Glottometrics 37
2017

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

External academic peers for Glottometrics

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 37 (2017), Lüdenscheid: RAM-Verlag, 2017. 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 37 (2017) **ISSN 2625-8226**

Contents

Glottometrics 37, 2017, 1-12

Random Crossings in Dependency Trees

Ramon Ferrer-i-Cancho[1]

Abstract. It has been hypothesized that the rather small number of crossings in real syntactic dependency trees is a side-effect of pressure for dependency length minimization. Here we answer a related important research question: what would be the expected number of crossings if the natural order of a sentence was lost and replaced by a random ordering? We show that this number depends only on the number of vertices of the dependency tree (the sentence length) and the second moment about zero of vertex degrees. The expected number of crossings is minimum for a star tree (crossings are impossible) and maximum for a linear tree (the number of crossings is of the order of the square of the sequence length).

Keywords: syntactic dependency trees, syntax, distance, crossings, planarity.

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 vertices are words and arcs indicate syntactic dependencies between these words (Fig. 1). Here we focus on the crossings between dependencies due to the linear arrangement of the vertices of a tree (Hays 1964, Holan et al. 2000, Hudson 2000, Havelka 2007).

Imagine that $\pi(v)$ is the position of vertex v in linear arrangement of the vertices of a tree, a number between 1 and n, with n being the length of the sequence. Imagine that we have two pairs of linked vertices: (u,v) and (s,t), such that $\pi(u) < \pi(v)$ and $\pi(s) < \pi(t)$. The arcs (or edges) defined respectively by (u,v) and (s,t) cross if and only if

$$\pi(u) < \pi(s) < \pi(v) < \pi(t) \quad (1)$$

or

$$\pi(s) < \pi(u) < \pi(t) < \pi(v). \quad (2)$$

C is defined 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).

According to crossing theory, C cannot exceed C_{pairs}, the number of edge pairs that can potentially cross, which is (Ferrer-i-Cancho 2013)

$$C_{pairs} = \frac{n}{2}\left(n - 1 - \left\langle k^2 \right\rangle\right), \quad (3)$$

[1] Complexity and Quantitative Linguistics Lab. Departament de Ciències de la Computació, LARCA Research Group, Universitat Politècnica de Catalunya (UPC). Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona, Catalonia (Spain). Phone: +34 934134028. Fax: +34 934137787.
E-mail: rferrericancho@cs.upc.edu

where n is the sequence length (the number of words/vertices) and $\left\langle k^2 \right\rangle$ is the second moment about zero of the degree, defined as

$$\left\langle k^2 \right\rangle = \frac{1}{n}\sum_{i=1}^{n} k_i^{\ 2}, \tag{4}$$

where k_i is the degree of the i-th vertex of the tree. As the first moment of the degree of a tree of n vertices is constant, i.e. $\left\langle k \right\rangle = 2 - 2/n$ (Noy 1998), the degree variance of a tree is fully determined by $\left\langle k^2 \right\rangle$ and n.

For the dependency tree of Fig. 1, Eq. 3 gives C_{pairs} = 18 since n = 9 and $\left\langle k^2 \right\rangle = 4$.

It has been argued that the small amount of crossings in real sentences (Liu 2010) could be a side-effect of a principle of dependency length minimization (Ferrer-i-Cancho 2006, Ferrer-i-Cancho 2013). A challenge for this hypothesis is that the number of crossings that is expected by chance (by ordering the vertices at random) is about the same value that is obtained in real sentences. Thus, a theoretical analysis of E[C], the expected number of crossings in a random linear arrangement of vertices is needed to shed light on the statistical significance of the rather low number of crossings in real sentences (Liu 2010). This is the goal of the next sections: Section 2 reviews previous results on the maximum value of C and Section 3 derives E[C] = C_{pairs}/3, and related results, e.g., the probability that two edges cross when arranged linearly at random. If the edges share no vertex the probability is 1/3 and it is zero otherwise. Section 4 discusses some applications of these results.

2. CROSSING THEORY

u~v is used to refer to the edge defined by the pair of vertices (u,v). The edges u~v and s~t, such that $u < v$ and $s < t$, cannot cross if they have a vertex in common, i.e. $u \in \{s,t\}$ or $v \in \{s,t\}$. Therefore $C > 0$ requires that there is at least a pair of edges that are formed by four different vertices. Thus $C = 0$ if $n < 4$ and $C > 0$ needs $n \geq 4$.

The structure of a tree, e.g., a syntactic dependency tree, can be defined by means of an adjacency matrix $A = \{a_{uv}\}$, where $a_{uv} = 1$ if the pair of vertices (u,v) is linked and otherwise $a_{uv} = 0$. The matrix is symmetric $a_{uv} = a_{vu}$ (the direction of a dependency is neglected). Loops are not allowed ($a_{uu} = 0$). $a_{uv} = 1$ and u~v are equivalent.

The number of crossings induced by the linear arrangement of the vertices can be defined as

$$C = \frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv} C(u,v), \tag{5}$$

where $C(u,v)$ is the number of different edges that cross with the edge u~v. By symmetry, $C(u,v) = C(v,u)$. The factor 1/4 of Eq. 5 comes from the fact that the same crossing is counted four times in that formula:

- Two times due to the double summation of Eq. 5, i.e. the target edge u~v is counted first through the pair (u, v) and second through its symmetric pair (v, u).
- Two times more due to the fact the edges of the form s~t with which the edge u~v crosses are counted twice, first through $C(u,v)$ and second through $C(s,t)$.

$C(u,v)$ can be defined in turn as

$$C(u,v)=\frac{1}{2}\sum_{\substack{s=1\\ s\neq u,v}}^{n}\sum_{\substack{t=1\\ t\neq u,v}}^{n}a_{st}C(u,v;s,t), \quad (6)$$

where $C(u,v;s,t)$ = 1 if the edge u~v crosses the edge s~t and $C(u,v;s,t)$ = 0 otherwise. The factor 1/2 in Eq. 6 comes from the fact that an edge is encountered twice in the double summation, first by the pair of vertices (s,t) and second by the pair (t, s).

It has been argued that $C(u,v)$ cannot exceed $C_{pairs}(u,v) = n - k_u - k_v$ where k_x is the degree of vertex x (Ferrer-i-Cancho 2013; see Appendix A of the present article for a derivation of $C_{pairs}(u,v)$). Thus the total number of crossings of the linear arrangement of a tree cannot exceed (Ferrer-i-Cancho 2013)

$$C_{pairs}=\frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n}a_{uv}C_{pairs}(u,v)=\frac{n}{2}\left(n-1-\left\langle k^2\right\rangle\right). \quad (7)$$

A star tree is a tree with a vertex of maximum degree while a linear tree is a tree where the maximum vertex degree is two (Fig. 3). Linear and star trees are important trees for crossing theory as they determine the range of variation of $\left\langle k^2\right\rangle$ in Eq. 7. $\left\langle k^2\right\rangle$ is minimized by a linear tree (Ferrer-i-Cancho 2013) and that tree is indeed the only minimum (Appendix B). Similarly, $\left\langle k^2\right\rangle$ is maximized by a star tree (Ferrer-i-Cancho 2013) and that tree is indeed the only maximum (Appendix B).

A very simple case to demonstrate Eq. 7 is a linear tree with n = 4. That tree has three edges and two leaves (a leaf is a vertex of degree one). Imagine that the two leaves are labeled with 1 and 4 and the other edges are labeled with 2 and 3. The only pair of edges that can cross are 1~2 and 3~4 (the two different edges formed by each of the two leaves), since they are the only pair of edges that do not share vertices. Thus C_{pairs} = 1 and C is binary, i.e. C = 1 (edges 1~2 and 3~4 cross) or C = 0 (edges 1~2 and 3~4 do not cross). Accordingly, applying n = 4 and $\left\langle k^2\right\rangle=(1+1+4+4)/4=5/2$ to Eq. 7 yields C_{pairs} = 1 for that linear tree.

3. RANDOM CROSSINGS

According to Eq. 5, the expected number of crossings induced by a random linear arrangement of the vertices is

$$E[C]=\frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n}a_{uv}E[C(u,v)] \quad (8)$$

while the expectation of $C(u,v)$ is in turn

$$E[C(u,v)]=\frac{1}{2}\sum_{\substack{s=1\\ s\neq u,v}}^{n}\sum_{\substack{t=1\\ t\neq u,v}}^{n}a_{st}E[C(u,v;s,t)]. \quad (9)$$

As $C(u,v;s,t)$ is an indicator variable, E[$C(u,v;s,t)$] = $p_c(u,v;s,t)$, the probability that the edges u~v and s~t cross knowing that $s \notin \{u,v\}$ and $t \notin \{u,v\}$. By the definition of crossing in Eqs. 1 and 2, it follows that $p_c(u,v;s,t)$ = 0 if the edges u~v and s~t have at least one vertex in common, i.e. $u \in \{s,t\}$ or $v \in \{s,t\}$. Otherwise, $p_c(u,v;s,t)$ = 1/3. To see the latter, notice that the random linear arrangement of two edges is equivalent to:

- Generating four different vertex positions with the only constraint that they are random numbers between 1 and *n* and positions that are not taken yet are equally likely.
- Sorting the four positions increasingly giving π_1, π_2, π_3 and π_4 such that $1 \leq \pi_1 < \pi_2 < \pi_3 < \pi_4 \leq n$. It is said that π_i has rank *i*.
- Assigning each of these four positions to a different vertex of the pairs of edges involved. Eqs. 1 and 2 mean that the two edges cross if and only if (u,v) is assigned (π_1, π_3) or (π_2, π_4).

Therefore the probability that $u{\sim}v$ and $s{\sim}t$ cross is the probability of assigning two of the four positions whose ranks are not consecutive to the vertices of $u{\sim}v$ with $u < v$, i.e. (a) $\pi(u) = \pi_1$ and $\pi(v) = \pi_3$ or (b) $\pi(u) = \pi_2$ and $\pi(v) = \pi_4$. Therefore,

$$p_c(u,v;s,t) = \frac{2}{\binom{4}{2}} = \frac{1}{3}. \qquad (10)$$

Interestingly, the probability that two edges cross does not depend on the sequence length *n* once it is known whether they share vertices or not (if the two edges share vertices the probability is zero regardless of *n*; if they do not share any vertex then $n \geq 4$ and the probability is 1/3). Furthermore, the identity of vertices involved is irrelevant for the probability that they cross once it is known if the edges share vertices or not. Thus, Eq. 9 becomes

$$E[C(u,v)] = \frac{C_{pairs}(u,v)}{3}. \qquad (11)$$

Applying Eq. 11 to Eq. 8 and recalling the definition of C_{pairs} in Eq. 7, we obtain

$$E[C] = \frac{1}{6}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv} C_{pairs}(u,v) = \frac{C_{pairs}}{3}. \qquad (12)$$

The combination of Eq. 11 and Eq. 10 yields

$$E[C] = \frac{C_{pairs}}{3}. \qquad (13)$$

A simple case is a linear tree with $n = 4$, as $C_{pairs} = 1$ transforms Eq. 13 into $E[C]=1/3$.

Applying Eq. 3 to Eq. 13, one finally obtains

$$E[C] = \frac{n}{6}\left(n - 1 - \left\langle k^2 \right\rangle\right) \qquad (14)$$

for $n \geq 4$.

For the dependency tree of Fig. 1, $n = 9$ and $\left\langle k^2 \right\rangle = 4$ gives $E[C] = 6$.

According to Eq. 14, $E[C] = 0$ for a star tree as $\left\langle k^2 \right\rangle = n - 1$ for that tree while

$$E[C] = \frac{n(n-5)}{6} + 1 \qquad (15)$$

for a linear tree as $\langle k^2 \rangle = 4 - 6/n$ in that case (Ferrer-i-Cancho 2013).

4. DISCUSSION

It has been shown that E[C] is determined exclusively by n and $\langle k^2 \rangle$ (Eq. 14). Given n, the range of variation of E[C] is then given by $\langle k^2 \rangle$, which is minimum for a linear tree and maximum for a star tree, i.e. (Ferrer-i-Cancho 2013)

$$4 - \frac{6}{n} \leq \langle k^2 \rangle \leq n - 1 \tag{16}$$

for a finite tree with $n \geq 2$ and thus giving

$$0 \leq E[C] \leq \frac{n(n-5)}{6} + 1 \tag{17}$$

thanks to Eq. 15 for any tree of at least four vertices (E[C] = 0 if $n < 4$).

Fig. 4 shows the upper bound of E[C] provided by a linear tree (Eq. 17), which obviously grows asymptotically as n^2 for sufficiently large n. Thus the possibility that the rather small number of crossings of real sentences (Liu 2010) is the outcome of some sort of optimization processes, possibly a side-effect of the minimization of dependency lengths (Ferrer-i-Cancho 2006, Ferrer-i-Cancho 2013) cannot be denied. Future research on the significance of the small amount of crossings of real sentences should consider the real value of C in sentences versus estimates of E[C] obtained through Eq. 14 with real values of $\langle k^2 \rangle$. Thus, investigating the scaling of $\langle k^2 \rangle$ as a function of n in real sentences from dependency treebanks (e.g., Civit *et al.* 2006, Böhmová *et al.* 2003, Bosco *et al.* 2000) is an important question for future research.

The results presented above can also help to shed light on the actual relationship between dependency length and crossings (Ferrer-i-Cancho 2006, 2013, Liu 2008). Imagine that $\langle d \rangle$ is the mean dependency length of the linear arrangement of vertices. The possibility of a natural correlation between C and $\langle d \rangle$ can be demonstrated starting from an actual sentence such as the one in Fig. 1 and swapping the position of pairs of vertices chosen at random. Fig. 5 shows that both C and $\langle d \rangle$ start from $\langle d \rangle$ = 11/8 = 1.375 and C = 0 for the sentence in Fig. 1 and then both increase as the number of these swaps increases till they converge to their values in a random linear arrangement, respectively, $E[C]$ = 6 (computed above) and $E[\langle d \rangle]$ = E[d] = (n+1)/3 = 10/3 ≈ 3.33 (Ferrer-i-Cancho 2004, 2013, Zörnig 1984). Notice that our swapping of vertex positions is a randomization procedure that preserves the dependency tree (i.e. the adjacency matrix of the tree), and thus preserves the degree's 2nd moment and the connectedness of the dependency network. Other research on dependency networks has employed procedures to generate random dependency structures that do not warrant that vertex degrees or connectedness are maintained (as needed by a tree) or forbid dependency crossings (Liu & Hu 2008).

Fig. 5 suggests that C and $\langle d \rangle$ are positively correlated, which is consistent with the hypothesis that the low frequency of dependency crossings could be a side effect of dependency length minimization (Ferrer-i-Cancho 2006). Future research could extend this kind of analysis to more sentences with the help of dependency treebanks (e.g., Civit *et al.* 2006, Böhmová *et al.* 2003, Bosco *et al.* 2000).

Final note: the mathematical results presented in this article have been applied in a series of articles: Ferrer-i-Cancho (2014), Ferrer-i-Cancho (2016a,b), Esteban *et al.* (2016) and Gómez-Rodríguez & Ferrer-i-Cancho (2016).

ACKNOWLEDGEMENTS

We are grateful to O. Jiménez for helpful discussions and G. Wimmer, J. Baixeries, H. Liu and R. Cech for helpful comments on this manuscript. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya, the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from MICINN (Ministerio de Ciencia e Innovación), the grant APCOM (TIN2014-57226-P) from MINECO (Ministerio de Economía y Competitividad) and the grant 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya).

REFERENCES

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

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

Bosco, C., Lombardo, V., Vassallo, D. & Lesmo. L. (2000). Building a treebank for Italian: a data-driven annotation schema. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation LREC 2000, Athens, p. 99-105.*

Civit, M., Martí, M.A. & Bufí, N. (2006). 'Cat3LB and Cast3LB: from Constituents to dependencies'. In: *Advances in Natural Language Processing/* (LNAI, 4139), pp. 141-153. Berlin: Springer Verlag.

Esteban, J.L., Ferrer-i-Cancho, R., & Gómez-Rodríguez, C. (2016). The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics, 063401.*

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.*

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

Ferrer-i-Cancho, R. (2014). A stronger null hypothesis for crossing dependencies. *Europhysics Letters 108 (5), 58003.*

Ferrer-i-Cancho, R. (2016a). Non-crossing dependencies: least effort, not grammar. In: Mehler, A., Lücking, A., Banisch, S., Blanchard, P. & Job, B. (eds.). *Towards a theoretical framework for analyzing complex linguistic networks.* Berlin: Springer, pp. 203-234.

Ferrer-i-Cancho, R. & Gómez-Rodríguez, C. (2016). Crossings as a side effect of dependency lengths. *Complexity 21 (S2), 320-328.*

Gómez-Rodríguez, C. & Ferrer-i-Cancho, R. (2016). The scarcity of crossing dependencies: a direct outcome of a specific constraint? http://arxiv.org/abs/1601.03210.

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).* Prague, Czech Republic: Association for Computational Linguistics, pp. 608-615.

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

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

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

Liu, H. (2010). Dependency direction as a means of word-order typology a method based on dependency treebanks. *Lingua. 120 (6), 1567-1578.*

Liu, H. & Hu, F. (2008). What role does syntax play in a language network? *Europhysics Letters 83, 18002.*

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.

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

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

APPENDIX A: THE NUMBER OF POSSIBLE CROSSINGS OF AND EDGE

$C_{pairs}(u,v)$ can be derived from $C(u,v)$ assuming that $C(u,v;s,t)$=1 in any circumstance, which transforms Eq. 6 into

$$C(u,v) = C_{pairs}(u,v) = \frac{1}{2}\sum_{\substack{s=1\\ s\neq u,v}}^{n}\sum_{\substack{t=1\\ t\neq u,v}}^{n} a_{st} = \frac{1}{2}\sum_{\substack{s=1\\ s\neq u,v}}^{n}(k_s - a_{us} - a_{vs})\,. \tag{A1}$$

Applying

$$\sum_{\substack{s=1\\ s\neq u,v}}^{n} k_s = 2(n-1) - k_u - k_v \tag{A2}$$

and

$$\sum_{\substack{t=1\\ t\neq u,v}}^{n} a_{st} = k_s - a_{us} - a_{vs} \tag{A3}$$

to Eq. A1 yields

$$C_{pairs}(u,v) = \frac{1}{2}\left(2(n-1) - k_u - k_v - (k_u - a_{uv} - a_{uu}) - (k_v - a_{uv} - a_{vv}\right) \tag{A4}$$

and finally

$$C_{pairs}(u,v) = (n-1) - k_u - k_v + a_{uv} + a_{uu} + a_{vv} = n - k_u - k_v \tag{A5}$$

as a_{uu} = a_{vv} = 0 (loops are not allowed) and a_{uv} = 1 as u and v are linked by the definition of $C(u,v)$.

APPENDIX B: LINEAR AND STAR TREES HAVE UNIQUE DEGREE 2nd MOMENT

To simplify the arguments below, we define the degree 2nd moment as $\left\langle k^2 \right\rangle = K_2/n$, where $K_2(n)$, is the sum of squared degrees of a tree of n vertices, i.e.

$$K_2(n) = \sum_{i=1}^{n} k_i^{\ 2}. \tag{B1}$$

$K_2^{\text{linear}}(n)$ and $K_2^{\text{star}}(n)$ are defined, respectively, as the sum of squared degrees of a linear tree and a star tree of n nodes. $K_2^{\text{linear}}(n)$ = $4n$ - 6 and $K_2^{\text{star}}(n)$ = $n(n$-1) (Ferrer-i-Cancho 2013). Here it will be shown that a linear tree is the only tree for which $K_2(n)$ reaches $K_2^{\text{linear}}(n)$ while a star tree is the only tree for which $K_2(n)$ can reach $K_2^{\text{star}}(n)$. Before proving these properties, the concept of tree reduction and compact definitions of star and linear trees will be introduced.

Tree reduction

Any tree of at least two vertices has at least two leaves (Bollobás 1998, p. 11). Thus, any tree of n + 1 vertices ($n \geq 2$ is assumed) can be reduced to a tree of n vertices by removing one of its leaves. Notice that this reduction will never disconnect the tree as the leaf removed cannot be attached to another leaf unless $n = 2$ (a leaf attached to another leaf when $n>2$ would contradict that a tree is a connected graph). Consider that the leaf removed is attached to a vertex of degree k in the original tree (the tree of $n + 1$ vertices). Then

$$K_2(n + 1) = K_2(n) + k^2 - (k - 1)^2 + 1 \qquad \text{(B2)}$$

for the original tree and thus

$$K_2(n + 1) = K_2(n) + 2k. \qquad \text{(B3)}$$

A star tree is a tree with a vertex of maximum degree.

A star tree of n vertices is a tree with a vertex of degree n-1 and n-1 leaves (Fig. 3). Indeed, a star tree of n vertices can simply be defined as a tree with a vertex of maximum degree (i.e. degree n-1). The point is that the fact that a vertex has degree n-1 implies that there are n-1 leaves. To see it, recall that the degree sequence of a graph of n vertices satisfies

$$\sum_{i=1}^{n} k_i = 2(n-1)\,. \qquad \text{(B4)}$$

Assuming without any loss of generality that the *n-th* vertex has maximum degree (i.e. $k_n = n-1$), Eq. B4 gives

$$\sum_{i=1}^{n-1} k_i = n-1 \qquad \text{(B5)}$$

As a tree is a connected graph, any vertex has degree greater than zero and Eq. B5 gives $k_1=...=k_i=...=k_{n-1}=1$, i.e. n-1 leaves, the number of leaves of a star tree.

