$. *p*(*d*) also defines the probability that the vertices forming an edge are at distance *d* (independently from the length of other edges). Thus, *E*[*d*], the expected value of the distance *d* separating two linked vertices is

$$E[d] = \sum_{d=1}^{n-1} p(d)d\ . \tag{B3}$$

Table 3
A summary of summations of powers of consecutive natural numbers
(Spiegel & Liu 1999)

a	$\sum_{x=1}^{n} x^a$	$\sum_{x=1}^{n-1} x^a$
1	$\frac{n(n+1)}{2}$	$\frac{n(n-1)}{2}$
2	$\frac{n(n+1)(2n+1)}{6}$	$\frac{(n-1)n(2n-1)}{6}$
3	$\frac{n^2(n+1)^2}{4}$	$\frac{(n-1)^2 n^2}{4}$

Applying Eq. B2 and Table 3 to Eq. B3, it is obtained

$$E[d] = \frac{2}{n(n-1)}\left(n\sum_{d=1}^{n-1} d - \sum_{d=1}^{n-1} d^2 \right) = \frac{n+1}{3}. \tag{B4}$$

for $n \geq 2$ after some algebra. Notice that $E[d]$ (Eq. B4) is the expected length of a single edge. $E[\langle d \rangle]$ is the expected mean arc length over all the edges of a tree (in which vertices have been randomly placed). It is easy to see that $E[d] = E[\langle d \rangle]$ for any tree because the expectation of a sum of random variables (independent or not) is the sum of the expectations of each of the variables (DeGroot 1989). Recalling the definition of $\langle d \rangle$ in Eq. 6, one has

$$E[< d >] = E\left[\frac{1}{n-1}\sum_{i=1}^{n-1} d_i\right] = \frac{1}{n-1}\sum_{i=1}^{n-1} E[d_i] = E[d]. \tag{B5}$$

as we wanted to prove.

$V[d]$, the variance of d of a single edge, is

$$V[d] = E[d^2] - E[d]^2. \tag{B6}$$

Firstly, we calculate $E[d^2]$. Applying Eqs. B2 and B3 to

$$E[d^2] = \sum_{d=1}^{n-1} p(d)d^2 , \tag{B7}$$

it is obtained

$$E[d^2] = \frac{2}{n(n-1)}\left(n\sum_{d=1}^{n-1} d^2 - \sum_{d=1}^{n-1} d^3\right). \tag{B8}$$

The application of Table 3 yields finally

$$E[d^2] = \frac{n(n+1)}{6} \tag{B9}$$

for $n \geq 2$ after some algebra.

Secondly, replacing the r.h.s. of Eqs. B4 and B9 into Eq. B6 one finally obtains

$$V[d] = \frac{(n+1)(n-2)}{18}, \tag{B10}$$

with $n \geq 2$ after some work.

As for $E[d_0]$, $E[d_0^2]$ and $V[d_0]$, knowing that $E[x - 1] = x$ and $V[x - 1] = V[x]$ (DeGroot 1989) and $d_0 = d - 1$, one obtains

$$E[d_0] = \frac{n-2}{3}, \tag{B11}$$

and $E[d_0^2] = E[d^2] - 2E[d] + 1 = n^2/6 + n/2 + 1/3$ and $V[d_0] = V[d]$. Eqs. B10 and B11 have also been derived in the context of the distance between not necessarily consecutive repeats in a sequence (Zörnig 1984).

B.2. The maximum mean dependency length.

We aim to calculate or bound above $\langle d \rangle_{\max}$, the maximum value that $\langle d \rangle$ can reach in a linear arrangement of a tree without crossings. Two procedures to arrange the vertices linearly will be presented: one for star trees and another for linear trees. Then it will be shown that value of $\langle d \rangle$ achieved by those procedures is actually maximum.

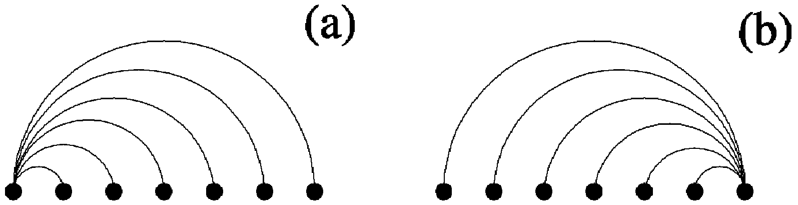

Figure 6. Two symmetric ways of arranging the vertices of a star tree in a way that the mean dependency length is $\langle d \rangle = n/2$.

Imagine that the hub of a star tree is placed at one of the extremes of the sequence of vertices (the hub is placed first or last) as in Fig. 6. In that case, the mean dependency length is

$$\langle d \rangle = \frac{1}{n-1}\sum_{i=1}^{n-1} d_i = \frac{1}{n-1}\sum_{i=1}^{n-1} d \ . \tag{B12}$$

Knowing Table 3, Eq. B12 yields

$$\langle d \rangle = \frac{n}{2} \tag{B13}$$

and

$$\langle d_0 \rangle = \langle d \rangle - 1 = \frac{n}{2} - 1. \tag{B14}$$

It is tempting to think that star trees are the only trees that can achieve this mean dependency length. Indeed, it easy to see that linear trees arranged linearly as in Fig. 7 also achieve the same mean dependency length than star trees with hub first or last as those arrangements of linear trees also obey Eq. B12.

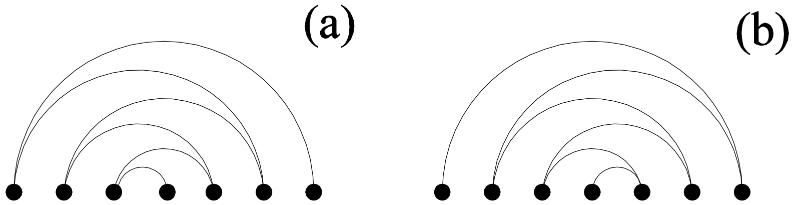

Figure 7. Two symmetric ways of arranging the vertices of a linear tree in a way that the mean dependency length is $\langle d \rangle = n/2$.

D is defined as the sum of dependency lengths, i.e. $D = (n-1)\langle d \rangle$ and $\Delta(x) = x(x-1)/2$. Next it will be shown by induction on *n* that a non-crossing tree with $D = \Delta(n)$ (and thus $\langle d \rangle = n/2$) has the maximum *D* that a non-crossing tree can achieve. The base of the induction is $n = 2$, where only a non-crossing tree can be formed. In that case $D = 1$ is maximum. The induction hypothesis is that any non-crossing tree of $n' < n$ vertices with $D = \Delta(n')$ has maximum *D*. It will be shown that a non-crossing tree of *n* vertices (n ≥ 3) and $D = \Delta(n)$ also has maximum *D*. To see it, consider that any non-crossing tree of *n* vertices can be constructed in two ways (Yuret 2006):

a) Concatenating two non-crossing subtrees that share the *v*-th vertex of the sequence (Fig. 8 (a)). That vertex is the last vertex of the first subtree and the first vertex of the second subtree. One subtree has *v* vertices and the other subtree has n-v+1 vertices. $2 \leq v \leq n - 1$ is required for being a true decomposition of a non-crossing tree of *n* vertices (each subtree having less than *n* vertices). For instance, the tree in Fig. 1 can be constructed by concatenating the subtree induced by words from ‘She’ to ‘for’ (both included) and the one induced by words from ‘for’ to ’passed’ (both included).

b) Concatenating two non-crossing subtrees that do not share any vertex, one with *v* vertices and the other with the following n-v vertices, and linking the first vertex of the first subtree with the last vertex of the second subtree (Fig. 8 (b)). $1 \leq v \leq n - 1$ is reqiured for being a decomposition of a non-crossing tree of *n* vertices. The non-crossing tree in Fig. 1 has not been constructed in this fashion but this is the case of the subtree induced by the words ‘for’, ‘the’ and ‘dangers’.

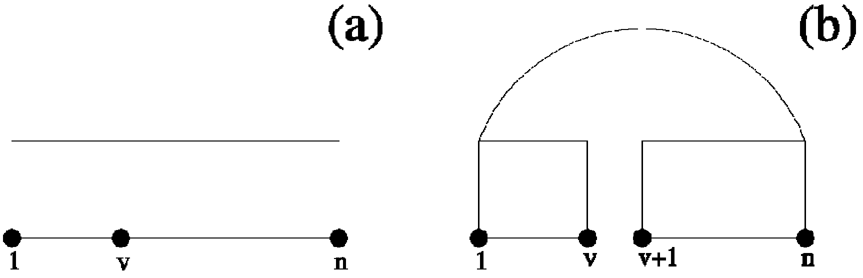

Figure 8. Schemes of two decompositions of a non-crossing tree. Rectangles indicate non-crossing subtrees. Circles indicate the first and the last vertex of each rectangle. In (a), the last vertex of the first subtree and the first vertex of the second subtree overlap. In (b), the subtrees are joined by a link between the first vertex of the first subtree and the last vertex of the second subtree.

$D_a(v)$ and $D_b(v)$ are defined as the maximum sum of arc lengths for construction a) and b), repectively, as a function of *v*, the position of the last vertex of the first non-crossing subtree. As for construction of type a), the maximum sum of dependency lengths that can be reached is

$$D_a = \max_{2 \leq v \leq n-1} \left(D_a(v) \right). \tag{B15}$$

By the hypothesis of induction, $D_a(v)$ is

$$D_a(v) = \Delta(v) + \Delta(n - v + 1) = v^2 - (n+1)v + \frac{n(n+1)}{2}. \tag{B16}$$

As for constructions of type b), the maximum sum of dependency lengths that can be reached is

$$D_b = \max_{1 \le v \le n-1} \left(D_b(v) \right). \tag{B17}$$

By the hypothesis of induction, $D_b(v)$

$$D_b(v) = n - 1 + \Delta(v) + \Delta(n - v) = v^2 - nv + \frac{n(n+1)}{2} - 1. \tag{B18}$$

If is easy to show that construction a) produces smaller sums of arc lengths than construction b) because

$$D_b(v) = D_a(v) + v - 1. \tag{B19}$$

for $2 \le v \le n - 1$ and then $D_b(v) > D_a(v)$ within that range of v.

Using $dD_b(v)/dv = 2v - n = 0$ it is easy to see that $D_b(v)$ has only one critical point within the interval $[1, n - 1]$, i.e. $v = n/2$. As $d^2D_b(v)/dv = 2 > 0$, $D_b(v)$ has a minimum at $v = n/2$ and therefore $D_b(1)$ and $D_b(n - 1)$ are equal maxima within that interval (by symmetry, $D_b(1) = D_b(n - 1)$, recall Eq. B18). Therefore the maximum D is

$$D_b(1) = D_b(n-1) = \frac{n(n-1)}{2} = \Delta(n) \tag{B20}$$

as we wanted to prove.

B.3. The minimum mean dependency length.

We aim to find a lower bound for $\langle d \rangle$ given the degree of each vertex. τ_i is defined as the sum of the lengths of the links formed with the i-th vertex. $\langle d \rangle$ can be written in terms of τ_i, i.e.

$$\langle d \rangle = \frac{1}{2(n-1)} \sum_{i=1}^{n} \tau_i . \tag{B21}$$

k_i is defined as the degree of the i-th vertex. We aim to find the minimum value of τ_i. This is equivalent to finding the minimum value of $\langle d \rangle$ for the star tree of $n = k_i + 1$ vertices defined by the i-th vertex and its k_i adjacent vertices (notice $\langle d \rangle = \tau_i/(k_i + 1)$ in that case).

If k_i is an even number, the minimum τ_i is obtained by placing $k_i/2$ of the adjacent vertices immediately before vertex i and $k_i/2$ of the remaining vertices immediately after, that is,

$$\tau_i \ge 2 \sum_{j=1}^{\frac{k_i}{2}} j . \tag{B22}$$

If k_i is an odd number, the minimum τ_i is obtained by placing $k_i/2+1$ of the adjacent vertices immediately before vertex *i* and $k_i/2$ of the remaining adjacent vertices immediately after it or by the symmetric configuration (i.e. placing $k_i/2$ of the adjacent vertices immediately after vertex *i* and $k_i/2+1$ of the remaining adjacent vertices immediately after it). Therefore,

$$\tau_i \geq 2\sum_{j=1}^{\frac{k_i-1}{2}} j + \frac{k_i+1}{2}. \tag{B23}$$

Merging Eqs. B22 and B23, one obtains

$$\tau_i \geq 2\sum_{j=1}^{\left\lfloor \frac{k_i}{2} \right\rfloor} j + \frac{k_i+1}{2}(k_i \bmod 2), \tag{B24}$$

being *x* mod *y* is the modulus of the division of *x* by *y*.

It is easy to see that this kind of arrangement of adjacent vertices around the *i*-th vertex is optimal (minimizes τ_i). If the *i*-th vertex is placed at position π, the nearest placements for an adjacent vertex are either positions π - 1 or π+1. If these two positions are already taken by adjacent vertices, the nearest positions available are π-2 and π+2, and so on.

Replacing Eq. B24 into Eq. B21, one gets

$$\langle d \rangle_{\min} \geq \frac{1}{2(n-1)}\sum_{i=1}^{n}\left(\left\lfloor \frac{k_i}{2} \right\rfloor \left(\left\lfloor \frac{k_i}{2} \right\rfloor + 1 \right) + \frac{k_i+1}{2}(k_i \bmod 2) \right). \tag{B25}$$

A lower bound of $< d >_{\min}$ that is simpler than that of Eq. B25 can be obtained. When k_i is even, Eq. B22 is equivalent to

$$\tau_i \geq \frac{k_i}{2}\left(\frac{k_i}{2} + 1 \right) = \frac{k_i^2}{4} + \frac{k_i}{2}. \tag{B26}$$

When k_i is odd. Eq. B23 is equivalent to

$$\tau_i \geq \left(\frac{k_i+1}{2} \right)^2 = \frac{k_i^2}{4} + \frac{k_i}{2} + \frac{1}{4}. \tag{B27}$$

Regardless of whether k_i is even or not, τ_i is bounded below by Eq. B26 and then Eq. B21 becomes

$$\langle d \rangle_{\min} \geq \frac{1}{2(n-1)}\sum_{i=1}^{n}\left(\frac{k_i^2}{4} + \frac{k_i}{2} \right). \tag{B28}$$

After some algebra, one obtains

$$\langle d\rangle_{\min} \geq \frac{n}{4(n-1)}\left(\frac{\langle k^2\rangle}{2} + \langle k\rangle\right). \tag{B29}$$

Replacing $\langle k\rangle$=2-2/*n* (Eq. 4), into Eq. B29 it is obtained finally

$$\langle d\rangle_{\min} \geq \frac{n\langle k^2\rangle}{8(n-1)} + \frac{1}{2}. \tag{B30}$$

If we consider a linear tree, there are *n*-2 vertices where k_i=2 and 2 vertices where k_i = 1, so Eq. B25 gives $\langle d\rangle_{\min}$ = 1, which is indeed the actual minimum for this kind of tree. We could also consider a star tree, where all vertices have k_i = 1 except the hub, which has k_i = *n* - 1. It is tempting to use Eq. B25 to bound $\langle d\rangle_{\min}$ below but the contribution of vertices of degree 1 will be underestimated. For this reason, it is convenient to consider

$$\langle d\rangle_{\min} = \tau_h/(n - 1), \tag{B31}$$

where τ_h is the true minimum value of τ_i that the hub can achieve. Eqs. B26 and B27 indicate that

$$\tau_h = \frac{n^2}{4}. \tag{B32}$$

if *n* is even (the hub has ood degree) and

$$\tau_h = \frac{n-1}{2}\left(\frac{n-1}{2} + 1\right). \tag{B33}$$

if *n* is odd (the hub has even degree). Applying Eqs. B32 and B33 to Eq. B31, it is obtained that a star tree has

$$\langle d\rangle_{\min} = \frac{n^2}{4(n-1)} \tag{B34}$$

if *n* is even and

$$\langle d\rangle_{\min} = \frac{n+1}{4} \tag{B35}$$

if *n* is odd.

APPENDIX C: CROSSING THEORY

We aim to bound above *C*, the number of link crossings. *C*=0 for *n*≤3 (if *n*≤2, the number of edges does not exceed 1 and thus crossings are impossible; if *n*=3, the two edges cannot cross as they have a vertex in common). Hereafter, $n>3$ is assumed. We do not aim to calculate C_{max}, the actual maximum number of crossings that a sentence can reach, but upper bounds of C_{max}.

C.1. A simple upper bound for the number of crossings.

If a sentence has n vertices, then C_{max} cannot exceed the number of different pairs of edges, i.e.

$$C_{\max} \leq \binom{n-1}{2} = \frac{(n-1)(n-2)}{2} \tag{C1}$$

for $n \geq 2$.

C.2. Upper bounds of the number of crossings from dependency lengths.

Since no crossing can be formed with edges of length 1 or n - 1, the actual number of edges that can be involved in a crossing is n - 1 - $N_e(1)$ - $N_e(n$ - 1) where $N_e(d)$ here is the actual number of edges whose length is d. Thus,

$$C_{\max} \leq \binom{n-1-N_e(1)-N_e(n-1)}{2}. \tag{C2}$$

Configurations where crossings are impossible can be derived imposing that the number of edges that can cross is at most 1, i.e.

$$n-1-N_e(1)-N_e(n-1) \leq 1, \tag{C3}$$

which means that crossings are impossible if (a) there is no arc of maximum length ($N_e(n$ - 1) = 0) and at most one arc has a length different than 1 ($n – 2 \leq N_e(1) \leq n$ - 1) or (b) there is an arc of maximum length ($N_e(n$ - 1) = 1) and at most one arc with a length between 1 and n - 1 ($n – 3 \leq N_e(1) \leq n$ - 2).

Upper bounds of C_{max} can be derived involving the length of each arc. Knowing that d - 1 is the number of vertices under an arc and $n – d$ - 1 is the number of vertices “off the arc”, the number of crossings with different arcs in which an arc of length d can be involved cannot exceed $c(d)=(d-1)(n-d-1)$. Notice that $c(d)$ could exceed n - 2, the maximum number of crossings in which an arc can be involved (e.g., take d=3 and n > 2), but $c(d)$ is exact when d = 1 or $d = n – 1$ ($c(d)$=0 in both cases). If d_i is the length of the i-th arc, one can write

$$C_{\max} \leq \frac{1}{2}\sum_{i=1}^{n-1} c(d_i), \tag{C4}$$

which applying $c(d)=nd-d^2-n+1$ becomes

$$C_{\max} \leq \frac{1}{2}\left(n\sum_{i=1}^{n-1} d_i - \sum_{i=1}^{n-1} d_i^{\,2} - (n-1)^2\right) \tag{C5}$$

and finally

$$C_{\max} \leq \frac{n-1}{2}\left(n\langle d\rangle - \langle d^2\rangle - n + 1\right). \tag{C6}$$

C.3. Upper bounds of the number of crossings from vertex degrees.

Upper bounds for C_{max} based on the structure of the tree will be derived next. It is convenient to write C as a function of the adjacency matrix $A = \{a_{ij}\}$,

$$C = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} C_{ij}\ , \tag{C7}$$

where C_{ij} is the number of crossings in which the pair of vertices (i,j) is involved (C_{ij}=0 if a_{ij}=0). Notice that the definition of link crossing given in Section 1 implies that an edge connecting the pair of vertices (i,j) cannot cross any edge formed with either i or j (including the edge under consideration itself). Thus the edge formed by the pair of vertices (i,j) cannot cross any of the $k_i + k_j - 1$ edges (being k_i the degree of the i-th vertex) formed involving vertex i or vertex j. The number of edges that can be crossed by the edge formed by (i,j) is thus $(n - 1) - (k_i + k_j - 1) = n - k_i - k_j$. Thus, $C_{ij} \leq n - k_i - k_j$. C_{pairs} is defined as the number of different edge pairs that can cross. Replacing C_{ij} by its upper bound, i.e., $n - k_i - k_j$, in Eq. C7, it is obtained

$$C \leq C_{pairs} = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(n - k_i - k_j). \tag{C8}$$

The previous Eq. gives after some work

$$C_{pairs} = \frac{1}{2}\left(n(n-1) - \sum_{i=1}^{n} k_i^{\ 2}\right) \tag{C9}$$

and finally

$$C_{pairs} = \frac{n}{2}\left(n - 1 - \langle k^2\rangle\right). \tag{C10}$$

Knowing that $\langle k^2\rangle = n-1$ in a star graph, Eq. C10 means that a star graph cannot have crossings (C=0) regardless of how its vertices are arranged linearly as $0 \leq C \leq C_{pairs} \leq 0$ in that case. A linear tree, which has minimum $\langle k^2\rangle$ (Appendix A), transforms Eq. C10 with $\langle k^2\rangle = 4 - 6/n$ into

$$C_{pairs} = \frac{n(n-5)}{2} + 3. \tag{C11}$$

Glottometrics 25, 2013, 22-29

Zur Verslänge im Altisländischen

Karl-Heinz Best

Abstract. In this contribution the distribution of word numbers in poetic texts in the Old Icelandic *Edda* is tested. The displaced binomial distribution seems to be the best model. But there are five cases in which the empirical findings deviate from this model. Four times other models could be fitted successfully.

Keywords: verse length, poetry, Old Icelandic

1. Zum Thema

Zwei Untersuchungen zum Deutschen waren der Hypothese gewidmet, dass die Häufigkeit, mit der Verslängen in Texten erscheinen, einem Sprachgesetz folgen. Die Theorie dazu wurde den Arbeiten von Altmann (1988a, b), Wimmer, Altmann u.a. (1994) sowie Wimmer & Altmann (1996) entnommen. Es hat sich herausgestellt, dass von den Verteilungen, die hierfür in Frage kommen, die verschobene Binomialverteilung am ehesten geeignet erscheint. Dies gilt auch für einen französischen Text, dessen Daten bei Muller (1972) zu finden sind.

Im vorliegenden Beitrag geht es nun darum, eine Überprüfung der gleichen Hypothese am Beispiel einer anderen Sprache fortzusetzen. Es handelt sich dabei um altisländische Lieder, die in der *Edda* enthalten sind. Es wurden nur solche Texte ausgewählt, die nur oder fast nur aus Verszeilen bestehen; Lieder mit größeren Prosaanteilen wurden nicht berücksichtigt. Die Lieder haben eine unterschiedlich lange mündliche Tradition hinter sich, bevor sie schriftlich aufgezeichnet wurden. Ihre Textform ist stellenweise problematisch, worauf der erste Herausgeber der benutzten Ausgabe hinweist: „Tiefer liegende verderbnisse der schriftlichen überlieferung und vollends alle störungen, die der mündlichen zeit zuzutrauen sind, habe ich grundsätzlich nicht angerührt...“ (Neckel, Gustav, Vorwort der ersten Auflage, in: *Edda* 1962, S. VI). Neckel verweist an gleicher Stelle außerdem auf „emendationen der herausgeber und kritiker“ sowie „konjekturen“. Hinzu kommen, wenn auch nur vereinzelt, Textlücken. Die Textgestalt der Lieder ist also alles andere als gesichert.

2. Bearbeitung der Texte

Für die Bearbeitung der Texte galten folgende Prinzipien: Das „Wort“ wird als ununterbrochene Graphemkette definiert; Bindestriche werden als Schriftzeichen gewertet, die Graphemketten zu einem Wort vereinen. Die Verszeile ergibt sich aus dem Druckbild der Gedichte. Es wurde immer das ganze Lied ausgewertet, aber ohne die Überschrift (also nur der laufende Text) und ohne die gelegentlichen Prosaeinleitungen und -enden.

3. Zur Frage nach einem Modell für die Verslängenverteilung

Anknüpfend an Best (2012a,b) wurde die Hypothese geprüft, dass auch im Fall der altisländischen Lieder der *Edda* die Binomialverteilung, definiert als

$$P_x = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots n \ ,$$

sich als geeignetes Modell für die Verteilung der Wörter auf Verszeilen erweisen wird. Die Binomialverteilung ist hier in der unverschobenen Form angegeben, muss aber immer in verschobener Form angewendet werden, da es keine Verse mit null Wörtern gibt. Allerdings kann auch keine einheitliche Form der verschobenen Binomialverteilung angegeben werden, da die kürzesten Verslängen in den Liedern unterschiedlich ausfallen. In zwei Fällen wurde statt der Binomialverteilung die verwandte Hyperbinomialverteilung

$$P_x = \frac{\binom{n}{x}}{\binom{m+x-1}{x}} q^x P_o, \quad x = 0, 1, \ldots n$$

mit $P_0 = [\,_2F_1(-n, 1; m; -q)]^{-1}$, an die Daten angepasst, wenn die Binomialverteilung keine akzeptablen Ergebnisse erbrachte. (Zu beiden Verteilungen und ihren Zusammenhang wird auf die entsprechenden Kapitel in Wimmer & Altmann 1999 verwiesen.)

Die Ergebnisse finden sich im folgenden Abschnitt 4.

4. Anpassung der verschobenen Binomialverteilung an die Gedichtdateien

Die Ergebnisse der Anpassung der Binomialverteilung und der Hyperbinomialverteilung an die Lieder der *Edda* finden sich in den folgenden Tabellen. Die Anpassungen wurden mit einer geeigneten Software, dem *Altmann-Fitter* (1997), durchgeführt.

In den Tabellen sind folgende Angaben enthalten: *x* ist die Zahl der Wörter pro Verszeile, n_x die Zahl der Verszeilen mit *x* Wörtern, NP_x die aufgrund der Binomialverteilung zu erwartende Anzahl der Verszeilen mit *x* Wörtern; *n* und *p* sind die Parameter der Binomialverteilung, *n*, *m* und *q* die der Hyperbinomialverteilung; X^2 ist das Chiquadrat, *P* die Überschreitungswahrscheinlichkeit für das berechnete Chiquadrat; *FG* gibt die Zahl der Freiheitsgrade an. Eine Anpassung mit $P \geq 0.05$ gilt als zufriedenstellend; Ergebnisse mit $0.05 \geq P > 0.01$ gelten nicht als zufriedenstellend, aber auch nicht als völlig misslungen. Diese Bedingungen sind in 17 von 20 Fällen erfüllt.

Die beiden Verteilungen werden, wie bereits erwähnt, in verschobener Form angepasst, da kein Vers mit $x = 0$ Wörtern existiert. In den angegebenen Formeln muss dazu lediglich statt *x* nun bei 1-verschobener Form, nämlich dann, wenn ein Vers nur ein Wort enthält, $x - 1$ gesetzt werden, bei 2-verschobener Form $x - 2$, wenn die Datei mit $x = 2$ beginnt, etc.

Nun die Ergebnisse der Anpassung der beiden Verteilungen an die Lieder der *Edda*; bei den Liedern Nr. 1 und Nr. 4 wurde die verschobene Hyperbinomialverteilung, in allen anderen Fällen die verschobene Binomialverteilung eingesetzt:

	1. *Vọlospá*, 1-15 *(Der Seherin Gesicht)*		2. *Grímnismál*, 57-68 *(Das Grimnirlied)*		3. *Hymisqviđa*, 88-95 *(Das Hymirlied)*		4. *Þrymsqviđa*, 111-115 *(Das Thrymlied)*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2			8	10.46				
3	1	1.08	53	32.61	1	5.62		
4	37	39.96	41	50.82	31	26.24	4	3.88
5	81	84.70	32	52.79	50	49.01	47	45.64
6	99	81.17	44	41.10	43	45.77	56	47.80
7	34	44.95	29	25.60	23	21.37	13	23.06
8	16	15.64	24	13.28	4	3.99	7	6.47
9	3	3.50	5	9.33			1	1.14
n =			3106		5			
p =			0.0010		0.4829			
n =	9						9	
m =	0.0689						0.1114	
q =	0.2832						0.1455	
X^2 =	7.048		34.727		4.973		5.899	
FG =	3		5		3		2	
P =	0.07		0.00		0.17		0.05	

Die Angaben im Kopf der Tabellen (Titel der Lieder mit Seitenangabe) beziehen sich auf die Ausgabe: *Edda. Die Lieder des Codex Regius nebst verwandten Denkmälern.* Herausgegeben von Gustav Neckel. I. Text. Vierte, umgearbeitete Auflage von Hans Kuhn. Heidelberg: Carl Winter Universitätsverlag 1962. Die deutschen Titel (in Klammern) sind der Ausgabe *Edda* (1963a,b) entnommen.

Anmerkung zu *Grímnismál*: das Lied besteht aus einem Wechsel von Lang- und Kurzzeilen; entsprechend zeigt die Verteilung der Wörter je Verszeile zwei Gipfel bei den 3- und den 6-Wort-Zeilen. Es scheint keine Verteilung zu geben, die man an diese Datei anpassen kann.

Zur Veranschaulichung diene die Graphik der Anpassung der verschobenen Hyperbinomialverteilung an Lied 1, *Vọlospá*, vgl. Abb. 1.

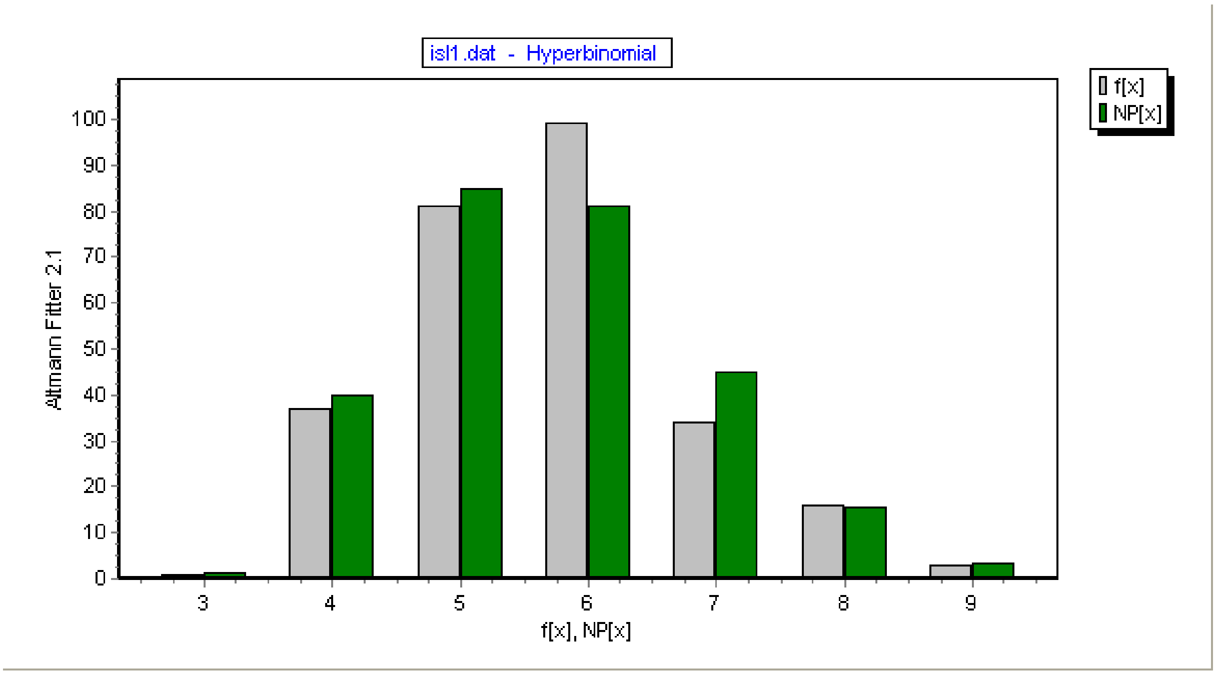

Abb. 1. Anpassung der Hyperbinomialverteilung an Lied 1, *Vǫlospá*

	5. *Helgaqviđa Hundingsbana in fyrri*, 130-139 *(Das jüngere Lied von Helgi dem Hundingstöter)*		6. *Grípisspá*, 164-172 *(Gripirs Weissagung)*		7. *Brot af Sigurđarqviđo*, 198-201 *(Das alte Sigurdlied)*		8. *Guđrúnarqviđa (in fyrsta)*, 202-206 *(Gudruns Gattenklage)*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
3	4	9.17					1	3.19
4	37	41.26	16	20.96	8	10.34	21	16.32
5	85	74.26	69	61.67	29	26.52	28	33.37
6	76	66.83	76	72.59	27	25.50	36	34.11
7	22	30.07	37	42.73	5	10.90|	20	17.44
8	3	5.41	11	12.57	6	1.75|	2	3.57
9			3	1.48				
n =	5		5		4		5	
p =	0.4737		0.3705		0.3906		0.5055	
X^2 =	9.407		4.728		1.065		4.880	
FG =	3		3		1		3	
P =	0.02		0.19		0.30		0.18	

Zu Lied Nr. 5: Eine zufriedenstellende Anpassung ist mit der erweiterten positiven Binomialverteilung mit *P* = 0.06 möglich.

Zu Lied Nr. 7: Die senkrechten Striche in der Datei zeigen eine Zusammenfassung der betreffenden Längenklassen an; dies gilt auch für die folgenden Tabellen.

	9. *Sigurđarqviđa in scamma,* 207-218 *(Das jüngere Sigurdlied; Das kurze Sigurdlied)*		10. *Guđrúnarqviđa (ọnnor),* 224-231 *(Gudruns Lebenslauf)*		11. *Guđrúnarqviđa (in Þriđia),* 232-233 *(Gudruns Gottesurteil)*		12. *Oddrúnargrátr,* 234-239 *(Oddruns Klage)*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
3	5	7.21						
4	32	39.06	33	27.54	3	3.69	9	12.93
5	98	84.63	49	58.28	12	11.12	41	37.13
6	91	91.68	53	52.87	13	13.40	46	42.63
7	48	49.66	35	26.64	9	8.07	21	24.48
8	9	10.76	4	8.05	1	2.43|	6	7.03|
9			1	1.61	1	0.29|	2	0.81|
n =	5		7		5		5	
p =	0.5200		0.2322		0.3760		0.3647	
X^2 =	4.415		7.461		0.511		2.364	
FG =	3		3		2		2	
P =	0.22		0.06		0.77		0.31	

	13. *Atlaqviđa in Grœnlenzca,* 240-247 *(Das alte Atlilied)*		14. *Atlamál in Grœnlenzco,* 248-263 *(Das grönländische Atlilied)*		15. *Guđrúnarhvọt,* 264-268 *(Gudruns Sterbelied)*		16. *Hamđismál,* 269-274 *(Das alte Hamdirlied*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2	1	0.65|						
3	0	4.88|			3	2.80	1	1.98
4	14	16.49	3	5.23	14	13.84	8	10.07
5	50	33.02	26	32.77	28	27.35	24	22.80
6	33	43.40	83	85.58	27	27.02	38	30.12
7	34	39.10	139	119.20	11	13.35	19	25.57
8	30	24.47	99	93.39	4	2.64	16	14.48
9	9	10.50	30	39.02			4	5.46
10	3	2.96|	2	6.79			0	1.33|
11	2	0.53|					2	0.20|
n =	10		6		5		9	
p =	0.4289		0.5109		0.4970		0.3615	
X^2 =	18.092		11.521		1.148		5.424	
FG =	5		4		3		5	
P =	0.00		0.02		0.77		0.37	

Zu Lied 14: Eine befriedigende Anpassung der Cohen-Binomialverteilung ist mit P = 0.70 möglich.

Anmerkung zu Text 15, *Guđrúnarhvọt*: Der Text ist an zwei Stellen unvollständig.

	17. *Baldrs Draumar*, 277-279 *(Balders Träume)*		18. *RígsÞula*, 280-287 *(Das Merkgedicht von Rig)*		19. *Hyndlolióđ*, 288-296 *(Das Hyndlalied)*		20. *Grottasọngr*, 297-301 *(Das Mühlenlied)*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2			1	0.88|				
3			8	7.61|				
4	2	3.31	18	27.34	29	27.05	15	18.46
5	11	13.25	72	52.41	62	63.28	40	33.74
6	28	19.88	51	56.50	57	61.68	22	25.69
7	9	13.25	22	32.48	40	32.06	9	10.43
8	3	3.31	13	7.78	6	9.37	5	2.69
9					1	1.56		
n =	4		6		6		6	
p =	0.5000		0.5897		0.2805		0.2335	
X^2 =	5.616		17.967		3.901		4.528	
FG =	2		3		3		2	
P =	0.06		0.00		0.27		0.10	

Die Graphik zu Lied 19, *Hyndlolióđ*, in Abb. 2 sieht wie folgt aus:

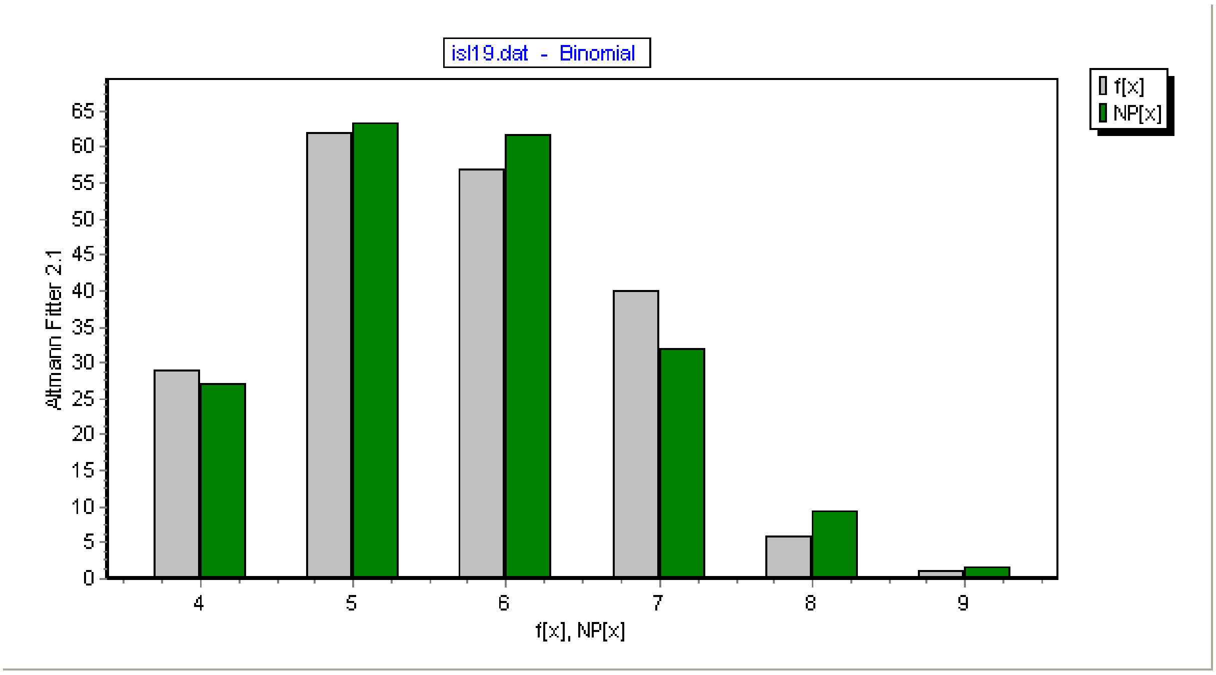

Abb. 2. Anpassung der Binomialverteilung an das Lied 19, *Hyndlolióđ*

5. Ergebnis und Perspektive

Die folgende Tabelle gibt eine Übersicht über die Ergebnisse der Anpassung der verschobenen Binomialverteilung an die altisländischen Lieder:

Lied	*P*	Lied	*P*	Lied	*P*	Lied	*P*
1	0.07*	6	0.19	11	0.77	16	0.37
2	0.00	7	0.30	12	0.31	17	0.06
3	0.17	8	0.18	13	0.00	18	0.00
4	0.05*	9	0.22	14	0.02	19	0.27
5	0.02	10	0.06	15	0.77	20	0.10

* Anpassung der Hyperbinomial-Verteilung

Es ist zu konstatieren, dass von den 20 Liedern der *Edda* 15 der verschobenen Binomialverteilung unterliegen; in zwei dieser Fälle ist das Ergebnis nicht wirklich befriedigend, muss aber auch nicht ganz verworfen werden. Bei zwei weiteren Liedern bewährt sich die Hyperbinomialverteilung als Modell. Für insgesamt 17 von 20 Liedern ist also eine Verteilung gefunden worden, der die Verslängen folgen. Die Tendenz ist damit sehr deutlich und stimmt mit der für deutsche Texte gefundenen weitgehend überein.

An Lied 2, *Grímnismál,* kann man keine der beiden Verteilungen anpassen; der Grund dürfte darin zu suchen sein, dass in diesem Lied ein systematischer Wechsel zwischen Lang- und Kurzzeilen enthalten ist, was zu zwei Häufigkeitsgipfeln führt und den hier verwendeten Verteilungen und anscheinend allen anderen, die sonst in Frage kommen, widerspricht. (Dies ist der Grund, weshalb auch ein anderes Lied der Sammlung, *Alvísmál*, bei den Auswertungen nicht berücksichtigt wurde.) Womöglich müssten Lang- und Kurzzeilen in solchen Fällen getrennt ausgewertet werden.

In zwei weiteren Fällen, den Liedern 13 und 18, lässt sich keine der beiden Verteilungen erfolgreich anpassen; dies gilt auch für alle anderen der 198 Verteilungen, die die benutzte Software, der *Altmann-Fitter* (1997), bearbeitet. Einen Grund für diese Abweichung anzugeben fällt schwer. Vielleicht spielt die besondere literarische Tradition dieser Lieder, auf die eingangs bereits hingewiesen wurde, dabei eine Rolle. Schaut man sich die Dateien der beiden Texte an, so wird außerdem schnell deutlich, dass in diesen beiden Fällen eine auffällige Bevorzugung einer bestimmten Verslänge zu erkennen ist: beide Male sind es die 5-Wort-Zeilen, die in ungewöhnlicher Häufigkeit auftreten.

Die Befunde bestätigen die Ergebnisse der Untersuchungen von Best (2011a,b) weitgehend. Die Binomialverteilung ist bisher am besten geeignet, die Verslängenverteilungen zu erfassen. Die Hyperbinomialverteilung muss jedoch ebenso in Betracht gezogen werden wie die erweiterte positive Binomialverteilung, die beide in Einzelfällen bessere Anpassungen ermöglichen. Es bleibt jedoch weiterhin die Frage, welche Rolle diese drei (und womöglich noch weitere) Verteilungen spielen werden, ob es sich also bei ihnen um Formen des Verteilungsgesetzes handelt, die sich bei beliebigen Verstexten bewähren. Denkbar wäre auch, dass sich außer den altisländischen Liedern 13 und 18 weitere Texte finden lassen, bei denen keine der bisher vorgeschlagenen Verteilungen als Modell dienen kann.

Bestimmt man die Verslänge anders, als es hier erfolgt ist, also durch die Zahl der Buchstaben, Laute, Moren, Morphe, Phoneme, Silben oder Versfüße, muss in jedem dieser Fälle damit gerechnet werden, dass das Sprachgesetz, das die Verteilung sprachlicher Einheiten regelt, andere Formen annimmt. Jede Ebene bedeutet andere Randbedingungen.

Als Ergebnis kann konstatiert werden, dass die Häufigkeit der Verslängen nicht chaotisch ist, sondern von einem Sprachgesetz gesteuert wird, das mit der Theorie der Satz- oder Wortlängen (und anderer sprachlicher Einheiten) übereinstimmt.

Literatur

Altmann, Gabriel (1988a). Verteilungen der Satzlängen. In: Schulz, Klaus-Peter (ed.), *Glottometrika 9* (S. 147-169). Bochum: Brockmeyer.

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

Best, Karl-Heinz (2012a). How many words are in a verse? An exploration. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G., (eds.), *Synergetic linguistics. Text and language as dynamic systems: 13-22.* Wien: Praesens.

Best, Karl-Heinz (2012b). Zur Verslänge bei G. A. Bürger. *Glottometrics 23, 56-61.*

Edda. Die Lieder des Codex Regius nebst verwandten Denkmälern. Herausgegeben von Gustav Neckel. I. Text. Vierte, umgearbeitete Auflage von Hans Kuhn. Heidelberg: Carl Winter Universitätsverlag 1962.

Edda (1963a). Übertragen von Felix Genzmer. Erster Band: *Heldendichtung.* Einleitungen und Anmerkungen von Andreas Heusler und Felix Genzmer. Revidierte Neuausgabe. Zweite Auflage. Düsseldorf, Köln: Eugen Diederichs Verlag.

Edda (1963b). Übertragen von Felix Genzmer. Zweiter Band: *Götterdichtung und Spruchdichtung.* Einleitungen und Anmerkungen von Andreas Heusler. Neuausgabe. Düsseldorf, Köln: Eugen Diederichs Verlag.

Muller, Charles (1972). *Einführung in die Sprachstatistik*. München: Hueber.

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

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

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

Software

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

Glottometrics 25, 2013, 30 - 42

Laws governing rank frequency and stratification in English texts

Róisín Knight, Lancaster University

Abstract. There are several laws that attempt to capture the regularities that seem to exist in the frequency structure of texts, by expressing the relationship between frequency and rank of words in a text. Within this field of research it has been found that stratification exists on many different levels, and the hypothesis proposed by Popescu, Altmann and Köhler (2010) allows for this to be explored further. This paper will use the method suggested by Popescu, Čech and Altmann (2011), to consider the presence of stratification in a new data set of English texts. Due to the fact that the study of this topic is a relatively new pursuit within linguistics, there is much confusion surrounding the question of what specific linguistic factors cause stratification. This paper attempts to answer this question, and tests Popescu, Mačutek and Altmann's (2009) theory that the number of strata in a text relates to the number of actors. However the results show that none of the texts studied, either containing a single actor or multiple actors, were found to be monostratal. Therefore the cause of stratification in texts is currently unknown, and until the mathematical representations of strata are able to shed light on this their application is limited.

Keywords: Rank-frequency distribution, stratification, English texts, PAK curve, Zipf

1. Introduction

Arguably the most influential work within the field of universal laws was carried out by George Zipf. Whilst he was not the first person to detect regularities in the frequency structure of texts, Altmann (2002: 25) proposes that his contribution to the field rivalled that of Newton to physics. Zipf (1935) systematically investigated the relationship between the frequency and frequency rank of words. As well as studying several languages, he also considered symbol frequencies at lower hierarchies of language, for example syllables and morphemes (Rousseau, 2002: 16). In all he found a stable relationship between rank and frequency, which he expressed through the following prototype of a power law function (Zipf 1935: 40):

[1] $$k = ab^2,$$

where *k* = constant, *a* = frequency, *b* = rank.

Zipf made several comments on his equation, attempting, to an extent, to interpret its parameters and suggest why it exists. For example, he acknowledged that values of the power not exactly equal to two do occur, with the power varying with different 'styles' and argued that the phenomenon existed due to the principle of least effort, i.e. due to the fact that people always choose the path that requires the least effort, shorter words are more common (Zipf, 1935: 222f.). It is important to note, however, that Zipf's impact extends beyond that of simply linguistics, as his law is established in many different disciplines, particularly sociology, and has been expanded to topics such as chaos theory, fractals and sand piles (Altmann, 2002: 22).

Many researchers have since built on Zipf's work, attempting to both explain it further and also find an equation that better expresses the relationship. This paper is primarily concerned with the work of Popescu, Altmann and Köhler (2010), who suggest that the following form (henceforth referred to as the PAK equation) should replace the Zipfian description:

[2] $$y = 1 + \sum_{i=1}^{k} W_i e^{\frac{-x}{v_i}}$$

where y = frequency, x = rank, k = number of terms used, W_i and v_i are parameters.

They propose that there are several advantages to this particular law. Firstly, they believe it provides a better fit. They compared their equation with the original Zipfian form for 100 texts in 20 languages and found that in the majority of cases the PAK curve provided a better fit (Popescu, Altmann and Köhler, 2010: 721). Secondly, they believe that their equation expresses the relationship in a simpler form (Popescu, Altmann and Köhler, 2010: 717-718). Thirdly, the lemmatised versions of short texts have been tested and it has been found that this does not bring new results in synthetic or analytic languages (Popescu, Čech and Altmann, 2011: 59).

Additionally, Popescu, Čech and Altmann (2011) reason that the PAK equation should replace the Zipfian form as it can account for heterogeneity within texts- a point which is central to this study. They argue that texts, partly due to characteristics of individual languages and partly due to language variability (cf. Croft, 2010), are composed of a number of components and therefore must be viewed as a mixture of statistical distributions (Popescu, Altmann and Köhler, 2010: 715). They carried out a study of 54 Slovakian poems and proposed a method for finding the number of strata present at the word-form level of a text. They defined stratification, which can be observed through the presence of strata, as the presence of different means of expression within a text (Popescu, Čech and Altmann, 2011: 54). They propose that the number of exponential components in the PAK curve signalises the number of strata and they therefore suggest the rule: 'if the constants in the exponents of two components are equal or almost equal, then one of the components is redundant and can be omitted' (Popescu, Čech and Altmann, 2011: 55). Applying this rule to the 54 poems, they found all of the texts to be monostratal.

However, whilst the PAK equation may be used to identify the level of stratification within a text, there is much confusion surrounding the question of what specific linguistic factors cause stratification. Several theories have been argued, however none are supported with evidence. Ziegler and Altmann (2003: 278) have postulated that stratification can arise automatically from the author's influence upon the text and also ad hoc from the reader, who sees the text from a certain cognitive perspective. Popescu, Mačutek and Altmann (2009: 14) have theorized that the two strata which can usually be observed represent autosemantics (i.e. content words) and synsemantics (i.e. function words). Popescu, Mačutek and Altmann (2009: 13) have also suggested that 'in a stage play there are as many parts as there are acts and as many strata as there are actors'. Popescu, Čech and Altmann's (2011) work seems to support this latter theory, as all of the texts they tested, and found to be monostratal, were written by a single author about a single topic.

When concluding their article, Popescu, Čech and Altmann (2011: 59) posed the following questions: firstly, can the same results be found in poetry of other languages? Also, can the same results be found in texts written in other languages? In this study, Popescu, Čech and Altmann's (2011) theory will be further explored through testing new data in order to answer these questions. Additionally, in order to test their theory further, texts with multiple actors will be tested, to consider whether these are multistratal.

2. **Data**

For this exploratory study, four text types were selected, with 36 texts being used in total. The data used, along with the source it was obtained from, can be observed in the table below:

Table 1
Data Used in this Study

Text Type	**Text Title**	**Text Source**
Keats Poems	Fancy	Project Guttenburg, www.gutenberg.org, accessed 10/01/12. Project Guttenburg is a source of free e-books, with all e-books having previously been published by bona fide publishers.
	Robin Hood to a Friend	
	Ode to a Grecian Urn	
	Ode to Nightingale	
	Ode to Psyche	
	To Autumn	
	To Charles Cowden Clarke	
	To George Felton Mathew	
	To Hope	
	To My Brother	
Friends Episodes	Season 1 Episode 1	www.friendscafe.org/scripts, accessed 10/01/12. Friends Cafe contains transcripts of all of the Friends episodes aired on TV. They are transcribed by fans, however the transcriptions used have been checked against the appropriate episodes, to ensure they are correct.
	Season 1 Episode 2	
	Season 1 Episode 3	
	Season 1 Episode 4	
	Season 1 Episode 7	
	Season 1 Episode 8	
	Season 1 Episode 9	
	Season 1 Episode 10	
	Season 1 Episode 11	
	Season 1 Episode 12	
Shakespeare Plays	Antony and Cleopatra	Project Guttenburg, www.gutenberg.org, accessed 10/01/12.
	As You Like It	
	Hamlet	
	Henry V	
	King Lear	
	Much Ado About Nothing	
	Measure for Measure	
	Merchant of Venice	
	Romeo and Juliet	
	Taming of the Shrew	
Conversations	Conversation 1	BNCWeb, bncweb.lancs.ac.uk, accessed 16/01/12. BNCWeb is a web-based client program for searching and retrieving data from the British National Corpus (BNC).
	Conversation 2	
	Conversation 3	
	Conversation 4	
	Conversation 5	
	Conversation 6	

It is important to note that there were two factors restricting my choice of text. Firstly, unlike Popescu, Čech and Altmann (2011) texts that contained less than 200 words were not analysed. In the cases of texts that contained multiple actors, texts were selected with the aim of ensuring that they had minimal proportions of speakers who said less than 200 words, in order to not largely skew the comparison between individual and whole texts. Other than this, selection was random. It is important to recognise that some linguists have criticised data selection from Project Guttenburg, due to the influence of the editor upon the texts (Lindquist, 2009: 22). However, it was seen as an acceptable source in regard to this study as the degree of editorial intervention at this level is often fairly minimal (Lindquist, 2009: 22) and this study only works with rank-frequency data, without reference to the actual words that were used.

3. Method

3.1. Organising the Data

When organising the Shakespeare and Friends texts, everything that was not speech was excluded from the analysis, for example stage directions and so on. Also, parts of texts that were repetitions of something another actor had already said or written were excluded. Such examples were excluded as they may have otherwise skewed the data, by not truly reflecting the actors the author is trying to create. Texts were then tested both as a whole, with all of the actors included, and in several parts, considering each actors lines individually.

When testing texts of individual actors, text that was not written in English was also excluded. This was only necessary in the cases of two actors in two of the Shakespeare plays. Including these parts of the text did not seem a particularly fair test of the theory that one actor equals one stratum, as it is added linguistic complexity.

3.2 . Fitting the PAK Equations

For each text, a complete rank-frequency list was created using the program AntConc. The default settings were used and all of the text was considered as lower case (cf. Wilson, 2009: 101 and Wilson, forthcoming). The PAK equation was then fitted, using the software R (for an introduction to the program R, see Chihara and Hesterberg, 2011). It is important to note at this juncture that the PAK equation was only tested to two terms, as research by Altmann (2008: 421) into texts from 20 languages found two components to be sufficient in capturing the rank-frequency distribution.

4. Results

The figures below show the value of the exponent of each of the first two strata given by equation [2] for each text. The results are shown in this way for ease of viewing, however the corresponding values, given to 4 decimal places, can be found in the appendix. A monostratal text would give identical values for v_1 and v_2 i.e. they would be seen to overlap.