A linear tree is a tree where all vertex degrees do not exceed two

A linear tree is a tree where all vertices have degree two except two leaves (Fig. 3). Indeed, a linear tree can simply be defined as a tree where all vertex degrees do not exceed two. Notice that in our last definition of linear tree we do not need to state the number of leaves and the number of vertices of degree two that we have. To understand our last definition of linear tree, suppose that a tree has n vertices and m leaves (then it has m – 2 vertices of degree 2). Then the sum of the degrees of leaves is m. If no vertex degree exceeds two then the sum of degrees of the vertices that are not leaves is $2(n - m)$. Then, Eq. B4 reduces to $m + 2(n - \mathrm{m}) = 2(n - 1)$ which gives $m = 2$. Thus, if no vertex degree exceeds two, one can be certain that the tree is linear.

A star tree is the only tree reaching $K_2^{star}(n)$

Next it will be shown that a star tree is the only tree for which $K_2(n) = K_2^{star}(n)$. If $n = 2$, then this is trivially true as the only possible tree is a star tree. Consider a tree of n+1 vertices (with $n > 2$) such that $K_2(n + 1) = K_2^{star}(n + 1)$. Thanks to Eq. B3, we know that

$$K_2(n) + 2k = K_2^{star}(n + 1) \qquad \text{(B6)}$$

for that tree. Adding that $K_2(n) \leq K_2^{star}(n)$ (Ferrer-i-Cancho 2013) to Eq. B6, it is obtained

$$k \geq \frac{K_2^{star}(n+1) - K_2^{star}(n)}{2} = \frac{(n+1)n - n(n-1)}{2} = n\,. \qquad \text{(B7)}$$

As k cannot exceed n in a graph of n + 1 vertices (without loops), Eq. B7 implies that $k = n$, which we have shown above to imply that the tree of n + 1 vertices is a star, as we wanted to prove.

A linear tree is the only tree reaching $K_2^{linear}(n)$

Next it will be shown that a linear tree is the only tree for which $K_2(n) = K_2^{linear}(n)$. If $n = 2$, then this is trivially true as the only possible tree is a linear tree. Consider a tree of n+1 vertices (with $n > 2$) such that $K_2(n + 1) = K^{linear}(n + 1)$. Thanks to Eq. B3, we know that

$$K_2(n) + 2k = K_2^{linear}(n + 1) \qquad \text{(B8)}$$

for that tree. Adding that $K_2(n) \geq K_2^{linear}(n)$ (Ferrer-i-Cancho, 2013) to Eq. B8, it is obtained

$$k \leq \frac{K_2^{linear}(n+1) - K_2^{linear}(n)}{2} = \frac{4(n+1) - 6 - (4n-6)}{2} = 2\,. \qquad \text{(B9)}$$

As k is the degree of a vertex that is not a leaf, if follows that any vertex in the original tree that is not a leaf has degree exactly 2, which we have shown above to imply that the tree of n+1 vertices is a linear tree, as we wanted to prove.

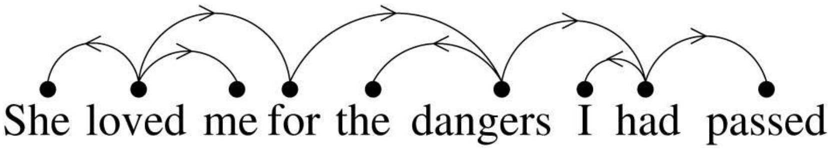

Figure 1. The syntactic structure of the sentence *'She loved me for the dangers I had passed'* following the conventions in (Mel'čuk 1988). *'she'* and the verb *'loved'* are linked by a syntactic dependency. Arcs go from governors to dependents. Thus, *'she'* and *'me'* are dependents of the verbal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

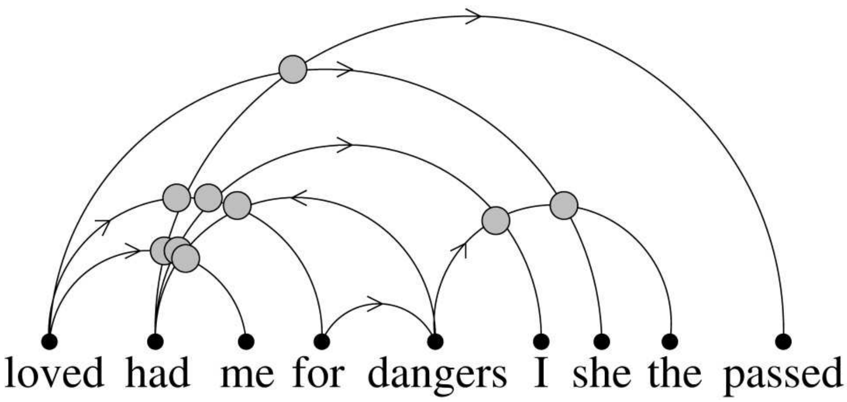

Figure 2. The structure of the sentence in Fig. 1 after a random linear rearrangement of its words. Gray circles indicate edge crossings.

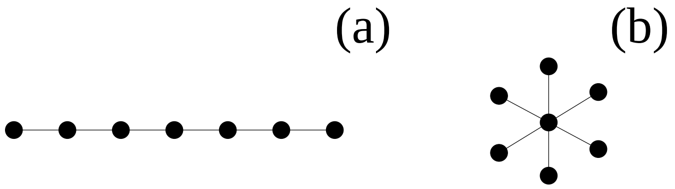

Figure 3. (a) a linear tree and (b) a star tree. A linear tree is a tree with the smallest possible number of leaves (only two leaves, Bollobás 1998, p. 11) while a star tree is the tree with the largest number of leaves.

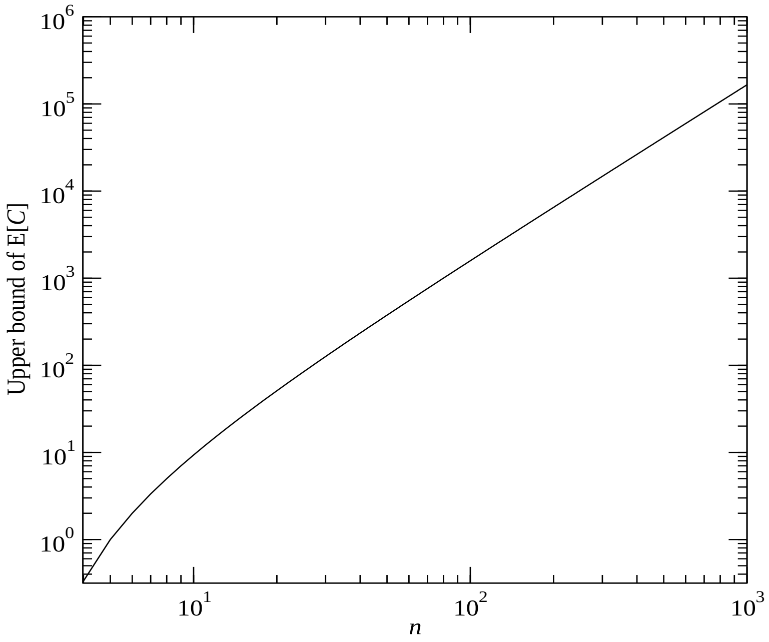

Figure 4. The upper bound of E[C] (the expectation of the number of crossings of a linear tree) as function of n, the number of vertices of the tree.

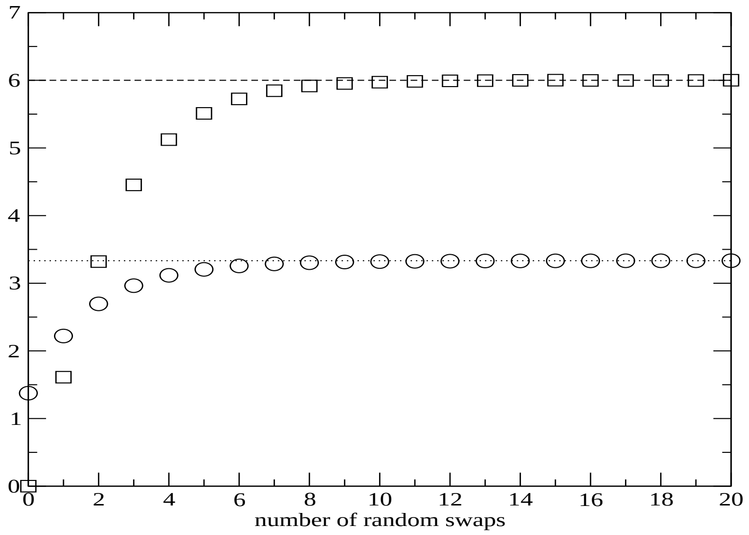

Figure 5. The evolution of $\langle d \rangle$, the mean dependency distance (circles), and *C*, the number of edge crossings (squares), versus the number of swaps of pairs of vertex positions for the sentence in Fig. 1. Each curve is the average over 10^6 replicas. $\langle d \rangle$ converges to $E[d] = 10/3$ (dotted line) while *C* converges to $E[C] = 6$ (dashed line).

Glottometrics 37, 2017, 13-33

The Distribution of Dependency Relations in *Great Expectations* and *Jane Eyre*

Jianwei Yan[1], *Siqi Liu*[2]

Abstract. This study explores features of specific literary works, *Great Expectations* and *Jane Eyre* (hereafter referred to as *GE* and *JE*), based on the theoretical framework of dependency grammar. Both works are masterpieces of critical realism in Victorian era. This study is a descriptive analysis, which investigates the dependency relations of both works, including dependency distance, dependency direction and dependency type. The results indicate that: 1) The difference of syntactic difficulty between *GE* and *JE* is not statistically significant in accordance with MDDs (mean dependency distances); 2) There is a similar trend in the distribution of ADDs (absolute dependency distances), but the differences between *GE* and *JE* in ADDs are highly significant; 3) there is no significant difference in the distribution of dependency directions between *GE* and *JE*; 4) Both *GE* and *JE* have forty-three same dependency types; Meanwhile, although the differences of the distribution of dependency types are highly significant, there is no significant difference between *GE* and *JE* in MDDs of dependency types.

Keywords: Dependency distance, Dependency direction, Dependency type, *Great Expectations*, *Jane Eyre*

1. Introduction

Dependency grammar is originated from the works of Lucien Tesnière (Tesnière, 1959). This basic approach to syntax seems to have been seized upon independently by many other dependency-based grammars since those early works, such as Word Grammar, Meaning-text Theory, Functional Generative Description, etc. It is well suited for the analysis of languages with free word order, such as Czech, Turkish, and Warlpiri. As a descriptive approach, dependency grammar not only provides theoretical framework for computational linguistics but also facilitates the applications on natural language processing and machine translation (Liu, 2009). In fact, dependency grammar, as every grammar, is of great significance for all areas of linguistic research.

Different from constituent grammar, which breaks sentences into constituents, dependency grammar connects individual words which have grammatical functions with respect to

[1] *College of Foreign Languages, Civil Aviation University of China, Tianjin, 300300, China. E-mail:* yanjianwei@aliyun.com

[2] *Department of English Language and Literature, Korea Maritime and Ocean University, Busan, 49112, South Korea*

each other in a sentence (Covington, 2001). Dependency grammar analyzes sentence structures by using the dependency relations between words in a sentence (Tesnière, 1959; Hudson, 2007; Nivre, 2006; Liu, 2009). This relation, which connects a governor with a dependent, is featured as binary, asymmetric and labeled. Usually the verb in a sentence is the governor which is the structural center of the whole sentence. Dependents are the other syntactic units either directly or indirectly connected with the verb, the governor.

The term "dependency distance" is the linear distance between the governor and the dependent, measured in terms of intervening words (Hudson, 1995). The greater the dependency distance, the more difficult it is to analyze the syntactic structure of the sentence (Gibson, 1998; Gibson and Pearlmutter, 1998; Hiranuma, 1999; Liu, 2008). As for the term "syntactic difficulty", there are also many scholars having spent efforts on this area from different perspectives. For instance, the length of linguistic constructs is a very important measurement, such as word and sentence length and their interrelations with other linguistic components (Menzerath, 1928; Köhler, 1982; Altmann, 1980, 1988; Wimmer et al., 1994; Wimmer and Altmann, 1996; Köhler, 2005; Grzybek et al., 2008; Fan et al., 2010; Levitsky and Melnyk, 2011). Hence, dependency distance is one of the measurements of syntactic difficulty, which is applied in this study. It is of great use for predicting syntactic difficulty, explaining the mechanisms of children language learning and designing better parsing algorithms for natural language processing (Liu et al., 2009a). Meanwhile, dependency distance also indicates the linear order of governor and dependent, which can be reflected in the term "dependency direction". Dependency direction shows whether the dependency relation is governor-initial or governor-final. When a governor precedes a dependent, the dependency direction is negative (governor-initial). Otherwise, the dependency direction is positive (governor-final). Measuring dependency direction of a language can indicate expressly the classification of the language typology (Liu, 2010). De Marneffe et al. (2008) designed Sandford Parser for the description of the dependency relationships in a sentence that can easily be understood. Typed dependency relations outputted by Stanford Parser make use of the Penn Treebank part-of-speech tags and phrasal labels, and contain approximately fifty grammatical relations. In this study, the typed dependency relations are generated and utilized for further analysis.

Regarding dependency relations, most previous studies focused on the cross-linguistic investigation. Hiranuma (1999) compared the dependency distances and dependency directions between Japanese and English, 1.43 and 1.386 respectively. Eppler (2005) calculated the mean dependency distances of the English and German, 0.49 and 0.87 respectively. Temperley (2007) examined the question whether language production reflects a preference for shorter dependencies based on a corpus of written English. Liu (2008) investigated the dependency distances of Chinese. Liu (2009) explored the probability distributions of the dependency relations extracted from a Chinese dependency treebank. Liu (2010) investigated twenty languages using treebanks with different sizes from 16,000 to one million dependencies. Oya (2011) focused on the average dependency distance of each sentence taken from three different sentence sets and presented the differences and similarities in the average dependency distances among these sentence sets. Wang (2015) analyzed the distribution of dependency distances in the nine domains of written English in the BNC. Most recently, Jiang and Liu (2015) explored the effects of sentence length on dependency distance, dependency direction and the implications, based on a parallel English-Chinese dependency treebank. Wang and Liu (2017) used quantitative methods to examine the distribution of dependency distances in written English from the BNC across genres controlled for sentence length.

All these studies on dependency relations are of great academic value for future studies, most of which focused on the cross-linguistic investigation of dependency distances or dependency directions. The innovation of this study is that it intends to investigate the features of specific literary works, *GE* and *JE*, based on the theoretical framework of dependency grammar. This study attempts to explore the dependency relations, including dependency distance, dependency direction and dependency type, within specific literary masterpieces of critical realism in Victorian era.

2. Materials and Methods

The materials employed in this study are the plain texts of *GE* and *JE*, both of which were created in the Victorian era. The number of tokens of *GE* and *JE* is 187,696 and 186,135 respectively. In terms of the number of sentences, there are 9,732 sentences and 9,774 sentences respectively. Therefore, both the number of tokens and the number of sentences are comparable.

As for their writers, both Charles Dickens (1812-1870) and Charlotte Bronte (1816-1855) are representatives of English critical realism in the 19th century. Charles Dickens's *GE* has attracted attentions of many scholars because of its achievement on narrative techniques and stylistic traits. Meanwhile, researchers have explored the value of Charlotte Bronte's *JE* from the perspective of feminism, which has inspired women to pursue their independence and freedom. Although both works have been studied for a long period, a descriptive study based on the framework of dependency grammar has never been done.

In this study the plain texts of *GE* and *JE* downloaded from Guttenberg (http://www.gutenberg.org/files/1400; http://www.gutenberg.org/cache/epub/1260) were used as corpora, which can serve as basis for linguistic analyses and descriptions (Kennedy, 1998). To obtain the data required, several instruments and software, including the Stanford Parser, R, Excel and SPSS, were employed. In terms of Stanford Parser, it was one of the biggest breakthroughs in the natural language processing in the 1990s. It attained the highest confidence-weighted score of all entrants in the 2005 competition by a significant margin (De Marneffe et al., 2006). The Stanford dependencies scheme (De Marneffe et al., 2006) has gained popularity throughout various natural language processing tasks (Banko et al., 2007; Meena and Prabhakar, 2007; Jason and Kessler, 2008). However, as a statistical parser, it still makes some errors. One issue that should be noted is that when a dependency type is labeled as *dep*, this means the software is unable to determine the precise dependency type between two words. This may be caused by a weird grammatical construction, a limitation in the Stanford dependency conversion software, a parser error, or an unresolved long distance dependency (De Marneffe et al., 2008). In this study the frequencies of the type *dep* are 4,647 and 5,018 respectively, accounting for 2.48% and 2.68% of all dependency relations. The proportions of errors occupy only a small amount of the whole data, and all of them were excluded during the analysis of dependency types.

First, the descriptions without a full stop in the plain texts of *GE* and *JE*, such as author, headlines of a text, lists etc. were deleted. Then, Stanford Parser was used to output the typed dependency relations of the two corpora. This was followed by the processing of an R program, which was written to generate data of dependency distances, dependency directions and dependency types.

For computing dependency distances for large corpora, Liu et al. (2009a) proposed a method for measuring the mean dependency distance (hereafter referred to as MDD) of a sentence with a sample of a treebank (a corpus with syntactic annotation). Formally, let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_x and W_y ($x \geq 1$, $y \leq n$), if W_x is a governor and W_y is its dependent, then the dependency distance (hereafter referred to as DD) between them is defined as the difference $x - y$; by this measure, the DD of adjacent words is 1. When x is greater than y, the DD is a positive number, which means the head follows the dependent; when x is smaller than y, the DD is a negative number and the head precedes the dependent. However, in measuring DD the relevant measure is the absolute value of DD.

The MDD of an entire sentence can be defined as:

$$\text{MDD (the sentence)} = \frac{1}{\mathrm{n-1}}\sum_{i=1}^{n-1} |\mathrm{DD_i}| \qquad (1)$$

Here *n* is the number of words in the sentence and DD*ᵢ* is the dependency distance of the *i*-th syntactic link of the sentence. Usually in a sentence there is one word (the root verb) without a head, whose DD is defined as zero.

This formula can also be used to calculate the MDD of a larger collection of sentences, such as a treebank:

$$\text{MDD (the sample)} = \frac{1}{\mathrm{n-s}}\sum_{i=1}^{n-s} |\mathrm{DD_i}| \qquad (2)$$

In this case, *n* is the total number of words in the sample, *s* is the total number of sentences in the sample and DD_i is the dependency distance of the *i*-th syntactic link of the sample.

Another formula can be used to calculate the MDD for a specific type of dependency relation in a sample:

$$\text{MDD (dependency type)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{DD_i} \qquad (3)$$

In this case, *n* is the number of examples of that relation in the sample. DD_i is the dependency distance of the *i*-th dependency type.

3. Results and Discussions

3.1 MDDs in *GE* and *JE*

Distance is an essential property of a dependency relation because of its implications for the cognitive cost of processing dependency (Liu, 2008). Likewise, MDD is also an important measure for predicating syntactic difficulty, which reflects the cognitive demands of the language concerned (Hudson, 1995). In accordance with Gibson (1998), the greater the dependency distances, the harder the sentence or the text concerned. In this study, MDDs of *GE* and *JE* were computed by an R program based on formula (2).

Table 1
The distribution of MDDs in *GE* and *JE*

Title of Work	**MDD**
GE	2.740
JE	2.746

As shown in Table 1, MDDs of *GE* and *JE* are 2.740 and 2.746 respectively. The MDD of *GE* is slightly shorter than that of *JE*. This was then followed by statistical tests. There are two tests to choose from. One is the t-test and the other the Mann-Whitney U test. The fact is that, apart from the type of the data (ratio or interval), the former also requires the normality of the data, which the test data do not meet. Therefore, the latter was chosen instead to test whether the difference between MDDs in *GE* and *JE* is significant or not.

The result of the Mann-Whitney U test (listed in Appendix II(a)) shows that the difference of MDDs in Table 1 is not statistically significant (p = 0.150 > 0.05). This means that the MDD of *JE* is slightly longer than that of *GE*, but the text of *GE* is not significantly easier than that of *JE*. In other words, the difference of the syntactic difficulty between *GE* and *JE* is not significant.

3.2 ADDs in *GE* and *JE*

There are 187,629 and 187,498 dependencies respectively in *GE* and *JE*. In the following section, the dependency distance is measured in terms of the number of intervening words, rather than as the difference between the words' position-number, to be comparable to other projects (Liu et al., 2009a). This means that the adjacent words in both texts have a dependency distance of 0, rather than 1, in this section. By this way, the frequencies of absolute dependency distances (hereafter referred to as ADDs), which ignore directions, are shown in Fig. 1.

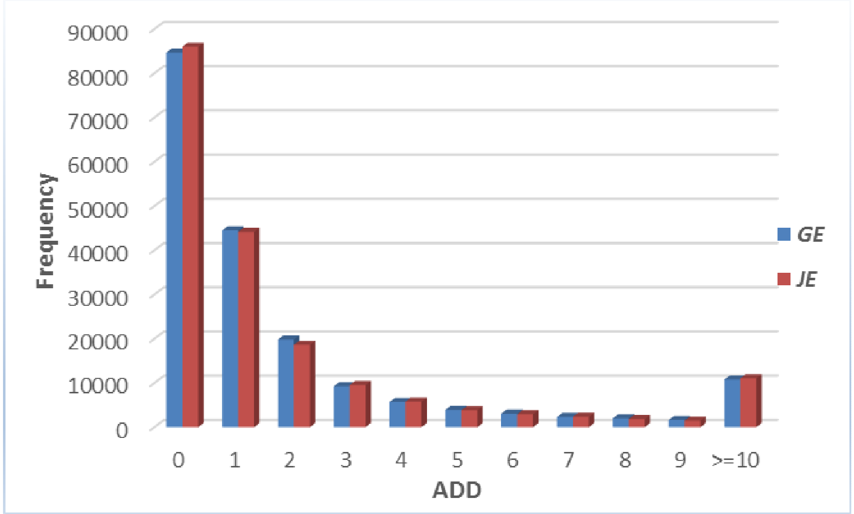

Fig. 1. The frequency of ADDs in *GE* and *JE*

From the figure above, the frequencies of both ADDs peak when ADDs equal zero. This is followed by gradual declines as the number of ADDs increase. When ADDs exceed 10, the sum of the frequencies is around 10,000, which is even smaller than the frequencies of ADDs equaling 2. The Mann-Whitney U test was then used to test whether the difference is significant or not between *GE* and *JE* in the frequency of ADDs. The result (listed in Appendix II(b)) indicates that there exist highly significant differences on ADDs in *GE* and *JE* (p = 0.000158 < 0.001). Hence, the differences between *GE* and *JE* on ADDs are still significant.