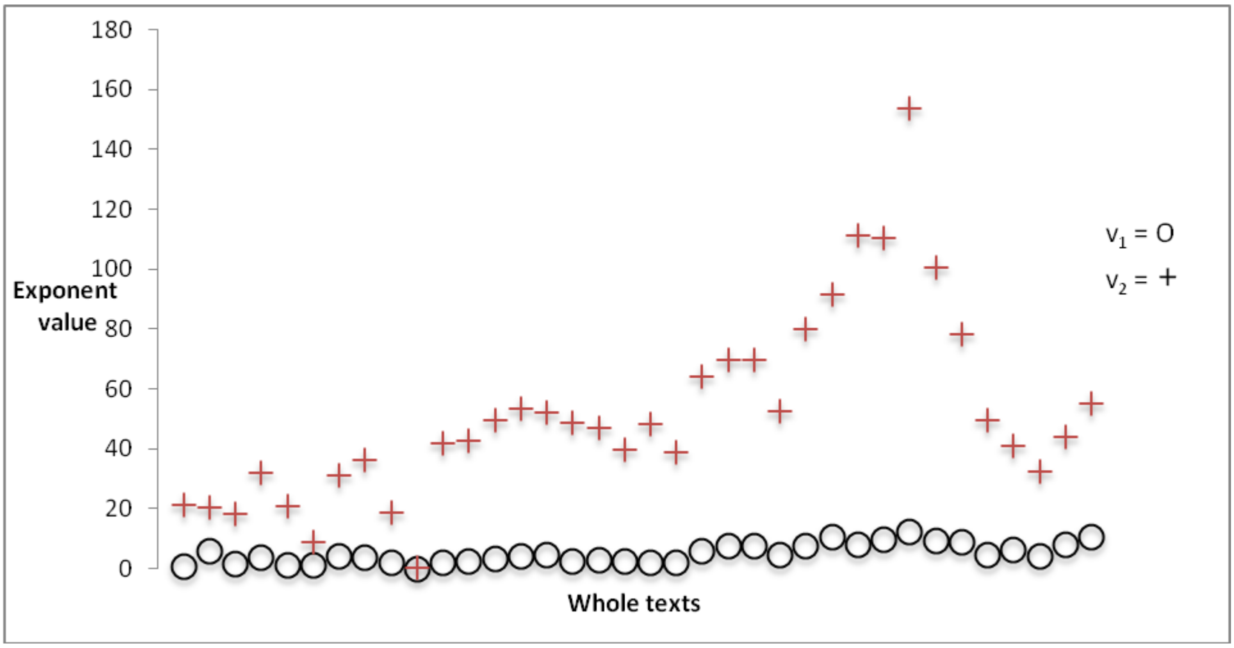

Figure 1. Exponent values showing stratification for all whole texts

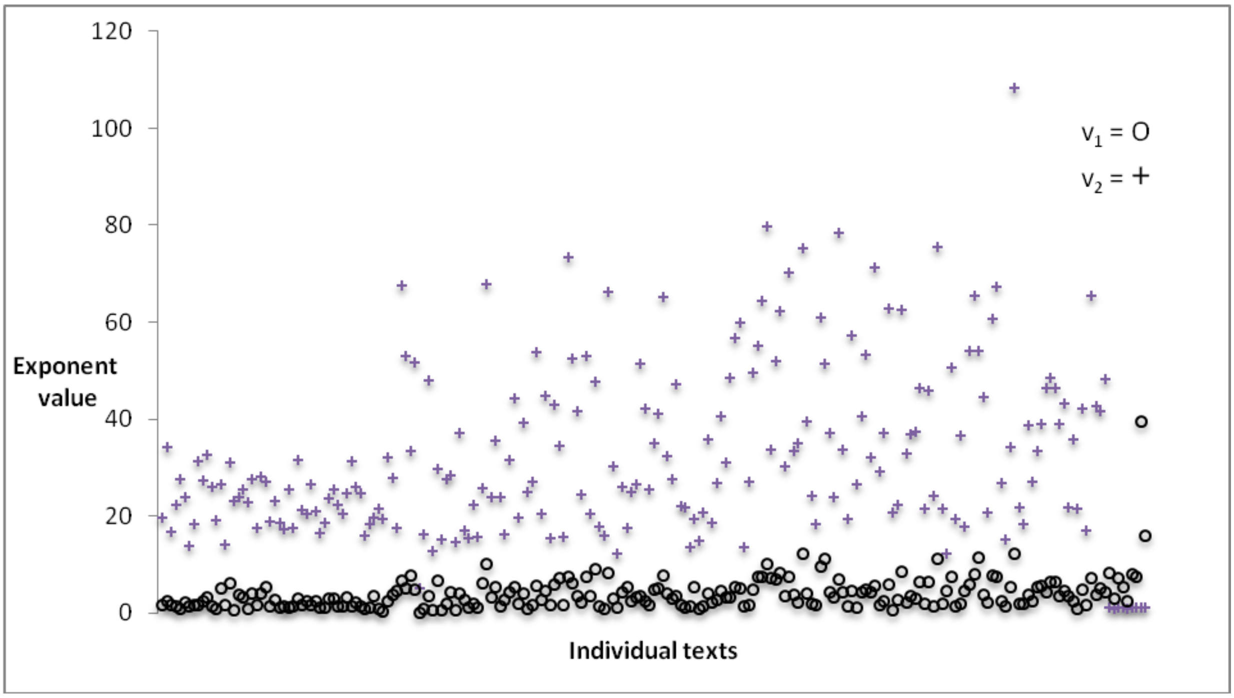

Figure 2. Exponent values showing stratification for all individual texts

5. Discussion

Figure 1 shows that all whole texts, including Keat’s poems, were found to contain more than one stratum. It is important to recognise that whilst in one instance there appears to be matching values for v_1 and v_2, in this case both the figures are very near to zero and are actually proportionally quite different. This appears to answers the questions posed by Popescu, Čech and Altmann (2011: 59), stated in the introduction, of whether their results would hold for other poets and poems in other languages, however perhaps not in the way they would have hoped. It seems clear that their findings do not similarly apply to all other poets. However forms of poetry can be very different, therefore the findings do provide new avenues to pursue. It would be interesting to consider a wider range of poets, perhaps just within the Slovakian language, to see how representative their original findings were. Additionally, it seems apparent that their results do not hold true for all poets writing in other languages. However again this is an area for further investigation - which set of results is more representative? Do different languages provide different findings?

Furthermore, the figure 2 shows that every text tested was found to contain more than one stratum. Therefore Popescu, Mačutek and Altmann’s (2009: 13) theory that each stratum corresponds to an actor in a text appears invalid. This result seems particularly surprising in the case of the individual speakers from the natural conversations. Such speakers are free from the influences of an author, whose writing may not exactly reflect that of the natural conversation they are attempting to portray. This had previously seemed to me to be the situation which was least likely to contain multiple strata.

For example, one of the conversations studied was a music lesson. It contained conversation between an instructor and a young child. The instructor took control of the conversation, directing the topic. The pupil seemed to have a minimal role, and mostly only spoke to offer feedback of agreement to his instructor. It may therefore seem more likely that this young boy’s text would be monostratal, however the was not found to be the case. The findings however suggest that the conversation is more complex than seems immediately obvious.

The fact that stratification was found to occur in natural conversations highlights that stratification is not simply due to the influence of an author in creating a situation. This perhaps hints that the occurrence of stratification is so complex it will be difficult for researchers to pin point the causes. However it may also be a source for further study, and it may be interesting to consider conversations where the participants are extremely comfortable around one another, so that any ‘agenda’ in the conversation is minimised.

The findings of this study also seem to suggest that stratification is not linked to text length. Whilst none of the texts tested were quite as small as those that Popescu, Čech and Altmann (2011) used, stratification was still found in texts that only just surpassed the 200 word limit that was imposed. However for this conclusion to be argued more comprehensively, it would be necessary for a further study test texts with the same minimum word count as Popescu, Čech and Altmann and additionally consider texts over the maximum word count tested in this study.

It seems unfortunate that the investigation found all of the texts to contain stratification, as it limits the amount that can be explored with regard to the parameters of strata in this analysis. The findings perhaps suggest that monostratal texts are not as easy to come by as Popescu, Čech and Altmann’s (2011) study suggests. However, due to the fact that both the previous research and this investigation were based on a limited selection of texts, it seems unclear what findings with regard to the frequency of monostratal texts should be expected, and further research is therefore needed on this point.

This investigation has aptly highlighted the difficulties researchers face in separating strata. Clearly, trying to establish which texts contain stratification is, at least in part, a game of trial and error. Establishing which types of texts contain stratification and which don't is a necessary step before the more specific question of how the parameters of strata vary can be fully answered. As few conclusions can be drawn about what types of texts may be monostratal, more research is needed into this area. It may be advantageous to explore several different genres of texts, to establish whether one particular style of writing is more likely to be monostratal than others. Until this is done successfully, the usefulness of the PAK equation is limited.

References

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 287-298.*

Altmann, G. (2002). Zipfian Linguistics. *Glottometrics 3, 19-26.*

Anthony, L. (2011). *AntConc* (Version 3.2.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Chihara, L. and Hesterberg, T. (2011). *Mathematical Statistics with Resampling and R.* Hoboken, NJ: Wiley.

"Friends: Kevin Bright". (2005). In: USA Today. http://www.usatoday.com/community/chat/2002-04-23-friends.htm. Retrieved 11 February 2012.

Ihaka, R. and Gentleman, R. (2011). *R* (Version 2.13.0) [Computer Software]. Auckland, New Zealand: Auckland University. Available from http://cran.r-project.org/.

Popescu, I.-I., Altmann, G. and Köhler, R. (2010). Zipf's Law: Another View. *Quality and Quantity 44(4), 713-731.*

Popescu, I.-I., Čech, R. and Altmann, G. (2011). On Stratification in Poetry. *Glottometrics 21, 54-59.*

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

Rousseau, R. (2002). George Kingsley Zipf: Life, Ideas, His Law and Informetrics. *Glottometrics 3, 11-18.*

Wilson, A. (forthcoming). *Random Categories and Zipfian Distributions*. Lancaster: Lancaster University.

Wilson, A. (2009). Vocabulary Richness and Thematic Concentration in Internet Fetish Fantasies and Literary Short Stories. *Glottotheory 2(2), 97-107.*

Zipf, G. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: M.I.T. Press. Second Edition 1968 [First Edition: Boston, Houghton-Mifflin, 1935].

Appendix

The tables below show the parameters of the first two strata given by the PAK equation. The results are shown to four decimal places, as in the work of Popescu, Čech and Altmann (2011):

Table 2
PAK parameter values for all whole texts

Text Type	Text Title	W_1	v_1	W_2	v_2
Keats	Fancy	96.3024	0.9108	10.1503	21.3670
	Ode to a Grecian Urn	7.9634	5.9327	5.7713	20.2090
	Ode to Nightingale	58.3399	1.5411	9.7880	18.2649
	Ode to Psyche	22.0114	3.8068	4.1565	32.1136
	Robin Hood to a Friend	50.3669	1.3325	5.2486	20.7412
	To Autumn	30.0822	1.0647	7.8960	8.7752
	To Charles Cowden Clarke	43.4000	3.9392	13.2708	31.1206
	To George Felton Mathew	38.3747	3.7147	6.3188	36.3109
	To Hope	10.5133	2.1752	9.0883	18.4997
	To My Brother	4.0351	0.0045	51.3482	0.1529
Friends	Season 1 Episode 1	284.0298	1.8375	47.5776	41.9167
	Season 1 Episode 2	121.3956	2.3942	36.2686	42.6731
	Season 1 Episode 3	125.7999	3.1545	31.3054	49.5574
	Season 1 Episode 4	95.9535	4.0958	26.7527	53.5521
	Season 1 Episode 7	79.6810	4.4578	22.3295	51.9668
	Season 1 Episode 8	145.7207	2.4156	30.9130	48.5522
	Season 1 Episode 9	148.7769	2.8250	31.2420	46.8643
	Season 1 Episode 10	139.7504	2.5106	40.5163	39.7148
	Season 1 Episode 11	189.0340	2.1030	35.0872	48.4428
	Season 1 Episode 12	143.6395	2.0364	38.1832	38.7845
Shakespeare	Antony and Cleopatra	636.6304	5.7545	223.2602	64.0634
	As You Like It	644.0289	7.4389	162.1732	69.8178
	Hamlet	368.0557	7.4561	79.2392	73.4126
	Henry V	972.0994	4.3736	242.1001	52.7452
	King Lear	678.5434	7.4612	163.2451	79.8722
	Measure for Measure	559.7344	10.5829	116.7651	91.6614
	Merchant of Venice	621.8242	7.7950	97.0473	111.0429
	Much Ado About Nothing	558.5773	9.7449	96.9707	110.4711
	Romeo and Juliet	508.0405	12.4186	79.6042	153.7024
	Taming of the Shrew	524.1986	9.1031	102.7972	100.4950
Conversations	Conversation 1	342.4882	8.8963	53.1287	78.2716
	Conversation 2	275.2425	4.5877	63.6365	49.4924
	Conversation 3	34.5035	6.0852	13.9995	40.7916
	Conversation 4	33.4622	4.0780	6.3038	32.5677
	Conversation 5	30.1265	7.9091	8.8473	43.8979
	Conversation 6	73.0797	10.4138	18.5184	54.9996

Table 3
PAK parameter values for all single-actor texts

Text Type	**Text Title**	**Actor**	$\mathbf{W_1}$	$\mathbf{v_1}$	$\mathbf{W_2}$	$\mathbf{v_2}$
Friends	Episode 1	Chandler	33.1355	1.5709	8.9398	19.5350
		Monica	39.4359	2.3823	10.1485	34.1325
		Phoebe	23.0869	1.5919	6.1870	16.6741
		Ross	45.8004	1.3432	10.6487	22.2958
		Rachel	210.3814	0.7394	15.8825	27.5641
		Joey	30.8900	2.1146	8.5626	23.7561
	Episode 2	Chandler	13.8339	1.4793	6.3956	13.7127
		Monica	10.3350	1.6870	6.1480	18.2170
		Ross	38.9061	1.6928	13.6491	31.3677
		Rachel	32.0456	2.4098	6.2552	27.2671
	Episode 3	Chandler	23.1323	3.1267	7.1853	32.4952
		Monica	43.2155	1.3566	10.9950	25.9324
		Phoebe	64.6008	0.8946	13.1988	19.1362
		Ross	5.0743	4.9971	5.7522	26.4463
		Rachel	8.4232	1.6174	6.5546	13.9648
	Episode 4	Chandler	10.4724	6.1996	3.6767	31.1318
		Monica	76.2925	0.6478	7.9523	23.0236
		Phoebe	10.7299	3.8601	5.4351	23.8762
		Ross	13.2797	3.2608	6.8003	25.3525
		Rachel	117.9950	0.7248	12.8103	22.9227
	Episode 7	Chandler	9.5144	4.0377	7.2565	27.7024
		Phoebe	16.1442	1.7357	7.9811	17.4587
		Ross	16.6669	4.0575	8.1737	28.1588
		Rachel	8.9565	5.2871	3.7604	27.0856
		Joey	27.6949	1.4901	4.9031	18.8548
	Episode 8	Chandler	17.0541	2.6963	7.1429	23.1364
		Monica	14.2438	1.0900	5.3339	18.5892
		Mrs Geller	19.1759	1.4723	5.6133	17.1449
		Ross	47.1092	1.1615	9.3556	25.4078
	Episode 9	Chandler	22.7411	1.2493	9.2691	17.6500
		Monica	21.7786	3.0576	8.0018	31.5247
		Phoebe	10.3305	1.6038	4.5511	21.1323
		Ross	20.5194	2.5346	10.7392	20.5165
		Rachel	52.1082	1.6363	7.9353	26.6172
		Joey	15.1919	2.5026	4.0487	21.0330
	Episode 10	Chandler	36.0917	1.0090	12.2933	16.5357
		David	50.6253	1.0154	10.0001	18.5572
		Phoebe	29.7010	2.8320	13.0345	23.5875
		Ross	13.4906	2.8640	7.4758	25.5522
	Episode 11	Chandler	26.3339	1.4321	7.3119	22.3044
		Monica	28.5531	1.4881	7.0519	20.5113
		Phoebe	15.9892	3.1437	6.8663	24.6605
		Ross	52.3026	1.3207	10.4046	31.4134
		Mrs Bing	31.3876	2.0824	5.2140	26.1225

		Joey	22.7087	1.3893	6.9773	24.6089
	Episode 12	Chandler	11.0703	0.9005	7.8088	15.9944
		Monica	20.7216	1.1277	5.1459	18.3128
		Phoebe	10.9409	3.3891	7.5399	19.6052
		Ross	115.4088	0.7163	15.1415	21.4679
		Rachel	351.1485	0.3973	11.1996	19.4450
		Joey	17.5463	2.3172	3.7937	32.1401
Shakespeare	Antony and Cleopatra	Agrippa	7.9636	3.6939	4.6171	27.7836
		Alexas	3.6752	4.6411	2.6432	17.6207
		Antony	150.1090	6.6524	48.3283	67.6190
		Caesar	86.8183	5.0312	28.4364	52.9967
		Charmian	10.2090	7.7317	7.1404	33.3741
		Cleopatra	116.1145	4.7154	52.9185	51.7738
		Clown	10.1100	0.0769	823.0656	5.1369
		Dolebella	14.5117	0.8347	7.1147	16.2174
		Enobarbus	80.4681	3.4767	26.3132	47.9371
		Eros	37.5450	0.6488	8.6015	12.6609
		Lepidus	8.5630	6.7850	4.4030	29.7320
		Maecenas	38.0891	0.5173	4.7961	15.1184
		Menas	13.4479	1.8307	7.5420	27.4477
		Messenger	15.7779	4.3194	6.2623	28.4340
		Octavia	47.9578	0.6023	6.2767	14.7190
		Pompey	23.4110	4.1047	11.7848	37.1559
		Scarus	8.8840	2.7353	4.3389	17.0335
		Soldier	13.5528	1.0147	5.1628	15.3466
		Soothsayer	9.1401	1.8835	6.1414	22.2202
		Ventidius	5.6049	1.0476	5.1012	15.6387
	As You Like It	Adam	11.4360	6.1500	8.4880	25.8360
		Celia	47.6814	10.1097	14.5144	67.9230
		Charles	10.9062	3.2596	7.1950	23.9201
		Corin	25.7898	5.2406	5.3984	35.4523
		Duke Senior	40.3351	1.3142	16.3192	23.8040
		First Lord	17.1159	2.7107	6.4213	16.2471
		Frederick	9.3162	4.2059	8.3281	31.5007
		Jaques	60.1792	5.2146	19.2069	44.2127
		Le Beau	24.3899	1.9876	7.8499	19.6652
		Oliver	38.1761	3.9117	15.0360	39.1947
		Orlando	258.8373	0.7121	58.1416	24.8834
		Phebe	36.1700	1.6344	14.9752	27.0379
		Rosalind	184.0942	5.7311	59.7612	53.7071
		Silvius	4.2886	2.7367	15.9854	20.4084
		Touchstone	81.9517	4.6644	28.6203	44.9114
	Hamlet	Bernardo	8.9643	1.5905	4.6326	15.3499
		Clown	20.6937	5.9914	8.0575	42.8827
		Ghost	21.3333	7.2180	6.0432	34.5484
		Guildenstein	18.7390	1.6560	9.0239	15.6366
		Hamlet	368.0557	7.4561	79.2392	73.4126
		Horatio	70.8023	6.1028	16.9538	52.5853

		King	128.9817	3.5035	54.5228	41.5730
		King Player	7.1907	2.2927	5.0085	24.5213
		Laertes	42.8778	7.4519	11.6602	52.8823
		Marcellus	10.2943	3.5801	8.1327	20.3590
		Ophelia	36.3920	9.1709	9.7487	47.6659
		Osirix	9.1509	1.3407	10.1873	17.7843
		PLayer1	54.8955	0.7337	8.4541	16.0537
		Polonius	82.8069	8.1726	19.0108	66.1311
		Queen	12.7550	2.8944	16.7522	30.2596
		Queen Player	3.8791	1.1208	7.9632	12.1293
		Rosencrantz	29.4226	4.2258	10.4344	26.0486
	Henry V	Bardolph	4.3950	5.2430	4.9310	17.4830
		Boy	28.6669	2.3959	10.5396	24.9835
		Burgundy	16.6255	3.2408	7.6334	26.5048
		Canterbury	101.6999	3.5939	13.0553	51.4408
		Chorus	133.3941	2.3002	16.1152	42.0487
		Constable	28.8095	1.6638	17.8506	25.5220
		Dauphin	34.2883	4.7746	9.7772	35.0750
		Exeter	50.3879	5.1914	7.3554	41.1331
		Fluellen	124.9577	7.6054	15.3445	65.1858
		French King	32.8324	4.0191	7.5501	32.3643
		Gower	27.2522	2.8635	7.4777	27.5678
		Henry	298.2390	3.5732	97.9114	47.1347
		Hostess	21.1732	1.5081	6.6444	22.1050
		Katherine	68.0565	1.0111	9.3996	21.8668
		Mac Morris	11.7439	1.4213	8.3288	13.6673
		Monjoy	6.3700	5.4327	7.7088	19.4755
		Nym	31.5223	0.8881	15.6343	14.7473
		Orleans	16.2457	1.4958	6.0680	20.7997
		Pistol	49.8568	4.0679	13.0887	35.7972
		Westmorland	11.6279	2.2307	2.7453	18.5891
		Williams	20.2974	2.6044	11.9375	26.9035
	King Lear	Alb	25.3881	4.5934	12.6622	40.4698
		Cornwall	24.1074	3.3415	12.7242	30.9044
		Edg	124.1456	3.2510	26.6761	48.4818
		Edm	93.2008	5.4716	18.9252	56.6541
		Fool	61.4186	5.2005	15.3716	59.9845
		France	6.5551	1.4081	6.2499	13.5967
		Gentleman	36.7527	1.7006	11.8384	26.9420
		Gloucester	75.6889	4.6826	27.1327	49.4959
		Goneril	34.7401	7.5128	10.8928	55.1487
		Kent	72.4426	7.4311	18.7697	64.3995
		Lear	133.6181	9.9956	34.8839	79.6786
		Oswald	17.6114	7.2708	4.4331	33.7596
		Reg	37.0091	6.8400	13.3323	51.9089
	Measure for	Angelo	60.1393	8.2447	18.9669	62.3914
		Claudio	27.2717	3.4262	10.9770	30.1378

	Measure	Duke	193.8398	7.4327	46.1844	70.3107
		Elbow	16.7057	3.7996	10.0172	33.3692
		Escalus	77.2969	2.2521	22.8091	34.9939
		Isabella	73.2197	12.3184	18.0916	75.2209
		Lucio	72.3262	4.0237	32.7541	39.3814
		Mariana	30.9927	1.9338	9.5916	24.0891
		Peter	10.4503	1.5083	6.3818	18.3799
		Pompey	36.2768	9.6783	7.5974	60.9938
		Provost	19.0340	11.2916	8.7550	51.4226
	Merchant of Venice	Antonio	40.0274	4.3185	17.0602	37.2154
		Arragon	10.6391	3.1436	9.2669	23.9421
		Bassanio	75.2478	6.9928	13.4669	78.4176
		Duke	13.0801	4.3873	4.2593	33.7508
		Gobbo	10.3696	1.4598	7.3190	19.4626
		Gratiano	40.0604	4.6287	11.2397	57.1903
		Jessica	39.3363	1.2086	11.7770	26.6134
		Launcelot	46.8791	4.2944	19.1477	40.6127
		Lorenzo	37.3135	4.8492	11.5656	53.1744
		Nerissa	18.5658	4.2189	9.3045	32.0796
		Portia	144.6836	5.5176	33.6032	71.2591
		Prince of Morocco	38.8794	1.5457	10.7848	29.2062
		Salarino	31.0770	2.3620	10.1874	37.1376
		Shylock	99.5434	5.8353	22.0396	62.7767
		Solanio	100.9333	0.6113	11.8413	20.5739
	Much Ado About Nothing	Antonio	12.4936	2.5810	6.9302	22.3021
		Beatrice	79.2899	8.4077	18.0762	62.4575
		Benedick	160.9169	2.1931	63.7606	32.7501
		Borachio	37.1601	3.6008	13.5916	36.8870
		Claudio	54.9166	2.8616	31.1557	37.4208
		Dogberry	41.4007	6.5265	17.5030	46.2757
		Friar	9.1330	1.7771	13.6564	21.5183
		Hero	23.3000	6.4402	10.2837	45.7167
		Don John	58.6470	1.3993	17.9098	24.1172
		Leonato	66.6995	11.0352	16.6520	75.4024
		Margaret	26.5386	1.8814	12.2022	21.3669
		Messenger	1.5200	4.4470	8.0380	12.3290
		Don Pedro	55.7348	7.5567	27.0872	50.6037
		Ursula	24.3088	1.2740	8.7168	19.3941
	Romeo and Juliet	Benvolio	48.1489	1.9736	16.3697	36.5124
		Chorus	3.7848	4.5352	4.1465	17.7689
		Friar	70.5746	5.9692	25.3374	54.1734
		Juliet	109.0679	7.8944	35.2564	65.3257
		Lady Capulet	14.9599	11.5347	6.1929	54.1374
		Mercutio	85.9426	3.8204	21.9038	44.6478
		Montague	7.9220	2.0949	5.7500	20.7809
		Capulet	53.6106	7.6568	19.9058	60.7473

		Nurse	57.5999	7.5824	15.7571	67.3986
		Paris	13.2837	2.5589	10.3192	26.8785
		Peter	19.5276	1.3174	7.8261	15.2062
		Prince	17.7457	5.4700	6.4128	34.1614
		Romeo	120.7897	12.3153	18.8824	108.4104
		Sampson	19.0783	1.7911	5.2537	21.8609
		Tybalt	4.9557	1.8313	5.6093	18.2740
	Taming of the Shrew	Baptista	42.3024	3.7526	16.6336	38.8009
		Bianca	34.4814	2.5397	6.7574	27.0105
		Biondello	27.3687	5.3930	10.3635	33.5170
		Gremio	35.0088	5.4982	16.6778	38.9996
		Grumio	47.2798	4.2799	17.1323	46.2500
		Hortensio	44.0772	6.4037	15.9306	48.4390
		Katherina	57.6972	6.3527	17.5631	46.4743
		Lord	36.9571	3.4825	12.6598	38.9966
		Lucentio	43.2045	4.5916	15.7969	43.1063
		Pedant	11.4734	3.4685	7.3585	21.7983
		Petruchio	175.1686	2.5603	74.0471	35.8513
		Sly	63.9986	0.9306	11.4632	21.6055
		Tranio	70.5963	4.7861	27.7535	42.1686
		Vincentio	4.1093	1.6858	10.1831	16.9201
Conversations	1	James	227.3406	7.3266	42.0963	65.5239
		Patricia	227.3680	3.8134	38.6934	42.6364
	2	James	103.0227	5.1465	31.1274	41.5708
		Margaret	166.0350	4.3920	36.3620	48.1140
	3	Barry	4.7966	8.3716	24.0751	0.9889
		Kevin	0.1255	2.9483	0.0117	0.9486
		Peter	5.1768	7.0824	21.0225	0.9878
	4	Kevin	2.0898	5.3332	12.1875	0.9797
		Peter	0.1899	2.5344	0.0095	0.9363
	5	Eddie	0.9828	7.9557	21.0187	0.9798
		None	6.7779	7.5754	32.5550	0.9870
	6	Eddie	1.7848	39.4484	25.3206	0.9867
		None	5.1030	15.8491	21.3239	0.9905

Glottometrics 25, 2013,43 - 53

Arc length, vocabulary richness and text size[1]

Ioan-Iovitz Popescu, Bucharest
Peter Zörnig, Brasilia
Gabriel Altmann, Lüdenscheid

Abstract. The article describes the behaviour of arc length computed from the ranked frequencies of some text units and strives for constructing an indicator which is independent of text size. Such an indicator may be used for text characterization, text comparison and classification**,** even for language comparisons.