To have a detailed look at the differences, the distribution of ADDs ranging from 0 to 9 and ADDs no less than 10 is shown in Table 2.

Table 2
The distribution of ADDs in *GE* and *JE*

	GE		***JE***	
	Frequency	**Percentage**	**Frequency**	**Percentage**
ADD = 0-9	176852	94.26%	176451	94.10%
ADD ≥ 10	10777	5.74%	11047	5.90%

The data presented in Table 2 shows that in *GE* and *JE* the smaller dependency distances account for a dominant proportion, 94.26% and 94.10% respectively, while the percentages of ADDs no less than 10 in *GE* and *JE* are less than 6%. Since the non-parametric test known as the Chi-Square test is especially useful when comparing the frequencies that we observed in a linguistic context and that are grouped into categories, the Chi-Square test was employed to test whether the difference between ADDs' distribution is significant or not. The result (listed in Appendix II(c)) indicates that there is no significant difference of the distribution of ADDs ranging from 0 to 9 and ADDs no less than 10 in *GE* and *JE* (p = 0.053 > 0.05). Therefore, the smaller dependency distances ranging from 0 to 9 in *GE* and *JE* occupy a dominant proportion.

When attention is paid to adjacent words, the distribution of adjacent dependencies and non-adjacent ones in *GE* and *JE* is presented in Table 3.

Table 3
The distribution of adjacent dependencies and non-adjacent dependencies in *GE* and *JE*

	GE		*JE*	
	Frequency	**Percentage**	**Frequency**	**Percentage**
ADD = 0	84683	45.13%	86009	45.87%
ADD > 0	102946	54.87%	101489	54.13%

It can be figured out that in *GE* and *JE* the adjacent dependencies roughly account for half proportion of the whole texts, 45.13% and 45.87% respectively. In accordance with Eppler (2005), who carried out a comparative study of English and German in dependency distances, there are about 78% of dependencies belonging to the category of adjacent words, which is quite different from the results above. This may due to the different size of samples chosen by the two studies since there are only 596 dependencies of English in his study. However, the results in this study correspond to the study carried out by Wang (2015), who made a comparison of the nine English domains and found that the distribution of adjacent dependencies ranges from 48.04% to 50.20% in different domains. The Chi-Square test was then used to compare the significance of the difference between *GE* and *JE* in the distribution of adjacent dependencies and non-adjacent dependencies. The result (listed in Appendix II(d)) shows that there is a highly significant statistical difference on the distribution of adjacent dependencies and non-adjacent ones in *GE* and *JE* (p = 0.000006 < 0.001). To conclude, *JE* comparatively has more adjacent dependencies.

3.3 Dependency Directions in *GE* and *JE*

Dependency directions, also known as positive dependencies and negative dependencies, are discussed in this section, and the tables of all dependency distances with directions in *GE* and *JE* are listed in Appendix I(a) and Appendix I(b). In Fig. 2, the distributions of dependency directions in *GE* and *JE* are presented. The abscissa of the figure is dependency distances; negative numbers indicate that the dependencies are governor-initial, while positive ones mean that the dependencies are governor-final.

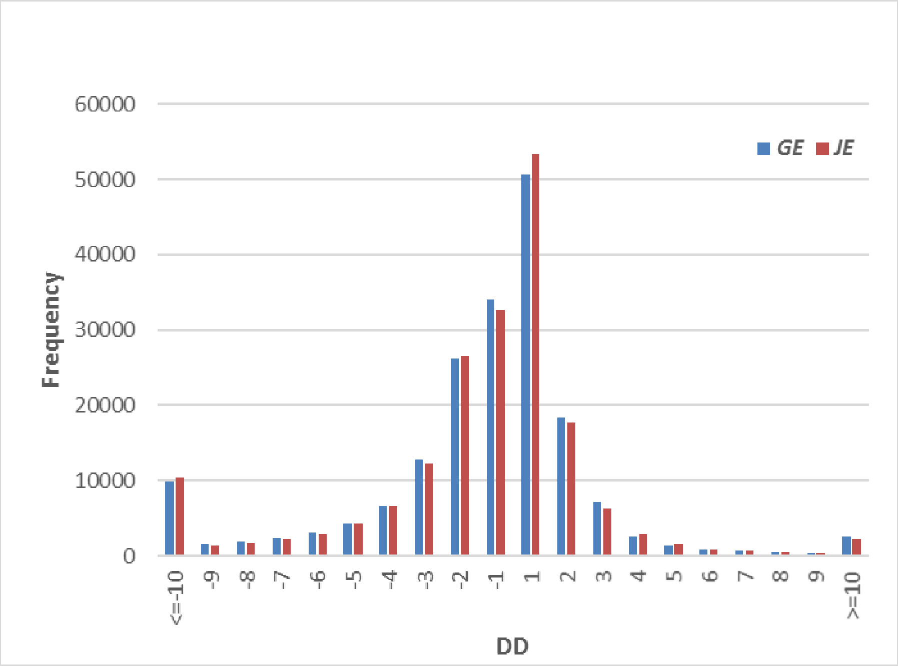

Fig. 2 The frequency of dependency distances in *GE* and *JE*

As shown in Fig. 2, the positive and negative dependencies in *GE* and *JE* share the same tendency: the smaller the absolute dependency distances, the higher the frequencies. The Mann-Whitney U test was then used to test whether the difference between *GE* and *JE* in the frequency of dependency distances is significant or not. The result (listed in Appendix II(e)) shows that there is no significant difference on the distribution of dependency directions in *GE* and *JE* (p = 0.806 > 0.05).

To have a detailed look at the differences, the distribution of overall positive and negative dependencies in *GE* and *JE* is shown in Table 4, the distribution of adjacent positive and adjacent negative dependencies in Table 5, and non-adjacent dependency directions in Table 6.

Table 4
The distribution of overall positive and negative dependencies in *GE* and *JE*

	GE		***JE***	
	Frequency	**Percentage**	**Frequency**	**Percentage**
DD ≥ 1 (Positive)	85133	45.37%	86502	46.13%
DD ≤ -1 (Negative)	102496	54.63%	100996	53.87%

Table 5
The distribution of adjacent positive and negative dependencies in *GE* and *JE*

	GE		***JE***	
	Frequency	**Percentage**	**Frequency**	**Percentage**
DD = 1 (Positive)	50680	59.85%	53386	62.07%
DD = -1(Negative)	34003	40.15%	32623	37.93%

Table 6
The distribution of non-adjacent positive and negative dependencies in *GE* and *JE*

	GE		***JE***	
	Frequency	**Percentage**	**Frequency**	**Percentage**
DD > 1 (Positive)	34453	33.47%	33116	32.63%
DD < -1(Negative)	68493	66.53%	68373	67.37%

Hiranuma (1999) pointed out that English is a language where the dependent tends to occur on either side of the head. The results in Table 4 show that about half of the dependencies are positive (governor-final) and the other half negative (governor-initial) confirms Hiranuma's findings. This was followed by a Chi-Square test to compare the significance of the difference between *GE* and *JE* in the distribution of overall positive and negative dependencies. The result (listed in Appendix II(f)) shows that there exists a highly significant difference between *GE* and *JE* as to the frequency and proportion of positive and negative dependencies (p = 0.000003 < 0.001); *JE* has a larger proportion of positive dependencies.

As for adjacent dependencies in Table 5, positive dependencies in *GE* and *JE* account for 59.85% and 62.07% respectively. A Chi-Square test was also used to test whether the difference between *GE* and *JE* in the distribution of adjacent positive and negative dependencies is significant or not. The result (listed in Appendix II(g)) shows that the difference between *GE* and *JE* in adjacent dependencies is also highly significant (p = 0.00 < 0.05).

Correspondingly, the non-adjacent dependencies of *GE* and *JE* tend to be negative (governor-initial). To be specific, *GE* tends to be governor-initial with a proportion of 66.53%, and negative dependencies in *JE* account for 67.37%, which means that *GE* and *JE* have more negative (governor-initial) dependencies in terms of non-adjacent dependencies. This was followed by a Chi-Square test to compare the significance of the difference between *GE* and *JE* in the distribution of overall positive and negative dependencies. The result (listed in Appendix II(h)) indicates that the difference between *GE* and *JE* in Table 6 is still highly significant (p = 0.000058 < 0.05).

3.4 Dependency Types in *GE* and *JE*

In this study the frequencies of the type *dep* are 4,647 and 5,018 respectively, accounting for 2.48% and 2.68% of all dependency relations, and all of them were excluded during the analysis of dependency types. Subsequently, the dependency types for *GE* and *JE* and all the frequencies for each type are presented in Fig. 3 and Table 7 below.

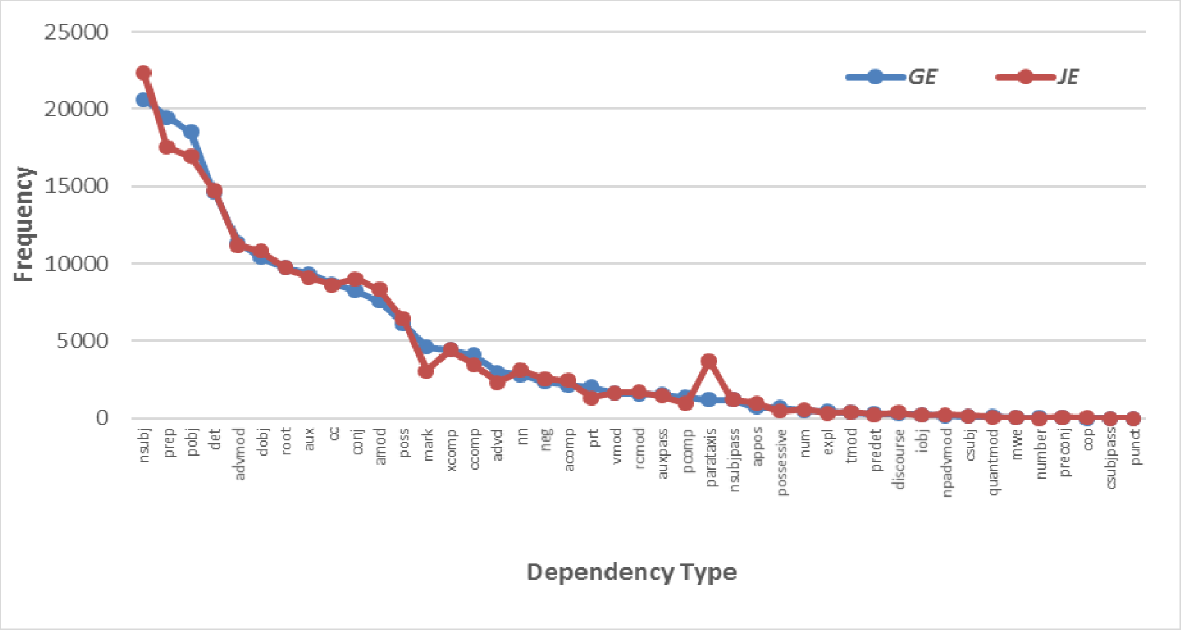

Fig. 3 The frequency of all dependency types in *GE* and *JE*

Table 7
The distribution of frequencies for each dependency type in *GE* and *JE*

	Frequency			**Frequency**	
Dependency Type	***GE***	***JE***	**Dependency Type**	***GE***	***JE***
nsubj	20651	22324	***auxpass***	1555	1434
prep	19471	17510	***pcomp***	1372	949
pobj	18526	16951	***parataxis***	1222	3700
det	14654	14765	***nsubjpass***	1191	1218
advmod	11339	11171	***appos***	725	988
dobj	10420	10798	***possessive***	702	478
root	9720	9744	***num***	490	598
aux	9363	9127	***expl***	440	322
cc	8716	8602	***tmod***	372	413
conj	8264	9016	***predet***	288	269
amod	7598	8319	***discourse***	279	356
poss	6128	6434	***iobj***	236	205
mark	4621	3082	***npadvmod***	183	237
xcomp	4449	4426	***csubj***	137	160
ccomp	4109	3492	***quantmod***	110	66
advcl	2973	2329	***mwe***	37	60
nn	2813	3105	***number***	26	8
neg	2382	2568	***preconj***	26	72
acomp	2174	2465	***cop***	15	26

prt	2013	1325	***csubjpass***	3	5
vmod	1610	1652	***punct***	1	1
rcmod	1578	1710			

From the figure and table above, both *GE* and *JE* have 43 same dependency types with similar frequencies in overall distribution. Among all these dependency types, the most dominant types in *GE* and *JE* are *nsubj*, *prep*, *pobj*, *det*, *advmod*, *dobj*, *root*, *aux*, *cc* and *conj*. What is more, the data displayed in Fig. 3 show that the most frequent type is *nsubj* for both with the frequencies of 20,651 and 22,324 respectively. This result confirms the research carried out by Wang (2015) that the most frequent dependency type in IMA (the imaginative domain) is *nsubj* while in the other domains the most frequent dependency type is *prep*. To be specific, the *nsubj* is a noun phrase which is the syntactic subject of a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or a noun (De Marneffe et al., 2008). For instance, in the sentences from *GE but the child is small, and the world is small*, the relation between the first *small* and *child* is *nsubj*, and the relation between the second *small* and *world* is *nsubj* as well. The reason why the *GE*, *JE* and even the imaginative domain having a large amount of *nsubj* may be that there are plenty of sentences that describe people, events, views, etc. in *GE* and *JE*. A Chi-Square test was then used to test whether the difference between *GE* and *JE* on the distribution of frequencies for each dependency type is significant or not. It is noted that the Chi-Square test is unreliable when the expected frequencies in any cell fall below 5, and it is advisable to apply Yate's correction, Likelihood Ratio, Fisher Exact test, etc., to get a reliable statistic. Since Yates's correction is only applied to 2 by 2 tables, the Likelihood Ratio was chosen instead. The result (listed in Appendix II(i)) shows that the differences between *GE* and *JE* in Fig. 3 are highly significant ($p = 0.00 < 0.05$).

The MDDs of each dependency type in *GE* and *JE* are shown in Fig. 4 and Table 8 below based on formula (3).

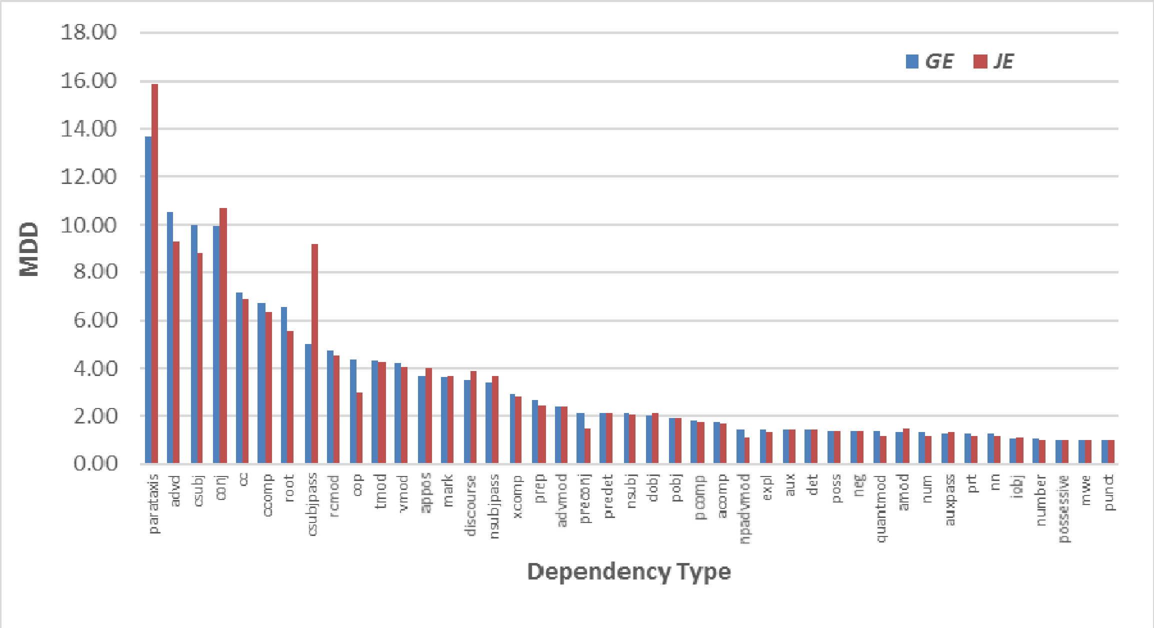

Fig. 4 The MDDs of all dependency types in *GE* and *JE*

Table 8

The distribution of MDDs for each dependency type in *GE* and *JE*

	MDDs			**MDDs**	
Dependency Type	**GE**	**JE**	**Dependency Type**	**GE**	**JE**
parataxis	13.71	15.89	***dobj***	2.02	2.14
advcl	10.52	9.31	***pobj***	1.91	1.88
csubj	10.01	8.79	***pcomp***	1.77	1.72
conj	9.95	10.69	***acomp***	1.71	1.65
cc	7.17	6.9	***npadvmod***	1.42	1.14
ccomp	6.75	6.38	***expl***	1.41	1.3
root	6.57	5.54	***aux***	1.4	1.42
csubjpass	5	9.2	***det***	1.4	1.42
rcmod	4.75	4.51	***poss***	1.34	1.33
cop	4.4	2.96	***neg***	1.34	1.37
tmod	4.35	4.31	***quantmod***	1.33	1.21
vmod	4.24	4.07	***amod***	1.3	1.46
appos	3.68	3.99	***num***	1.29	1.18
mark	3.61	3.69	***auxpass***	1.27	1.31
discourse	3.51	3.93	***prt***	1.26	1.17
nsubjpass	3.39	3.7	***nn***	1.22	1.19
xcomp	2.93	2.84	***iobj***	1.09	1.13
prep	2.64	2.46	***number***	1.08	1
advmod	2.36	2.36	***possessive***	1.03	1.01
preconj	2.12	1.46	***mwe***	1	1
predet	2.11	2.09	***punct***	1	1
nsubj	2.09	2.07			

As the figure shows, the MDDs of *npadvmod, expl, aux, det, poss, neg, quantmod, amod, num, auxpass, prt, nn, iobj, number, possessive, mwe,* and *pinct* are all around 1. It means that there are almost no words between the governors and the dependents in these types since the calculation of MDDs employs 1 as the reference value. In other words, these 17 dependency types are all adjacent dependencies. To be specific, dependency types from *GE,* such as *amod* (*lady, young*) in *young lady, aux* (*betrayed, have*) in *have betrayed*, *det* (*the, dog*) in *the dog* and *nn* (*wall, stone*) in *stone wall* are all belonging to the category of adjacent dependencies. Meanwhile, Fig. 4 shows that the dependency types, *parataxis, advcl, csubj, conj* and *cc* in *GE* and *JE,* have apparently larger MDDs. In other words, these dependency types cost more cognitive efforts than the adjacent ones.

Although the dependency types, *parataxis, conj* and *csubjpass* in *JE* have the longer MDDs than those in *GE*, there are also some types in *GE* having longer MDDs than that in *JE*, such as *advcl*, *csubj* and *cc*. The Chi-Square test was then conducted and the result is listed in Appendix II(j). Since the expected frequencies in some cells fall below 5, the Likelihood Ratio was applied, which indicates that the differences between *GE* and *JE* on MDDs of dependency types are not significant (p = 1.000 > 0.05). This may be attributed to the fact that both *GE* and *JE* belong to the category of imaginative works.

In sum, there are significant differences on the distribution of dependency types in *GE* and *JE,* but the differences on MDDs of dependency types in *GE* and *JE* are not significant.

4. Conclusion

The study attempts to investigate specific literary works, *GE* and *JE*, from the perspective of dependency grammar and explores MDDs, ADDs, dependency directions and dependency types of both works. The major findings are stated as below:

The MDD of *GE* is slightly shorter than that of *JE*, but the difference of the syntactic difficulty between *GE* and *JE* is not statistically significant. In other words, the text of *GE* is not significantly easier than that of *JE*.

There is a similar trend in ADDs' distribution, but the differences between *GE* and *JE* on ADDs are highly significant. On the one hand, there is no significant difference of the distribution of ADDs ranging from 0 to 9 and ADDs no less than 10 in *GE* and *JE*, and the smaller dependency distances ranging from 0 to 9 in *GE* and *JE* occupy a dominant proportion. One the other hand, there is a highly significant difference of the distribution of adjacent dependencies and non-adjacent ones in *GE* and *JE*: *JE* comparatively has more adjacent dependencies.

The differences of the distribution of dependency directions in *GE* and *JE* are not significant, and both works share the same distribution tendency: the smaller the absolute dependency distances, the higher the frequencies. However, there exist highly significant differences between *GE* and *JE* in the frequency and proportion of positive and negative dependencies, the adjacent dependencies, and the non-adjacent dependencies.

Both *GE* and *JE* have 43 same dependency types with similar frequencies in overall distribution. However, the differences between *GE* and *JE* are highly significant. To be specific, there are 17 dependency types belonging to the category of adjacent dependencies with MDDs around 1. On the other hand, there are some dependency types in *JE* are slightly longer than that of *GE,* but the differences on MDDs of dependency types in *GE* and *JE* are not significant.

References

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

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

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O. (2007). Open information extraction for the web. In: *IJCAI* 7, 2670-2676.

Covington, M.A. (2001). A fundamental algorithm for dependency parsing. In: *Proceedings of the 39th Annual ACM Southeast Conference*, 95-102.

De Marneffe, M.C., MacCartney, B., & Manning, C.D. (2006). Generating typed dependency parse from phrase structure parses. In: *Proceedings of LREC* 6, 449-454.

De Marneffe, M.C., Manning, C.D. (2008). Stanford typed dependencies manual. URL *http://nlp.stanford.edu/software/dependencies_manual.pdf*.

Eppler, E.M. (2005). *The syntax of German-English code-switching*. Doctoral dissertation. University of London.

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

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76.

Gibson, E., Pearlmutter, N.J. (1998). Constraints on sentence comprehension. *Trends Cognitive Sciences* 2(7), 262–268.

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

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

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper. URL *http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf*.

Hudson, R. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Kessler, J.S. (2008). Polling the blogosphere: a rule-based approach to belief classification. In *ICWSM*, 68-75.

Jiang, J., Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications – Based on a parallel English – Chinese dependency treebank. *Language Sciences* 50, 93-104.

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

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In: Lehfeldt, W., Strauss, U. (eds.), *Glottometrika* 4. Bochum:Brockmeyer, 103-113.

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

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

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9 (2), 159–191.

Liu, H. (2009). *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.

Liu, H. (2009). Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics* 16(3), 256-273.

Liu, H. (2010). Dependency direction as a means of word-order typology: a method based on dependency treebanks. *Lingua* 120 (6), 1567–1578.

Liu, H., Hudson, R., Feng, Z. (2009a). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory* 5 (2), 161–174.

Liu, H., Zhao, Y., Li, W. (2009b). Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznan Studies in Contemporary Linguistics* 45(4), 509–523.

Meena, A., Prabhakar, T.V. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In: G. Amati, C. Carpineto, G. Romano (eds.), *ECIR* 2007, *LNCS* 4425. Springer Berlin Heidelberg, 573 -580.

Menzerath, P. (1928). Über einige phonetische probleme. In: *Actes du Premier Congrès International des Linguistes*, Leiden: Sijthoff, 570-571.

Nivre, J. (2006). *Inductive Dependency Parsing*. Dordrecht: Springer.

Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In: *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 42–53.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition* 105(2), 300–333.

Tesnière, L. (1959). *Eléments de Syntaxe Structurale*. Paris: Libraire C. Klincksieck.

Wang, Y. (2015). *A Study on the Distribution of Dependency Distances in Different Domains of Written English in the BNC*. Unpublished master dissertation. Dalian Maritime University.

Wang, Y., Liu H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences* 59, 135-147.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics* 1, 98-106.

Appendices

Appendix I(a). Table of all dependency distances with directions in *GE*

Distance	Frequency	Distance	Frequency	Distance	Frequency
144	1	31	19	-39	53
142	1	30	28	-40	42
139	1	29	30	-41	33
130	1	28	35	-42	39
128	1	27	40	-43	33
125	1	26	42	-44	30
124	1	25	41	-45	21
121	1	24	52	-46	26
117	1	23	67	-47	20
113	1	22	48	-48	14
112	1	21	65	-49	10
108	1	20	81	-50	13
104	1	19	90	-51	17
103	1	18	110	-52	13
99	1	17	104	-53	13
98	1	16	132	-54	12
96	1	15	140	-55	22
90	1	14	176	-56	12
89	1	13	166	-57	5
88	2	12	226	-58	11
84	2	11	258	-59	8
83	1	10	327	-60	7

82	3	9	402	-61	4
81	1	8	490	-62	1
79	1	7	665	-63	9
78	1	6	874	-64	8
74	1	5	1362	-65	5
73	1	4	2639	-66	7
72	2	3	7066	-67	4
71	1	2	18372	-68	3
70	2	1	50680	-69	8
69	2	-1	34003	-70	4
68	2	-2	26133	-71	4
67	3	-3	12739	-72	1
66	2	-4	6630	-74	1
65	4	-5	4313	-75	4
64	3	-6	3022	-76	2
63	2	-7	2394	-77	3
62	2	-8	1835	-78	1
61	3	-9	1613	-79	1
60	3	-10	1293	-80	5
59	3	-11	1068	-81	2
58	4	-12	921	-82	3
57	3	-13	738	-83	4
56	3	-14	652	-84	2
55	3	-15	539	-85	3
54	1	-16	475	-86	2
53	3	-17	428	-87	3
52	5	-18	346	-88	3
51	2	-19	334	-89	2
50	5	-20	269	-90	3
49	8	-21	253	-91	1
48	2	-22	217	-92	2
47	7	-23	213	-93	2
46	10	-24	198	-94	3
45	6	-25	149	-95	2
44	3	-26	142	-98	1
43	7	-27	128	-99	2
42	8	-28	107	-102	1
41	14	-29	103	-103	1
40	15	-30	98	-104	2
39	16	-31	89	-105	1
38	15	-32	74	-108	1
37	17	-33	88	-113	1
36	11	-34	74	-115	1

35	11	-35	73	-118	2
34	17	-36	65	-122	1
33	20	-37	59	-125	1
32	31	-38	39	-129	1

Appendix I(b). Table of all dependency distances with directions in *JE*

Distance	Frequency	Distance	Frequency	Distance	Frequency
105	1	12	239	-56	16
103	1	11	262	-57	13
97	1	10	349	-58	21
94	1	9	404	-59	16
91	2	8	598	-60	21
88	1	7	671	-61	9
81	1	6	941	-62	16
80	1	5	1462	-63	11
78	3	4	2912	-64	22
77	3	3	6298	-65	7
74	1	2	17649	-66	11
73	1	1	53386	-67	9
71	2	-1	32623	-68	11
69	3	-2	26486	-69	10
68	2	-3	12318	-70	8
66	2	-4	6635	-71	8
65	1	-5	4285	-72	9
64	2	-6	2874	-73	10
63	3	-7	2262	-74	6
62	1	-8	1764	-75	7
61	2	-9	1423	-76	5
59	3	-10	1111	-77	4
57	2	-11	941	-79	1
56	2	-12	802	-80	5
55	3	-13	757	-81	8
54	4	-14	642	-82	5
53	3	-15	553	-83	3
52	1	-16	488	-84	4
51	1	-17	356	-85	2
50	5	-18	360	-86	3
49	3	-19	336	-87	3
48	5	-20	314	-88	2
47	3	-21	244	-89	2
46	3	-22	236	-90	3
45	5	-23	226	-91	1
44	4	-24	210	-92	4

43	7	-25	198	-93	4
42	5	-26	175	-94	3
41	6	-27	176	-95	2
40	9	-28	132	-96	2
39	9	-29	142	-98	1
38	5	-30	118	-99	1
37	11	-31	118	-100	1
36	8	-32	111	-101	1
35	12	-33	109	-102	1
34	10	-34	92	-103	1
33	12	-35	85	-104	2
32	13	-36	71	-105	1
31	15	-37	92	-106	1
30	16	-38	78	-108	3
29	15	-39	69	-110	2
28	22	-40	52	-112	1
27	28	-41	67	-113	1
26	29	-42	54	-114	2
25	31	-43	57	-116	1
24	29	-44	47	-117	1
23	45	-45	46	-120	1
22	52	-46	42	-122	1
21	45	-47	31	-125	1
20	67	-48	35	-126	2
19	64	-49	28	-128	1
18	71	-50	33	-131	1
17	80	-51	32	-138	2
16	90	-52	35	-141	1
15	130	-53	35	-153	1
14	142	-54	19	-191	1
13	164	-55	31		

Appendix II(a). Comparison between *GE* and *JE* on MDDs

Test Statistics[a]

	MDD
Mann-Whitney U	4.699E7
Wilcoxon W	9.476E7
Z	-1.440
Asymp. Sig. (2-tailed)	.150

a. Grouping Variable: Work

Appendix II(b). Comparison between *GE* and *JE* on the frequency of ADDs

Test Statistics[a]

	ADD
Mann-Whitney U	1.747E10
Wilcoxon W	3.505E10
Z	-3.778
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Work

Appendix II(c). Comparison between *GE* and *JE* on the distribution of ADDs

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.750[a]	1	.053		
Continuity Correction[b]	3.723	1	.054		
N of Valid Cases[b]	375127				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10908.19.

b. Computed only for a 2x2 table

Appendix II(d). Comparison between *GE* and *JE* on the distribution of adjacent dependencies and non-adjacent dependencies

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	20.639[a]	1	.000		
Continuity Correction[b]	20.609	1	.000		
N of Valid Cases[b]	375127				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 85316.20.

b. Computed only for a 2x2 table

Appendix II(e). Comparison between *GE* and *JE* on the frequency of dependency distances

Test Statistics[a]

	DD
Mann-Whitney U	1.758E10
Wilcoxon W	3.518E10
Z	-.246
Asymp. Sig. (2-tailed)	.806

a. Grouping Variable: Work

Appendix II(f). Comparison between *GE* and *JE* on the distribution of overall positive and negative dependencies

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	21.931[a]	1	.000		
Continuity Correction[b]	21.900	1	.000		
N of Valid Cases[b]	375127				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 85787.53.

b. Computed only for a 2x2 table

Appendix II(g). Comparison between *GE* and *JE* on the distribution of adjacent positive and negative dependencies

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	88.651[a]	1	.000		
Continuity Correction[b]	88.558	1	.000		
N of Valid Cases[b]	170692				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 33054.21.

b. Computed only for a 2x2 table

Appendix II(h). Comparison between *GE* and *JE* on the distribution of non-adjacent positive and negative dependencies

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16.178[a]	1	.000		
Continuity Correction[b]	16.140	1	.000		
N of Valid Cases[b]	204435				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 33543.72.

b. Computed only for a 2x2 table

Appendix II(i). Comparison between *GE* and *JE* on the distribution of frequencies for each dependency type

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.462E3[a]	42	.000
Likelihood Ratio	2.527E3	42	.000
N of Valid Cases	365462		

a. 4 cells (4.7%) have expected count less than 5. The minimum expected count is 1.00.

Appendix II(j). Comparison between *GE* and *JE* on the distribution of MDDs for each dependency type

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	2.549[a]	42	1.000	1.000		
Likelihood Ratio	2.575	42	1.000	1.000		
N of Valid Cases	284					

a. 68 cells (79.1%) have expected count less than 5. The minimum expected count is 1.00.

Glottometrics 17, 2017, 34-53

The Study of Adverbials in Czech

Kateřina Pelegrinová[1]*, Gabriel Altmann*

Abstract. Each well defined linguistic concept can be studied quantitatively. Though this way has no end, one must perform the study stepwise. Here we analyze the behavior of adverbs and adverbial expressions and apply the models to Czech texts. The adverbials are classified in 13 classes and we study the class size, the length in individual classes, the placing of adverbials, the runs of left and right adverbials and the gaps between right adverbials. Further problems are sketched in the Introduction.

Keywords: Adverbials, Czech, classes, sequences, runs, gaps, placing, models

1. Introduction

In a „normal“ sentence there is always something (subject, topic, theme) one speaks about and something one says about this "subject" (rheme, comment). Sentences not fulfilling this criterion contain ellipses which can mostly be reconstructed on the basis of the context.

If a certain noun is the "subject" of information, then the rest of the sentence consists of predicates. The (logical) predicates of the first order are adjectives and verbs. The adjective is mostly – but not always - part of the noun phrase, the verb is the head of the verb phrase. Everything else is a predicate of the second order, e.g. those parts of sentence which belong to the set of verb valency. One part of the second order predicates are adverbials whose identification and classification may differ from researcher to researcher and from language to language. Adverbial expressions need not consist of one word only. One may consider also adverbial phrases and clauses – depending on definition and interest. Adverbial expressions may contain various parts of speech and further predicates of higher order. Adverbial expressions may be distinguished from simple adverbs. There may be expressions which do not contain an adverb, e.g. *the house on the mountain;* but an adverbial expression may contain one or more adverbs or words.

Adverbial expressions have various properties, all of which must be strictly operationalized if one wants to find a textual or stylistic regularity. Let us enumerate at least some of them.

(1) Length measured in terms of word numbers. Unfortunately, this is not quite simple because one must decide whether clitics (like e.g. the Slavic zero-syllabic prepositions, reflexive pronouns with verbs, Indonesian "-kah", "-lah", Japanese "ka", French persons) are integral parts of words or one follows the official way of writing when counting. The problem is still more serious in non-alphabetic languages, e.g. Chinese. In Japanese, all "prepositions" stay behind the noun, and have the same role as affixes in Hungarian; or should one consider names as one word or more, etc. As can be seen, the decision is nothing "objective", one works with a given definition in order to find some regularity. A secondary way to decide

[1] [1] Katarina Pelegrinova (pelegrinovak@seznam.cz), University Ostrava, Czech Republic:; Gabriel Altmann (RAM-Verlag@t-online.de)

which of the definitions is "better" is the best agreement of the obtained data with one of the possible models.

(2) The affiliation of the adverbial expression to a given class, cf. those found in Czech by Čech and Uhlířová (2014): Place, Time, Manner, Means, Aspect, Condition, Measure, Cause, Result, Origin, Purpose, Concession, Originator. Different authors use a smaller set of classes, e.g. Yesypenko (2008) distinguishes adverbs of I. Repetition and frequency, II. Place and direction, III. Condition and consequence, IV. Manner, V. Degree and quantity, VI. Question adverbs. Quite other classifications may be found in various works (cf. Internet: http://hypermedia.ids-mannheim.de/call/public/gruwi.ansicht?v_id=525). But even here, sometimes decisions are necessary, e.g. in the first line of "Erlkönig" by Goethe we have "Wer reitet so spät durch Nacht und Wind?" where the first adverbial "so spät" belongs evidently to the Time-class, but the second, "durch Nacht und Wind" (*through night and wind*) is not easy to classify. As a matter of fact it belongs to two different classes (Time and Manner?) but the given part "so spät durch Nacht und Wind" can be considered a unique adverbial expression. Needless to say, each class can be further subdivided according to various criteria, hence the number and quality of classes is different with every researcher. It is to be remarked that methodologically those definitions and classifications are "better" which lead to the establishing of some regularity. They do not represent "truth". Any classification in linguistics is a striving for finding some elementary order.

(3) Frequency in the text which can easily be stated if one considers individual words (adverbs) or individual classes (adverbial expressions). Since one speaks about classes, each occurrence even of the same adverbial must be counted. The frequency in the text yields a rank-frequency distribution which can be used for text(type) characterization. However, if the adverbials were scaled in some way, one could obtain also a kind of special spectrum of frequencies. Without scaling one must consider very long texts in order to obtain a reasonable spectrum. One can see that the first steps in the research, namely definition, identifications, segmentation, scaling and counting are the most "uncertain" activities in any kind of research. The development of scientific disciplines, new paradigms, scientific revolutions, etc. are the best witnesses of the change of our view of reality.

(4) Polysemy which may cause the attribution of the same adverbial to different classes, e.g. the German "gerade" may have a "manner" and a "temporal" meaning. Frequently, even the context does not always allow making a definite decision. Complex adverbial expressions are in each case a problem. However, the problem exists in all POS classes; well known is the problem of the adjective "hard" which may belong to many classes. In order to study the polysemy of adverbials the text alone is not sufficient, one must take into consideration also a large dictionary of the given language. In general, one may suppose that the shorter the adverbial, the stronger may be its polysemy – because it has few predicates of higher order which accompany and specify it.

(5) The number of parts of speech which are present in the adverbial, e.g. the expression "durch Nacht und Wind" contains 4 words but only 3 parts of speech. The first way of measurement concerns length (in terms of word numbers), the second one concerns complexity. Up to now, no scaling of complexity in this domain has been proposed. The problem gets very complex if the adverbial expression is a whole clause. Evidently, typology should devote more attention to this phenomenon. The complexity in strongly analytic languages may differ from that in strongly synthetic languages.

(6) The number of grammatical categories present in all words of the adverbial expression. In some languages they contain case, number, gender, time, mood, position, etc. This depends rather on the prevailing type of language hence this way of investigation is rather typological. It can be expected that the longer is the adverbial, the more grammatical categories will be contained in it.

(7) Psycholinguistic properties of the adverbial expressions (dogmatism, polyanna, emotionality, imagery, etc.), their possible scaling, distribution, etc.

(8) Discourse properties (cf. e.g. Jørgensen, Phillips 2002).

(9) A possible attribution to a kind of speech act. Aspects (7) to (9) represent a special discipline for which adverbials are merely one of the possibilities.

(10) Position in relation to its head (if there is some), which can be either dichotomized (e.g. before – behind or left – right) or measured in form of distance from the head of the sentence. The distance can be measured in various ways, usually one considers the number of words lying in between. In the poetic language one may subdivide the adverbial expression in two parts first of which is placed in front of the verse, the second behind it. We shall distinguish here merely left and right position and study their frequencies, runs and gaps.

(11) Mean predicate value of the complete adverbial expression. Of course, the scaling must be determined before the analysis. If a verb is a predicate of first order, then its adverb is a predicate of second order. But there may be still prepositions, conjunctions and other parts which must obtain a special degree, too. Hence every sentence may be presented as a sequence of predicative degrees and one may obtain a new field of investigation. Here, the dependence (or other) grammar could excellently serve our strivings. One could state the order/degree of predication from the place of the word in the graph representing the dependencies in the sentence.

(12) Is the adverbial a predicate of a noun, of a verb, of an adjective, or of another adverbial, etc., i.e. there are different possible weightings yielding new vistas.

(13) A number of various properties of adverbials can be found in books or articles dedicated to them (cf. e.g. Ney 1982; Hoye 1997; Rijkhoff 2002; Thompson, Longacre 1985; Diessel 2005; Ford 1993).

One cannot study all the properties because we still do not even know what is relevant. The relevance of a property can be judged by its involvement in the control cycle as proposed by Köhler (1986, 2005) and by the state of its theoretical substantiation. We adhere to two principles: First, no property of language is completely isolated, all are parts of some self-regulation cycle. Second, if we want to obtain laws, we must derive a hypothesis from a background theory, test it in many languages and find its links to other properties.

Grammarians study adverbials from different points of view: syntactic rules, meaning, form, placing, etc. However, the construction of any theory must transcend this practical level. Here, we shall restrict ourselves to some few points because a very broad examination opens a door into a separate discipline. We shall consider only Czech texts but the methods and the results can (must) be generalized.

2. Simple adverbs and adverbial expressions

Simple adverbs yield us the first image of the “verbosity” of a text. The more adverbs there are, the more precisely the entities are described, the deeper are the sentences syntactically. Considering simple adverbs, one can omit prepositions, conjunctions, pronouns, interjections. However, in German, there are prefixes, identical with prepositions; if they are detached (e.g. *ich sehe mich **vor***), one considers them adverbs.

The simple count of adverbs yields a different picture than the size of classes set up according to meaning, even if both may follow some variant of Zipf’s law. In the same way, adverbial expressions and their classes may display a quite different image.

In order to scrutinize some properties of adverbials and show some problems we consider the situation in Czech texts.

For the sake of exemplification we show the adverbials belonging to the class “Time” in the Czech text T 1.

Time		
nikdy	potom	už
dávno	dneska	pak
v životě	v polovině	najednou
stále	znovu	ihned
nikdy	při akci	po události
zároveň	v den	měsíc
v letech	teď	pak
v dubnu	pořád	při natáčení
hned	v roce	nedávno
někdy	na léta	v době
v letech	o devatenáct let	opět
na chvíli	později	dlouho
většinou	v roce	

Čech and Uhlířová (2014) measured the size/frequency of individual classes and stated that the rank-frequency distribution abides by the Zipf-Alekseev function. If we count the numbers of adverbial expressions in classes we obtain the results presented in Table 1. The model fitted to the data may be either a discrete distribution or a sequence or even a continuous function. The model itself merely shows that there is some regularity which can be captured mathematically and subsumed under a background theory (e.g. Wimmer, Altmann 2005). The next step of justification is its link to other properties of texts.

Since the observed ranked distribution of classes studied here is short and simple, for the given numbers one can find more than 30 discrete distributions; however, it may happen that the theoretical distribution is bell shaped because the first two values are equal. Here we shall use merely the Zipf-Alekseev continuous function defined as

(1) $$f(x) = cx^{a+b\ln x}$$

resulting from the differential equation

(2) $$\frac{dy}{y} = \frac{K + M * \log x}{Rx} dx \, ,$$

which is based on the unified theory (cf. Wimmer, Altmann 2005) and yields (1) after reparametrization. It has been applied to data of various kinds and it seems to be a good extension of the power function proposed by G.K. Zipf for ranking the frequencies.

One adheres to the given model as long as no or only a small number of exceptions appear. In that case, one tries to explain the exceptions as boundary conditions; one may modify some classes, add a parameter, pool some small classes, etc. but in the end one should find a general model which fits to the majority of data. This cannot be done analyzing only one language, but one must begin somewhere.

The results of ranking the individual classes in Czech texts are presented in Table 1. The ranks are not ascribed to the same classes in all texts but depend on the frequency.