Keywords*: arc length, rank-frequency, vocabulary richness, text size*

For a rank-frequency distribution with frequencies f_1, f_2,…,f_V, satisfying $f_1 \geq \ldots \geq f_V$, the arc length L is defined as

$$(1) \qquad L(f_1, f_2,\ldots,f_V) = \sum_{i=1}^{V-1} \sqrt{(f_i - f_{i+1})^2 + 1},$$

where V denotes the maximal rank (see e.g. Popescu, Mačutek, Altmann 2009: 49). Representing the distribution graphically, one can interpret L as the sum of Euclidean distances between points, corresponding to consecutive frequencies f_{i} and f_{i+1}. The definition can easily be extended to an arbitrary numeric sequence (discrete or numeric time series) x_1, x_2,…by substituting f_i in (1) for x_i.

In quantitative linguistics the arc length can be considered an elementary indicator of vocabulary richness. In this article we will define some variants of this concept which are (more or less) independent of the text size N and which are useful for text characterization and comparison.

Since the relative arc length L/N of the rank-frequency distribution proposed in Popescu, Čech, Altmann (2011) is still dependent on the text size N, a further modification has been introduced, namely

$$(2) \qquad \Lambda = (L/N)\log_{10}N$$

and its variance has been used to set up an asymptotic test for the difference of two texts. It could be shown that in spite of this modification the indicator still displays a week dependence on N, as can be seen in Figure 1.

[1] Address correspondence to G. Altmann: ram-verlag@t-online.de

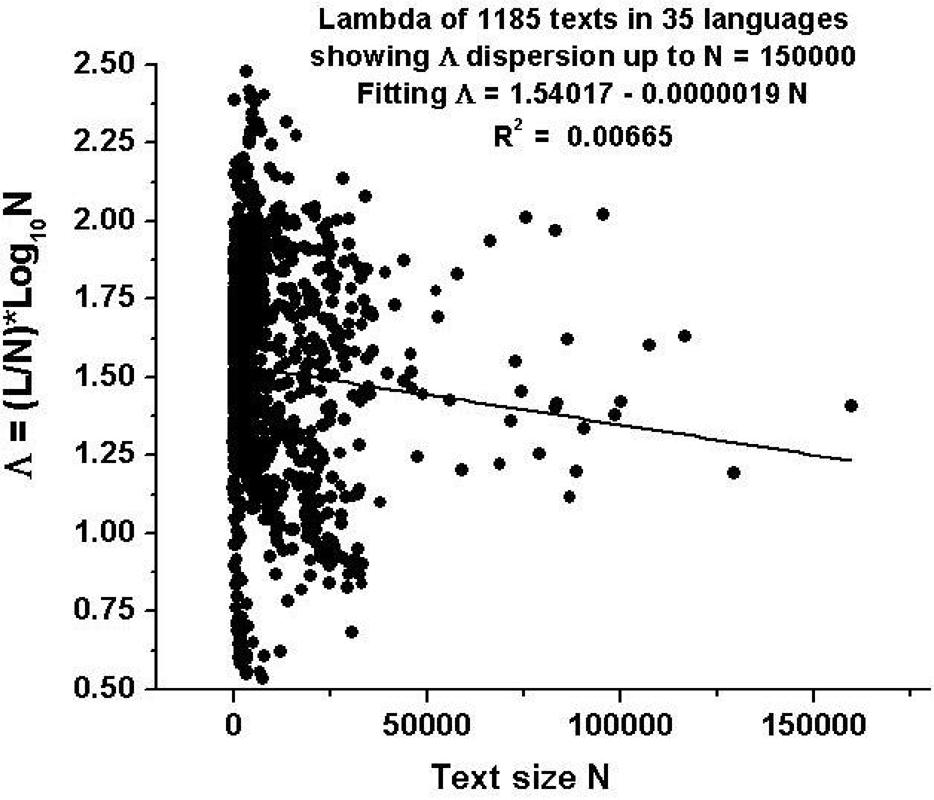

Figure 1. Slight dependence of Λ on N (reproducing Figure 2.1c from Popescu, Čech, Altmann (2011) showing the Λ dispersion)

Of course, the dispersion is enormous because different texts, languages and text-sorts are involved, hence the determination coefficient cannot be satisfactory. The regression coefficient is very small but with increasing N it can get greater values. We used 1185 texts in 35 languages, a sufficient background for analysis. As a matter of fact, the points lie in a triangle. In order to get the points on a horizontal straight line, we modified the given indicator by changing the constant $\log_{10}N$ into $(\log_{10}N)^{1.14282575}$. Using a quite inductive approach we attained an ideal horizontal positioning by a modified lambda as

(3a) $\Lambda_{mod} = (L/N)(\log_{10}N)^{1.14282575}$,

or alternatively, starting from the original Λ,

(3b) $\Lambda_{mod} = \Lambda(\log_{10}N)^{0.14282575}$

yielding the results presented in Figure 2 where a logarithmic scale is used for the N axis.

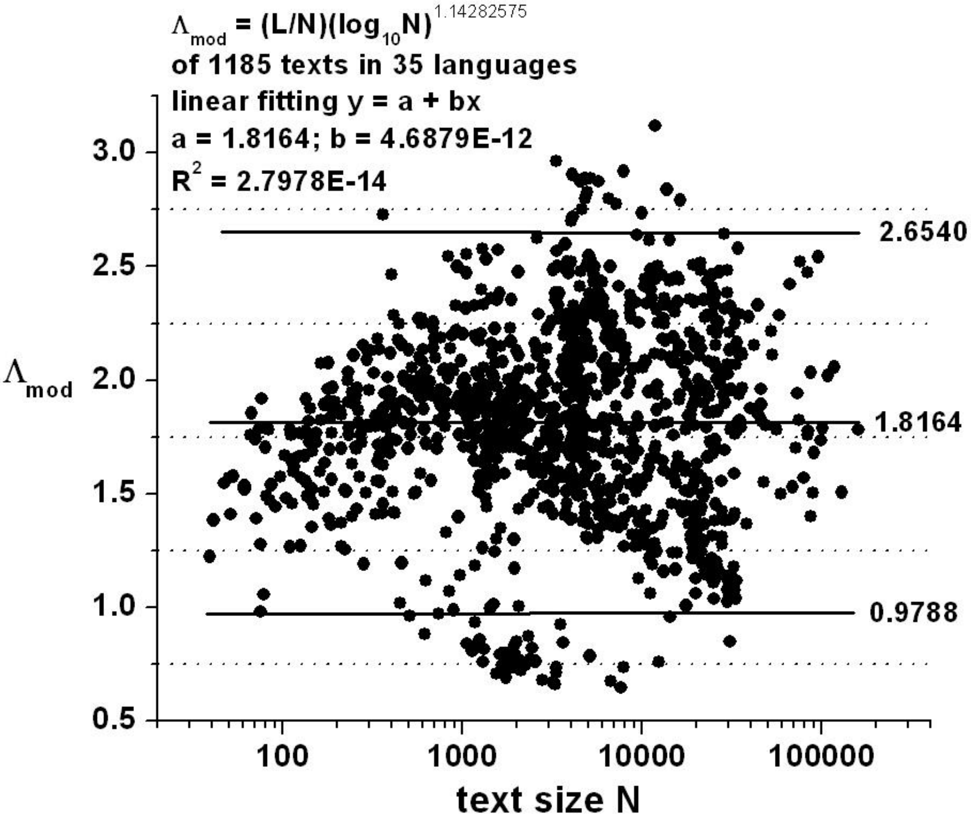

Figure 2. Almost horizontal positioning of Λ_{mod} points.

As can be seen, the regression coefficient has a non-zero value only on the 12th decimal place and even with N = 150000 it remains to be quite small. This time we obtain a non-biased, neutral cloud of points in which we can search for the position of texts, languages or text sorts.

The variance of Λ_{mod} can be written in terms of N and L as

$$(4) \qquad Var(\Lambda_{mod}) = \frac{(\log_{10} N)^{2,2856}}{N^2} Var(L) .$$

$Var(L)$, i.e. the variance of the arc of the distribution, has a very complex formula presented in Popescu, Mačutek, Altmann (2009: 52f.). But since $Var(\Lambda)$ is known for 1185 texts, $Var(\Lambda_{mod})$ can be obtained from it by the simple transformation $Var(\Lambda_{mod}) = (\log_{10}N)^{0,2858} Var(\Lambda)$.

Another way of relativizing the arc length is dividing it by its maximum. The maximum value of the arc length for a given text size N is the optimal solution of the nonlinear optimization problem

$$(5) \qquad \text{Maximize} \sum_{i=1}^{V-1} \sqrt{(f_i - f_{i+1})^2 + 1}$$

subject to $f_1 + f_2 + \ldots + f_V = N$, $f_1 \geq f_2 \geq \ldots \geq f_V \geq 1$, where $f_1, \ldots, f_V$ are integer variables with positive values.

The optimal solution of (5) is $f_1 = N - V + 1$, $f_2 = \ldots = f_V = 1$, and the corresponding optimal value is

(6) $$L_{\max} = \sqrt{(N-V)^2 + 1} + V - 2$$

We will not go into details of the proof of formula (6) which can be performed by using the Karush-Kuhn-Tucker conditions.

For not too small text sizes we get

(7) $$L_{max} = N - V + V - 2 = N - 2.$$

Since vocabulary richness strongly depends on the tail of the rank-frequency distribution, the text is the richer, the greater is

(8) $$L_{rel} = L/L_{max}.$$

This concept was introduced earlier under the name B_1 (cf. Popescu et al. 2009: 50, 57-61). This indicator is adequate for comparisons, too, because its variance is

(9) $$Var(L_{rel}) = \frac{Var(L)}{L_{\max}^2},$$

since for the given data L_{max} is a constant. However, as will be shown below, neither this relativization does stabilize the arc length. Actually, for not too small text sizes, according to (7), we have $L_{max} \approx$ N, hence

(2a) $$\Lambda = (L/N)\log_{10}N = L_{rel}\log_{10}N$$

and

(3c) $$\Lambda_{mod} = L_{rel}(\log_{10}N)^{1.14282575}$$

A third way of characterizing vocabulary richness is the consideration of only those frequencies representing autosemantics. As shown in different places, the fuzzy boundary between synsemantics and autosemantics is the h-point, hence $N_h = \sum_{x=h}^{V} f_x$ is the sum of all frequencies having ranks equal or greater then h.

Similarly, only the autosemantic tail of the arc length, $L_h = \sum_{x=h}^{V} \sqrt{(f_x - f_{x+1})^2 + 1}$ will be considered. In this way synsemantics will be omitted as far as possible. Thus we obtain the indicator

(10) $$\Lambda_h = \frac{L_h(log_{10}(N_h))}{N_h} .$$

A further modification would follow from adding to N_h the autosemantic correction $h^2/2$ yielding

(11) $$\Lambda_{auto} = \frac{L_h(log_{10}(N_h + h^2/2))}{N_h + h^2/2} .$$

The $h^2/2$ area correction has been used before in the definition of the vocabulary richness indicator R_1 (cf. Popescu, et. al. 2009: 33).

Let us return to Figure 2 containing an almost horizontal trend of lambdas. The question is now, how can we define our problems adequately. (1) If we want to compare two texts, we must necessarily take into account the tedious computation of the variance as has been shown in Popescu, Mačutek, Altmann (2009: 52f.). (2) If we want to classify the texts in several classes, e.g. extremely great lambda, moderate lambda, extremely small lambda, we have an easier task. We place e.g. a 95% confidence interval around Λ_{mod} which is made simply by the fact that the regression coefficient can be considered zero (as seen in Figure 2 it is b = 4.6879E-12). Hence the interval for Λ_{mod} can be easily constructed.

Let N be x and Λ_{mod} be y. Here y is a horizontal straight line $y = a$, where a = 1.8164, because b is approximately zero. That means that $\bar{y}$ = 1.8164 = a. The 95% confidence interval around the y yields

$$P(a + bx - A < y < a + bx + A) = 0.95$$

Since b ≈ 0, we can omit it. Thus we obtain

(12) $$P(a - A < y < a + A) = 0.95.$$

Because our enormous sample contains n = 1185 texts, the value A can be written as $u_{\alpha/2}s_y$ where s^2 is the variance of the lambda values, i.e. in our case where the sum of squared deviation of x (= N) is enormous, it reduces to

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2$$

(for an exact *A* see e.g. Mood, Graybill 1963) which in the present case yields

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(\Lambda_{\text{mod}} - a)^2 = \frac{1}{1185}\sum_{i=1}^{1185}(\Lambda_{\text{mod}} - 1.8164)^2$$

The computation using our data yields s^2_y = 0.182626, hence s_y = 0.427348, and we have $u_{0.025}$ = 1.96. Inserting these values in (9) we obtain the interval

$$1.8164 -1.96(0.427348) < y < 1.8164 + 1.96(0.427348)$$

yielding finally

$$0.9788 < y < 2.6540$$

Thus the lower straight line in Figure 2 is y = 0.9788, the upper one is y = 2.6540. In this way we obtain either four groups: one over 2.6540, one between 2.6540 and 1.8164; one between 0.9788 and 1.8164; and the last below 0.9788; or three groups, if we pool the two middle ones.

The interval can be made smaller or larger - with different confidences - and if we add further texts, it will change automatically. The present result shows that the upper interval (> 2.6540) contains 1 Czech, 2 Hungarian, 1 Romanian and 16 Latin texts. One can say that this is the domain of strongly synthetic languages. The lower interval (< 0.9788) contains 2 Dutch, 1 Maori, 1 Rarotongan, 3 Samoan, 5 Marquesan and 34 Hawaiian texts. This is the domain of strongly analytic languages. In any case, Λ_{mod} is also an indicator of morphological simplicity of a language.

In order to illustrate the procedure of measurement of richness and perform a classification of texts, text-sorts and languages we use the data from Popescu, Čech, Altmann (2011) and present the mean Λ_{mod} separately for text-sorts in individual languages, as shown in Table 1. We subdivide the texts into four groups: those above the upper confidence boundary containing only Latin prose texts; those above the median line, those below the median line; and those below the lower confidence boundary. The values in each interval are presented in decreasing order.

Table 1
Mean Λ_{mod}-s for text-sorts in individual languages

Language	**Genre**	**mean Λ_{mod} (descending)**
Latin	prose	2.8102
	Upper interval line	2.6540
Hungarian	newspaper texts	2.4602
Hungarian	poetry	2.4498
Latin	poetry	2.4467
Polish	prose, translation	2.3807
Marathi	aesthetics	2.3275
Belorussian	prose, translation	2.3184
Kannada	social sciences	2.3030
Ukrainian	prose, translation	2.2992
Czech	prose, translation	2.2947
Marathi	official and media	2.2834
Slovak	prose, translation	2.2711
Czech	poetry	2.2698
Hungarian	prose	2.2682
Marathi	poetry	2.2324
Sorbian	prose, translation	2.1767
Slovenian	prose, translation	2.1687
Romanian	poetry	2.1520
Russian	prose	2.1471
Croatian	prose, translation	2.1468
Serbian	prose, translation	2.1436
Czech	prose	2.0738
Latin	history and philosophy	2.0520
Slovak	poetry	2.0250
Marathi	social sciences	1.9869
Russian	poetry	1.9802
Romanian	prose	1.9685
Finnish	prose	1.9589
Bulgarian	prose, translation	1.9477

Kannada	commerce	1.9172
Slovenian	prose	1.9095
Marathi	commerce	1.8806
Turkish	prose	1.8758
Slovak	prose	1.8606
German	poetry	1.8483
Macedonian	prose, translation	1.8467
Marathi	natural and professional sciences	1.8366
Latin	rhetorics	1.8272
	Median Λ_{mod} line	1.8164
Bulgarian	5 private letters	1.8134
Italian	EoY Presidential speeches	1.7758
English	poetry	1.7454
German	prose	1.7314
Italian	poetry	1.7229
Czech	stories by children	1.7044
Indonesian	newspaper texts	1.6883
Italian	prose	1.6651
Tagalog	poetry	1.6455
Czech	scientific texts	1.6246
French	poetry	1.5712
Tagalog	prose	1.5565
English	Nobel lectures	1.5555
English	prose	1.5537
Lakota	tape-recorded texts	1.4364
English	stories by children	1.4223
French	prose	1.3620
Swedish	prose	1.2844
English	scientific texts	1.2822
Japanese	prose	1.2106
Dutch	prose	1.1850
Spanish	prose	1.1843
Maori	folk narratives	1.0580
Rarotongan	prose	1.0303

	Lower interval line	0.9788
Samoan	prose	0.9638
Marquesan	folklore texts	0.8878
Hawaiian	prose	0.7748

If several text sorts have been analyzed in one language, one can set up the order of text sorts for each language separately. Thus we obtain:

Latin: prose - poetry - history and philosophy - rhetorics
Hungarian: newspaper - poetry - prose
Russian: prose - poetry
Marathi: aesthetics - official and media - social sciences - commerce - natural and professional sciences
Czech: prose translation - poetry - prose - //stories by children - scientific texts
Kannada: social sciences - commerce
Slovak: prose translation - poetry - prose
Romanian: poetry - prose
Bulgarian: prose translation - //private letters
//Italian: presidential speeches - poetry - prose
//English: Nobel lectures - poetry - prose -stories by children - scientific texts
//Tagalog: poetry - prose
//French: poetry - prose

The double slant lines (//) show the respective part of the confidence interval. As can be seen, prose translation is always richer than the original prose because the translator must follow the text in the original language but since (s)he cannot perform a word for word translation, many synonyms and paraphrases must be used. The above order is, however, valid only for the texts used; possibly the study of other translations would yield other results.

In general, poetry is richer than prose but this preliminary statement must be further scrutinized.

If one wants to perform a morphological classification of languages, it is sufficient to take simply the mean of means for a given language and set up an order from strongly synthetic to strongly analytic languages. Using the data in Table 1 we obtain Table 2.

Table 2
Mean of means of Λ_{mod}-s by languages

Language	**mean Λ_{mod} (descending)**
Hungarian	2.3927
Polish	2.3807
Belorussian	2.3184
Ukrainian	2.2992
Latin	2.2840
Sorbian	2.1767
Croatian	2.1468
Serbian	2.1436
Kannada	2.1101
Marathi	2.0912
Russian	2.0637
Romanian	2.0603
Slovak	2.0522
Slovenian	2.0391
Czech	1.9935
Finnish	1.9589
Bulgarian	1.8806
Turkish	1.8758
Macedonian	1.8467
German	1.7899
Italian	1.7213
Indonesian	1.6883
Tagalog	1.6010
English	1.5118
French	1.4666
Lakota	1.4364
Swedish	1.2844
Japanese	1.2106
Dutch	1.1850

Spanish	1.1843
Maori	1.058
Rarotongan	1.0303
Samoan	0.9638
Marquesan	0.8878
Hawaiian	0.7748

We can conclude that the Λ-indicators presented above yield a "preliminarily" solid way of characterizing a special property of text, enable us to compare texts, perform a kind of classification of texts in text sorts, and last but nor least, show a ranking of languages with regard to their synthetism/analytism. Needless to say, many further texts must be processed in order to make the results more stable. Since counting of word form frequencies can be performed mechanically and the computation of the given formulas is a matter of simple programming, it is to be hoped that some time it will be possible to process all texts in a given corpus.

Another two problems are to be solved: (i) The given indicators should be compared with other ones that capture the same property (vocabulary richness, synthetism, text-sort indicators). (ii) It should be shown how indicators of other text properties are related to the indicators given above. This tedious problem should aim at the setting up of a control cycle of text properties and at last at the formulation of at least a part of text theory.

References

Mood, A.M., Graybill, F.A. (1963). *Introduction to the theory of statistics*. New York: McGraw-Hill.

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

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

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style, and typology. *Glottotheory 3(1), 89-96.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). The golden section in texts. *ETC - Empirical Text and Culture Research 4, 30-41.*

Glottometrics 25, 2013,54 -68

A continuous model for the distances between coextensive words in a text

Peter Zörnig, Brasília

1. Introduction

The present paper is a contribution to the study of dynamical properties of text generation, which aims to discover the regularities underlying the repetition of text units. This is a relatively new research topic, in contrast to investigations of frequency distributions (see Zörnig (2010, 2013), Tuzzi et al. (2012)).

We continue the research of Zörnig (2013) where the distribution of distances between words of equal length has been modelled for 22 texts of different languages. While in the aforementioned article a discrete probabilistic model, namely the mixed negative binomial distribution (MNB) is employed, the present paper makes use of a continuous model, namely the Zipf-Alekseev function (ZAF). The study is motivated by a related research (Tuzzi et al. (2012), section 3) which proved that for modelling the distances between equal parts-of-speech in Italian texts, the ZAF was more adequate than one of the most appropriate discrete distributions. In fact, in the following we show that the ZAF is adequate for all the 22 texts of Zörnig (2013) including 10 additional texts which could not be modelled by any discrete distribution.

Another comparison between discrete and continuous models can be found in Kelih and Zörnig (2012), and general relations between these different approaches of modelling are discussed in Mačutek and Altmann (2007).

2. Basic concepts

As in Zörnig (2010, 2013) we interpret a real text as a sequence $S = (s_1,\ldots,s_n)$ of length n, consisting of elements chosen from the set $\{1,\ldots,m\}$, where the element r occurs exactly k_r times for $r = 1,\ldots,m$ ($k_1 +\ldots+ k_m = n$).

The distance between two consecutive elements of type r is defined as the number of elements $\neq r$, lying between them. For a given sequence S, we denote by $f_d^{(r)}$ the number of occurrences of the distance d between two consecutive elements of type r. The total frequency of the distance d is defined by

$$f_d = f_d^{(1)} +\ldots+ f_d^{(m)}. \qquad (2.1)$$

We are interested in the distribution of the distances $f_0, f_1,\ldots,f_{n-2}$ in a given sequence S.

In the present application, the element s_j of the sequence S represents the length of the j-th word of the real text, measured by the number of syllables.

Example 2.1: Consider the title of the article Zörnig (2010). Writing down the length of the n = 14 words we obtain the sequence

$$S = (4,4,1,1,4,1,3,2,4,3,1,1,2,2). \qquad (2.2)$$

The frequencies of the text elements are

$$k_1 = 5,\ k_2 = 3,\ k_3 = 2,\ k_4 = 4 \ \ (m = 4).$$

Between the consecutive elements of type 1 we encounter the distances 0, 1, 4 and 0. Thus we have two occurrences of the distance 0, one occurrence of the distance 1, and one occurrence of the distance 4, i.e. $f_0^{(1)} = 2$, $f_1^{(1)} = f_4^{(1)} = 1$. In the same way we find for the other text elements $f_0^{(2)} = f_4^{(2)} = 1$, $f_2^{(3)} = 1$, $f_0^{(4)} = f_2^{(4)} = f_3^{(4)} = 1$. Thus the overall distance frequencies are, see (2.1):

$$f_0 = 4,\ f_1 = 1,\ f_2 = 2,\ f_3 = 1,\ f_4 = 2.$$

In general it holds (see Zörnig (2013, section 2))

$$\sum_{d=0}^{n-2} f_d = n\text{-}m. \qquad (2.3)$$

3. Fitting data by means of the Zipf-Alekseev function

We study the distribution of the distances between words of equal length in 32 texts of different languages with length between 280 and 3140 words, see the following tables. For each text given in form of a sequence (2.2), the observed distance frequencies f_0, $f_1,\dots, f_{19}$ have been determined with the aid of a MAPLE program (see the columns f_d in the following tables). The theoretical frequencies $\bar{f}_d$ have been determined by using the ZAF which will be defined and justified in the following section.

The first 22 texts (Tab. 1-6) are the same as in Zörnig (2013). Additional 10 texts which could not be fitted by a discrete model are presented in Tab. 7-9.

For each of the 32 texts the following additional information is given in the tables:

n: length of the sequence S
m: number of different text elements of S
k_i: frequency of the element i
$N = n - m$: sample size, see (2.3)

The lower boxes in the table contain the results of the fitting:
a, b, C are the optimal parameter values, and R^2 is the coefficient of multiple determination, defined by

$$R^2 = 1 - \frac{\sum_{d=0}^{19} (f_d - \bar{f}_d)^2}{\sum_{d=0}^{19} (f_d - f_{mean})^2}\ , \qquad (3.1)$$

where $f_{mean} = \frac{1}{20} \sum_{d=0}^{19} f_d = \frac{n-m}{20}$ is the mean value of the observed frequencies f_d.

The value (3.1) serves as a criterion for the "goodness of fit", and a fit is considered very good, if $R^2 > 0.9$ (Altmann 1997). Hence for all 32 languages the ZAF fits the data very well.

4. Justification of the use of the Zipf-Alekseev function

In modelling frequency distributions in linguistics, one always starts from the assumption that there is an attractor value *a*, prescribed by the given language which is steadily changed by the speaker or writer depending on diverse conditions like style, aim, text sort, etc. (see e.g. Tuzzi et al. (2012, Section 3)). This results in a "speaker force" *g(x)* which may assume different forms. In several applications this force is assumed to be linear, i.e. $g(x) = a + bx$. Assuming that g(x) changes only slowly in dependence of *x*, one can also assume that $g(x) = \alpha + \beta\ ln\ x$, which we will do in the following. Since language must be in equilibrium, the hearer controls the speaker changes to avoid that the text gets incomprehensible. The hearer applies a "force" *h*(*x*) which usually is assumed to increase proportionally with *x*, i.e. $h(x) = \gamma\ x$.