Table 1
Adverbials in Czech texts: ranking of class sizes
(Zipf-Alekseev function)

	T 1		T 2		T 3		T 4		T 5	
Rank	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
1	56	55.46	77	76.72	77	76.66	65	64.47	100	100.16
2	52	54.57	64	65.64	55	57.00	34	37.39	79	78.28
3	38	33.20	37	32.92	34	30.26	26	22.34	34	36.13
4	19	18.76	14	15.43	16	15.71	17	14.20	18	15.75
5	6	10.64	4	7.35	7	8.40	9	9.50	6	7.04
6	5	6.17	3	3.63	2	4.67	7	6.62	3	3.29
7	4	3.68	2	1.87	1	2.69	2	4.77	3	1.60
8	3	2.25	1	1.00	1	1.60	1	3.53	3	0.82
9	2	1.42	1	0.55	1	0.99	1	2.67	2	0.43
10	1	0.91	1	0.31	1	0.62				
11	1	0.60								
	a = 0.7345 b = -1.0940 c = 55.4763 R^2 = 0.9881		a = 0.7072 b = -1.3447 c = 76.7159 R^2 = 0.9955		a = 0.2885 b = -1.0329 c = 76.6617 R^2 = 0.9954		a = -0.4805 b = -0.4408 c = 64.4728 R^2 = 0.9859		a = 0.6241 b = -1.4126 c = 100.1160 R^2 = 0.9981	
	T 6		T 7		T 8		T 9		T 10	
Rank	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
1	128	127.94	66	65.77	36	35.39	44	44.28	61	61.05
2	46	46.93	58	58.79	35	38.06	28	26.15	40	39.34
3	24	20.98	31	31.62	30	22.50	13	15.36	18	20.81
4	9	10.75	23	15.89	7	12.02	8	9.52	15	11.17
5	4	6.05	2	8.09	4	6.39	8	6.21	6	6.26
6	4	3.65	2	4.25	3	3.47	4	4.21	3	3.66
7	3	2.32	1	2.30	3	1.94	3	2.96	1	2.22
8	2	1.54	1	1.30	2	1.11	3	2.14	1	1.40
9	1	1.06			1	0.66	2	1.58	1	0.91
10							2	1.19		
11							2	0.91		
	a = -1.1074 b = -0.4898 c = 127.9379 R^2 = 0.9987		a = 0.7008 b = -1.2447 c = 65.7705 R^2 = 0.9809		a = 0.9889 b = -1.2751 c = 35.3864 R^2 = 0.9474		a = -0.4107 b = -0.5034 c = 44.2796 R^2 = 0.9904		a = -0.0431 b = -0.8526 c = 61.0523 R^2 = 0.9929	
	T 11		T 12		T 13		T 14		T 15	
Rank	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
1	72	70.92	100	99.74	38	37.87	91	90.12	95	94.55
2	47	53.16	53	55.25	37	37.52	44	50.80	54	57.68
3	41	31.63	32	26.31	22	21.51	40	28.28	37	26.52
4	22	18.70	11	13.02	13	11.32	19	16.69	6	12.29
5	6	11.37	4	6.82	4	5.98	4	10.40	2	5.97
6	3	7.15	3	3.76	2	3.24	4	6.78	1	3.05
7	2	4.63	1	2.17	2	1.81	2	4.58	1	1.64

8	2	3.09	1	1.31	1	1.04	2	3.20	1	0.91
9	2	2.11			1	0.62	1	2.29		
10	1	1.47			1	0.38				
	a = 0.1295 b = -0.7870 c = 70.9231 R^2 = 0.9662		a = -0.2360 b = -0.8892 c = 99.7409 R^2 = 0.9941		a = 0.8436 b = -1.2368 c = 37.8732 R^2 = 0.9954		a = -0.4377 b = -0.5618 c = 90.1202 R^2 = 0.9667		a = 0.0459 b = -1.0949 94.5520 R^2 = 0.9784	
	T 16		T 17		T 18					
Rank	f_x	f_t	f_x	f_t	f_x	f_t				
1	71	70.73	113	112.94	43	43.22				
2	26	29.27	39	39.34	33	32.04				
3	23	14.07	19	19.55	18	18.51				
4	3	7.59	12	11.48	9	10.62				
5	2	4.45	11	7.43	5	6.27				
6	2	2.78	5	5.14	5	3.83				
7	1	1.82	2	3.73	4	2.42				
8	1	1.23	1	2.80	4	1.57				
9	1	0.87	1	2.17	2	1.05				
10	1	0.62								
11	1	0.46								
	a = -0.9362 b = -0.4861 c = 70.7315 R^2 = 0.9745		a = -1.3936 b = -0.1846 c = 112..9437 R^2 = 0.9980		a = 0.1488 b = -0.8378 c = 43.2178 R^2 = 0.9980					

Here, the parameter *c* depends on the first frequency; parameter *a* is some constant dictated both by the given language and by the hearer/reader who wants to maintain the equilibrium satisfying his needs. Parameter *b* may be considered the result of the effort of the writer. It may differ with different authors, styles, text types but it may also develop. Its analysis and description will require very extensive investigations. In the above texts it is always negative.

In case of text T 8, we obtain a good fitting but the theoretical function has its maximum at x = 2. A monotonous decrease can be attained by applying a simpler function but the fitting is, in any case, very satisfactory. The number of texts, text types (here merely journalistic and poetic) and the number of languages do not allow generalization but one can consider the result as a good basis for further investigations.

3. Length of adverbials

Though the above data are very restricted because of the shortness of some texts, one can order the classes according to the average length of adverbial expressions in them. Length is measured in terms of the number of words in the adverbial. Adding the lengths of adverbials in a given class and dividing the sum by their number we obtain the mean length of an adverbial class, as displayed in Table 2. However, one could apply also the number of morphemes but not the number of syllables.

It would not be fruitful to study directly the distribution of lengths in each class separately because there are usually few length classes or there are too few adverbials in some classes. However, averages are a sufficient background.

Table 2
Mean length of adverbials in individual classes in Czech texts

	T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9
Place	2.02	2.23	2.09	1.99	2.23	2.42	2.21	3.05	1.93
Time	1.88	2.49	1.65	1.64	2.22	1.91	2.58	2.00	1.15
Manner	1.58	1.54	1.71	2.35	1.35	1.67	1.50	1.35	1.57
Means	1.33	1.00	1.00	-	1.00	1.00	1.50	1.33	1.00
Aspect	3.80	2.00	2.00	2.00	2.00	2.25	3.00	2.00	2.00
Condition	4.75	-	7.00	5.00	-	2.50	-	2.00	2.00
Measure	1.32	1.14	1.13	1.24	1.20	1.44	1.00	1.14	1.00
Cause	5.00	8.50	6.14	6.22	6.20	5.00	3.50	4.00	6.38
Purpose	4.00	3.67	5.50	5.14	4.00	5.25	5.00	11.33	7.00
Concession	5.00	10.00	9.00	7.00	-	-	-	-	3.33
Originator	-	-	-	-	-	-	-	-	-
Result	2.00	2.00	-	-	-	-	-	-	-
Origin	-	-	-	-	2.00	-	-	-	2.00

	T 10	T 11	T12	T 13	T 14	T 15	T 16	T 17	T 18
Place	1.95	1.71	1.84	1.96	1.83	1.65	1.82	1.88	1.93
Time	2.33	1.77	1.70	2.68	1.98	1.68	2.78	1.26	1.28
Manner	1.28	1.71	1.32	1.11	1.90	1.41	1.31	1.64	1.39
Means	1.00	-	1.00	1.00	1.00	-	1.00	1.00	1.00
Aspect	2.00	-	-	-	-	-	2.00	-	2.00
Condition	2.00	3.00	2.00	-	-	4.00	3.00	7.50	-
Measure	1.16	1.09	1.45	1.31	1.05	1.00	6.00	1.18	2.00
Cause	7.67	6.00	3.00	4.50	3.50	2.00	3.00	8.00	6.40
Purpose	-	4.33	3.00	5.00	7.50	3.00	2.50	3.40	4.00
Concession	-	4.00	-	3.75	4.50	-	5.00	-	-
Originator	-	-	-	-	-	-	-	-	-
Result	2.00	2.00	-	2.00	-	-	-	2.00	2.00
Origin	-	2.00	-	2.00	2.00	2.00	2.00	-	-

In Table 2 one can observe some outliers, e.g. in some texts the class "Concession" contains the longest adverbials. This may be caused by the fact that there is e.g. solely one adverbial "clause" and nothing else; however, there may be affixes in a language expressing "concession" and making the "concessive" adverbials much shorter.

One may ask whether the given mean length is a property of text type, a property of the language, that of the language of the author, etc. The more general question is: does a general scaling exist that holds true for all languages or is there at least a tendency that can be discovered? To this end the data must be made a little bit smoother, e.g. by ranking the values in each text separately, and the samples may be compared using Kendall's coefficient of concordance. The results of ranking for Czech texts are shown in Table 3. The individual categories having the same value obtained the mean rank (building ties) and the sum of ties is taken into account in the evaluation [http://www.real-statistics.com /reliability/kendalls-w/]. In Table 3, the means are replaced by ranks found in individual texts.

Table 3
Kendall's concordance test for ranked means of lengths of adverbials (in Czech)

	T 1	**T 2**	**T 3**	**T 4**	**T 5**	**T 6**	**T 7**	**T 8**	**T 9**
Place	6	5	5	7	3	4	5	3	7
Time	8	4	8	8	4	6	4	5	9
Manner	9	8	7	5	7	7	6.5	7	8
Means	10	10	10	11.5	9	9	6.5	8	10.5
Aspect	5	6.5	6	6	5.5	5	3	5	5
Condition	3	12	2	4	11.5	3	11	5	5
Measure	11	9	9	9	8	8	8	9	10.5
Cause	1.5	2	3	2	1	2	2	2	2
Purpose	4	3	4	3	2	1	1	1	1
Concession	1.5	1	1	1	11.5	11.5	11	11.5	3
Originator	12.5	12	12	11.5	11.5	11.5	11	11.5	12.5
Result	7	6.5	12	11.5	11.5	11.5	11	11.5	12.5
Origin	12.5	12	12	11.5	5.5	11.5	11	11.5	5

	T 10	**T 11**	**T 12**	**T 13**	**T 14**	**T 15**	**T 16**	**T 17**	**T 18**
Place	6	8.5	4	7	7	6	9	5	7
Time	2	7	5	4	5	5	5	7	9
Manner	7	8.5	7	9	6	7	10	6	8
Means	9	12	8	10	9	11	11	9	10
Aspect	4	12	11	12	11.5	11	7.5	11.5	4
Condition	4	4	3	12	11.5	1	3.5	2	12.5
Measure	8	10	6	8	8	8	1	8	4
Cause	1	1	1.5	2	3	3.5	3.5	1	1
Purpose	11.5	2	1.5	1	1	2	6	3	2
Concession	11.5	3	11	3	2	11	2	11.5	12.5
Originator	11.5	12	11	12	11.5	11	12.5	11.5	12.5
Result	4	5.5	11	5.5	11.5	11	12.5	4	4
Origin	11.5	5.5	11	5.5	4	3.5	7.5	11.5	12.5

We apply the formulas

$$W = \frac{12QSR}{m^2(N^3 - N) - m\sum_{j=1}^{m} V_j}$$

where m is the number of texts (here 18.), N is the number of adverbial classes /categories (here 13), T_i is the sum of the i^{th} row (sum of ranks of an adverbial class),

$$QSR = \sum_{i=1}^{N} (T_i - \bar{T})^2$$

is the square of the deviations of the row sums from their mean. Since we take ties of ranks into consideration, we compute for them

$$V_j = \sum_{h=1}^{S_j} (v_k^3 - v_k)$$

where S_j is the number of ties in the given text. One can obtain the chi-square as

$$X^2 = \frac{12 * QSR}{mN(N+1) - \frac{1}{N-1}\sum_{j=1}^{m} V_j},$$

or, having computed W, one takes $X^2 = m(N\text{-}1)W$, with N-1 DF. Without presenting the individual numbers and computations we state that the chi-square with 12 degrees of freedom yielding $X^2 = 112.71$ and is highly significant. That means, the representation of adverbials in this group of texts is not unique.

However, this phenomenon may be tested individually using further texts in various languages. Further, the degree of dependence must be measurable, too, in order to find a quantitative expression of the dependence. In order to test whether the mean lengths of two classes significantly differ, one may apply the normal test for difference of two means according to the formula

$$z = \frac{\bar{x}_{place} - \bar{x}_{time}}{\sqrt{Var(\bar{x}_{place}) + Var(\bar{x}_{time})}}$$

where $Var(\bar{x}) = Var(x)/n$, where the means in the formula are the means of values in Table 2, i.e. the values in Table 2 are considered x. Consider for example the mean lengths in "Place" and "Time" in Table 2. We obtain

Place = [2.02 + 2.23 + 2.09 + 1.99 + 2.23 + 2.42 + 2.21 + 3.05 + 1.93 + 1.95 + 1.71 + 1.84 + 1.96 + 1.83 + 1.65 + 1.82 + 1.88 + 1.93]/18 = 2.04

Time = [1.88 + 2.49 + 1.65 + 1.64 + 2.22 + 1.91 + 2.58 + 2.00 + 1.15 + 2.33 + 1.77 + 1.70 + 2.68 + 1.98 + 1.68 + 2.78 + 1.26 + 1.28/18 = 1.94

The variances can be obtained from the usual formulas. For the above categories we obtain z = (2.04 – 1.94)/0.1375 = 0.72 which is not significant, hence using this data one cannot confirm Zipf's conjecture, "…adverbs of time are on the average less independent and therefore shorter than adverbs of place" (1935/68: 242). Nevertheless one can ask the very general question: Are there some tendencies concerning length of adverbials in individual texts, in text types, in languages, in different epochs, etc.? The research needs a very extensive investigation.

But the difference can easily be seen in considering inductively the ranking of class means (Table 3), i.e. the sum of ranks of a class is a characteristic of the given text type or writer, or language.

4. Placing

As a matter of fact, there are 3 kinds of possible places of an adverbial: in front of the specified word (L = left), behind the specified word (R = right) and in form of a symploke, one part in front of, the second part behind the word (LR = left-right). The third possibility can be found especially in poetic language, e.g. in Slovak “na vysokom stál kopci” (*on a high he stood hill*). The place of adverbials may be characteristic for a text, text sort or language. There are languages using only R-adverbials. A simple indicator may characterize the given text. There is also the possibility that a special class of adverbials has an opposite tendency than the other classes. Hence we seek an indicator that captures all tendencies and expresses numerically the state of affairs.

The numbers are too small to use the chi-square criterion but one may test the individual classes considering L + R = n, p = 0.5, and computing the binomial probability that the numbers of L and R are not equal. Since we perform a two-sided test, the critical probability is 0.05. Consider e.g. the distribution of Time adverbials in Text 1 where we have L + R = 96 + 91 = 187 = n. Hence we compute for the L adverbials

$$P(L \geq 96) = \sum_{i=96}^{187} \binom{187}{i} 0.5^{187}$$

or for the right ones

$$P(R \leq 91) = \sum_{i=0}^{91} \binom{187}{i} 0.5^{187}$$

yielding the same result (because p = 0.5) and obtain P(L) = 0.3850. For both types we obtain 2(0.3850) = 0.7790 which is greater than 0.05, hence there is no tendency to prefer L or R types. As can be seen in Table 4, where the one-sided probabilities are given, one finds also asymmetries, e.g. in Text 2 or Text 6.

Table 4
(A)symmetry of Left-Right placing (P = binomial probability)

Text	L	R	P		Text	L	R	P
T 1	96	91	0.3850		T 10	73	77	0.4033
T 2	121	83	0.0050		T 11	93	105	0.2172
T 3	94	101	0.3338		T 12	91	114	0.0621
T 4	73	89	0.1193		T 13	77	44	0.0017
T 5	114	134	0.1138		T 14	119	85	0.0103
T 6	92	129	0.0076		T 15	112	85	0.0318
T 7	104	80	0.0448		T 16	80	52	0.0092
T 8	64	57	0.2928		T 17	102	101	0.5000
T 9	60	57	0.4267		T 18	43	80	0.0005

In texts T 2, T 7, T 13, T 14, T 15, T 16,…. the adverbials tend to stay preferably in front (left) of the described entity. The contrary tendency can be found in texts T 6, T 18.

5. Runs of L and R

In some languages (texts, text types) it may be grammatically prescribed which position must be occupied by an adverbial. In other ones, style may require a regular position which may variegate. In order to state the facts one may perform tests for runs either globally, i.e. for the complete text, or individually, for each adverbial class separately. Here we shall restrict ourselves to the global testing. Our results are restricted to one language, hence one cannot draw consequences for setting up a general law-like hypothesis. Nevertheless, one can use the results to draw consequences concerning the given language (here Czech), the given text type or the given time period. The respective formulas can be found e.g. in Bortz, Lienert, Boehnke (1990, Ch. 11). For our information only the result of the normal test (z) is interesting. If z is in interval <-1.96, 1.96>, there is no tendency. If z < -1.96, the number of runs is too small, one can suppose a structural prescription; if z > 1.96, there are too many runs and one can suppose a stylistic treatment of adverbials. However, the interpretations must be done according to the text type.

Table 5
Runs of R and L

Text	**n**	L	R	**r**	**E(r)**	**σ_r**	**z**
T 1	187	96	91	91	94.4331	6.8141	-0.50
T 2	204	121	83	98	99.4608	6.8754	-0.21
T 3	195	94	101	84	98.3744	6.9551	-2.07*
T 4	162	73	89	86	81.2099	6.2819	0.76
T 5	248	114	134	114	124.1935	7.8067	-1.30
T 6	221	92	129	110	108.4027	7.2073	0.22
T 7	184	104	80	85	91.4348	6.6481	-0.97
T 8	121	64	57	67	61.2975	5.4586	1.04
T 9	117	60	57	58	58.4615	5.3814	-0.27
T 10	150	73	77	58	775.9467	6.0988	-0.32
T 11	198	93	105	86	99.6363	6.9918	-1.95
T 12	205	91	114	107	102.2097	7.0510	0.68
T 13	121	97	44	47	57.0000	5.0662	-1.97*
T 14	207	119	88	83	102.1787	7.0145	-2.73*
T 15	197	112	85	75	97.6497	6.8677	-3.30*
T 16	132	80	52	52	64.0303	5.4631	-2.20*
T 17	203	102	101	94	102.4975	7.1061	-1.20
T 18	123	43	80	51	56.9350	5.0286	-1.18

As can be seen, the asterisk in the last column indicates the surplus or the deficiency of the number of runs. That means, in the given text there is a tendency either to place the subsequent adverbials at the same position (left or right) or change many times the position.

6. Gaps

Another aspect of the sequences of L and R can be obtained by considering the gaps between placings of the identical element. According to Skinner's (1939, 1957) hypothesis the probability of a small distance (gap) between identical entities in text is greater than the probability of greater distances. This is associated with the reinforcement of a stimulus evoked by elements of whatever kind. The gap may be counted in two ways: as a number of elements of other sort between identical elements, or as the number of steps between an element and the next identical element. The second way yields a gap which is greater (+1) then those won by the first kind of counting.

The first discoveries of this phenomenon can be ascribed to G.K. Zipf (1935, 1937a,b, 1945, 1946, 1949), later on many linguists scrutinized the phenomenon and brought a number of possible models (Spang-Hanssen 1956; Yngve 1956; Herdan 1966; Uhlířová 1967; Brainerd 1976; Králík 1977; Altmann 1984; Zörnig 1984a,b, 1987; Strauß, Sappok, Diller, Altmann 1984; Altmann 1988; Chen 1988; Chen, Cheng, Kim 1992; Prün 1997; Altmann, Köhler 2015). Here we shall adhere to the conjecture that the positions of adverbials are very abstract entities and the increase of the size of the gap is simply proportional to that of the smaller gap, i.e.

$$P_x = qP_{x-1} \quad x = 0,1,2,\ldots$$

where q ε (0,1) is constant. Solving the equation we obtain the simple geometric distribution

$$P_x = qp^x, x = 0,1,2,\ldots$$

where $p = 1 - q$. Computing the gaps between the Rs in individual texts and fitting the above formula to the frequencies of gap sizes we obtain the results presented in Table 6. Needless to say, the distributions of which the geometric is a special case would yield still better results but our aim is to simplify the modeling as far as possible. In other languages perhaps one of the other models must be applied. We used, if it was necessary, various poolings of classes.