Assuming that *y* is the theoretical frequency and *g*(*x*)/*h*(*x*) its relative rate of change, we obtain the differential equation

$$\frac{y'}{y} = \frac{g(x)}{h(x)} \qquad (4.1)$$

which is equivalent to

$$\frac{y'}{y} = \frac{\alpha + \beta \ln x}{\gamma x}. \qquad (4.2)$$

By using the notations $a := \frac{\alpha}{\gamma}$ and $b := \frac{\beta}{2\gamma}$ one can rewrite (4.2) as

$$\frac{y'}{y} = \frac{a + 2 * b \ln x}{x} \qquad (4.3)$$

Table 1

	a) Bulgarian N. Ostrovskij, Kak se kaljavaše stomanata, Chapter 1 n = 926, m=6 k_1 = 336 k_2 = 269 k_3 = 213 k_4 = 78 k_5 = 27 k_6 = 3		**b) Hungarian** press: A nominalizmus forradalma n = 1314 , m=9 k_1 = 392 k_7 = 9 k_2 = 304 k_8 = 8 k_3 = 266 k_9 = 2 k_4 = 159 k_5 = 128 k_6 = 46		**c) Hungarian** press: Kunczekolbász n = 458, m= 9 k_1 = 122 k_7 = 8 k_2 = 129 k_8 = 1 k_3 = 81 k_9 = 1 k_4 = 68 k_5 = 34 k_6 = 14		**d) Macedonian** N. Ostrovskij, Kako se kaleše čelkiot, Chapter 1 n = 1123, m=6 k_1 = 426 k_2 = 280 k_3 = 217 k_4 = 123 k_5 = 56 k_6 = 21	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	201	204.05	232	232.46	79	76.86	202	204.17
1	223	212.62	230	229.50	73	80.91	251	247.93
2	140	148.24	182	178.33	70	63.69	189	184.18
3	88	96.61	129	133.63	55	47.64	114	124.48
4	63	62.81	95	100.35	31	35.50	80	82.77
5	44	41.45	84	76.25	24	26.70	51	55.42
6	31	27.89	55	58.75	21	20.33	35	37.65
7	27	19.14	41	45.89	15	15.69	34	26.01
8	16	13.39	41	36.32	12	12.26	26	18.27
9	12	9.53	34	29.08	5	9.70	20	13.03
10	8	6.89	26	23.53	5	7.75	9	9.44
11	9	5.05	20	19.23	5	6.26	11	6.93
12	9	3.76	16	15.85	8	5.10	11	5.15
13	5	2.83	14	13.17	6	4.19	3	3.87
14	0	2.15	15	11.04	6	3.47	9	2.94
15	2	1.65	4	9.30	0	2.89	9	2.26
16	2	1.28	5	7.88	4	2.43	5	1.75
17	4	1.00	7	6.72	2	2.05	7	1.37
18	1	0.79	2	5.77	3	1.74	3	1.07
19	5	0.63	4	4.97	3	1.48	1	0.85
	a = 0.6581 b = -0.8638 C = 204.05 R^2 = 0.9950		a = 0.3623 b = -0.5494 C = 232.46 R^2 = 0.9973		a = 0.4933 b = -0.6046 C = 76.86 R^2 = 0.9804		a = 0.9110 b = -0.9146 C = 204.17 R^2 = 0.9950	

Table 2

	a) Romanian O, Paler, Aventuri solitare, excerpt n = 891 , m=7 k_1 = 392 k_2 = 220 k_3 = 151 k_4 = 92 k_5 = 22 k_6 = 13 k_7 = 1		**b) Romanian** N, Steinhardt, Jurnalul fericirii, Trei soluții n = 1511, m=7 k_1 = 706 k_2 = 375 k_3 = 220 k_4 = 142 k_5 = 51 k_6 = 13 k_7 = 4		**c) Russian** Ostrovskij , Kak zakaljalas stal' n = 792, m=7 k_1 = 264 k_2 = 265 k_3 = 168 k_4 = 70 k_5 = 17 k_6 = 7 k_7 = 1		**d) Serbian** N. Ostrovskij, *Kako se kalio čelik*, Chapter 1 n = 1001, m=6 k_0 = 7 k_1 = 359 k_2 = 328 k_3 = 198 k_4 = 81 k_5 = 28	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	200	204.40	408	412.27	208	206.19	260	255.10
1	240	223.72	367	348.85	152	158.85	185	204.90
2	118	140.36	200	217.70	111	108.64	152	141.91
3	75	79.93	117	131.78	79	74.94	123	98.25
4	53	45.22	80	81.14	63	53.01	67	69.48
5	29	26.05	63	51.31	37	38.47	48	50.32
6	26	15.38	38	33.34	24	28.57	26	37.25
7	15	9.32	25	22.23	24	21.65	29	28.12
8	17	5.79	31	15.16	11	16.70	20	21.60
9	12	3.68	18	10.56	8	13.08	10	16.84
10	11	2.39	14	7.49	6	10.38	9	13.31
11	7	1.58	11	5.40	9	8.34	8	10.65
12	7	1.06	16	3.95	3	6.78	5	8.61
13	8	0.73	12	2.93	6	5.56	3	7.03
14	6	0.50	7	2.20	2	4.60	1	5,79
15	8	0.35	7	1.67	6	3.83	3	4.80
16	6	0.25	3	1.28	2	3.22	2	4.02
17	2	0.18	7	1.00	3	2.72	1	3.38
18	6	0.13	8	0.78	0	2.31	1	2.86
19	2	0.10	2	0.61	2	1.98	2	2.44
	a = 0.9378 b = -1.1651 C = 204.41 R^2 = 0.9818		a = 0.3408 b = -0.8393 C = 412.27 R^2 =0.9933		a = -0.0226 b = -0.5104 C = 206.19 R^2 = 0.9951		a = 0.0560 b = -0.5369 C = 255.10 R^2 = 0.9863	

Table 3

	a) Slovak Bachletová, Moja Dolná zem n = 873, m=9 k_1 = 232 k_6 = 3 k_2 = 325 k_7 = 0 k_3 = 204 k_8 = 0 k_4 = 87 k_9 = 1 k_5 = 21		**b) Slovak** Bachletová, Riadok v tlačive: nezamestnaný n = 924, m=7 k_1 = 258 k_6 = 11 k_2 = 258 k_7 = 1 k_3 = 233 k_4 = 120 k_5 = 43		**c) Slovenian** N. Ostrovskij, Kako se je kalilo jeklo, Chapter 1 n = 977, m=6 k_1 = 426 k_2 = 300 k_3 = 172 k_4 = 61 k_5 = 17 k_6 = 1	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	217	215.80	210	205.90	300	299.90
1	169	174.80	157	169.67	203	205.33
2	129	121.93	119	123.62	141	130.35
3	86	84.93	114	90.15	73	85.09
4	58	60.38	73	66.98	52	57.59
5	42	43.93	45	50.81	46	40.27
6	38	32.66	46	39.28	33	28.97
7	24	24.75	24	30.88	30	21.34
8	13	19.08	19	24.64	13	16.05
9	13	14.93	17	19.92	6	12.29
10	18	11.83	11	16.29	10	9.56
11	7	9.49	14	13.46	10	7.53
12	7	7.70	7	11.23	2	6.01
13	3	6.30	6	9.44	2	4.85
14	4	5.20	6	8.00	8	3.95
15	6	4.33	3	6.82	5	2.24
16	3	3.62	2	5.85	1	2.68
17	2	3.06	4	5.05	3	2.24
18	4	2.59	4	4.38	1	1.88
19	0	2.21	2	3.82	3	1.59
	a = 0.0647 b = -0.5319 C = 215.80 R^2 = 0.9968		a = 0.0374 b = -0.4568 C = 205.90 R^2 = 0.0943		a = -0.1844 b = -0.5225 C = 299.90 R^2 = 0.9956	

Table 4

	a) Sundanese Aki Satimi (Online) n = 1283, m=5 k_1 = 308 k_2 = 593 k_3 = 284 k_4 = 81 k_5 = 17		**b) Sundanese** Agustusan (Salaka Online) n = 416, m=6 k_1 = 97 k_2 = 203 k_3 = 74 k_4 = 36 k_5 = 5 k_6 = 1		**c) Indonesian** Pengurus PSM terbelah (press) n = 345, m=6 k_1 = 35 k_2 = 139 k_3 = 109 k_4 = 56 k_5 = 5 k_6 = 1		**d) Indonesian** Sekolah ditutup (press) n = 280, m = 6 k_1 = 40 k_2 = 94 k_3 = 105 k_4 = 33 k_5 = 5 k_6 = 3	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	342	344.16	121	121.91	110	109.07	73	74.49
1	303	293.17	93	88.75	59	63.84	72	64.58
2	179	192.36	52	56.37	43	40.63	28	40.41
3	126	123.19	38	36.22	32	27.71	28	24.38
4	77	80.18	15	24.00	25	19.88	14	14.93
5	56	53.46	20	16.40	9	14.81	12	9.38
6	37	36.53	13	11.52	13	11.36	7	6.06
7	24	25.54	8	8.29	6	8.92	9	4.01
8	17	18.22	11	6.09	11	7.15	3	2.72
9	21	13.24	7	4.56	6	5.81	2	1.88
10	17	9.78	8	3.47	0	4.80	5	1.32
11	9	7.33	0	2.68	2	4.00	3	0.95
12	11	5.57	3	2.09	3	3.38	0	0.69
13	9	4.28	4	1.65	1	2.87	2	0.51
14	5	3.33	3	1.32	1	2.47	2	0.38
15	5	2.61	2	1.06	3	2.13	1	0.29
16	5	2.97	0	0.86	1	1.85	0	0.22
17	2	1.65	0	0.71	2	1.62	1	0.17
18	0	1.33	2	0.58	0	1.43	0	0.13
19	1	1.08	0	0.49	0	1.26	1	0.10
	a = 0.2784 b = -0.7354 C = 344.16 R^2 = 0.9974		a = -0.0407 b = -0.6022 C = 121.91 R^2 = 0.9899		a = -0.5572 b = -0.3109 C = 109.07 R^2 = 0.9879		a = 0.3936 b = -0.8650 C = 74.49 R^2 = 0.9691	

Table 5

	a) Bamana Masadennin n = 2616, m=8 k_1 = 1680 k_2 = 535 k_3 = 231 k_4 = 100 k_5 = 50 k_6 = 10 k_7 = 9 k_8 = 1		**b) Bamana** Sonsanin n = 2393, m=7 k_1 = 1515 k_2 = 575 k_3 = 159 k_4 = 89 k_5 = 43 k_6 = 11 k_7 = 1		**c) Bamana** NamakƆrƆba halakilen n = 1407, m=5 k_1 = 893 k_2 = 384 k_3 = 97 k_4 = 24 k_5 = 9		**d) Bamana** Bamak' sigicoya n = 1138, m=6 k_1 = 695 k_2 = 255 k_3 = 126 k_4 = 43 k_5 = 18 k_6 = 1	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	1227	1229.08	1086	1087.58	706	706.36	461	462.39
1	501	484.78	508	495.05	266	261.71	248	238.43
2	230	239.37	223	242.48	121	130.03	116	126.90
3	105	134.97	124	130.46	76	75.06	63	72.75
4	76	83.04	73	75.60	48	47.55	49	44.40
5	52	54.36	56	46.41	39	32.11	26	28.48
6	48	37.28	34	29.83	22	22.72	19	19.03
7	44	26.52	22	19.90	16	16.67	21	13.13
8	24	19.43	21	13.70	15	12.58	23	9.32
9	29	14.59	20	9.67	6	9.72	10	6.77
10	18	11.18	24	6.99	11	7.66	5	5.02
11	22	8.71	12	5.14	8	6.14	10	3.78
12	22	6.90	15	3.85	2	4.99	10	2.90
13	22	5.53	10	2.93	3	4.10	4	2.25
14	18	4.49	9	2.26	5	3.41	8	1.77
15	12	3.68	5	1.76	1	2.86	3	1.40
16	10	3.04	6	1.39	2	2.43	3	1.13
17	10	2.54	5	1.10	1	2.07	3	0.91
18	4	2.14	6	0.89	6	1.78	1	0.74
19	5	1.81	8	0.72	2	1.54	4	0.61
	a = -1.0909 b = -0.3625 C = 1229.08 R^2 = 0.9980		a = -0.7412 b = -0.5688 C = 1087.58 R^2 = 0.9988		a = -1.2478 b = -0.2664 C = 706.36 R^2 = 0.9996		a = -0.5770 b = -0.5461 C = 462.39 R^2 = 0.9969	

Table 6

	a) Vai Mu ja vaa I (T. Sherman) n = 3140, m=5 k_1 = 1893 k_2 = 1033 k_3 = 186 k_4 = 86 k_5 = 2		**b) Vai** Sa'bu Mu'a'… n = 495, m=4 k_1 = 281 k_2 = 189 k_3 = 21 k_4 = 4		**c) Vai** Vande bɛ Wu'u n = 426, m=4 k_1 = 270 k_2 = 124 k_3 = 29 k_4 = 3	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	1496	1494.89	233	232.44	176	176.35
1	670	681.86	104	109.26	124	121.71
2	369	341.14	65	55.09	44	49.90
3	183	188.00	33	30.40	20	19.59
4	101	111.56	13	18.01	13	7.98
5	55	70.07	5	11.28	10	3.43
6	43	46.02	6	7.38	4	1.55
7	39	31.33	2	5.01	2	0.74
8	20	21.98	5	3.50	5	0.36
9	15	15.81	0	2.50	0	0.19
10	11	11.62	4	1.83	1	0.10
11	6	8.70	2	1.37	4	0.05
12	13	6.62	0	1.04	6	0.03
13	6	5.11	1	0.80	2	0.02
14	4	3.99	3	0.62	2	0.01
15	3	3.16	1	0.49	0	0.01
16	8	2.52	1	0.39	0	0.00
17	7	2.03	1	0.31	2	0.00
18	8	1.65	0	0.25	0	0.00
19	3	1.35	0	0.21	0	0.00
	a = -0.7694 b = -0.5239 C = 1494.89 R^2 = 0.9994		a = -0.7108 b = -0.5460 C = 232.44 R^2 = 0.9961		a = 0.5151 b = -1.5150 C = 176.35 R^2 = 0.9950	

Table 7

	a) Vai Siika n = 1662, m=6 k_1 = 857 k_2 = 366 k_3 = 263 k_4 = 122 k_5 = 47 k_6 = 7		**b) Tagalog** Rosales n = 1958, m=8 k_1 = 718 k_2 = 642 k_3 = 317 k_4 = 182 k_5 = 74 k_6 = 19 k_7 = 4 k_8 = 2		**c) Tagalog** Hernandez Limang n = 1738, m=8 k_1 = 659 k_2 = 552 k_3 = 317 k_4 = 127 k_5 = 65 k_6 = 15 k_7 = 2 k_8 = 1		**d) Tagalog** Hernandez Magpinsan n = 1466, m=8 k_1 = 498 k_2 = 496 k_3 = 250 k_4 = 147 k_5 = 57 k_6 = 14 k_7 = 2 k_8 = 2	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	480	481.10	288	303.07	344	348.96	226	235.98
1	368	367.14	592	558.98	440	425.90	389	362.47
2	244	232.09	332	371.03	284	295.54	234	267.73
3	138	146.52	169	196.71	175	183.79	174	168.51
4	72	94.97	128	98.80	110	112.39	96	102.12
5	66	63.39	85	49.55	69	69.42	69	61.86
6	40	43.49	55	25.28	49	43.67	49	37.97
7	41	30.58	37	13.21	35	28.04	33	23.71
8	28	21.97	28	7.08	33	18.37	22	15.09
9	24	16.09	28	3.89	25	12.27	26	9.78
10	19	11.98	19	2.19	16	8.34	13	6.45
11	11	9.06	24	1.26	11	5.77	6	4.33
12	17	6.94	16	0.74	17	4.05	13	2.95
13	16	5.38	12	0.45	8	2.88	11	2.04
14	8	4.22	10	0.27	7	2.07	11	1.43
15	7	3.34	11	0.17	10	1.51	7	1.01
16	2	2.67	11	0.11	5	1.11	4	0.72
17	10	2.15	11	0.07	5	0.83	6	0.52
18	2	1.74	10	0.04	5	0.62	4	0.38
19	11	1.43	5	0.03	3	0.47	3	0.28
	a = 0.0776 b = -0.6746 C = 481.10 R^2 = 0.9957		a = 2.0781 b = -1.7239 C = 303.07 R^2 = 0.9766		a = 1.0374 b = -1.0820 C = 348.96 R^2 = 0.9957		a = 1.4812 b = -1.2437 C = 235.98 R^2 = 0.9857	

Table 8

	a) Romanian Popescu n = 1002, m=6 k_1 = 504 k_2 = 275 k_3 = 149 k_4 = 60 k_5 = 12 k_6 = 2		**b) German** Assads Familiendiktatur n = 1415, m=10 k_1 = 612 k_2 = 380 k_3 = 243 k_4 = 103 k_5 = 43 k_6 = 17 k_7 = 7 k_8 = 6 k_9 = 2 k_{10} = 2		**c) German** ATT00012 n = 1146, m=9 k_1 = 517 k_2 = 296 k_3 = 170 k_4 = 96 k_5 = 37 k_6 = 17 k_7 = 6 k_8 = 5 k_9 = 2	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	293	293.45	400	398.58	320	320.14
1	247	247.41	270	277.96	225	225.16
2	159	149.65	191	182.07	151	147.46
3	76	87.46	129	122.63	93	99.08
4	40	52.01	78	85.48	72	68.84
5	38	31.82	65	61.49	48	49.32
6	24	20.04	43	45.35	36	36.27
7	18	12.96	31	34.22	20	27.28
8	8	8.60	16	26.33	29	20.91
9	19	5.83	31	20.59	19	16.30
10	8	4.03	15	16.33	11	12.89
11	5	2.84	12	13.12	11	10.32
12	6	2.03	9	10.66	9	8.36
13	3	1.47	5	8.75	11	6.84
14	7	1.08	12	7.24	5	5.64
15	8	0.81	4	6.04	3	4.70
16	2	0.61	5	5.08	4	3.94
17	1	0.46	3	4.30	2	3.33
18	1	0.35	4	3.66	4	2.82
19	3	0.27	5	3.14	2	2.41
	a = 0.3808 b = -0.9046 C = 293.45 R^2 = 0.9945		a = -0.1897 b = -0.4765 C = 398.58 R^2 = 0.9975		a = -0.1694 b = -0.4881 C = 320.14 R^2 = 0.9984	

Table 9

	a) German Die Stadt des Schweigens n = 1567, m=10 k_1 = 737 k_2 = 417 k_3 = 227 k_4 = 104 k_5 = 45 k_6 = 18 k_7 = 6 k_8 = 10 k_9 = 1 k_{10} = 2		**b) German** Terror in Ost-Timor n = 1398, m=9 k_1 = 638 k_2 = 399 k_3 = 214 k_4 = 90 k_5 = 36 k_6 = 11 k_7 = 5 k_8 = 4 k_9 = 1		**c) German** Unter Hackern... n = 1363, m=9 k_1 = 637 k_2 = 345 k_3 = 181 k_4 = 125 k_5 = 38 k_6 = 22 k_7 = 9 k_8 = 4 k_9 = 2	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	465	466.80	409	420.24	398	300.73
1	337	326.21	305	307.72	277	268.14
2	184	203.62	193	188.95	162	171.31
3	138	129.64	109	119.14	110	113.24
4	94	85.42	73	77.38	84	77.75
5	46	58.16	49	51.85	46	55.19
6	45	40.77	36	35.73	39	40.28
7	27	29.30	29	25.25	31	30.11
8	13	21.51	19	18.24	30	22.96
9	15	16.10	18	13.42	18	17.81
10	18	12.24	22	10.05	14	14.03
11	14	9.45	13	7.64	14	11.19
12	16	7.39	8	5.88	12	9.04
13	11	5.84	9	4.58	8	7.37
14	13	4.67	4	3.61	10	6.07
15	9	3.77	5	2.87	13	5.04
16	7	3.06	8	2.31	8	4.22
17	4	2.51	5	1.87	5	3.55
18	7	2.07	6	1.52	5	3.01
19	7	1.72	1	1.25	4	2.57
	a = -0.1098 b = -0.5874 C = 466.80 R^2 = 0.9958		a = 0.0059 b = -0.6475 C = 410.14 R^2 = 0.9980		a = -0.2422 b = -0.4816 C = 399.73 R^2 = 0.9977	

This differential equation has the solution

$$y(x) = Cx^{a + b \ln x}, \tag{4.4}$$

representing the Zipf-Alekseev function.

We have made use of model (4.4) to predict the distance frequencies. Since *y* is not defined for *x* = 0, we fitted the observed values f_d by *y*(*d*+1); i.e. the theoretical frequencies $\bar{f}_d$ in Tables 1-6 are given by

$$\bar{f}_d = C(d+1)^{a + b \ln (d+1)}, \tag{4.5}$$

for *d* = 0, 1,...,19, where *a, b, C* are the optimal parameter values in the lower parts of the tables, which have been determined iteratively.

5. Correlation between parameters

We finally investigate the question whether there is a correlation between the optimal parameters *a* and *b*, listed in the lower boxes of Tables 1 to 9. In Fig. 1 the parameter pairs (*a, b*) are graphically illustrated as points in the plane, showing that *b* tends to decrease linearly if *a* increases. In fact, the coefficient of linear correlation is *R* =-0.85. Note that the square of *R* is the coefficient of determination (see Section 3). The sign of *R* is positive or negative if the regression line is increasing or decreasing, respectively. We have fitted the linear model *b* = *c* + *da* to these data, which resulted in the optimal values *c* = -0.6646 and *d* = -0.3990. Table 10 represents the previously calculated parameter values of *a* and *b* and the computed value of *b*, i.e. the value -0.6646 - 0.3990*a* (which corresponds to the straight line in Fig. 1). The last column contains the residuals *b* - (*c*+*da*). A large residual indicates an outlier. The largest residual (with absolute value 0.6448) was obtained for the text Vai in Table 6.c (see Fig.1). This might indicate that this text is essentially different from the other 31 texts studied above. Relative high residuals (i.e. with absolute value ≥ 0.23) can also be observed for the texts in Table 1b, c (Hungarian) and in Table 7b (Tagalog).

Fig. 1: Linear regression

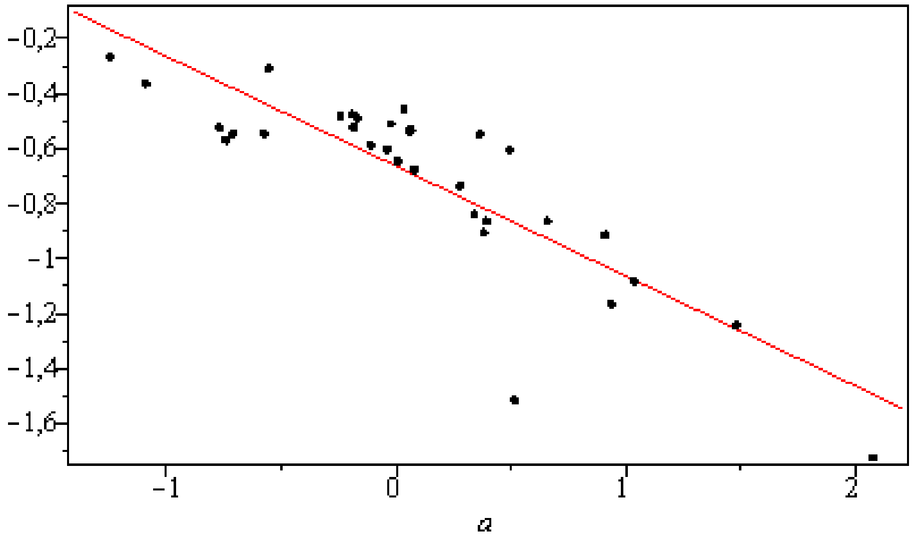

Table 10
Relationship between the parameters

Text	*a*	*b*	*c+da*	residual
Tab.1a	0.6581	-0.8638	-0.9272	0.0634
Tab.1b	0.3623	-0.5494	-0.8092	0.2598
Tab.1c	0.4933	-0.6046	-0.8615	0.2569
Tab.1d	0.9110	-0.9146	-1.0282	0.1136
Tab.2a	0.9378	-1.1651	-1.0388	-0.1263
Tab.2b	0.3408	-0.8393	-0.8006	-0.0387
Tab.2c	-0.0226	-0.5104	-0.6556	0.1452
Tab.2d	0.0560	-0.5369	-0.6870	0.1501
Tab.3a	0.0647	-0.5319	-0.6904	0.1585
Tab.3b	0.0347	-0.4568	-0.6796	0.2228
Tab.3c	-0.1844	-0.5225	-0.5910	0.0685
Tab.4a	0.2784	-0.7354	-0.7757	0.0403
Tab.4b	-0.0407	-0.6022	-0.6484	0.0462
Tab.4c	-0.5572	-0.3109	-0.4423	0.1314
Tab.4d	0.3936	-0.8650	-0.8217	-0.0433
Tab.5a	-1.0909	-0.3625	-0.2293	-0.1332
Tab.5b	-0.7412	-0.5688	-0.3689	-0.1999
Tab.5c	-1.2478	-0.2664	-0.1667	-0.0997
Tab.5d	-0.5770	-0.5461	-0,4344	-0.1117
Tab.6a	-0.7694	-0.5239	-0.3576	-0.1663
Tab.6b	-0.7108	-0.5460	-0.3810	-0.1650
Tab.6c	0.5151	-1.5150	-0.8702	-0.6448

Tab.7a	0.0776	-0.6764	-0.6956	0.0210
Tab.7b	2.0781	-1.7239	-1.4993	-0.2300
Tab.7c	1.0374	-1.0820	-1.0786	-0.0034
Tab.7d	1.4812	-1.2437	-1.2557	0.0120
Tab.8a	0.3808	-0.9046	-0.8166	-0.0880
Tab.8b	-0.1897	-0.4765	-0.5889	0.1124
Tab.8c	-0.1694	-0.4881	-0.5970	0.1089
Tab.9a	-0.1098	-0.5874	-0.6208	0.0334
Tab.9b	0.0059	-0.6475	-0.6670	0.0195
Tab.9c	-0.2422	-0.4816	-0.5680	0.0864

Acknowledgements

I would like to thank Gabriel Altmann, Emmerich Kelih, Sven Naumann, Anja Overbeck, Ioan-Iovitz Popescu and Andrij Rovenchak for providing me with files containing the word lengths of the analyzed texts.

References

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

Kelih Emmerich; Zörnig, Peter (2012). Models of morph lengths: Discrete and continuous approaches. *Glottometrics* 24, p. 70-78.

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

Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Zörnig, Peter; Altmann, Gabriel (2012). Aspects of the behaviour of parts-of-speech in Italian texts. *Glottometrics 24, 41-69.*

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

Zörnig, Peter (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.

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions*. Lüdenscheid, RAM-Verlag.

Glottometrics 25, 2013, 69 - 84

The lexical-semantic fields of verbs in English texts

Olha Pavlyshenko[1]

Abstract. In this paper the lexical-semantic groups of verbs in texts of English fiction have been considered. It is shown that the frequency distribution of lexical-semantic groups of verbs in texts of English fiction makes it possible to characterize the lexical-semantic structure of author's idiolect. The strongest characterization potential is concealed in the frequency distribution of lexical-semantic fields that are formed by the verbs. The area of high-frequency words contains the words of nominative, stylistically neutral type, and the area of author's idiolect is located on the periphery of the lexical-semantic field. The constants of semantic distances that characterize the area of author's idiolect in the structure of lexical-semantic fields do not depend on the quantity and quality of authors' texts and represent the fundamental lexical-semantic regularities of author's style.