Table 6
Fitting the geometric distribution to the frequencies of gaps between subsequent Rs of size x in Czech texts

	T 1		T 2		T 3		T 4	
Gap size	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
0	37	38.31	33	32.81	59	50.60	47	48.79
1	22	20.41	23	19.52	20	25.00	27	22.04
2	10	10.88	8	11.61	7	12.35	7	9.96
3	5	5.80	8	6.91	6	6.10	4	4.50
4	4	3.09	3	4.11	3	3.01	2	2.03
5	1	1.65	2	2.45	3	1.49	2	1.67
6	0	0.88	0	1.46	1	0.74		
7	0	0.47	2	0.87	0	0.36		
8	3	0.53	0	0.52	1	0.35		
9			1	0.31				
10			1	0.45				
	p = 0.4672 X^2 = 1.54 DF = 5 P = 0.91		p = 0.4050 X^2 = 2.76 DF = 6 P = 0.84		p = 0.5060 X^2 = 6.45 DF = 5 P = 0.26		p = 0.5482 X^2 = 2.18 DF = 4 P = 0.70	

	T 5		T 6		T 7		T 8	
Gap	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
0	77	72.32	74	73.97	38	36.24	24	24.78
1	26	33.00	34	30.89	16	19.40	13	13.62
2	18	15.05	13	12.90	16	10.39	13	7.48
3	5	6.87	1	5.39	3	5.56	1	4.11
4	4	3.13	3	2.25	1	2.98	3	2.26
5	2	1.43	1	0.94	0	1.59	1	2.75
6	0	0.65	0	0.39	0	0.85		
7	0	0.30	1	0.28	1	0.46		
8	0	0.14			0	0.24		
9	0	0.06			0	0.13		
10	1	0.05			2	0.07		
11					0	0.04		
12					1	0.04		
	p = 0.5438 X^2 = 3.37 DF = 5 P = 0.64		p = 0.5824 X^2 = 4.22 DF = 4 P = 0.38		p = 0.4646 X^2 = 6.30 DF = 4 P = 0.18		p = 0.4506 X^2 = 7.84 DF = 4 P = 0.10	
	T 9		T 10		T 11		T 12	
Gap	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
0	28	26.59	40	39.87	62	54.57	60	60.46
1	14	13.96	19	18.95	15	26.21	30	28.11
2	4	7.34	9	9.01	18	12.59	14	13.07
3	6	3.86	4	4.28	4	6.04	6	6.08
4	2	2.03	2	2.04	3	2.90	0	2.83
5	1	1.07	0	0.97	1	1.39	3	2.46
6	1	1.18	1	0.46	0	0.67		
7			1	0.42	2	0.62		
8								
9								
	p = 0.4744 X^2 = 2.82 DF = 5 P = 0.73		p = 0.5245 X^2 = 0.03 DF = 4 P = 0.9999		p = 0.5198 X^2 = 9.33 DF = 5 P = 0.0965		p = 0.5350 X^2 = 3.15 DF = 4 P = 0.5340	

	T 13		T 14		T 15*		T 16	
Gap	f_x	f_t	f_x	f_t	f_x	f_t	f_x	f_t
0	20	16.65	46	35.89	41	38.75	26	20.06
1	11	10.20	14	21.08	17	19.50	12	12.17
2	3	6.25	9	12.39	7	9.81	3	7.38
3	1	3.83	7	7.28	1	4.94	2	4.48
4	3	2.35	2	4.27	2	2.48	1	2.72
5	1	1.44	3	2.51	4	1.25	3	1.65

6	2	0.88	4	1.48	1	0.63	1	1.00
7	0	0.54	1	0.87	4	0.32	0	0.61
8	1	0.33	0	0.51	0	0.16	1	0.37
9	0	0.20	0	0.30	1	0.16	1	0.22
10	0	0.12	0	0.18			0	0.14
11	0	0.06	0	0.10			0	0.08
12	0	0.05	1	0.15			1	0.13
13	0	0.03						
14	0	0.02						
15	0	0.01						
16	0	0.01						
17	0	0.0040						
18	0	0.0025						
19	0	0.0015						
20	1	0.0024						
	p = 0.3873 X^2 = 6.16 DF = 5 P = 0.40		p = 0.4125 X^2 = 10.53 DF = 6 P = 0.10		p = 0.4988 X^2 = 2.20 DF = 2 P = 0.33		p = 0.3934 X^2 = 8.89 DF 5 P = 0.11	

	T 17		T 18	
Gap	f_x	f_t	f_x	f_t
0	59	53.02	54	50.58
1	18	25.99	15	18.20
2	17	12.74	5	6.55
3	3	6.25	3	2.35
4	3	3.06	1	0.85
5	2	1.50	1	0.48
6	1	0.74		
7	0	0.36		
8	1	0.35		

	p = 0.5098
	X^2 = 6.62
	DF = 5
	P = 0.25
			p = 0.8403
			X^2 = 1.68
			DF = 3
			P = 0.64	

- = pooling to 5

In T 11 one can see that x = 2 is smaller than x = 3. Preliminarily we may conjecture that there is some boundary condition but one can fit also another distribution which captures this deviation, e.g. the Gegenbauer distribution which is a generalization of the geometric distribution (cf. Wimmer, Altmann 1999).

Gaps can be considered not only as an expression of stimulus strength but also as a characteristic of a property of the entities taken into account. However, up to now we do not know what a property is involved. Thus, a very extensive investigation at all levels of a language would be necessary in order to determine the properties. One may conjecture that the mechanism has something to do with our cerebral mechanisms, education, inclinations but up to now only Skinner's very general hypothesis is known.

7. Further problems

As any linguistic unit, the adverbials have an infinite number of properties. We studied here merely their length, placing, runs and gaps but one can imagine that the research will continue. Some hints at the possible vistas: If we abbreviate the classes using some letters, e.g. P = place, T = time, etc. then we obtain a sequence of abbreviations. The frequencies have been scrutinized but the sequences of letters can be further segmented in Köhlerian motifs which have many properties as already shown (Köhler 2015, Köhler, Naumann 2008). Their frequencies, lengths, etc. will be different from text to text and also from unit to unit other than adverbials. Further, if we perform the ranking according to the frequency of the given units, and replace the abbreviations by their ranks (here 1 to 13), we obtain a sequence of numbers which may again be considered a sequence of motifs.

Another possibility is to consider an adverbial as a (logical) predicate of noun, verb, adjective or another adverbial. Replacing the adverbials by the entities of which they are predicates we again obtain a sequence of abbreviations which have their frequencies, can be transformed in motifs, etc. If we replace the adverbials by the degree of their predicativity, we obtain a new numerical sequence whose properties can be scrutinized.

The individual classes of adverbials may be subdivided in several classes according to the grammar of language, e.g. time adverbials may be subdivided in past, present and future subclasses; place adverbials can be subdivided according to the nearness to the object (e.g. in-out, right-left, above-below, in front of-behind, near-far, etc.), etc.

As a matter of fact, the way into the depth of adverbials is infinite. Here we merely tried to show some aspects. In order to obtain stronger confirmations, not only more Czech texts must be scrutinized but also their development in history, similar samples from as many languages as possible. The degree of confirmation may change in the course of the project and new models may appear. This is a normal way of science.

References

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

Altmann, G., Köhler, R. (2015). *Forms and degrees of repetition in texts. Detection and analysis*. Berlin/Munich/Boston: de Gruyter Mouton.

Bortz, J., Lienert, G.A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin/Heidelberg/New York: Springer-Verlag**.**

Brainerd, B. (1976). On the Markov nature of text. *Linguistics 176, 5-30*

Čech, R., Uhlířová, L. (2014**).** Adverbials in Czech: Models for their frequency distribution. In: Altman, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.). *Empirical Approaches to Text and Language Analysis.* Lüdenscheid: RAM-Verlag, 2014, 49-49

Chen, Y.-S. (1988). An exponential recurrence distribution in the Simon-Yule model of text. *Cybernetics and Systems: An International Journal 19, 521-545.*

Chen, Y.-S., Chong, P.P., Kim, J.-S. (1992). A self-adaptive statistical language model for speech recognition. *Cybernetica 35(2), 103-127*.

Diessel, H. (2005). Competing motivations for the ordering of main and adverbial clauses. *Linguistics 43, 449-470.*

Ford, C.E. (1993). *Grammar in Interaction. Adverbial Clauses in American English Conversation.* Cambridge: Cambridge University Press.

Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin: Springer (p. 127-130).

Hoye, L. (1997). *Adverbs and Modality in English*. London/New York: Routledge.

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

Jørgensen, M., Phillips, L. (2002). *Discourse Analysis as Theory and Method*. London: Sage Publications.

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

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

Köhler, R. (2015). Linguistic motifs. In: Mikros, G., Mačutek, J. (eds.), *Sequences in language and text: 89-108.* Berlin/Boston: de Gruyter Mouton.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T- segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin/Heidelberg: Springer.

Králík, J. (1977). An application of exponential distribution law in quantitative linguistics. *Prague Studies in Mathematical Linguistics 5, 223-235.*

Ney, J.W. (1982). The order of adjectives and adverbs in English. *Forum Linguisticum 6, 217-257.*

Prün, C. (1997). A text linguistic hypothesis of G.K. Zipf. *J. of Quantitative Linguistics 4, 244-251.*

Rijkhoff, J. (2002). *The Noun Phrase*. Oxford: Oxford University Press.

Skinner, B.F. (1939). The alliteration in Shakespeare's sonnets: A study in literary behavior. *Psychological Record 3, 186-192.*

Skinner, B.F. (1957). *Verbal behavior*. Acton: Copley.

Spang-Hanssen, H. (1956). The study of gaps between repetitions. In: Halle, M. (Ed.), *For Roman Jakobson: 497-502*. The Hague: Mouton.

Strauß, U., Sappok, Ch., Diller, H.J., Altmann, G. (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7, 73-100*.

Thompson, S.A., Longacre, R.E. (1985). Adverbial clauses. In: Shopen, T. (ed.), *Language Typology and Syntactic Description Vol. II: 171-234*. Cambridge: Cambridge University Press.

Uhlířová, L. (1967). Statistics of word order of direct object in Czech. *Prague Studies in Mathematical Linguistics 2, 37-49*.

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

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

Yesypenko, N. (2008). An integral qualitative-quantitative approach to the study of concept realization in text. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 308-327*. Chernivci, ChNU.

Yngve, V. (1956). Gap analysis and syntax. *IRE Transactions PGIT-2, 106-112*

Zipf, G.K. (1935/68). *The Psycho-biology of Language. An Introduction to Dynamic Philology*. Cambridge, Mass: MIT.

Zipf, G.K. (1937a). Observations on the possible effect of mental age upon the frequency-distribution of words from the viewpoint of dynamic philology. *Journal of Psychology 4, 239-244.*

Zipf, G.K. (1937b). Statistical methods in dynamic philology (Reply to M. Joos). *Language 132, 60-70*.

Zipf, G.K. (1945). The repetition of words, time-perspective and semantic balance. *The J. of General Psychology 32, 127-148.*

Zipf, G.K. (1946). The psychology of language. In: Hariman, P.L. (ed.), *Encyclopedia of Psychology: 332-341*. New York: Philosophical Library.

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

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

Zörnig, P. (1984b). The distribution of the distance between like elements in a sequence II. *Glottometrika 7, 1-14.*

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

Texts:

1. **ŠKODA, Jan.** Nejslavnější padouch komiksových příběhů zastíní i hrdinného Batmana. In: *Reflex* [online]. [cit. 22. 12. 2016]. Accessible from: http://www.reflex.cz/clanek/causy/75224/nejslavnejsi-padouch-komiksovych-pribehu-zastini-i-hrdinneho-batmana.html
2. **BYSTROV, Michal**. Producent George Martin: Pátý Beatle se nebál odhazovat hudební konvence. In: *Reflex* [online]. [cit. 22. 12. 2016]. Accessible from: http://www.reflex.cz/clanek/causy/75229/producent-george-martin-paty-beatle-se-nebal-odhazovat-hudebni-konvence.html
3. **ŠKODA, Jan.** Eliška Junková. In: *Reflex* [online]. [cit. 22. 12. 2016]. Accessible from: http://www.reflex.cz/clanek/causy/76003/eliska-junkova. html
4. **ŠAFRÁNEK, Šimon.** Hollywood v Číně. In: *Reflex* [online]. [cit. 22. 12. 2016]. Accessible from: http://www.reflex.cz/clanek/causy/75999/ hollywood-v-cine.html
5. **BYSTROV, Michal** 2016: Accessible from: http://www.reflex.cz/clanek/ causy/74555/nejslavnejsi-cesky-loupeznik-vaclav-babinsky-dozil-v-klastere.html
6. **MATYÁŠOVÁ, Judita** 2016, Accessible from: http://www.lidovky.cz/ udelam-tam-tropickou-lasku-ddg-/kultura.aspx?c=A161223_132845_ ln_kultura_hep
7. **VÁŇA, Jan** 2016, Accessible from: http//www.h7o.cz/zit-technologie
8. **KLÍČOVÁ, Eva** 2016, Accessible from: http://www.h7o.cz/uzitecny-idiot-z-prazske-kavarny/
9. **PUSKELY, Martin** 2016, Accessible from: http://www.h7o.cz/obrana-kritiky/ #sthash.vczHeCO2.dpuf
10. **VYBÍRAL, Zbyněk** 2016, Accessible from: http://www.h7o.cz/kriticky- ano-ale-proc-vulgarne/#sthash.n9ob6oBO.dpuf
11. **ŠRÁMEK, Fráňa** 1916, *Splav*. Praha: Mladá fronta, 1972, p. 431-436.
12. **ŠRÁMEK, Fráňa** 1916, *Romance*. Praha: Mladá fronta, 1972, p. 413-417.
13. **ČAPEK, Josef** 1946. *Na spánek*. From: ČAPEK, Josef, *Básne z koncentračního tábora*, p. 19-26. Praha: Fr. Borový.
14. **WOLKER, J.** *Svatý kopeček*. From: Wolker, J. *Host do domu*. [online]. V MKP 1. vyd. Praha: Městská knihovna v Praze, 2011. [cit. 7. 2. 2017] Accessible from: http://web2.mlp.cz/koweb/00/03/37/00/11/host_do_domu.pdf
15. **ERBEN, K. J.** *Svatební košile*. From: Erben, K. J. *Kytice*. Praha: Albatros, 1965, p. 31-43.
16. **ČELAKOVSKÝ, F.L.** *Prokop Holý*. From: Čelakovský, F. L. *Ohlas písní českých.* [online]. V MKP 1. vyd. Praha: Městská knihovna v Praze, 2011, p. 23-26. [cit. 7. 2. 2017] Accessible from: http://web2.mlp.cz/koweb/00/03/37/00/52/ohlas_pisni_ceskych.pdf

17. **BŘEZINA, O.** *Se smrtí hovoří spící…* From: Březina, O. *Stavitelé chrámů.* [online]. V MKP 1. vyd. Praha: Městská knihovna v Praze, 2011, p. 24-29. [cit. 7. 2. 2017] Accessible from: https://web2.mlp.cz/koweb/00/03/59/20/93/stavitele_chramu.pdf
18. **KAINAR, J.,** *Lazar a píseň.* Praha: SNKLU 1960.

Appendix

Sequences of left-right adverbials in texts

T 1
[L,R,R,L,R,L,R,R,L,R,R,L,R,R,L,R,L,L,R,L,R,R,L,L,R,L,L,L,L,R,R,L,R,R,R,R,L,L,L,R,R,R,R,L,R,L,R,R,L,L,R,L,R,L,L,L,L,R,L,L,L,L,L,L,L,L,R,R,R,L,R,L,R,L,L,R,R,R,R,L,L,L,R,R,L,R,L,L,L,R,L,L,L,R,R,R,R,L,L,L,R,R,R,L,R,L,R,L,L,L,L,L,R,R,R,R,R,L,R,R,L,L,L,L,R,R,L,L,L,L,R,L,L,R,L,L,R,L,R,L,L,L,L,L,L,L,L,R,L,R,R,R,R,L,R,R,L,L,R,R,L,R,R,L,R,L,L,R,R,L,L,R,R,L,L,R,L,R,R,R,R,R,R,R,R,R,L]

T 2
[L,L,L,R,L,L,L,L,L,L,L,L,L,R,L,R,R,L,R,L,L,R,R,L,R,R,R,L,L,L,L,R,R,R,R,L,R,R,L,R,L,R,L,L,L,L,L,L,L,L,L,L,R,R,L,L,L,R,R,L,L,L,R,L,L,R,R,R,R,L,R,L,L,L,L,R,R,L,L,R,L,L,L,R,L,R,L,R,L,R,R,L,L,R,L,L,R,L,L,L,L,L,L,L,R,R,L,R,L,R,R,R,L,L,L,L,R,R,R,L,R,R,R,L,L,L,L,L,L,L,R,L,L,L,R,L,L,L,L,L,R,L,R,L,R,R,L,L,R,L,R,R,L,R,L,R,L,L,L,R,R,R,R,L,L,L,L,L,R,L,L,L,R,R,L,R,L,R,L,R,L,L,R,R,R,L,R,R,L,L,L,R,R,R,R,L,L,L,R,L,L,R,L,R]

T 3
[L,R,L,R,R,L,L,L,R,L,L,L,R,R,R,R,R,R,R,L,L,L,L,L,L,L,L,R,L,R,L,R,L,R,R,R,R,L,R,R,R,R,R,R,L,R,L,L,R,L,L,L,R,R,R,R,L,R,L,R,L,R,R,R,L,L,R,R,R,R,R,L,L,L,L,L,L,R,L,L,L,R,R,L,R,R,R,L,L,R,L,R,R,R,R,L,L,L,L,L,R,L,L,L,R,L,L,R,L,R,R,L,L,L,L,R,L,R,R,R,R,L,L,L,L,L,R,R,R,R,L,R,L,L,R,L,L,L,L,R,L,R,R,R,L,L,L,R,R,R,R,R,L,L,L,L,L,R,R,L,L,R,L,L,R,R,R,L,R,R,R,L,L,L,L,R,R,R,R,R,L,R,L,R,R,R,R,L,R,L,R,R,R,R,R]

T 4
[L,L,R,L,R,R,L,R,R,L,R,R,R,L,R,R,R,R,R,R,L,R,L,R,L,L,R,L,R,L,R,R,R,L,L,L,L,L,R,R,L,R,L,L,R,R,L,R,R,L,R,L,R,L,L,L,L,L,R,L,L,R,R,L,R,R,L,R,L,R,R,L,L,L,R,R,R,R,R,R,R,L,L,L,L,R,R,L,L,L,R,L,R,R,R,L,L,R,R,R,R,R,L,L,L,R,L,R,L,L,R,R,R,L,R,L,R,R,L,R,R,L,L,L,L,R,R,L,L,L,R,R,R,L,R,L,R,L,L,R,L,R,R,R,L,R,R,R,R,L,R,R,R,R,L,R,R,L,L,R,L,R]

T 5
[R,R,L,L,R,R,R,R,L,L,R,L,R,L,L,R,L,R,R,R,R,R,R,R,R,R,L,R,L,L,L,L,R,R,R,L,R,R,L,R,R,L,L,R,R,R,L,R,L,R,R,L,L,R,L,R,R,L,L,R,R,R,L,L,L,R,L,L,R,R,L,L,R,R,L,L,L,R,R,L,L,R,L,L,L,L,L,L,L,L,L,L,R,L,R,L,R,R,L,R,L,L,L,L,L,R,R,L,L,R,R,R,L,L,L,L,R,R,L,L,L,L,R,R,L,R,L,R,R,R,R,R,R,R,R,L,R,R,R,L,R,R,L,R,L,L,L,R,R,R,L,L,R,L,L,R,L,L,L,R,L,L,L,L,L,R,R,R,L,L,R,R,R,L,L,R,R,L,R,L,R, R,R,L,L,R,L,L,R,R,R,R,R,L,R,L,R,L,R,R,R,L,L,R,R,L,R,L,L,L,R,R,R,R,R,R,R,L,R,R,L,R,L,R,R,R,R,R,R,R,L,L,R,L,R,R,R,R,R,L,L,L,L,R,R,R,R,L]

T 6
[L,R,L,R,R,R,R,L,R,R,R,L,R,R,L,R,L,R,R,L,L,R,R,R,L,R,R,L,R,L,L,R,R,L,R,R,R,R,L,R,L,L
,L,R,L,L,R,R,R,L,R,R,R,R,L,R,L,L,R,L,R,R,L,R,R,R,R,R,L,L,R,R,L,R,L,L,R,L,R,R,R,L,R,L
,R,R,L,R,R,R,R,R,R,R,R,R,R,R,R,L,L,R,L,R,R,R,L,R,L,R,R,L,R,R,R,L,L,R,R,L,R,R,R,L,R,
L,L,L,L,R,L,R,R,R,R,R,R,R,L,L,L,L,L,R,R,R,R,L,R,R,L,R,R,L,R,L,L,R,L,L,R,R,R,L,L,L,L,
R,L,R,R,R,R,L,L,R,L,R,R,R,R,R,L,L,R,L,R,L,L,R,L,L,L,L,L,L,L,R,R,L,R,L,R,L,L,L,L,R,R,
R,L,L,L,L,R,R,R,L,R,R,R]

T 7
[L,L,R,L,R,L,R,L,R,L,R,L,L,L,L,R,R,L,R,L,R,L,L,L,R,R,R,R,L,L,L,L,L,L,L,R,L,R,R,R,L,L,
L,L,R,L,L,R,R,R,R,R,L,L,R,L,L,L,R,R,R,R,R,L,R,L,L,R,R,L,R,L,R,R,R,L,R,L,R,L,L,L,L,L,
L,L,L,L,L,R,L,R,R,L,L,R,R,R,R,L,L,R,L,L,R,R,L,L,L,R,R,R,L,L,R,R,L,R,R,L,L,R,L,R,L,R,
L,L,R,L,L,L,L,L,L,L,L,L,L,L,L,R,R,R,R,R,R,R,L,L,R,R,L,L,L,L,L,L,L,L,L,L,R,L,R,L,R,L,R
,L,R,R,R,R,L,L,R,R,L,L,R,R,L,L,]

T 8
[L,R,L,L,L,R,L,L,L,L,L,R,L,R,R,L,L,L,L,R,L,L,R,L,R,R,R,R,R,R,L,R,L,L,R,R,R,L,L,R,L,R,
R,L,R,R,R,L,R,L,R,L,L,R,L,L,R,L,R,R,L,L,R,R,L,R,L,L,R,R,L,R,L,L,L,R,R,L,L,R,L,L,R,L,
L,R,L,R,R,R,L,R,R,R,L,R,L,L,L,L,R,R,R,L,L,R,L, L,L,L,R,L,L,R,R,R,L,L,R,R,L]

T 9
[L,R,L,L,L,R,L,L,L,L,L,R,L,R,R,L,L,L,L,R,L,L,R,L,R,R,R,R,R,R,L,R,L,L,R,R,R,L,L,R,L,R,
R,L,R,R,R,L,R,L,R,L,L,R,L,L,R,L,R,R,L,L,R,R,L,R,L,L,R,R,L,R,L,L,L,R,R,L,L,R,L,L,R,L,
L,R,L,R,R,R,L,R,R,R,L,R,L,L,L,L,R,R,R,L,L,R,L,L, L,L,R,L,L,R,R,R,L,L,R,R,L]