Key words: *semantic fields, semantic distance, author's idiolect.*

1. Introduction

There are several possibilities to determine a semantic field – we have chosen the following one: a semantic field is a set of words grouped by meaning referring to a specific subject [Jackson 2000]. The basis of defining the semantic fields is a lexical-semantic paradigm that is a set of words which are determined by a set of semantic features. The core of a semantic field is formed by the words, the dominant values of which constitute the main features of the semantic field. The periphery of a semantic field is formed by the words which contain the basic concepts of the semantic field indirectly through a series of differential characteristics that are related to the basic concept that establishes the semantic field (Verdieva 1986). The specifying and differentiating semantic links within a semantic field determine the structure of the field (Kuznetsova 1989). The German linguist Trier, one of the founders of the theory of semantic fields (Corson 1995), paritioned the considered structure of words to verbal and conceptual fields. He also believed that semantic fields are continuous, i.e. the words of a semantic field embrace its conceptual area continuously and the composition of a dictionary covers the whole range of language concepts (Ufimtseva 1962). The paper by Gliozzo (2009) proposes the concept of semantic domains, which complements the concept of semantic fields. The definition of semantic domains is similar to the methods of computer analysis of texts, and it is based on corresponding text collections belonging to the domain under analysis and characterizing the semantic concepts that distinguish this domain. The lexical composition of semantic fields is defined in various ways (Gol'dberg 1988). One of them is to single out the general concept on the base of which a lexical-semantic field is formed. Another way is to determine a word or group of words and then find their respective synonyms. Semantic fields can also be determined on the basis of expert simultaneous occurrences of words in given contexts. The description of semantic fields can also be found in (Crow and Quigley 1985; Fisiak 1985). An example of lexicographic computer system representing the semantic network of links between words is a WordNet system (Fellbaum 1998), developed at Princeton University. This system is built on the expert lexicographic analysis of semantic structural relationships that reflect the denotative and connotative characteristics of a lexemic composi-

[1] Ivan Franko National University of Lviv (Ukraine), e-mail: pavlsh@yahoo.com

tion of a dictionary. The semantic fields in the WordNet are represented as lexicographic files. Nouns, verbs, adjectives and adverbs are organized in synsets – the sets of synonyms. Nouns and verbs are grouped according to the semantic fields.

In this paper we study the distribution of semantic fields in texts of English literature. We also define the quantitative structure of a semantic field as the core and the periphery, i.e. as a rough set, and analyze the markers of author's idiolect in the semantic fields.

2. The Frequency of a Semantic Field

Let us consider one of the typical distributions of English verbs to lexical-semantic fields. Such a distribution was taken as the basis for the e-linguistic dictionary WordNet. It was offered by the scientists at Princeton University (USA) (Fellbaum 1998). The semantic fields in the WordNet network (http://wordnet.princeton.edu) are presented as lexicographic files. We selected for our study the following lexicographic files of verbs: body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather.

We calculate the frequency of a word *j* as follows:

$$P_{tj} = \frac{n_j}{N_{text}} \tag{1}$$

where n_j is a number of occurrences of the word *j* in the text sample containing N_{text} words. The probability that the word, randomly met in the text, refers to the lexical-semantic field *f* is equal to the sum of words frequencies that belong to the given field *f*

$$P_f = \sum_{i=1}^{N_f} P_{ti} \tag{2}$$

where N_f is a number of words in the field *f*. It is obvious that our consideration does not involve all English verbs, but only some of them, hence it is reasonable to calculate the word frequency of the lexical-semantic field *f* in the spectrum of considered fields. It can be evaluated by the formula

$$P_{sf} = \frac{P_f}{\sum_{i=1}^{N_{sf}} P_{fi}} \tag{3}$$

where N_{sf} is a number of the semantic fields under study. The value P_{sf} describes how frequently the accidentally met word in the text is a part of the field *f*, provided this word belongs to the studied range of verbs. Obviously the sum P_{sf} by all the semantic fields is equal to 1. The set of P_{sf} values characterizes the frequency structure of verbal lexical-semantic fields in the analyzed texts. The calculated frequency distribution of P_{sf} by the semantic fields in the analyzed texts of English fiction is given in Appendix. The verbs identified in the texts were distributed by semantic fields and placed in descending order of textual frequency within each semantic field.

3. The Quantitative Core and Periphery of Lexical-Semantic Fields

The hierarchical structural organization is typical for frequency dictionaries of lexical-semantic fields. Let us introduce the quantitative and frequency definition of core and periphery of lexical-semantic fields. We suppose that the core of the lexical-semantic field is formed by the words, the total frequency of which is at least 0.5. In other words, the total use of words that belong to the core of a lexical-semantic field in texts is 50% of all words of the given field. The close periphery is offered to consist of the words representing 40% of occurrences in texts, and far periphery is assumed to consist of the words representing the last 10% of occurrences. The words for the core, close and far periphery are placed in the frequency row of the lexical-semantic field in descending order of frequency. That is, the axis of words on the graph of frequencies can be divided by two points into three frequency areas – core, close periphery, far periphery. Since different lexical-semantic fields contain a different number of words, it is advisable to introduce a new variable that would characterize numerically the semantic distance of a word to the core of the field. If we assume that the semantic distances of words in descending frequency row of the lexical-semantic field vary from 0 to 1 regardless of the number of words in the field, then the semantic distance of the word S_j from the core of the lexical-semantic field can be calculated as:

$$S_j = \frac{j-1}{N_f - 1} \qquad (4)$$

where j is the rank of the word, N_f is the number of words in the lexical-semantic field. This means that the first word of the frequency range of the lexical-semantic field corresponds to 0, and the last one corresponds to 1. In order to find the $S_{0.5}$ value, which divides the ranks axis of the frequency curve into the core and the periphery, one must solve the equation

$$\sum_{i=1}^{k_{0.5}} P_{fi} = 0.5; \qquad S_{0.5} = \frac{k_{0.5} - 1}{N_f - 1} \qquad (5)$$

where $k_{0.5}$ is the rank of the last word in the initial part of the words row that is built in the descending order by frequencies, and the sum of words frequencies of this part is equal to 0.5. A similar equation is to be solved to find the value of $S_{0.9}$ which divides the axis of words ranks into close and far peripheries of the semantic field.

For our analysis we use the electronic text sample of English literature totaling to about 800 million words, which consists of about 10,000 works of 1000 various authors. This text sample is formed with the use of electronic databases of English literary works. To study the frequency distribution of lexical-semantic fields of verbs we selected the works by Arthur Conan Doyle (33 works), Jack London (38 works), Herbert Wells (26 works), Charles Dickens (52 pieces), Mark Twain (44 works), Oscar Wilde (18 works). The style of these writers is characterized by artistic and stylistic expressiveness and originality. The total amount of the authors’ text sample is about 15 million words. In total the calculations of the frequency structure of lexical-semantic fields of verbs were done for 998 works of 32 authors. The composition of described above lexical-semantic fields was formed using dictionaries definitions of the electronic thesaurus WordNet. The total list of obtained verb infinitives is about 5000 words. Additionally, the verb forms for the third person singular, the past tense, the present and past participles, and gerund were included into the structure of lexical-semantic fields. Thus, the total scope of verbs under study is about 20,000 words.

As a result of the equation (5) solving for all the semantic fields under study, it was figured out that limit of the core and close periphery separation is characterized by the value

$$S_{0.5} = 0.05 \pm 0.02 \qquad (6)$$

and the limit of the close and far peripheries separation is characterized by the value

$$S_{0.9} = 0.3 \pm 0.1 \qquad (7)$$

The words of the descending frequency range in the lexical-semantic fields for which S_j <0.05, are not less than 50% of all words occurrences of given field; the words for which S_j <0.3 are not less than 90% of all occurrences, and the words for which S_j> 0.3 are not more than 10% of all words occurrences of certain lexical-semantic field. The values (6) and (7) are obtained by means of averaging the values found for the frequency distribution of considered 15 verbal semantic fields in English fiction. Within the limits of the accuracy obtained the data values do not depend on the quantity and quality of the lexical-semantic field, and they are the constants of the words system organization into lexical-semantic fields of verbs, along with the constant of Zipf law for frequency distribution.

4. The Distribution of Words of Verbal Lexical-Semantic Fields in Authors' Texts

Let us consider the words distribution in the verbal lexical-semantic field of verbs of communication in the texts of English literature. For a comparative analysis we have selected the works of Jack London, Mark Twain, and Oscar Wilde. To characterize the words of the semantic field under study in authors' texts, we introduce a value of D_j, which shows how many times a particular word *j* occurs more frequently in the authors' texts in comparison with those of linguo-stylistic norm:

$$D_j = \frac{P_{aj}}{P_{tj}} \qquad (8)$$

where P_{aj} is the word frequency, calculated by the formula (1) in the text sample of a certain author; P_{tj} is the word frequency in the whole text sample of all authors, i.e. in the approximation to the linguo-stylistic norm. In Table 1 there are some examples of words of different lexical-semantic fields with the coefficient D>1 in the texts of three authors: Jack London, Mark Twain, Oscar Wilde. The words are placed in the order of descending value of coefficient D_j. For each word we calculated the value of semantic distance S_j and marked the number of the lexical-semantic field. These words can be regarded as the markers of author's idiolect.

Table 1
Lexical markers of the author's idiolect

Herbert Wells				**Jack London**				**Mark Twain**			
Word	*Lexical-semantic field*	*D_j*	*S_j*	*Word*	*Lexical-semantic field*	*D_j*	*S_j*	*Word*	*Lexical-semantic field*	*D_j*	*S_j*
muddle	Contact	23,3	0,56	slather	Contact	82,8	0,88	finger-print	Creation	43,6	0,76
bogey	Contact	21,3	0,8	sled	Motion	74,9	0,62	powwow	Communication	23,7	0,81
obsess	Emotion or Psych	17,4	0,87	grubstake	Possession	73,9	0,86	shuck	Change	15,6	0,56
punt	Contact	15,5	0,62	unlash	Contact	70,9	0,88	chaw	Consumption	12,7	0,79
gesticulate	Communication	15	0,6	tauten	Change	42,2	0,74	resurrect	Bodily Functions and Care	10,9	0,77

profiteer	Possession	13	0,79	electroplate	Contact	34,8	0,86	splotch	Contact	10,3	0,8
wallpaper	Contact	11,8	0,78	snowshoe	Motion	33,9	0,76	teethe	Bodily Functions and Care	9,4	0,79
disentangle	Contact	11,5	0,62	mush	Motion	29	0,71	cowhide	Contact	8,7	0,76
camouflage	Perception	11,1	0,83	bunk	Possession	26	0,58	drowse	Bodily Functions and Care	8,2	0,65
whack	Contact	10,7	0,63	doss	Bodily Functions and Care	25,3	0,83	swap	Motion	7,7	0,74
clamber	Motion	10	0,52	riffle	Contact	24,8	0,8	cooper	Creation	7,5	0,43
individualize	Cognition	9,5	0,73	frazzle	Bodily Functions and Care	23,2	0,84	boomerang	Motion	7,1	0,73
impact	Contact	8,7	0,58	resurrect	Bodily Functions and Care	22,1	0,77	fart	Bodily Functions and Care	7	0,76
attenuate	Change	8,6	0,54	gouge	Contact	22,1	0,71	lynch	Social Interaction	7	0,63
interlude	Creation	8,6	0,59	hunch	Motion	21,1	0,67	simplify	Change	6,7	0,48
disconnect	Contact	8,3	0,6	befuddle	Consump-tion	19,7	0,83	whoop	Communication	6,7	0,53
throb	Perception	8,3	0,57	hike	Motion	19,7	0,76	quarantine	Change	6,3	0,51
unify	Change	8	0,56	relive	Cognition	19,4	0,78	duplicate	Creation	6,2	0,51
goggle	Perception	7,9	0,75	gibber	Communication	19,3	0,73	shovel	Contact	6	0,46
corrugate	Contact	7,7	0,71	prod	Contact	18,8	0,72	roost	Change	5,8	0,47
superpose	Contact	7,7	0,77	swat	Contact	17,5	0,81	swag	Motion	5,8	0,71
foreshorten	Change	7,4	0,62	hoodoo	State	16,5	0,9	slouch	Motion	5,8	0,61
yelp	Communication	7,1	0,61	clutter	Change	16,4	0,61	alligator	Change	5,6	0,49
crescendo	Change	7	0,6	impact	Contact	15,1	0,58	calendar	Cognition	5,6	0,48
readjust	Change	6,9	0,56	recuperate	Bodily Functions and Care	15	0,74	crick	Bodily Functions and Care	5,5	0,74
indurate	Change	6,8	0,64	orate	Communication	14,9	0,82	skip	Motion	5,5	0,58
flare	Weather	6,8	0,56	yelp	Communication	14,9	0,61	starboard	Motion	5,5	0,6
underline	Communication	6,6	0,77	yaw	Motion	14,1	0,75	swelter	Bodily Functions and Care	5,2	0,69
slum	Social Interraction	6,5	0,65	disrupt	Change	13,9	0,63	tally	Communication	5,1	0,61
collide	Contact	6,3	0,7	burgeon	Change	13,8	0,69	suds	Contact	4,8	0,74

disavow	Communication	6,1	0,69	sunburn	Bodily Functions and Care	13,7	0,67	autograph	Communication	4,5	0,6
boo	Communication	6	0,69	sublimate	Motion	13,7	0,73	shred	Contact	4,5	0,51
subserve	Social Interaction	5,9	0,76	clam	Contact	13,5	0,58	drivel	Bodily Functions and Care	4,4	0,75
rearrange	Change	5,7	0,53	collide	Contact	13,2	0,7	solidify	Change	4,4	0,58

The divergences of word frequencies in authors’ texts are partly due to the author's style. Every author has his unique set of words the frequencies of which exceed significantly (i.e. 5-10 times) the ones summarized by the whole textual base. The sets of such words found in certain semantic fields in the samples of authors’ texts, can be considered as a peculiarity of author’s style. Our investigations show that all the words describing the semantic aspect of the author’s style belong to the far periphery due to the value S_j.

We have calculated the average S_j values for the words which meet the following conditions: $D_j > 1$; $D_j > 2$; $D_j > 4$, i.e. we considered the words that occur in the texts simply more frequently, twice more frequently, four times more frequently as compared with the approximation to the linguo-stylistic norm. As a result of the calculations conducted, the following values were obtained:

$$\begin{gathered} S_{D>1} = 0.39 \pm 0.14; \\ S_{D>2} = 0.5 \pm 0.15; \\ S_{D>4} = 0.59 \pm 0.12. \end{gathered} \tag{9}$$

The values obtained are averaged for the texts of all six authors under study. These values are the constants that characterize the area of the semantic field, where the author's lexicon is located. According to our distribution of lexical-semantic fields, the area of author’s idiolect of verbs is in the far periphery of the semantic fields. It follows from (9) that with the increasing value D_j, the value S_j increases as well, i.e., the more frequently the word occurs in the authors’ texts in comparison with the linguo-stylistic norm, the farther from the core of the lexical-semantic field it is located. The analysis of dictionary definitions of authors’ words showed that the specifying differentiating specialized and rarely used semes are typical for them.

5. The Examples of the Use of Author’s Idiolect Markers that Occur in Literary Texts

Here we give some examples of the use of markers of author’s idiolect in the texts of English prose.

The use of author’s idiolect markers in the works of Herbert Wells:

Word	Lexical-Semantic Field	D	S
Unify	Change	8	0,56

*…He was established now in the sure conviction of God's reality, and of his advent to **unify** the lives of men and to save mankind…*

(The Soul of a Bishop by H. G. Wells)

…And so, just as I cling to the belief, in spite of hundreds of adverse phenomena, that the religious and social stir of these times must ultimately go far to ***unify*** *mankind under the kingship of God, so do I cling also to the persuasion that there are intellectual forces among the rational elements in the belligerent centres, among the other neutrals and in America, that will co-operate in enabling the United States to play that role of the Unimpassioned Third Party, which becomes more and more necessary to a generally satisfactory ending of the war…*

(War and the Future by H. G. Wells)

…Even those who have neither the imagination nor the faith to apprehend God as a reality will, I think, realize presently that the Kingdom of God over a world-wide system of republican states, is the only possible formula under which we may hope to ***unify*** *and save mankind…*

(War and the Future by H. G. Wells)

Word	Lexical-Semantic Field	D	S
Throb	Perception	8,3	0,57

…My blood-vessels began to ***throb*** *in my ears, and the sound of Cavor's movements diminished. I noted how still everything had become, because of the thinning of the air…*

(The First Men in the Moon by H. G. Wells)

…Confronted they were, and there was no getting away from it. He would make this appalling viscus beat and ***throb*** *before the shrinking journalists – no uncle with a big watch and a little ever baby ever harped upon it so relentlessly; whatever evasion they attempted he set aside…*

(The War in the Air by H. G. Wells)

… You could see his muscles ***throb*** *and jump, and he twisted about…*

(When the Sleeper Wakes by H. G. Wells)

Word	Lexical-Semantic Field	D	S
Subserve	Social Interaction	5,9	0,69

… The New Republican is a New Republican, and he tests all things by their effect upon the evolution of man; he is a Socialist or an Individualist, a Free Trader or a Protectionist, a Republican or a Democrat just so far, and only so far, as these various principles of public policy ***subserve*** *his greater end…*

(Mankind in the Making by H. G. Wells)

… In the initiative of the individual above the average, lies the reality of the future, which the State, presenting the average, may ***subserve*** *but cannot control…*

(A Modern Utopia by H. G. Wells)

… But that's only to be done by concentrating one's life upon one main end. We have to plan our days, to make everything ***subserve*** *our scheme…*

(The New Machiavelli by H. G. Wells)

The use of author’s idiolect markers in the works of Jack London:

Слово	Lexical-Semantic Field	D	S
Hike	Motion	19,7	0,76

…"An' of course the dogs can ***hike*** *along all day with that contraption behind them," affirmed a second of the men…*

(The Call of the Wild by Jack London)

…"Look here, Smoke, I ain't travelin' no more with a ornery outfit like this. Right here's where I sure jump it. You an' me stick together. Savve? Now, you take your blankets an' ***hike*** *down to the Elkhorn. Wait for me. I'll settle up, collect what's comin', an' give them what's comin'. I ain't no good on the water, but my feet's on terry-fermy now an' I'm sure goin' to make smoke…"*

(Smoke Bellew by Jack London)

… We got enough money for a month's grub an' ammunition, an' we ***hike*** *up the Klondike to the back country. If they ain't no moose, we go an' live with the Indians. But if we ain't got five thousand pounds of meat six weeks from now, I'll – I'll sure go back an' apologize to our bosses. Is it a go?.."*

(Smoke Bellew by Jack London)

Word	Lexical-Semantic Field	D	S
Recuperate	Bodily Functions and Care	15	0,74

… I often doubt, I often doubt, the worthwhileness of reason. Dreams must be more substantial and satisfying. Emotional delight is more filling and lasting than intellectual delight; and, besides, you pay for your moments of intellectual delight by having the blues. Emotional delight is followed by no more than jaded senses which speedily ***recuperate****. I envy you, I envy you…"*

(The Sea Wolf by Jack London)

… Dogs on vacation, boarding at the Cedarwild Animal School, were given every opportunity to ***recuperate*** *from the hardships and wear and tear of from six months to a year and more on the road…*

(Michael, Brother of Jerry by Jack London)

… We parted at Papeete. I remained ashore to ***recuperate****; and he went on in a cutter to his own island, Bora Bora. Six weeks later he was back. I was surprised, for he had told me of his wife, and said that he was returning to her, and would give over sailing on far voyages…*

(South Sea Tales by Jack London)

Word	Lexical-Semantic Field	D	S
Collide	Contact	13,2	0,7

… The hunter, in turn, was in a quandary. His rifle was between his knees, but if he let go the steering-oar in order to shoot, the boat would sweep around and ***collide*** *with the schooner. Also he saw Wolf Larsen's rifle bearing upon him and knew he would be shot ere he could get his rifle into play…*

(The Sea Wolf by Jack London)

… He stood up abruptly, towering to such height that Daughtry looked to see the crown of his head ***collide*** *with the deck above…*

(Michael, Brother of Jerry by Jack London)

… The wide rooms seemed too narrow for his rolling gait, and to himself he was in terror lest his broad shoulders should ***collide*** *with the doorways or sweep the bric-a-brac from the low mantel…*

(Martin Eden by Jack London)

The use of author's idiolect markers in the works of Mark Twain:

Word	Lexical-Semantic Field	D	S
Roost	Change	5,8	0,47

… There were two powerful parties at Court; therefore to make a decision either way would infallibly embroil them with one of those parties; so it seemed to them wisest to ***roost*** *on the fence and shift the burden to other shoulders…*

(Personal Recollections of Joan of Arc by Mark Twain)

… He is a terror; and not just in this vicinity. His mere name carries a shudder with it to distant lands--just he mere name; and when he frowns, the shadow of it falls as far as Rome, and the chickens go to ***roost*** *an hour before schedule time…*

(Personal Recollections of Joan of Arc by Mark Twain)

… We spent one pleasant day skirting along the Isles of Greece. They are very mountainous. Their prevailing tints are gray and brown, approaching to red. Little white villages surrounded by trees, nestle in the valleys or ***roost*** *upon the lofty perpendicular sea-walls…*

(The Innocents Abroad by Mark Twain)

Word	Lexical-Semantic Field	D	S
Resurrect	Bodily Functions and Care	10,09	0,77

… The adoption of cremation would relieve us of a muck of threadbare burial-witticisms; but, on the other hand, it would ***resurrect*** *a lot of mildewed old cremation-jokes that have had a rest for two thousand years…*

(Life on the Mississippi by Mark Twain)

… I will dig up the Romans, I will ***resurrect*** *the Greeks, I will furnish the government, for ten millions a year, ten thousand veterans drawn from the victorious legions of all the ages--soldiers that will chase Indians year in and year out on materialized horses, and cost never a cent for rations or repairs…*

(The American Claimant by Mark Twain)

… An effort was made to ***resurrect*** *it, with the proposed advantage of a telling new title, and Mr. F. said that The Phenix would be just the name for it, because it would give the idea of a resurrection from its dead ashes in a new and undreamed of condition of splendor…*

(Roughing It by Mark Twain)

Word	Lexical-Semantic Field	D	S
Tally	Communication	5,1	0,61

… He sat down and puzzled over these things a good while, but kept muttering, "It's no use; I can't understand it. They don't ***tally*** *right, and yet I'll swear the names and dates are right, and so of course they OUGHT to* ***tally****. I never labeled one of these thing carelessly in my life. There is a most extraordinary mystery here…"*
(The Tragedy of Pudd'nhead Wilson by Mark Twain)

… Do they ***tally****?"*
The foreman responded: "Perfectly."
"Now examine this pantograph, taken at eight months, and also marked A. Does it ***tally*** *with the other two?"*
The surprised response was: "NO – THEY DIFFER WIDELY!"
"You are quite right. Now take these two pantographs of B's autograph, marked five months and seven months. Do they ***tally*** *with each other?"*
"Yes – perfectly."
"Take this third pantograph marked B, eight months. Does it ***tally*** *with B's other two?"*
"BY NO MEANS!.."
(The Tragedy of Pudd'nhead Wilson by Mark Twain)

6. Conclusions

Calculated frequency distribution of lexical-semantic groups of verbs in authors' texts of English prose makes it possible to select the lexical-semantic structure of author's idiolect. The frequencies of some lexical-semantic fields may vary considerably for different authors, due to divergences in the author's idiolect, and this is a linguo-stylistical characteristic of author's texts. The largest classification potential is given by the frequency distribution of lexical-semantic fields that are formed by the verbs, for which the ratio of the words frequency in the author's texts and the ones of linguo-stylistic norm exceeds a certain threshold. The change of the frequency distribution of words in the semantic field of verbs for different authors concerns the words of both high and low frequency. However, in the area of low-frequency words the variation of the same words for the texts of different authors is several times more pronounced. Thus, the area of high-frequency words contains the words of nominative, stylistically neutral type, and the area of author's idiolect is located on the periphery of the lexical-semantic field. The constants of semantic distances that characterize the area of author's idiolect in the structure of lexical-semantic fields do not depend on the quantity and quality of author's texts and represent the fundamental lexical-semantic regularities of author's style.

In our further studies we plan to explore in more detail the markers of author's idiolect and the stylometric potential of lexical-semantic fields.

References

Brinton, Laurel J. (2000). *The structure of modern English: a linguistic introduction.* Amsterdam/Philadelphia: John Benjamins Publishing Company

Corson, David (1995). *Using English Words.* Dordrecht-Boston: Kluwer.

Crow, J.T., Quigley, J.R.A. (1985) Semantic Field Approach to Passive Vocabulary Acquisition for Reading Comprehension. *TESOL Quarterly 19(3), 497–513.*

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fisiak, J. (ed.) (1985). *Historical Semantics - Historical Word-Formation*. Berlin-New York-Amsterdam: Mouton

Gliozzo, A., Strapparava, C. (2009). *Semantic Domains in Computational Linguistics*. Berlin: Springer.

Gol'dberg, V.B. (1988). *Kontrastivnyj analiz leksiko-semanticheskih grup (na materiale angliyskogo, russkogo i nemetskogo yazykov)* [Contrastive analysis of lexical-semantic groups (on the base of English, Russian, and German languages]. Tambov: TGPI.

Jackson,H., Amvela, E.Z. (2000). *Words, meaning and vocabulary: An introduction to modern English lexicology*. New York: Continuum.

Kuznetsova, E.V. (ed.) (1989). *Leksiko-semanticheskie gruppy russkih glagolov* [Lexical-semantic groups of English verbs]. Irkutsk: Izdatel'stvo Irkutskogo Universiteta.

Popov, Z.D. (ed.) (1989). *Polevye struktury v sisteme yazyka* [Field structures in language system]. Voronezh.: Izdatel'stvo Voronezhskogo Universiteta.

Ufimtseva, A.A. (1962). *Opyt izucheniya leksiki kak sistemy (na materiale angliyskogo yazyka)* [The experience of studying language as a system (on the base of English language)]. Moscow: Izdatel'stvo Akademii nauk SSSR.

Verdieva, Z.N. (1986). *Semanticheskie polya v sovremennom angliyskom yazyke* [Semantic fields in modern English language]. Moscow: Vysshaya shkola.