T 10
[L,L,R,R,L,L,R,L,R,R,L,R,L,R,R,L,L,,L,R,R,R,L,L,L,L,L,L,R,L,R,R,R,L,R,R,L,L,R,R,L,R,R
,L,L,R,R,L,R,R,R,R,R,L,R,L,R,L,R,R,L,L,R,R,L,,L,L,L,R,L,L,R,R,L,R,L,R,R,R,R,R,L,R,R,
R,L,L,R,R,R,L,L,L,R,L,L,L,L,R,L,R,R,L,L,R,L,L,R,R,R,R,R,L,R,R,L,R,L,R,R,L,L,R,R,R,R,
L,L,L,R,R,L,L,L,L,L,L,L,R,L,L,R,L,R,L, L,L,R,R,R,R]

T 11
[L,R,L,L,R,R,R,R,R,R,L,L,L,R,L,R,L,L,R,L,L,L,R,L,L,L,R,R,R,R,L,R,L,R,L,L,L,R,R,R,L,L,
R,L,L,L,L,R,L,L,L,L,L,L,L,R,R,L,L,R,R,L,L,R,L,R,R,R,L,L,R,L,R,L,L,L,L,L,R,R,L,L,R,L,L
,R,R,R,R,R,R,R,R,L,L,R,L,L,L,L,L,L,L,R,R,R,R,R,R,R,L,R,R,L,L,R,R,L,L,R,R,R,L,L,R,R,
R,R,L,L,R,L,L,L,L,R,R,L,L,R,R,L,R,L,L,R,R,L,R,L,L,R,L,R,R,R,R,R,R,R,L,R,R,L,R,R,L,
L,L,L,R,R,R,L,R,R,L,L,R,L,R,R,R,L,R,R,R,R,L,R,R,R,R,R,R,R,R,R]

T 12
[R,R,L,R,L,L,L,L,L,R,R,L,L,R,R,R,R,L,R,L,L,L,R,L,R,R,R,L,R,L,L,L,L,L,R,R,L,L,R,R,R,R,
R,L,L,R,R,R,R,L,L,R,R,L,R,R,R,L,R,L,L,L,R,L,R,R,L,R,R,R,L,R,L,L,R,R,L,L,R,R,L,R,L,R,
L,R,L,L,L,R,L,L,L,L,L,R,L,L,L,R,R,R,R,R,R,L,R,L,R,R,R,R,L,R,L,L,L,R,L,L,R,R,R,L,R,R,
R,L,R,L,L,R,R,L,L,R,L,L,L,R,L,R,L,R,L,R,R,R,R,L,R,L,L,R,R,R,R,L,L,R,L,R,R,L,R,L,R,L,
L,R,R,R,L,R,R,R,R,L,R,L,R,R,R,R,R,R,L,R,L,R,R,L,R,L,L,R,R,R,R,R,R,R,L,L,R]

T 13
[R,R,L,L,L,L,L,L,R,R,R,R,R,L,L,L,L,L,L,L,L,R,L,R,L,L,L,L,R,L,L,R,R,L,R,L,R,L,R,L,L,L,
L,R,R,L,R,R,L,L,L,L,R,R,R,L,R,R,R,R,L,R,R,L,R,L,R,L,L,L,L,L,R,L,L,L,R,R,R,R,L,L,L,L,
L,L,L,L,L,L,L,L,L,L,L,L,L,L,L,L,R,R,L,R,R,L,L,R,L,L,R,L,L,L,L,L,L,R,L,R,R]

T 14
[R,R,L,L,L,L,L,L,R,R,L,R,R,L,R,L,L,L,R,L,L,L,L,L,L,L,R,R,L,R,R,R,L,L,L,L,L,L,R,L,L,L,
R,R,R,L,R,L,L,L,L,L,R,L,R,L,L,R,L,L,L,L,L,L,R,L,L,L,R,L,R,L,L,R,R,L,L,R,L,L,L,R,L,L,L
,L,R,R,L,L,L,L,R,R,R,R,R,R,R,L,L,R,R,L,L,L,L,L,R,L,L,R,R,R,R,R,L,R,R,R,L,L,L,L,L,L,L,
L,L,L,L,L,R,R,R,R,L,L,R,L,R,R,R,L,L,L,L,L,R,L,L,L,L,L,L,R,L,R,R,R,R,L,L,R,R,R,R,L,R,
L,L,L,R,L,L,R,R,R,R,R,R, L,R,L,L,L,R,R,R,L,L,R,R,L,R,L,R,R,R,R,L,L,L,R,L,R,R]

T 15
[L,L,L,L,R,R,L,L,L,L,L,L,L,L,L,R,L,L,L,L,L,R,L,L,L,L,L,L,L,R,L,L,R,R,R,L,R,L,R,L,L,L,L
,L,R,L,L,R,R,R,R,L,L,L,L,R,R,L,L,R,L,R,L,R,L,L,R,R,R,R,L,R,R,R,R,R,R,L,L,L,L,L,L,L,R,
L,R,L,L,L,L,L,L,R,R,R,R,R,R,R,R,L,L,R,L,L,L,L,R,L,R,L,R,R,R,L,L,L,L,L,R,R,R,L,R,R,R,
L,L,L,R,L,R,R,R,R,R,R,R,L,R,L,R,R,L,R,R,R,R,L,R,L,L,R,L,R,R,L,L,L,L,L,R,R,R,L,R,L,R,
L,L,R,R,R,R,R,R,R,R,L,L,L,L,L,L,L,R,L,L,L,L,L,L,L,L,L,L]

T 16
[L,L,R,L,L,L,L,L,R,R,L,R,L,R,L,L,L,L,L,R,R,L,L,L,L,L,R,L,L,L,R,L,R,R,R,L,R,R,R,L,L,L,
L,L,L,R,L,L,L,L,R,L,R,L,R,R,L,L,R,R,R,R,R,L,R,R,R,R,L,L,R,R,R,R,L,R,L,L,R,R,L,L,L,L,
L,L,L,L,L,R,L,R,R,L,L,L,R,R,R,L,R,L,L,L,L,L,L,L,L,L,L,L,R,R,L,L,L,L,L,L,L,L,R,R,L,R,
R,R,L,R,R]

T 17
[L,R,L,L,L,R,L,L,R,L,L,L,L,L,R,R,L,R,L,L,R,R,R,L,R,L,L,R,R,R,R,R,R,L,L,L,L,R,L,L,R,R,
R,R,R,R,R,R,R,L,L,R,L,R,R,R,L,L,R,R,R,R,R,L,L,L,L,L,L,R,R,L,L,R,L,L,L,L,R,L,R,R,L,R,
R,R,L,R,L,R,R,R,L,R,L,R,R,L,R,L,R,R,L,L,L,L,L,R,R,R,R,R,R,L,L,R,L,L,L,L,R,L,L,R,R,R,
R,L,R,L,L,R,L,L,R,L,L,R,R,L,L,L,R,R,L,R,R,R,L,L,R,R,L,R,L,L,L,L,R,L,L,R,L,L,R,L,R,L,
R,R,R,R,L,L,L,L,L,L,L,L,R,R,L,L,R,R,R,R,L,L,R,L,R,L,R,R,R,R,L,L,L,R,R]

T 18
[R,L,R,L,R,L,R,R,L,L,R,R,R,R,R,R,R,L,R,R,R,R,R,R,L,R,R,L,L,L,R,R,L,R,R,R,L,L,L,R,L,
R,R,R,R,L,R,R,R,R,R,R,L,L,L,L,R,R,L,L,R,L,R,R,R,R,R,R,L,R,R,L,L,R,R,L,R,L,R,R,R,R,
R,R,L,R,R,L,R,R,R,R,R,R,R,R,R,L,L,L,L,L,R,L,R,R,L,L,R,R,R,R,R,L,L,L,R,R,R,L,L,R,R]

Glottometrics 37, 54-81

Steppe Homeland of Indo-Europeans Favored by a Bayesian Approach with Revised Data and Processing

Hans J. Holm[1]

Abstract

Despite dozens of hypotheses, the origin and development of the Indo-European language family are still under debate. A well-known glottochronological approach to this problem using Bayesian computation of language divergence dates claims to have provided evidence for the period of Neolithic expansion known as the "Anatolian hypothesis." The dates have met with considerable criticism from other disciplines. I decided to investigate the evidence for these dates by replicating and analyzing the approach. During this process, a further approach located a date of origin from between 3950 – 4740 BC. One of the insights of this study was that previous results were significantly disrupted by poorly attested languages, which were consistently removed step by step.

This paper supports this finding using data from the previous approaches and my own updated dataset. The resulting date is around 4800 BC. However, the topology of the trees differed considerably over the course of several hundreds of tests. This problem was avoided in previous approaches by rigorous topological forcing. Here we apply a west–east dichotomy from a previous purely lexicostatistical (i.e. without times) approach based on the best available Indo-European dataset of approx. 1,100 verbal roots, which produces dates around 4100 BC. These dates reflect the most recent state of knowledge in linguistics, archeology and genetics in favor of the Steppe hypothesis. A new synopsis of the wheel problem, a primary argument for the divergence date, shows that not one but three different Indo-European denotations coincide in different areas with different types of wheel–axle constructions. Archeological cultures likely to have been affected by the migrations are presented visually at the end of this paper.

Keywords: Indo-European, glottochronology, Urheimat, Bayes‘ reasoning, Swadesh list.

1. Introduction

Indo-European (henceforth "IE") is a family of languages defined by commonly inherited words and grammar. IE was spoken in prehistorical times from western Europe to the Indian subcontinent reaching as far east as Xinjiang in modern northwest China. Since the discovery of this language family 200 years ago, IE's prehistoric homeland (or formation area) has been widely debated with linguists still in disagreement over its genealogical development (cf. e.g. Ringe, Warnow & Taylor (2002), Meier-Brügger (2010), Fortson (2010)).

[1] Address correspondence to: [hjjaholm@arcor.de]

Among the dozens of proposed origin locations and dates, the two most favored were Anatolia in the seventh millennium BC and the Eurasian (Forest) Steppes in the fifth millennium BC. The significant time difference between these two periods stimulated researchers to compute the time elapsed between known linguistic changes using a method known as “glottochronology” (GC). Attempts in this direction were first made during the 1930s (cf. Embleton 1986, passim; Holm 2005). With the advent of radioactive dating in the 1950s, linguists discovered that linguistic changes could also occur at computable rates, leading to the development of the initial GC. Soon after this, biologists began to detect some regularity in gene mutations and eventually equated these with linguistic changes (Holm 2007).

In contrast to these earlier approaches involving fixed rates of linguistic change, recent Bayesian approaches allow for a more realistic “relaxed clock” (Drummond 2006). In this manner Gray et al. (2003) calculated a primary divergence date of c. 6700 BC,[2] which roughly coincides with the onset of the so-called “Neolithic Revolution” in Anatolia around 7000 BC (cf. Renfrew, 1987). Bouckaert et al. (2012), henceforth Bou12, located the first split at c. 6500 BC using an impressive method for calculating the geographical area of origin and subsequent diffusion into their historical or modern territories. An input error prompted a correction resulting in a new median estimate of c. 5579 BC (Bou13). My recalculations based on the published input file resulted in a date of c. 8200 BC later revised to c. 5508 BC (see Tables 2 a and b below with more comparisons). Neolithic expansion had already penetrated far into central Europe by these revised dates. Prehistorians and linguists reject this for contradicting the evidence provided by traceable objects common in IE languages and datable archeological finds of the same objects throughout the Eurasian Steppe belt (Anthony 2007). This latter argument has been widely accepted, although Bou12/13 continues to maintain that it is “controversial,” notably citing Mallory & Adams (2006) as evidence to the contrary.

A recent approach taken by Chang et al. (2015) [henceforth Cha15] offers different root dates between 3930 and 4740 BC as proof for the Steppe hypothesis, although they fall on the outer edges for the era (4500–3500 BC) 4500–3500 BC presupposed by them for this hypothesis, let alone younger suggestions. The results provided by Cha15 could only be achieved by forcing eight extinct languages (including Latin, Old Irish, and Vedic Sanskrit) based on the assumption that they are direct and single ancestors of their modern linguistic relatives.

The aim of the present paper is to apply previously used methodology, in particular the phylogenetic software BEAST (Drummond 2012), to analyze the previously, often self-contradictory topological and chronological results in relation to the linguistic input, and paying particular attention to the gaps and loans included in the word lists. Both the new topological and chronological results should be interesting for Indo-Europeanists. Section 2 analyzes old and new input data while Section 3 analyzes the effects of gaps and loans. Section 4 briefly summarizes the arguments put forward by various disciplines in favor of the Steppe hypothesis before presenting the abstract topology and chronology overlaid with known periods of archeological cultures. Section 5 offers some concluding remarks.

[2] Because IE dispersal is a historical problem, we use the customary designator “BC” used in historical science. The calculations rely on word lists dated around 2000 CE and are converted accordingly.

2. The material - word lists

The basic assumption in GC is that every change in the relationship between a linguistic sign and its test meaning (concept referent) is related to an elapsed time period (cf. Embleton 1986, passim), Holm 2007). The test datasets[3] can be gathered in different ways. For ease of etymological assessments, data for GC purposes should be ordered in a matrix as demonstrated in Table 1 below.

GC wordlists demand a high philological and etymological standard because even slight mistakes have a considerable impact with statistically low sample sizes of 207 to 100 test meanings. Anatolian languages, which are of central interest because of their split time (Gray and Atkinson, 2003), are especially prone to calculation errors due to the extreme number of loans and even gaps (cf. Fig. 1**)**.

Out of various available word lists (see Holm 2007), this paper only starts utilizing the ones used in Bou12/13 and Cha15, which were generally based on the hastily gathered lists of Dyen (1997). No individual loans had been tagged in the 2011 version, and despite continued updating, even the 2014 version in Dunn (2015) contained considerable and obvious errors.[4] Several examples for Albanian alone are given in Holm (2011). Unacceptable gaps remained even in some living languages such as Kurdish (thus omitted in the narrow and medium datasets of Cha15). Further examples, such as the mis-cognation of the Cymric forms of the meaning “I” or the Kurdish and Albanian forms of “all,” reveal that the editing authors paid insufficient attention to their data. Additional data for extinct languages provided by Bou12/13 from Ringe, Warnow, & Taylor (2002) is also partially outdated (cf. Holm 2011, Cha15). The quantitative relationships are shown in Fig. 1.

The recent study of Cha15 also made use of the IE lists in Dunn (2013 version) in which some Iranian and Hittite data were amended. Aside from the above-mentioned cases, they cite, e.g., the Russian word *plod* for the concept “fruit” (Cha15). This does not comply with rules of GC sampling, which require the most common, unmarked translation - here the loan *frukti* - for the sake of comparability. The word *plod* is a modern biological term, a concept unlikely to have been in use in IEs. Furthermore, the authors appear not have consulted standard dictionaries resulting, for example, in a gap in the list for the Hittite concept of “feather,” although the Hittite translation *pattar* is available in Kloekhorst (2008) and Kassian/ Yakubovich (2011).

Due to the insufficient quality of the previous word lists, a completely new one was deemed necessary. The choice of meanings[5] for this new test set is based on the final proposal of Swadesh (1971), the founding father of GC. He reduced his first lists of over 200 meanings to 100[6] arguing for “quality over quantity” (Swadesh 1955: 124). This new list consists of 17 languages mainly representing one extinct and one recent language for each of the 12 primary branches of Proto-IE (hereafter PIE). This list is referenced in this article as H17 (Holm 2016).

[3] Note that for the sake of comparability, GC requires data to comprise “universal” concepts with the most common, unmarked translations available in as many tested languages as possible. It follows that these concepts are thus meant neither to be “basic” in the sense of second language acquisition nor particularly resistant against borrowing (Swadesh 1955).

[4] The dataset is continuously updated and improved.

[5] However, not his (unavailable) word lists, as a reviewer erroneously implied.

[6] Not 92, as cited in Cha15.

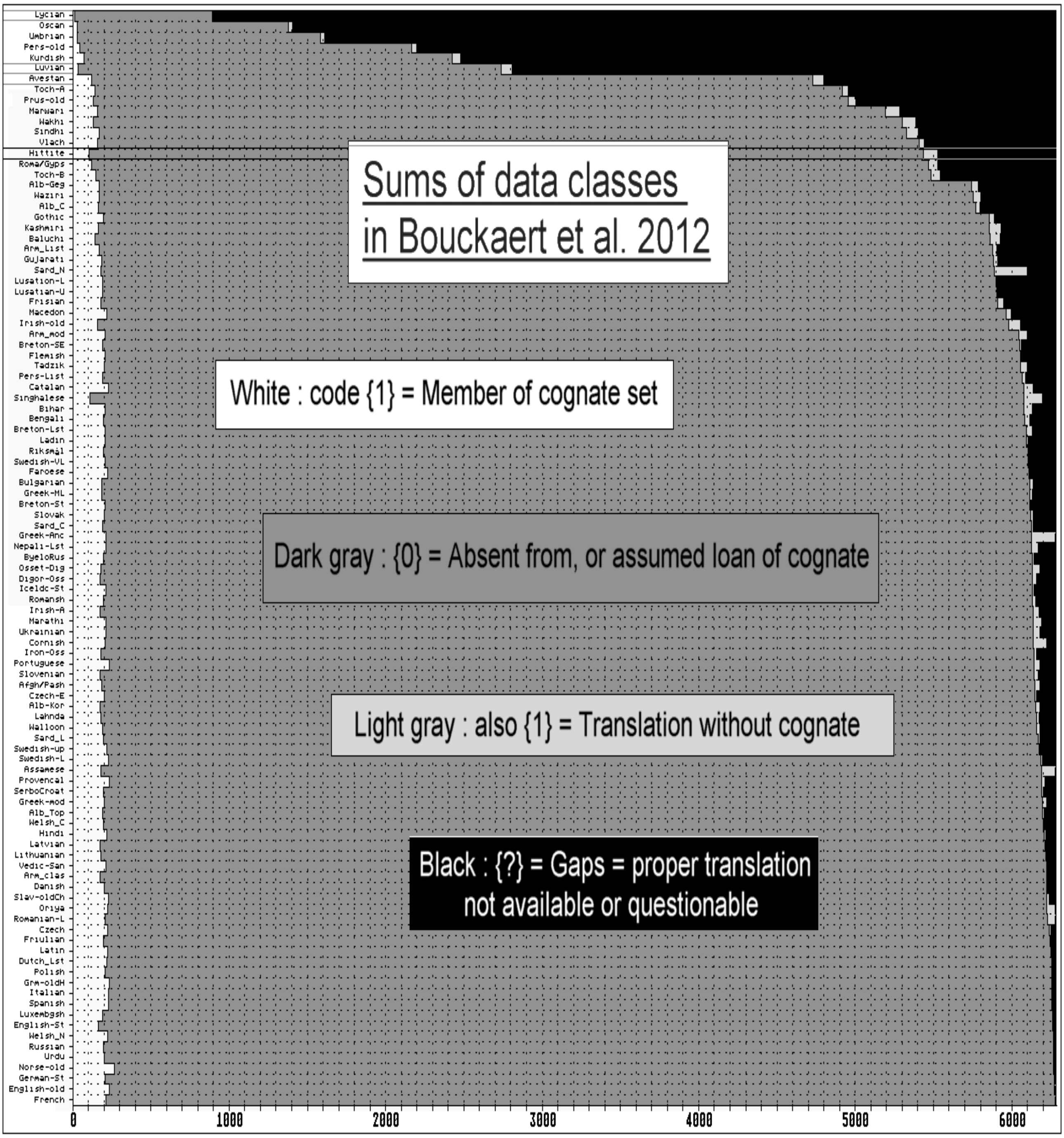

Fig. 1. Summed word classes per language (Bouckaert et al. 2012, 2013). White: {1} members of cognate sets; Dark gray {0}: absent from, or loan of {1};light gray: also {1}, for etymological orphan translations; black: {?} for gaps = no translation found.

3. Methods - data processing

3.1. Principles (GC, Bayes, models)

Recent Bayesian approaches construct family trees by computing tree-like phylogenies in which all given languages are "leaves" connected by branches (edges). These language leaves represent subgroups from the branching points (nodes) which stem back to a common "root" determined by a conventional algorithm. We tested both previously applied models starting with the Dollo model (Alekseyenko 2008), following Ryder 2010 and Bou12 (Supplementary materials p. 6) in which they argue that "The Stochastic Dollo process … applies what may be a more natural model of cognate evolution by postulating that a cognate can only arise once … ." By contrast, Cha15:217 claims7 that this model is "ill-suited to modeling RM traits."

In order to solve these seemingly contradictory attestations (see Holm 2007 for details), the BEAST software tentatively exchanges the branches of a "starting tree" in defined ways and amounts of MCMC[8]-chains—typically around 50–200 million times (see Bou12/13, Cha15 for technical details). Because strict clock models do not match the reality of language change, Bou12/13 use a "relaxed clock" model (Drummond et al. 2006). It must be noted that such a model also can only distribute locally calibrated rates, which do not necessarily have to be the true ones in the uncalibrated branches. The software finally computes the resulting posterior (logarithmic ∝- or shape) probability for every MCMC-run using an elaborated variant of Bayes' theorem, thus allowing the tree with the highest probability distribution given the data, model and test parameters to be selected.

3.2. Properties and coding of linguistic data

The linguistic translations can up to now only be represented by the very narrow codes {1}, {0} and an ambiguity code, here {?} for the mathematical process. The available translation is marked with the code {1} in the line for each language, which in the majority of cases is followed by {0, 0,…}, corresponding to unrelated traits {1} in other languages. Only then it is followed by the {1}-coded trait of the next meaning (Table 1).