Appendix

The frequency structure of the lexical-semantic fields of the authors' texts

№	Author	Lexical-semantic fields (the first number is the P_f value of the semantic field, the second number is the total amount of words of the semantic field found in the text)							
		Bodily Functions and Care	Cognition	Competition	Contact	Emotion	Perception	Social interaction	Weather
		Change	Communication	Consumption	Creation	Motion	Possession	Stative	
1.	Neutral style	0.0411 15131145	0.0754 27753548	0.0448 16499044	0.0864 31831553	0.0299 11012303	0.0519 19129659	0.1178 43377262	0.0033 1222883
		0.1142 42057787	0.0993 36549604	0.0207 7639096	0.0436 16059961	0.0879 32355365	0.0737 27140323	0.1094 40268327	
2.	Scott, Walter, Sir, 1771-1832	0.0411 76203	0.0810 150172	0.0458 84947	0.0807 149734	0.0272 50574	0.0467 86607	0.1248 231311	0.0022 4190
		0.1070 198409	0.1088 201659	0.0219 40622	0.0409 75893	0.0802 148649	0.0809 149995	0.1102 204336	
3.	Austen, Jane, 1775-1817	0.0443 17925	0.0870 35174	0.0311 12580	0.0598 24200	0.0375 15189	0.0493 19945	0.1294 52333	0.0010 425
		0.1024 41432	0.1078 43600	0.0207 8379	0.0428 17307	0.0737 29820	0.0863 34888	0.1262 51050	
4.	Lytton, Edward Bulwer, 1803-1873	0.0437 91324	0.0757 158184	0.0425 88853	0.0792 165438	0.0362 75575	0.0522 109091	0.1205 251694	0.0028 5903
		0.1121 234093	0.1057 220757	0.0189 39648	0.0409 85541	0.0825 172251	0.0783 163537	0.1078 225153	
5.	Disraeli, Benjamin, 1804-1881	0.0411 10126	0.0800 19722	0.0401 9882	0.0740 18251	0.0333 8208	0.0497 12263	0.1294 31889	0.0023 587
		0.1124 27697	0.1031 25400	0.0192 4744	0.0420 10355	0.0825 20328	0.0776 19141	0.1126 27759	
6.	Poe, Edgar Allan,	0.0430 8741	0.0720 14613	0.0382 7754	0.0939 19066	0.0267 5427	0.0553 11236	0.1106 22447	0.0041 844

	1809-1849	0.1226 24890	0.0963 19542	0.0148 3014	0.0455 9236	0.0860 17457	0.0770 15630	0.1134 23019	
7.	Gaskell, Elizabeth Cleghorn, 1810-1865	0.0444 36021	0.0819 66371	0.0355 28751	0.0790 63987	0.0367 29767	0.0573 46434	0.1154 93449	0.0020 1659
		0.1143 92629	0.1005 81393	0.0209 16977	0.0402 32609	0.0878 71163	0.0705 57121	0.1128 91402	
8.	Thackeray, William Makepeace, 1811-1863	0.0478 41555	0.0710 61612	0.0450 39116	0.0877 76109	0.0285 24740	0.0479 41605	0.1228 106584	0.0024 2154
		0.1065 92420	0.1062 92209	0.0234 20355	0.0429 37262	0.0880 76399	0.0729 63282	0.1063 92246	
9.	Dickens, Charles, 1812-1870	0.0475 135906	0.0725 207025	0.0414 118489	0.0882 252136	0.0273 78028	0.0552 157785	0.1154 329526	0.0027 7962
		0.1126 321533	0.1011 288827	0.0198 56739	0.0380 108524	0.0933 266651	0.0734 209790	0.1108 316558	
10.	Trollope, Anthony, 1815-1882	0.0396 91407	0.0870 200815	0.0378 87383	0.0648 149603	0.0296 68514	0.0452 104523	0.1327 306390	0.0011 2547
		0.1052 242856	0.1005 232152	0.0209 48395	0.0428 98830	0.0840 193875	0.0851 196523	0.1230 284025	
11.	Bronte, Charlotte, 1816-1855	0.0455 7079	0.0771 12005	0.0378 5881	0.0873 13584	0.0361 5619	0.0606 9435	0.1086 16897	0.0038 596
		0.1128 17562	0.1020 15869	0.0200 3121	0.0391 6083	0.0868 13518	0.0710 11052	0.1109 17264	
12.	Bronte, Emily Jane, 1818-1848	0.0456 3163	0.0725 5030	0.0375 2605	0.0864 5996	0.0358 2484	0.0559 3881	0.1096 7602	0.0037 259
		0.1101 7637	0.1071 7430	0.0243 1688	0.0398 2762	0.0897 6220	0.0711 4936	0.1102 7648	
13.	Eliot, George, 1819-1880	0.0468 30510	0.0836 54467	0.0374 24380	0.0781 50860	0.0349 22733	0.0582 37909	0.1148 74820	0.0017 1168
		0.1136 73985	0.0960 62549	0.0207 13535	0.0415 27026	0.0850 55361	0.0733 47793	0.1137 74100	
14.	Bronte, Anne, 1820-1849	0.0455 5826	0.0879 11248	0.0358 4585	0.0700 8961	0.0394 5044	0.0534 6836	0.1170 14964	0.0027 355

		0.1082 13843	0.1064 13607	0.0217 2787	0.0413 5281	0.0859 10984	0.0749 9577	0.1092 13965	
15.	Collins, Wilkie, 1824-1899	0.0432 93815	0.0806 174834	0.0384 83276	0.0735 159326	0.0299 64859	0.0573 124263	0.1214 263257	0.0017 3819
		0.1098 238086	0.1133 245688	0.0194 42136	0.0393 85282	0.0823 178549	0.0740 160466	0.1153 249973	
16.	Meredith, George, 1828-1909	0.0466 61737	0.0760 100661	0.0406 53822	0.0833 110253	0.0321 42497	0.0538 71292	0.1179 156038	0.0036 4862
		0.1068 141407	0.1088 144021	0.0205 27198	0.0435 57565	0.0859 113750	0.0739 97823	0.1058 140059	
17.	Carrol, Lewis, 1832-1898	0.0460 3818	0.0799 6627	0.0442 3669	0.0835 6926	0.0279 2315	0.0565 4684	0.1102 9135	0.0019 158
		0.1130 9368	0.1100 9123	0.0210 1744	0.0423 3513	0.0968 8025	0.0596 4943	0.1065 8831	
18.	Butler, Samuel, 1835-1902	0.0430 17976	0.0878 36676	0.0370 15446	0.0717 29935	0.0277 11566	0.0496 20714	0.1234 51502	0.0017 717
		0.1183 49407	0.0940 39257	0.0185 7721	0.0484 20209	0.0757 31607	0.0820 34252	0.1206 50355	
19.	Twain, Mark, 1835-1910	0.0388 52388	0.0787 106169	0.0439 59240	0.0916 123587	0.0291 39350	0.0527 71040	0.1147 154696	0.0035 4738
		0.1170 157758	0.0960 129528	0.0208 28166	0.0441 59576	0.0959 129388	0.0673 90851	0.1049 141495	
20.	Hardy, Thomas, 1840-1928	0.0412 50775	0.0742 91523	0.0438 53966	0.0837 103173	0.0287 35370	0.0546 67287	0.1160 142917	0.0031 3941
		0.1159 142859	0.0925 113986	0.0180 22237	0.0429 52881	0.0985 121463	0.0690 85084	0.1173 144511	
21.	Stevenson, Robert Louis, 1850-1894	0.0426 55764	0.0743 97101	0.0438 57359	0.0861 112513	0.0307 40212	0.0523 68401	0.1170 152899	0.0039 5154
		0.1140 149053	0.0954 124682	0.0199 26014	0.0441 57630	0.0918 120062	0.0737 96361	0.1097 143415	
22.	Wilde, Oscar, 1854-1900	0.0484 12244	0.0778 19663	0.0408 10308	0.0806 20352	0.0334 8452	0.0528 13348	0.1224 30920	0.0027 682

		0.1142 28859	0.0932 23537	0.0218 5515	0.0486 12282	0.0819 20681	0.0677 17115	0.1130 28543	
23.	Shaw, Bernard, 1856-1950	0.0444 2860	0.0832 53622	0.0425 27435	0.0763 49183	0.0318 20519	0.0502 32347	0.1255 80896	0.0015 1009
		0.1088 70101	0.0975 62843	0.0209 13523	0.0446 28770	0.0910 58666	0.0713 45939	0.1096 70633	
24.	Conrad, Joseph, 1857-1924	0.0436 47674	0.0713 77930	0.0442 48286	0.0950 103778	0.0310 33931	0.0590 64514	0.1084 118447	0.0044 4849
		0.1186 129543	0.0966 105488	0.0190 20748	0.0379 41441	0.0984 107519	0.0644 70384	0.1074 117356	
25.	Doyle, Arthur Conan, Sir, 1859-1930	0.0445 45279	0.0688 69929	0.0508 51651	0.0938 95318	0.0240 24479	0.0532 54135	0.1138 115654	0.0035 3643
		0.1166 118534	0.0921 93613	0.0186 18921	0.0399 40622	0.0967 98285	0.0724 73587	0.1106 112389	
26.	Kipling, Rudyard, 1865-1936	0.0432 36120	0.0691 57758	0.0488 40836	0.1057 88379	0.0276 23132	0.0501 41901	0.1092 91259	0.0040 3408
		0.1172 97962	0.0907 75784	0.0233 19510	0.0435 36412	0.1033 86368	0.0635 53111	0.0999 83474	
27.	Wells, Herbert George, 1866-1946	0.0393 38190	0.0788 76593	0.0445 43284	0.0892 86605	0.0305 29623	0.0572 55585	0.1077 104624	0.0037 3688
		0.1271 123400	0.0920 89370	0.0201 19561	0.0441 42846	0.0938 91084	0.0632 61413	0.1080 104927	
28.	Galsworthy, John, 1867-1933	0.0470 46617	0.0728 72164	0.0446 44179	0.0939 93037	0.0335 33213	0.0655 64959	0.1045 103588	0.0030 2985
		0.1155 114444	0.0853 84511	0.0225 22366	0.0386 38249	0.1008 99850	0.0620 61446	0.1098 108833	
29.	Dreiser, Theodore, 1871-1945	0.0385 11579	0.0871 26192	0.0383 11520	0.0778 23395	0.0331 9969	0.0538 16194	0.1175 35350	0.0024 738
		0.1119 33644	0.0989 29734	0.0221 6657	0.0425 12790	0.0880 26480	0.0741 22281	0.1134 34102	
30.	London, Jack, 1876-1916	0.0409 56009	0.0675 92251	0.0482 65854	0.1065 145603	0.0291 39808	0.0528 72200	0.1069 146138	0.0045 6257

		0.1166 159395	0.0886 121054	0.0230 31494	0.0410 56017	0.1064 145429	0.0643 87967	0.1029 140640	
31.	Woolf, Virginia, 1882-1941	0.0436 8755	0.0777 15607	0.0417 8384	0.0924 18544	0.0316 6352	0.0615 12347	0.1023 20528	0.0036 724
		0.1192 23917	0.1013 20328	0.0177 3559	0.0421 8463	0.0914 18336	0.0617 12392	0.1115 22371	
32.	Joyce, James, 1882-1941	0.0456 10434	0.0676 15468	0.0431 9861	0.1043 23849	0.0300 6861	0.0561 12828	0.1039 23758	0.0047 1089
		0.1192 27242	0.0992 22667	0.0222 5094	0.0398 9113	0.1008 23052	0.0632 14449	0.0994 22716	
33.	Lawrence, David Herbert, 1885-1930	0.0498 15739	0.0728 23007	0.0424 13407	0.0952 30071	0.0386 12198	0.0608 19208	0.1018 32172	0.0039 1238
		0.1161 36675	0.0853 26961	0.0262 8292	0.0384 12130	0.0977 30885	0.0590 18654	0.1115 35221	

Glottometrics 25, 2013,85 - 93

Stratification in texts

Gabriel Altmann, Lüdenscheid
Ioan-Iovitz Popescu, Bucharest[1]
Dan Zotta, Bucharest

Abstract. Stratification in texts is a process analogous to those in nature and culture. Though one cannot identify the individual strata in every case, it is possible to show the rise of this phenomenon in mathematical terms and apply the resulting formulas to examples from textology and music. It allows also to study the evolution of a writer, text sort, language or music.

***Keywords**: stratification, text, music, differential equations*

Stratification is a property inherent to all material things. Modern science, especially physics, has shown it in innumerable cases and the process of discovery continues incessantly. But even human artefacts have strata. Some of them are created by concept formation in order to give us orientation and a basis for analysis, other ones are necessary for the artificial thing itself in order to be considered as such, e.g. colour or grey strata for pictures and paintings; pitch height, length, intensity, rhythm and colour for music; words, blanks, punctuation for writing; segmental and suprasegmental strata for spoken language, etc. Long time ago linguists stated that an utterance is stratified, even if is it written: The text is not a homogeneous mass and even its simple understanding requires a multistratal analysis which is automatized in the mother tongue and must be learned laboriously in foreign ones. Strata like sentence, clause, phrase, word, morpheme, syllable, phoneme are taught even in the school and they have the agreeable property that each stratum is linked with the neighbouring (higher or lower) stratum by means of Menzerath's law. Though this is a stochastic law, its existence contributes to the good conscience of linguistics to be a science just like its great sister, the physics.

But it would be foolish to suppose that our way ends at this point. There are at least three directions in which we can continue our way of stratification research. The first is the zone between text and its components. There are some purposefully created layers like chapters, paragraphs, acts in the stage play, decided by the author; other ones have been discovered and can be captured only analytically: up to now there is the "hreb" or sentence aggregate discovered by Hřebíček (1997) represented by all sentences of a text containing a synonym, a reference or some other identifying semantic connection between sentences; and the motif discovered by Köhler (2006, 2008a,b) consisting of non-decreasing sequences of some measured entities. The motif is a formal entity, hreb is rather a semantic one.

The second possibility is the classification of different entities in many different subclasses - a speciality and final aim of qualitative linguistics: there are parts-of-speech, grammatical categories, different types of morphemes, phrases, clauses, sentences, i.e. even within one class - which are merely Menzerathian chain-links in the hierarchy -, there are different substrata that can be identified formally or semantically. Though the author may select them deliberately, it would be very courageous to suppose that (s)he does not act in agreement with a law. One of such laws is e.g. Zipf's law in all its forms.

[1] Address correspondence to: Ioan-Iovitz Popescu, e-mail: iovitzu@gmail.com.

A third research possibility is the investigation of the number of sub-strata that occur within one stratum. An analogy with nuclear physics or microbiology is evident. We “open” the atom (being an element of a stratum) to see whether and what kinds of entities are in its interior; we open the DNA to see what it consists of. In linguistics, we arrived at a point at which we can at least state *how many* substrata are contained in a homogeneous stratum, e.g. that of words. We can, of course, state the frequency of word classes and see that synsemantics are more frequent than autosemantics, that short words are more frequent than long words but this all are properties constructed conceptually by us and follow some laws known from synergetic linguistics. But even these classes are combined in such a way that no grammar or semantics can approach them. The substrata may arise stepwise: by change of theme, by pauses in writing, by the development of the story, etc., but they can also be eliminated: the author may correct the text, the editor may strive for uniformity, etc. The reader/hearer need not even perceive a difference and most probably none of these text creators (writer, editor, reader) is conscious of something like strata in text.

The discovery and identification of strata in text - with whatever unit - is a problem for the far future. Though in stage plays there is a manifest stratification represented by different persons, other kinds are not easy to be identified. In some other domains of language it is easier to find strata, for example in the monolingual dictionary where each word is defined in terms of words which have a more general meaning. E.g. a “revolver” is a “weapon”; the weapon is an “instrument”; the instrument is an “artefact”; the artefact is a “thing”. In this way one obtains strata of generality. Besides, it is evident that the more general the meaning, the fewer words are contained in the stratum. In the same way one can obtain strata of concreteness-abstractness, emotionality, metaphor, imagery, dogmatism, etc. known from psycholinguistics.

Nevertheless, there is a possibility of tracing down at least the existence of strata and their number in text using a mathematical reasoning. Unfortunately, it must be applied for each linguistic entity separately: if there is stratification in the vocabulary of the text, it need not exist e.g. for sentence length. In the second stage of the research it will also be necessary to substantiate the existence of strata linguistically.

We start from the following assumptions: The writer begins to write. At a certain (unknown) point in text he changes his strategy concerning certain units and continues with a slightly different strategy. Then somewhere he changes again to a new strategy that means, he performs a change of the change. In mathematical terms, the first change is $dy/dx = y'$; the change of this regime means simply a new change, i.e. $d^2y/dx^2 = y''$, etc. It is a matter of empirical fact that the function y and its derivatives obey a linear relationship, as will be shown in continuation.

Let us model a linguistic phenomenon which can be ranked, scaled or weighted. If the values converge to a constant (e.g. absolute frequencies converge to 1, relative frequencies to 0), we can always use the approach

(1) $f(x) = C + y(x),$

C being a real positive constant.

If we suppose the existence of stratification and restrict ourselves to two strata, we may express this assumption by

(2) $y(x) = A_1exp(k_1x) + A_2exp(k_2x)$

used successfully to rank-frequency sequences proposed as an alternative to Zipf’s law which does not capture stratification (cf. Popescu, Altmann, Köhler 2010). The derivatives of (2) are

$$y' = A_1k_1exp(k_1x) + A_2k_2exp(k_2x)$$

(3)

$$y'' = A_1\, k_1^2exp(k_1x) + A_2\, k_2^2exp(k_2x).$$

From (2) and (3) we have the following differential equation

(4) $y'' - (k_1 + k_2)y' + (k_1\, k_2)y = 0$

where $k_1 \neq k_2$ are real numbers. Denoting further by

$$p = - (k_1 + k_2)$$

$$q = (k_1k_2)$$

we get the standard form of the 2nd order linear homogeneous ordinary differential equation with constant coefficients

(5) $y'' + py' + qy = 0.$

Conversely, let's start from this equation

$$y'' + py' + qy = 0$$

where p and q are real numbers, and look for a solution

$$y = exp(kx).$$

Inserting it into the above equation we have

$$(k^2 + pk + q)exp(kx) = 0$$

or, because $exp(kx)$ is never zero, we obtain the so called *characteristic equation*

$$k^2 + pk + q = 0$$

with the *discriminant*

$$\Delta = p^2 - 4q$$

If $\Delta > 0$, the characteristic equation has two real and distinct solutions, k_1 and k_2, given by

$$k_1 = (-p + \sqrt{}\Delta) / 2$$
$$k_2 = (-p - \sqrt{}\Delta) / 2,$$

hence the corresponding solution of the considered differential equation is

$$y(x) = A_1 exp(k_1 x) + A_2 exp(k_2 x)$$

with A_1 and A_2 to be determined from initial conditions. Obviously,

$$p = -(k_1 + k_2)$$

$$q = k_1 k_2.$$

To conclude, the fitting function consisting of two exponential components represents the solution for the case $\Delta > 0$, of the 2nd order linear homogeneous ordinary differential equation with constant coefficients, see more, for instance, at http://www.efunda.com/math/ode/linearode_consthomo.cfm

The generalization is straightforward: the fitting function consisting of *n* exponential components represents the solution of the *n*th order linear homogeneous ordinary differential equation with constant coefficients, for the case when all solutions of the characteristic equation are real and distinct numbers.

The above solution of the stratification problem has the advantage of telling us the number of strata of the given unit in the given text (cf. Popescu, Altmann, Köhler (2010); Popescu, Čech, Altmann (2011); Popescu, Mačutek, Altmann (2009); Popescu, Martináková-Rendeková, Altmann (2012)). However, it does not enable us to identify the strata.

Take as an example the word form frequency in Goethe's poem *Erlkönig* ranked in decreasing order as shown in Table 1.

Table 1
Ranked word form frequencies in Erlkönig by Goethe

x	f_x	***x***	f_x	***x***	f_x	***x***	f_x
1	11	32	2	63	1	94	1
2	9	33	2	64	1	95	1
3	9	34	2	65	1	96	1
4	7	35	2	66	1	97	1
5	6	36	2	67	1	98	1
6	6	37	2	68	1	99	1
7	5	38	2	69	1	100	1
8	5	39	2	70	1	101	1
9	4	40	1	71	1	102	1
10	4	41	1	72	1	103	1
11	4	42	1	73	1	104	1
12	4	43	1	74	1	105	1
13	4	44	1	75	1	106	1
14	4	45	1	76	1	107	1
15	4	46	1	77	1	108	1
16	3	47	1	78	1	109	1
17	3	48	1	79	1	110	1
18	3	49	1	80	1	111	1

19	3	50	1	81	1	112	1
20	3	51	1	82	1	113	1
21	3	52	1	83	1	114	1
22	2	53	1	84	1	115	1
23	2	54	1	85	1	116	1
24	2	55	1	86	1	117	1
25	2	56	1	87	1	118	1
26	2	57	1	88	1	119	1
27	2	58	1	89	1	120	1
28	2	59	1	90	1	121	1
29	2	60	1	91	1	122	1
30	2	61	1	92	1	123	1
31	2	62	1	93	1	124	1

If we fit the data with a function having a sum of three exponential functions in its expression, that is with

(6) $f(x) = 1 + A_1 exp(k_1 x) + A_2 exp(k_2 x) + A_3 exp(k_3 x)$,

we obtain the results presented in Figure 1 with the determination coefficient $R^2 = 0.9824$.

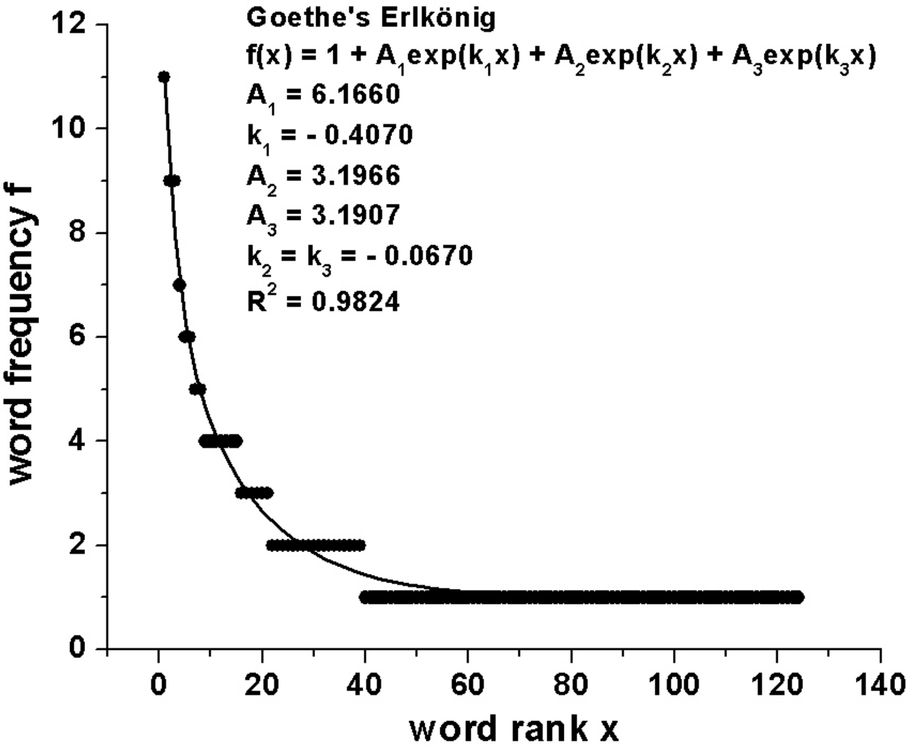

Figure 1. Fitting the word rank-frequencies in *Erlkönig* by Goethe with a function of type (6) indicates two strata

As can be seen, the parameters in the exponent k_2 and k_3 are equal hence we can omit one component and add the corresponding multiplicative constants $A_2 + A_3$. One obtains finally

$$f(x) = 1 + 6.1160_1 exp(-0.4070x) + 6.3872 exp(-0.0670x)$$

We can conclude that concerning word forms the poem has two strata. The function can be enlarged to more components - following from the differential equation of *n*-th order - but in case that some of the parameters yield non-realistic values, e.g. too great ones, one should omit them as outliers. It is to be noted that using the exponential function with one component we obtain still very good fitting results ($R^2 = 0.9648$) but we do not learn how many components there are. Hence the above method should be started always with several components. The next (qualitative) step would be the *identification* of the two strata, but this is more or less a philological affair.

This technique has been successfully used in many cases cf. e.g. Tuzzi, Popescu, Altmann, (2010: Ch. 5.1, 5.2), Nemcová, Popescu, Altmann (2010), Fan, Altmann (2010), Beliankou, Köhler (2010), Sanada, Altmann (2009), Laufer, Nemcová (2009), Kelih (2009), Knight (2013), etc. It is to be noted that this approach does not yield a “text model”, it is merely a means to find the number of strata. There are always functions which would yield better fittings but their interpretation is quite different.

Let us consider some musical examples in which we found different stratifications.

Consider first the pitch rank-frequencies in Stravinsky’s *The Firebird Suite*. Beginning with three components we obtain the result presented in Figure 2. As can be seen, all parameters in the exponent are identical, hence there is only one stratum and the computed rank-frequencies abide by $f_x = 1 + 265.9074 exp(-0.0686x)$ where the parameters A_i were summed up.

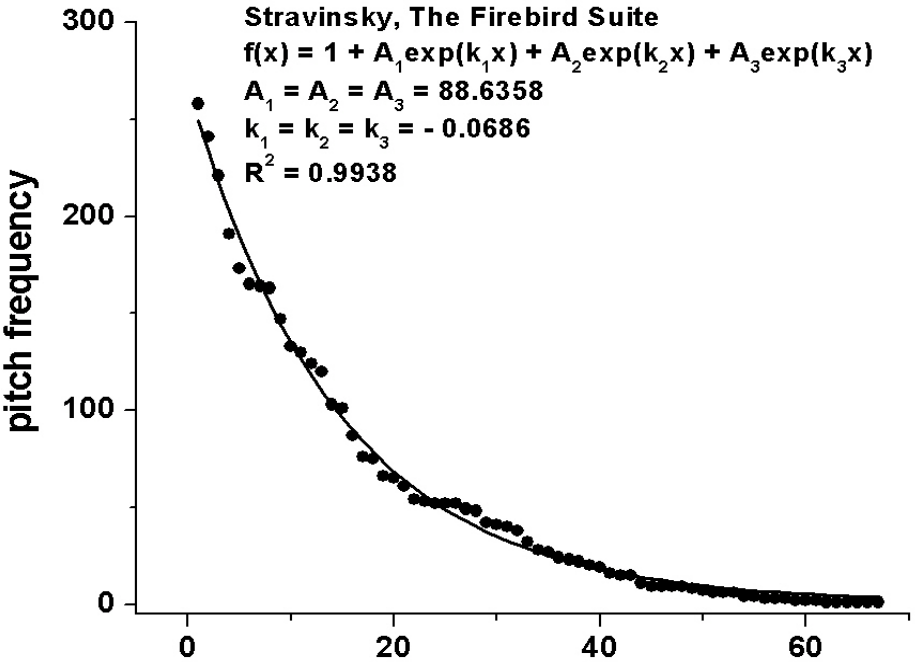

Figure 2. Fitting the pitch rank-frequencies in Stravinsky’s *The Firebird Suite* with a function of type (6) indicates a single stratum.

In Beethoven’s *Sonata No. 5*, presented in Figure 3, we find two strata because $k_2 = k_3$, hence $f_x = 1 + 93.1319 exp(-0.7054x) + 446.3417 exp(-0,0594x)$.

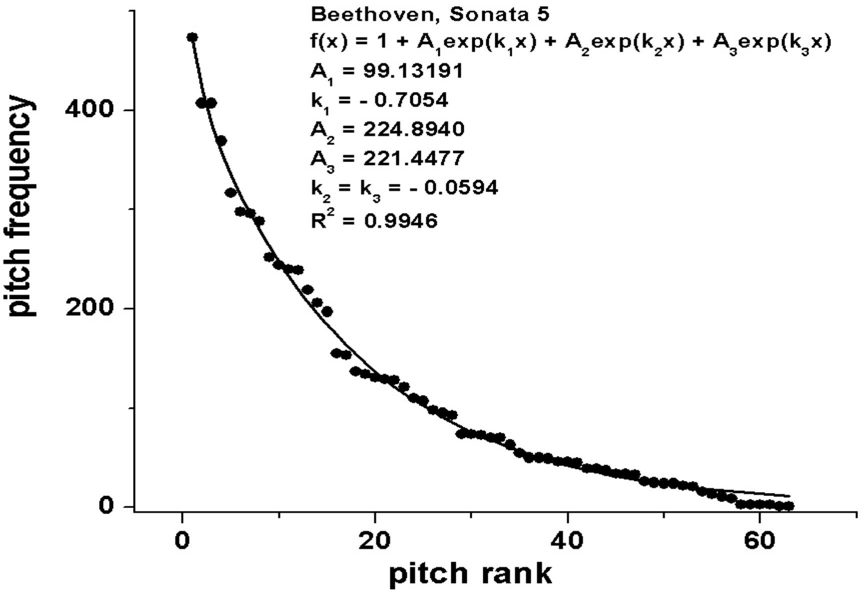

Figure 3. Fitting the pitch rank-frequencies in Beethoven's S*onata 5* with a function of type (6) indicates two strata.

A critical case is Mozart's *Sonata A major K.331* presented in Figure 4 indicating three strata but actually, there are only two strata because the excessively high multiplicative constant A_1 = 16869.3891 value corresponds to an outlier. If we compute directly two strata, we obtain f_x = 1 + 822.0111*exp*(-0.0853*x*) + 2322.1284*exp*(-2.2435*x*) with R^2 = 0.9942. But even here we have still A_2 = 2322.1284 which is more than twice the observed f_1 = 1002. If we consider it an outlier, we obtain the monostratal fitting in form f_x = 1 + 923.0682*exp*(-0.0951*x*) with R^2 = 0.9782 which is very satisfactory. This case shows that not all data can be satisfactorily checked; perhaps Mozart's Sonata had to be partitioned in three parts and all analyzed separately.

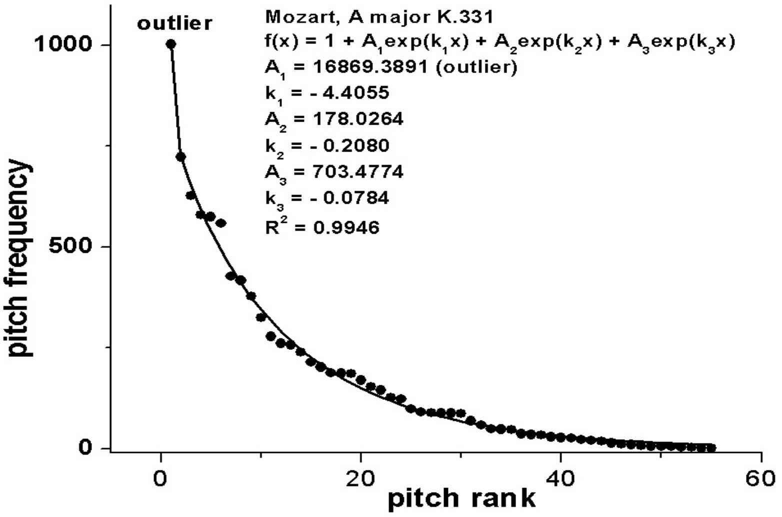

Figure 4. Fitting the pitch rank-frequencies in Mozart's *Sonata A major K.331* with a function of type (6) indicates three strata

Summary

Since Zipf's power function or the corresponding zeta distribution do not always capture satisfactorily the sequence of ranked frequencies, a more satisfactory solution is a sum of exponential expressions which at the same time gives information about the number of strata in the frequencies. The aim of this article was to show that the background linguistic hypothesis concerning changes in the strategy of text creation leads to a differential equation of n-th order. Usually a third order is sufficient but in many cases the fitting itself shows that the order can be reduced. If a text is monolithic, it contains only one stratum. Unfortunately, there are so many aspects of human artefacts - and their number increases with the progress of science - that an enormous number of analyses will be necessary in order to get a more solid basis in this research.

Stratification is, as a matter of fact, a special aspect of self-organization. If something evolves, it gets more complex. Languages and texts are no exceptions. In systems theoretical view, strata are sometimes subsystems evolving in the neighbourhood of and interdependence with other subsystems. For language it is a known fact but for texts it is not that evident because text is a ready product. However, text represents at least two entities: the entity created by the author and the entity interpreted by the reader. The second entity differs with every reader. It is not identical with the written entity - otherwise no "literary science" would exist - and it may change even with one reader. The interpreted text gets part of the mind of the reader and evolves as his mind evolves.

Stratification in language and text has some intersections with diversification, one of the Zipfian forces (cf. Köhler 2005). Everything diversifies in language; the language community and the hearer slow this process down, otherwise the communication would break down. But diversified entities create dialects, sociolects, idiolects, new languages, different presentations of stage plays, new vistas of texts, etc. As a matter of fact, the present article shows merely the stratification process but does not identify the strata.

Acknowledgments. The authors are most grateful to Professor Cristian Calude for his valuable review of this article.

References

Beliankou A., Köhler R. (2010). The distribution of parts-of-speach in Russian texts. *Glottometrics 20, 59-69.*

Fan, F., Altmann G. (2010). On meaning diversification in English. In: *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn*: *223-233.* Frankfurt: Lang.

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

Kelih, E. (2009). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle, *Glottometrics 18, 52-68.*

Knight, R. (2013). Laws Governing Rank Frequency and Stratification in English Texts. *Glottometrics 25 (present issue).*

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

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

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

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

Laufer, J., Nemcová E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25.*

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

Nemcová, E., Popescu, I.-I., Altmann, G. (2001). Word associations in French. In: Berndt, A., Böcker, J. (eds.), *Sprachlehrforschung: Theorie und Empirie: 223-237*. Frankfurt: Lang.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law — another view. *Quality and Quantity 44(4), 713-731.*

Popescu, I.-I., Čech, R., Altmann, G. (2011). On stratification in poetry. *Glottometrics 21, 54-59.*

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

Popescu, I.-I., Martináková-Rendeková Z., Altmann G. (2012). Stratification in musical texts based on rank-frequency distribution of tone pitches, *Glottometrics 24, 25-40.*

Sanada H., Altmann G. (2009). Diversification of postpositions in Japanese, *Glottometrics 19, 70-79.*

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

Glottometrics 25, 2013, 94 - 96

REPORT

Research activities at the Department of General Linguistics of the Philosophical Faculty of Palacký University in Olomouc, Czech Republic

Martina Benešová, Dan Faltýnek

In the following lines, we would like to give an account of the research at the Department of General Linguistics of the Philosophical Faculty of Palacký University (DGL) and the possibilities of its potential widening. The reason for the genesis and updating of this presentation is, last but not least, the motivation of those potentially interested in the participation in the research. We are interested in engaging scientists from other universities, specialists and even students looking for an opportunity to take part in the research. At the Department of General Linguistics of the Philosophical Faculty of Palacký University, there has been created a research team focusing on using quantitative methods in the fields where they have already been corroborated or where positive results of such an approach are expected. Under the leadership of Jan Andres and Jan Kořenský, the analysis of a text exhibiting the Menzerath-Altmann law is elaborated here; in that respect the hypothesis on the language fractality is tested. The team of DGL develops the input of Luděk Hřebíček into the theory of language/text fractality, follows the current foreign trends in the cooperation with other universities, and we attempt to test newly formulated hypotheses (e.g. the works of Radek Čech). In any case, the close contact of mathematicians and linguists at DGL already now reaps the harvest of interesting results:

(a) in the mathematical field, we elaborate the concept of these regularities as aspects of the fractal nature of the language/text in the relation to different ways of understanding fractality (cf. the series of papers by Martina Benešová and Jan Andres);

(b) from the point of view of the regularization of text segmentation – setting up the methodology for determining language units - we formulate hypotheses of the reasons for manifesting the above mentioned statistical tendencies in the text (semantic, neural etc.);

(c) the algorithm of the quantitative analysis of a linguistic sample has been elaborated to the above mentioned aim by Martina Benešová for both linguists and mathematicians.

Under the terms of quantitative approaches, the team gradually tests the hypotheses of the reasons for manifesting the Menzerath-Altmann law in the text. The approach of the team is based on the axiomatization of this law manifestation in naturally-produced texts. Using this axiomatization, it is possible to test the grammar adequacy: if the grammatical description is not adequate, it will determine its units (constructs and constituents – e.g. morphemes and words, words and sentences) so that the relation manifested by the Menzerath-Altmann law does not even appear after the text segmentation and in the following analysis. The team regards this procedure to be useful for assessing grammatical description of the language (but also e.g. of the genetic coding). The validity of one grammar of the Czech case has been tested this way, where the preposition is treated as a part of a complex case structure; if the preposition is not considered a part of speech (it is a constituent of the word), the Menzerath-Altmann law is manifested; otherwise not, cf. (Benešová, 2011), (Faltýnek, 2012).

From the point of view of methodological bases, the team follows several basic assumptions. They are, above all: a naturally-produced spoken or written text of a given length is used for the lexicostatistical analysis from the point of view of the statistical conclusiveness. Considering the unit determination (constructs and constituents), uniform criteria are strictly held on each level (in the opinion of the DGL team members, this fundamental condition has not been sufficiently met so far in similar researches). Under these conditions, the team wants to continue testing the following particular hypotheses:

(a) We set up the text units solely with respect to their sound/acoustic quality, their phonetic form. The units concerned are the sound, syllable, phonological word, intonation unit, replica and text. If a text segmented in this way behaves according to the Menzerath-Altmann law, it is reasonable to suppose that the reason of it must be the processes connected with producing the acoustic signal, or that above all the expression of speech characteristics are the main role players, etc. The answers can be detected preferably in relation with the activity of motoric neural correlates of speech/articulation organs, in relation with cognitive parts for planning activities etc.

(b) The hierarchy, on the other hand, allows us to falsify the hypothesis on semantic or systemic motivation for showing the Menzerath-Altmann law in the text. Quantitative researches performed in this field show that the Menzerath-Altmann law arises due to the text semanticity, cf. Hřebíček (1995, 1997, 2002, 2007), Andres (2010), Andres et al. (2011). If the units of phoneme, morpheme, word, utterance, text (and other) are used in the research, this hypothesis should be open for being tested stepwise. From the point of view of the systemic description above all between the level of the word and morpheme, number of approaches are available – including the zero morpheme, omitting it, setting up the unit thanks to the correlation in the paradigm, setting up the unit under the terms of the responsibility-competence approach to the syntagma with respect to paradigmizing the classes etc. The chosen approach can test the grammar descriptions using such concepts. Such an approach to testing the correctness of the grammar description has not yet been used in linguistics.

Research activities of the DGL team members are associated with researches in the field of biosemiotics and bioinformatics. DNA analyses have proved that the above mentioned lexicostatistical relations can appear even in this field. It can be justified by information characteristics of DNA or its semiotic character. The lexicostatistical team plans to apply the gained methods of the text analysis in the research of DNA. Generally, the capacities of the linguistic analysis of DNA and proteosynthesis are planned to be extended. The team members cooperate in this field with the Faculty of Science of Charles University in Prague, the guarantee of the cooperation is Anton Markoš (Department of Philosophy and History of Science, Charles University in Prague). Coping with DNA/RNA and their parts from the linguistic point of view will be regularized by the team members using linguistic approaches, which is closely connected with careful characterizing the proteosynthesis as a type of semiosis (cf. Markoš, 2010, 2011; Faltýnek 2012). Using quantitative methods will facilitate the testing of this semiotic modeling of the protesynthesis by means of its manifesting statistical dependencies.

The above mentioned approaches will then be gradually used even in the field of psychology or psycholinguistics. Under the terms of the current research at DGL, we prepare the research of the relation of neurophysiological correlates and showing lexicostatistical relations in texts. Fundaments of the mathematical theory of fractals enunciate the functional dependency of qualitative relations of the text on neuro- and psychological structures. This aspect of the relation between cognition and the text will be developed, and its potential and limits will be tested. The research is already at present aimed at aphasic patients. The DGL team wish to follow and to use the results of other researches in this field (e.g., Ferrer-i-Cancho, 2006) and to continue in testing the appearance of the Menzerath-Altmann law in the text. The team members would like in the future to elaborate methods of the disease prediction and diagnostics by means of the automatic analysis of a patient's text.

References

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

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

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

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

Faltýnek, D. (2012). *Semiotic Primitives in Grammar Construction.* Olomouc: Palacký University Olomouc.

Ferrer-i-Cancho, R. (2006). When language breaks into pieces. A conflict between communication through isolated signals and language. *Biosystems 84, 242-253.*

Hřebíček, L. (1995). Text Levels. *Language Constructs, Constituents and the Menzerath–Altmann Law*. Trier: Wissenschaftlicher Verlag Trier.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

Hřebíček, L. (2002). *Stories about Linguistic Experiments with the Text*. Prague: Academia (in Czech).

Hřebíček, L. (2007). *Text in Semantics. The Principle of Compositeness*. Prague: Oriental Institute.

Markoš, A., Faltýnek, D. (2011). Language Metaphors of Life. *Biosemiotics 2(4), 171–200.*

Markoš, A. (ed.) (2010). *Jazyková metafora živého*. Červený Kostelec: Pavel Mervart.

Book Review

James W. Pennebaker (2011): *The Secret Life of Pronouns - What our Words Say about Us*.
New York: Bloomsbury Press
Reviewed by **Jingqi Yan,** Zhejiang University, Hangzhou**.**

Summary

Numerous linguists, when unearthing the secrets of language, focus more on the content words, in particular, the verb and the noun. It is reasonable for them to pay more attention to these words since these words always play the role of head-words in a sentence or a clause and tend to convey more meanings and weight. Nevertheless, the significance of the function words cannot be overshadowed and be regarded as "junk words". The beauty of language lies in its simplicity in that no component in a language is rubbish. The book *The Secret Life of Pronouns - What our Words Say about Us*, written by James W. Pennebaker, intends to present to us, in an original perspective, the secrets of the function words. It is out of one's expectation that such "junk words" can extend to a great subject, raising 10 chapters of discussion. These "junk words" include pronouns, articles, prepositions, auxiliary verbs, negations, conjunctions, quantifiers and common adverbs and they are the words we use most frequently and thus reflect people's mental state, social relations, thinking patterns and personalities to a great extent. Through the author' observation, the individual's usage of these function words follows a constant pattern in terms of frequency and he hypothesizes that by active expressive writing, one can improve his mental health. Via plain but humorous writing, the author was able to attract outsiders as well as the insiders with his vivid and ample examples and testings. One can find this book full of excitement and adventure just like a detective novel, for both are seeing through the whole by simple and small details and particles.

It is common that we all wish to know what other people actually are thinking and whether what they have said is from genuine hearts. The issue of seeking the possible connection and psychological implications between language and the society, language and human beings has long puzzled human beings. This book can be a mystery revealer or on the contrary all the more perplex you. Of the ten chapters of the book, the first two chapters can be the introduction. It presents the reasoning as to why the author himself wrote this book, how to undergo word analysis on the function words, and how can the research be applied to practical use. It gives a short display of the connection between social psychology and language as well as the possible physiological and psychological influence of expressive writing. This statement was evidenced by the computer-based program - Linguistic Inquiry and Word Count (LIWC). With LIWC, the connection between the words people used and the emotive cues underneath these saying were able to form a clear clue. The author tries to identify the possible application of expressive writing into the mental therapy for people with traumatic experience, and build up typical forms of healthy writing. However, no matter what the way research goes, he emphasized that words usage can reflect psychology, but not influence or cause the psychological changes. As the title of the second chapter goes, the author tries to convince the readers to "ignore the content, celebrate the style". In the 20 most commonly used words in English according to the language bank, all of them are function words, taking up 30% of all words, with the word *I* ranking first on the list. Indeed, function words are highly socialized, restricted to the specific time, space and individuals, and they leave a hint on the relationship between the speakers and the listeners, our subtle perception to things and events in our life and on the connections to the culture.

In the following 8 chapters, more details and perspectives are given in the depiction.

Chapter 3 ("The Words of Sex, Age and Power"), Chapter 4 ("Personality: Finding the Person Within") and Chapter 5 ("Emotion Detections") discuss the language of who we are. In Chapter 3, socio-psychological characteristics are taken into account. By using the LIWC, the author has clarified the stereotypes about the language differences between men and women. Age and Social class differences in language behavior have also indicated some overlap with the gender characteristics. Pennebaker concludes that word differences among age, gender and social classes can be separated into two clusters, the "noun cluster", including "articles, nouns, prepositions and big words", and the "pronoun-verb cluster", composing of "personal and impersonal pronouns, auxiliary verbs and certain cognitive words frequently linked to hedge phrases" (p. 61[1]). People who were classified into the "noun-cluster" group are those men, older people and higher social classes, whereas women, younger people and lower social classes are classified into the "pronoun-verb cluster" group. The justification of the emergence of the dichotomy might be explained by power and status.

In Chapter 4, individual psychology about personality differences in words is examined. Here one question arises: Can various individual writing styles reflect personalities? To justify this question, Pennebaker collected thousands of stream-of-consciousness essay samples from people all writing on the same general topic. By analyzing these essays with the factor analysis to "see what clumps of function words emerged" (p. 66), the author has identified three different writing styles corresponding to three thinking patterns: formal, analytic and narrative thinking. Further, there are two psychological small experiments for the readers to try by themselves and get to know their own personality, both available on line. The first one is describing the picture of the bottle in written form and the second is describing a backyard party picture. These experiments had psychological theory foundation by Anna Freud, who claimed that "people naturally project their own thoughts and feelings onto other people and objects" (p. 78). The theme of this chapter seems to demonstrate that you are what you say. The words you say or write disclose your personality. Personality is something stable and your language style, in a certain sense, has a fixed model distinguished from others, reflecting your thinking patterns. Different emotions affect your thinking patterns. That's what Chapter 5 further discusses on the basis of the formal chapter about personality. People in a positive emotion tend to use high rate *we* words, more specific, concrete nouns and references to particular times and places (p. 87). Those in a negative emotion use more I-words, past- and future-tense verbs (p. 87), indicating the immersion of the past and the future and the self-introspection. Anger, different from negative one, is characterized in language by more attention to others, using high rate of second-person and third-person pronouns and more present tense verbs (p. 88). In practical significance, suicidal tendency can be probed through word analysis. In a broader perspective, the emotional fluctuation of a country on a certain major event can be detected.

Chapter 6 to 9 shift the topic to the social situations of people with the cues in function words. Chapter 6, entitled "Lying Words", considers the eternal question—how to detect lies? Research has found that lying required more efforts to justify itself. Several types of deception were explored, including self-deception and intentional deception. Despite the various reasons in motives, language patterns show no salient differences. One marker which best detects lies is the first-person singular pronouns. I-words indicate the self-awareness and self-attention, making people more honest. To be more specific, true statement may contain more "insight words such as realize, understand, think and the like" (p, 129), tend to be more specific with details and include more indicators to other people for verification and fewer verbs. Hierarchy in human society can be autonomous and required. Once a leader is nominated, there would be a natural shift of his language style, distinct from non-leader members. In inspecting

[1] The page number cited in this review is based on the digital book edition with the **epub** format.

function words, the book has suggested that leader language is featured by low use of *I*-words, high use of *we*- words and *you*- words. People's language has a property that it changes in the acting of different role in society. This assumption runs through the following chapters.

Another assumption is that usually, we unconsciously accommodate our language style to the speakers' style or mimic others' speech, causing a resemblance of thinking patterns. This can be evidenced by "the matching of function words" (p. 155), called the Language Style Matching (LSM), and such matching can be accomplished in the first 15 to 30 seconds as conversation begins. The purpose of the matching is to reduce the interaction friction by building up the same interactive framework. According to the variation of attention between both parties, frequent and automatic adjustment will promote the ongoing of the conversation. Such action is like dancing. To quantify LSM, a formula has been introduced. Studies indicate that when one speaker lies, the other party intuitively pays more attention and changes his speaking style more, resulting in a higher LSM, i.e., your brain somehow recognized the lying words and makes some corresponding reactions even though you are not aware of this change. Similar result is found in the multitasking conversation where the distracted pairs share higher LSM. The LSM approach can even assist in our speculation about love relationships. "The conversation dance", as the author called such behavior, can be expounded in the attention focus. The more people synchronize their function words use, the more they pay attention to each other. In consequence, they come to similar thinking patterns.

In chapter 9, the author goes beyond mutual relationship and explores the sense of a person's identity in a group, company and community. Here, *we*- words are taken as an important marker for social identity. The more we use *we* words, a stronger sense of belonging is established. Shift of *we*-words can also be observed in the conversation. In addition, LSM is regarded as a tool to reflect the cohesion of a group. The result is in accordance with the formal chapter, with higher consistency in function words use suggesting closer ties. The practical significance lies in the better tracing of the geographical region of groups

In the previous chapters, Pennebaker presents a panoramic view about the indication of function words in different social psychological phenomenon. In the final chapter, some interesting projects by Pennebaker and his students were introduced. Word analysis can be employed to answer innovative questions. Pennebaker tried using words to track authors, identify the real authorship of Shakespeare's works for example. He has also used word sleuthing to predict wars and terrorist attacks. Pennebaker, in conclusion, has foreseen future application of function word analysis to uncover historic mystery, predict future behavior, and assess students' proficiency.

Critic:

This book chooses words, or to be more specific, function words as the object for research in order to dig deep into people's social and psychological state. Word frequency count has been adopted in his many researches with the tool LIWC. The idea behind LIWC is that "the words people use would reflect their feelings and that by simple process of counting these words we can gain insights into their emotional states" (p. 4). This LIWC system is a "probabilistic system": the more words are counted, the more accurate the system will be. In LIWC, words from the texts are separately categorized into different word dictionaries which have some psychological marks. This computer program has provided basic statistical resources for Pennebaker's research, making his research transcend the spatial and time limitations. The analysis of word frequency has been an important research approach in quantitative and corpus linguistics. Pennebaker has applied word frequency count into psychological studies, concerning people's psychological secrets through function words. Beside word frequency count, some other mathematical calculation has also been utilized. Nevertheless, In Pennebaker's belief, qualitative and quantitative studies can "complement" each other to get better results. Such perception can be discovered throughout the whole book. The book has covered quite a few revealing and interesting research methods. Histograms, linear graph and other diverse tables and graphs are making a direct and accurate demonstration to the results. Several case studies, along with some authentic conversation recordings are conducted as well. With regards to the collection of data and material, a variety of online samples were employed from internet sources like blogs, twitter and Facebook, from the author's own lab experiments, from historical corpora, etc. It can be foreseeable that, as long as the delicate issue of privacy is ensured, more corpora will be drawn on from internet in the future since internet is the treasury of authentic texts. For instance, in order to discover how national traumatic event would have affected people's emotion and unity, the author has utilized people' blogs to compare the *I*- and *we*- words before and after 911 Attack occurred. In the whole book, the author often conducted several researches and experiments in different settings to verify one hypothesis. Such behavior has greatly strengthened his reliability of his research.

The merits of this book not only lie in its multiple dealing in researches, it is really a book unfolding an inter-disciplinary picture of the secrets of function words along with the indication on daily life. The topics Pennebaker selects are of high relevance to our daily life activities, which can grasp a large amount of common readers. In fact, Pennebaker has always kept the ordinary readers in mind in the whole book. He tries to avoid intricate and abstruse terms and writes in quite a humorous style which constantly amused me when I was reading. In addition, some delicate and popular topics have also been included in his researches. In explaining the speaking style differences between men and women, he has also considered the possibility of a person's speaking pattern transformation after gender reassignment is given. Although actual solutions have not yet been expounded, he has opened wide a broader and more innovative view for future researches. In his illustration of his hypothesis, he keeps a skeptical standing about his own research, and frankly points out his own limitations and suggestions for his researches. At the same time, he also combines his hypothesis to the practical application. He has made quite a few efforts to relieve patients' mental load with Post-Traumatic Stress Disorder (PTSD) through expressive writing. Finally, there's one thread running through the whole book, that is, he repeatedly stressed that, "language is a powerful reflection of a person but does not change the person on its own." Such view is in alignment with Sapir-Whorf Hypothesis that language can reflect thinking but cannot determine thinking. Therefore, there is no chance that by "changing the ways we use word, we can change our psychological state" (p 84).

As we extol the shiny parts of this book, its limitations cannot be ruled out either. First, this book has constantly attempted to make a distinction among people by comparing the frequency differences of *I*- and *we*- words or other function words, but what are the specific boundary between high and low use of these words? The book does not give any specific index for distinction, which may weaken the practical applicability and its scientific merits. Besides, no further statistical explanation about the data differences makes readers question about the data significance among different groups and they may criticize its data validity. The inferences in this book are based on the comparison among different groups, there's not an absolute division or a statistical correlation formula for precise distinction. For example, when we have one language material at hand, how can we decide the gender, power status and age of its author without any contrast to other material? Therefore, we had better make more researches on finding some mathematical regularity between *I*- and *we*- words for division. If certain mathematical equation or constant is found and proved, then it may better elucidate itself and predict more language phenomena. Second, this book has relied its researches largely on the LIWC program. However, this computer program has its own defect which is caused by its detachment from the context on a whole. By sorting words into the specific dictionary labeled with emotion indications one by one, the context is ignored. In this case, some vague words or highly context-dependent words cannot be identified or can be falsely classified. Some ironic and sarcastic texts are particularly confusing the machine. The phenomenon of polysemy is making the situation even more intricate. Pennebaker himself has recognized these defects of operability in his book, though the defects remain unsolved. With the computer program itself not fully matured, we might doubt its data output and its hypothesis. Therefore, we can add manual review following the computer classification. Yet, manual review will increase labor and time cost particularly when the corpora are large. Finally, in terms of the psychological application, Pennebaker suggest expressive writing as a therapy on PTSD patients. For those who suffered from traumatic experiences, they are encouraged to write down their own experiences to relieve their mental load, but the actual effectiveness of such therapeutic method remains pending. Promising as his research is, most of the applications of this hypothesis are still in infancy.

In conclusion, this book presents a promising and innovative research field focusing on the function words. It has led common readers into the interest of our daily language. Besides, the book has presented language in a multiple inter-disciplinary perspective. It has connected frequency of the function words with the analysis of psychology and cognition, which broadly expand the research possibilities of text-based statistical methods, the quantitative linguistics, corpus linguistics and quantitative stylistics for example, and hence enable us to perceive and explore human cognition and mental state through quantitative analysis on the text. In this way, this book successfully connects the dots of language with other domains of linguistics within its many branches. The researchers in psycholinguistics, sociolinguistics, quantitative linguistics and neurolinguistics can all find their research interest and receive sparkling insights in this book. With all these questions and limitations unsolved and raised by Pennebaker, it in another sense has created more possibilities of future investigation.