3.2.1. Cognates

Cognates are identified by linguists by means of sound laws that have developed from a hypothetical PIE root.[9] The example in Table 1 gives translations of the meaning "fish" with four probable IE roots in column (trait) 1 to 4. The cells of the languages included in each cognate set are coded as {1} and those that are not are coded as {0}. These agreeing {1}-codes of the cognate traits combine their languages assuming that these languages either are or have been related more closely than those not thus combined have been.[10] Note that this assumption

[7] In fact, the property of the Dollo model, namely, assuming traits that come into existence exactly once, "suits it to traits that cannot be homoplastic (Appendix C)." In other words, suits it to traits that are "homologous". The case described in their appendix, however, is exactly homologous because the different meanings evolved in Romance ("foot") vs. some Indo-Iranian languages ("leg") go back to a common root PIE *pe/od with a perhaps ambiguous meaning "foot", "leg."

[8] The Markov chain Monte Carlo is a stochastic algorithm for drawing samples from a posterior distribution to get an estimate of the distribution (http://beast.bio.ed.ac.uk/glossary#MCMC).

[9] This description must necessarily remain incomplete. For more information, see the linguistic textbooks or for the glottochronological cases in particular Holm (2007).

[10] Cha15's assumption that potentially common original roots for, e.g., the meanings "foot" and "leg" in some languages would cause the software to attract these languages is unconvincing in the light of GC

is relativized considerably in real languages by chance replacements, gaps and loans, often leading to contradicting combinations.

Table 1

Data matrix excerpt of the H17 data (Holm 2016) with translations of one of the test meanings in the test languages. Column (or trait) 1 contains a cognate set. Columns 2–4 contain branch orphans. Columns 5 contains a singleton, 6 a loan, 7 an inadequate translation, meaning “meat of cow,” and in columns 8 and 9, we have gaps (no translations) in two languages. This is also an example of the traps involved with the lists and translations. The founder of GC Morris Swadesh, initially wrote “meat,” however, after step-by-step changes copied by different authors, Swadesh (1971) finally clarified his intended concept by switching to “flesh” as a body part in contrast to “bone.” Note that not only these terms overlap in several languages.

a. Test meaning “flesh” and its translations										**b. Coding**									**c. Alternative**								
Trait No.	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
Russian	mjaso	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Lithuanian	mēsà	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Old Icelandic	0	0	0	0	hold	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0
Bokmål	0	0	0	0	0	[kjøtt]	0	0	0	0	0	0	0	0	**0/1**	0	0	0	0	0	0	0	0	**1**	0	0	0
Old Irish	0	feóil	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0
Mod. Irish	0	feoil	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0
Italian	0	0	carne	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
Latin	0	0	carō	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
Albanian	mīshi	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Anc. Greek	0	0	0	κρέας	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
Mod. Greek	0	0	0	κρέας	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
Mod. Armenian	mis	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Hittite	0	0	0	0	0	0	0	n/a	0	**?**	**?**	**?**	**?**	**?**	**?**	**?**	**?**	**?**	0	0	0	0	0	0	0	**1**	0
Tocharian-B	mīsa	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Avestan	0	0	0	0	0	0	(gav-)	0	n/a	**?**	**?**	**?**	**?**	**?**	**?**	**?**	**?**	**?**	0	0	0	0	0	0	0	0	**1**
Vedic Sanskr.	māṃsá	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
Hindi	mā̃	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0

We now analyze the properties of the etymological categories, “cognates,” “orphans,” “loans” and “gaps.”

stochastics because these form–meaning combinations form their own different, widely dispersed traits not distinguishable from different roots in the affected languages for the software. The software cannot conclusively prove that such meanings may have split from a common trait closer to the root. Though such meanings may partly complement one another, they merely do this by combining individual traits. This becomes very clear when we recognize that different traits generally complement each other when left undisturbed by linguistic orphans.

3.2.2. Orphans (also known as singletons, isolates, unique traits)

Orphans appear in glottochronological (or Swadesh) lists as translations without etymological connections within the list. The examples in Table 1 columns 2–4 could be referred to as branch orphans, and that in column 5 as an orphan. Orphans are also coded as {1} because they have substituted a meaning slot somewhere in history that may be regarded as a genealogical event (like a biological mutation) and as indicative of elapsed time in glottochronology. In the absence of a described etymological relation, orphans are given (trait) column of their own.

3.2.3. Gaps

Gaps appear in glottochronological test sets where a translation is unattested (or has been overlooked) for a test meaning in a language. Beside “flesh” in Table 1, other examples of gaps include “all” in Umbrian, and “bark (of trees)” in Hittite. As shown in Section b of Table 1, gaps are coded by filling the complete “meaning slot” with a row of {?}s. The BEAST software interpolates these {?} codes according to the {1}:{0} distribution in the affected language. This appears to be a reasonable approximation.

Astonishingly, any reduction of gap-affected languages significantly reduces the root ages as demonstrated by the examples in Table 2a. Test series 1 gives the mean of 12 replications of the published Bou12 input file in which 283 forgotten “empty” traits contained a considerable number of {?} codes. Series 2 shows that omitting these forgotten traits alone reduces the root age by c. 1000 years. A further reduction (test series 3) reduced the root by a further 460 years for the three most gap-affected languages. Series 4 replicates the Cha15 B2 test by itself using the Bou13 data minus 6 gap-affected languages following their indication “that empty slots have to be avoided” because empty slots caused their model to “underestimate the number of unique traits in the language.” Now the most-affected languages have been omitted, the omission of 52 less gap-affected languages in test series 5 cause a smaller reduction. Note that all these tests represent the middle of their test series (more data in Appendix 1) and therefore cannot be considered outliers. Note further that both posteriors and clade credibility improve with every reduction of gapped languages.

The reason for the enormous reduction between test series 1 and 2 cannot be ascribed to a loss of calibration points because they are identical. The severely gap-affected Hittite and Tocharian languages were kept in all datasets because obtaining their positions is one of the aims of all approaches.

The Bou13 revision mentioned above not only canceled the “empty” traits but also switched to the previously refuted covarion model. After obtaining slightly better Bayes factors with the corrected data, they revised their former position (cf. 3.1.) arguing that the “The covarion is a flexible model that allows cognates to transition from relatively fast to slow rates of change. This flexibility may allow the model to deal with homoplasy[11] due to borrowing better than the Stochastic Dollo model.” While allowing transition from relatively fast to slow rates of change appears at first glance to be advantageous, borrowing should not present a major problem for specialized historical linguists (see 3.2.4.).

[11] Cha15 made a special case out of this that I address in 3.1.

Table 2.a.
Effects of changes in previous data and models, with BEAST v. 1.7.5. Legend: "ln HCC"=ln[12] highest clade credibility. For more details, see App. 1.

Ser. #	Test Type	Data source	"empty" {0;?} columns	Handling of gap-affected languages	taxa	Age BC	-ln Posterior	-ln HCC
1	A Dollo	Bou12	kept	kept	103	**6500**± 80	52 230	n/a
2	A Dollo	Bou13	no	kept	103	**5508**±104	51 590	6 200
3	A Dollo	Bou13	no	3 extinct languages omitted	100	**5048**± 62	50 540	5 760
4	A Dollo	Cha15-B2	no	6 extinct languages omitted	97	**4835**± 15	48 750	6 690
5	A Dollo	Bou13	no	52 most gap-affected languages omitted (except Hittite and Tocharian B)	51	**4722**± 50	27 176	3 351
6	B Cov. publ.	Bou12	kept	kept	103	**8381**±192	51 994	17 000
7	C Cov. Publ.	Bou13	no	kept	103	**7870**±1612	c.52 400	24 500

Replications of the published data and covarion model in series 6 increased the root age by over 2000 years(!), yielding virtually unanalyzable results with the corrected alignment (series 7). Confronted with this extreme difference, a co-author informed me that they had applied an additional element in the input file[13] not contained in the publication of Bou13. This necessitated new calculations, the results of which are shown below in Table 2.b.

Only now do the results appear relatively consistent, test series 8 eventually shows the expected agreement with the in Bou13 published result. The different result of the replication in Cha15 can be explained by their data and parameter changes. Test series 9 shows that the omission of the six most gap-affected languages reduces the root age from a mean of c. 5580 BC (test series 8 with 103 languages) by a significant 730 years to c. 4854 BC with the amended parameter. The posteriors with the reduced data are improved by approx. 5 %[14] on Bou103 (Table 2.a, Ser. 2).

[12] Smaller figures in the negative natural logarithms (-ln) are better because they represent a higher probability.

[13] AllowIdenticals="true". I owe a debt of gratitude to Philippe Lemey for recommending and providing this latest version to me.

[14] Cha15, FN 28 notes "that the improvement with covarion was slight (a gain of 0.5% in the log marginal likelihood)…"

Table 2.b
Data unchanged, covarion model, amended with "allowIdentical" argument, BEAST v. 1.8.4.

Ser. #	Test Type	Data source x taxa		Age BC	-ln Posterior	-ln HCC
(Pub.)	Citation	Bou13 x103	Published result[15]	**5579**	47 769	n/a
8	D	Bou13 x103	Repetitions with amended version	**5588**±78	~49 000	~13 420
(Pub.)	Citation	Cha15-B1 x103 "Replication of Bou13" (with other considerable changes)		**5750**	~48 170?	n/a
(Pub.)	Citation	Cha15-B2 x97= 6 extinct languages omitted	as published	**4810**	~46 220?	n/a
9	E		My replication	**4898**	-46 256	15.240

The poor dataset may explain the described behavior in Bou13 and Cha15. We therefore tested both models with the linguistically updated H17 dataset introduced in chapter 2 accompanied by the Bayes factors provided by BEAST v. 1.8.4 as the "stepping-stone" model test. In addition, we tested a dataset where all meanings with missing translations (=gaps) had been cancelled:

Table 3
Data: H17; gaps kept vs. canceled; Dollo vs. covarion model

Ser. #	Test Type	Data source x taxa	Handling of gaps	Model	Age BC	-ln Posterior	-ln HCC	-ML SS[16]
10 (3 runs)	F	H17(?) x760	{?}-coded	Dollo	**5056** ± 9	3 906	1.120	**3 662** ± 0
11 (3 runs)	G	H17(-) x658	cancelled		**4793** ± 10	3 559	1.120	**3 330**±13.8
12 (3 runs)	H	H17(?) x760	{?}-coded	Covarion amended	**4229** ± 48.5	4 194	4.930	**3 876**±236
13 (3 runs)	I	H17(-) x658	cancelled		**4181** ± 63	3 854	5.672	**3 641**±45.3
				Assessment	Cov. lower	Dollo better	Dollo better	Dollo better

[15] Obtained from unpublished input file and therefore different from test 6 with published input file.

[16] Following the "model selection tutorial (Rambaut 2014) we calculated the marginal likelihoods by stepping stone sampling (SSML) provided in BEAST v. 1.8.4 (Baele et al. (2012) and Baele et al. (2013)).

The H17 comparison shows that the complete omission of gapped meanings (test series 11 and 13) lead to better posteriors than changing the model. Comparing the models, the covarion test series (12, 13) yields worse results in all criteria. In addition, the results of the covarion series are more inconsistent, and even the distribution of rates in the resulting phylogenies not shown here are self-contradictory. We therefore cannot advocate the results provided by the covarion model.

3.2.4. Loans[17] (borrowings)

Loans are linguistic substitutions originating outside (sometimes in older stages of) the language (family) in question. Historical linguists are aware of several properties that distinguish[18] loans from cognates, even if the words are linguistically related.

Early GC assumed that words were irreversibly substituted only once, loans or not, for every bilateral loan situation (Embleton 1986) or by subtracting loans from lateral computations (Starostin 2000). Ryder (2010) tried to solve this problem by incorporating "catastrophe events" into his otherwise clocklike approach.

Bou12,13 mentioned this problem in their SM asserting, "[w]e can therefore be confident that […] the binary coding of the cognate data allows accurate phylogenetic inference, […] not impaired by […] realistic rates of borrowing." This, along with the absence of any marked loans in the "2012 IE Wordlist," indicates that the authors were unaware of the "realistic rate" of loans in several languages. The reduced number of loans (Swadesh 1955) still make up 15–35% of the words in Bokmål, Celtic, Albanian, Armenian, Hittite, Hindi and others even in our smaller test dataset (see Fig. 1).

BEAST methodology, however, has no adequate answer to the irregular behavior of loans, and canceling all affected meanings would reduce the databases too much. Assuming that loans particularly affect the most versatile part of the lexicon and are thus prone to substitutions, the closest which led to an exponential model that correlated with radioactive decay. Soon, however, Bergsland et al. (1962) demonstrated that loans could in fact take every GC computation to the absurd[19] to a degree dependent upon their proportion. Consequently, most prominent glottochronologists applied methods to avoid the bias caused by loans either by calculating separate rates individually approximation would be handling them like orphans with their own {1}-coded trait as well as a {0} in the receiving language (e.g. mountain in English, cf. Table 4).

This is a clear argument against the approach in Bou12/13 ({0}-code only) and Cha15 (case "excluded" {0} in test A4 only) because the loans thereby lose their stochastic property. This also explains the misinterpretation of some loans as orphans by Bou12/13 (examples include the Albanian loans *qafë* "neck", *koske* "head", *qen* "dog" or the Hittite *šalli* "big" (cf. Holm 2011)) and the real reason why their coding with {1} had no adverse consequences[20] resulting falsely in "accurate phylogenetic inference."

[17] The usual terms "loan" and "borrowing" are a misnomer because such substitutions are not returned; they are rather "copies" (Starostin et al. 2000).

[18] E.g. by sound laws and meaning variety and distribution (see Anttila 1989).

[19] As they compared standard Swadesh lists of natural languages (five North Germanic, two Georgian, and two Armenian), these are a realistic choice, in contrast to the criticism of Ryder (2010).

[20] The same applies to the remark of one reviewer that I had misinterpreted the one or other Hittite word as loan.

Puzzling remains the choice of Cha15 who write “we follow Bouckaert et al. (2012) and put 0 in the cell of a tagged loanword […],” however, perform their three basic tests (A1, 2, and 3) with loans coded with {1} in a cognate set as demonstrated here in an excerpt of their test A3 with the two meanings “animal” and “mountain.” It is clearly visible that in this way they combine English with the Romance languages rather than the Germanic.

Table 4
Loans mistakenly coded with {1} in the receiving language (here English) erroneously combines it with the loan-giving family (here Romance)

Language	animal	mountain
Latin	00001000000000000000	000001000000000000000000000000
Romanian	00001000000000000000	000001000000000000000000000000
Catalan	00001000000000000000	000001000000000000000000000000
Portuguese	00001000000000000000	000001000000000000000000000000
Spanish	00001000000000000000	000001000000000000000000000000
French	00001000000000000000	000001000000000000000000000000
Provencal	00001000000000000000	000001000000000000000000000000
Ladin	00001000000000000000	000001000000000000000000000000
Romansh	00001000000000000000	000001000000001000000000000000
Friulian	00001000000000000000	000001000000001000000000000000
Italian	00001000000000000000	000001000000000000000000000000
Gothic	10000000000000000000	000000000000000000000000001100
OW_Norse	10000000000000000000	000000000000001000000000000000
Icelandic	10000000000000000000	000000000000001000000000000000
Faroese	10000000000000000000	000000000000001000000000000000
Norwegian	10000000000000000000	000000000000001000000000000000
Swedish	10000000000000000000	010000000000000000000000000000
Danish	10000000000000000000	010000000000000000000000000000
English !	0000***1***000000000000000	00000***1***000000000000000000000000
Frisian	10000010000000000000	010000000000000000000000000000

3.3. Topological alternative

In most tests (such as those in Table 2a**,** details in App. 1) Hittite evolved by splitting off first. However, the results are inconsistent because sometimes basal splits of Indo-Iranian (and sometimes Balkan languages) from the others (cf. App. 1., column 7c) appear. It is worth noting that in his approach that omitted orphans and inserted “catastrophe impacts,” Ryder (2010: Fig. 5.6) also obtained Indo-Iranian as the first to split followed by Albanian and the combined Hittite-Tocharian languages in third place. Perhaps owing to similar observations, Cha15 decided to use topological constraints to fix Hittite and Tocharian as first splits and thus forfeited the opportunity to test and prove this.

A closer inspection of the positional variations in several hundred tests (including those described in detail in App. 1 and 3) reveals that the primary branches do not vary randomly. Rather they tend to vary within two “main limbs:” an “eastern limb” that consists of the Anatolian–Tocharian, Indo–Iranian, and Balkan group, and a stable “western limb” that consists of the Balto–Slavonian, Germanic and Italo–Celtic groups. Exactly this dichotomy was obtained by the purely lexicostatistical approach of Holm (2008) based exclusively on the best available IE dataset (Rix et al. 2001) with around 1,140 verbal roots (which are also known to be much less prone to borrowing than nouns). This topology was obtained using a hypergeometric estimator for the number of original symplesiomorphies at the date of the split between any two branches after parsing the data according to the languages' Zipf[21]–Pareto distribution to avoid a possible bias.

Table 5
Data: H17, no gaps, west-eastern monophyly; "Allow Identical" argument; BEAST 1.8.4.22

Ser. #	Model\Results	Test Type	Age BC	-ln Post.	-ln HCC	-ln ML SS[23]
14	Dollo	GW, 3 runs	**4102** ±4.36	3,556	0.275	3314
15	Covarion	IW, 3 runs	**3524** ± 8	3,852	3.075	3340
	Assessment			Dollo much better!	Dollo much better!	Dollo better

As in tests series 10 to 13 (Table 3), the posteriors are better with the Dollo model except for the insignificant difference with the Bayes factors. The rate distributions again contradict each other from one covarion run to the other. Note that the consistent $\log_n$ posteriors around −3556 are decisively better than the approx. −50,000 obtained in previous approaches. For all of these reasons the choice can only be the series 14 dates for the first IE split at 4102 ± 4.36 BC with a 95% highest posterior density (HPD)[24] interval of c. 7230 to 5040 years b2k.[25]

4. Selected arguments[26] from other fields favoring a Steppe homeland for PIE

In B&a(2012, 2013), the arguments of many prominent researchers favoring an eastern European homeland have been dismissed as debatable. In the following we provide a brief review of these arguments, which our HPD interval for the first splits of PIE further substantiates, before presenting our chronology in Fig. 3 with the chronologically well-defined

[21] Zipf's law states that given any corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. It is related to the Pareto distribution (see any statistical textbook).

[22] I owe a debt of gratitude to Philippe Lemey for recommending and providing this latest version to me.

[23] We calculated the marginal likelihoods by stepping stone sampling (SSML), using the codes in Baele et al. (2012) and Baele et al. (2013) following the model selection tutorial (Rambaut 2014).

[24] Highest Posterior Density is the shortest interval in parameter space that contains the here 95 % of the posterior probability.

[25] Given that the databases are roughly attested around the year 2000 AD, "ago" equals "before 2 thousand (b2k)".

[26] This is not intended to be a full discussion of the issues surrounding this debate. Here, I recommend specialized literature e.g. Pereltsvaig & Lewis (2015) and the other cited sources.

prehistorical cultures in the possible migration areas, which does not necessarily indicate their Indo-European character.

4.1. The Neolithic expansion

The video accompanying the results in Bou12(SM[27]) attempts to parallelize Neolithic expansion with a computed diffusion process. IE expansion does not however match the known dates of the historic Neolithic Revolution (cf. e.g. Manning 2014). While Bou12/13 also repeatedly stressed that their approach supported the Anatolian Farming hypothesis (Renfrew 1987, passim), they immediately relativized this by adding that "[…] we think it unlikely that agriculture serves as the sole driver of language expansion […]," further arguing that the five major IE subfamilies emerged between 4000 and 2000 BC and were thus "contemporaneous with a number of later cultural expansions evident in the archeological record, including the Kurgan expansion." Whatever this might mean, the contrary would appear to be more convincing: Neolithic farmers can naturally be supposed to have brought their native language, which remained as a substrate upon the arrival of PIE from the Steppes.

4.2. Linguistic criteria

Bou12 (SM) claim "Our inferred outgroup (Anatolian) is consistent with the orthodox view in Indo-European linguistics (55[=Fortson 2010]), " However, this is not "the orthodox view," and it is marked as debated in the majority of the textbooks, none of which favors an Anatolian homeland for PIE. Moreover, Fortson (2010) also states with certainty "[t]hat they [the Hittites] or their ancestors did not originally inhabit Anatolia" and "[t]he Hittites, […], presumably came from the north."

The scholastic instruments used to detect historical neighborhoods in linguistics are loanwords and grammatical parallels. Although linguists are usually able to distinguish the direction of borrowing, as well as loans against common heritage from higher-level families, this is not irrefutable evidence for a neighborhood. Historical linguists have long viewed Uralic (UR) as the most probable neighbor for the PIE family. The results in Bou12/13 cannot explain the IE linguistic connections with the UR languages north of the Steppes at different stages. They imply that the arguments for this neighborhood "remain controversial" citing only two authors, neither of whom is particularly competent in this field, while simply dismissing the scholarly work of generations of well-known specialized historical linguists who favor (Proto)Uralic as the historical neighborhood for PIE. Uhlenbeck (1937) notes, "[...] the obvious possibility that the Indo-Germanic mother language might have been a mixed language with Uralic as one of its components." Seebold (1970) demonstrated in detail the agreements in the systems of personal pronouns. Anttila (1989) writes "[t]he Indo-Uralic hypothesis looks particularly strong, because the agreement is very good in pronouns and verbal endings, as well as in basic vocabulary." Campbell (1990) convincingly describes a "large number of similarities" among the names for trees in UR and IE which (independent of their character) point to an early neighborhood. Rédei's (1986) claim to have identified seven loanwords from PIE in Proto-Uralic has been accepted by several authors including Koivulehto (2001) and Mallory & Adams (2006). Helimski (2001) considered these and other words to be very good examples of communication between neighboring peoples in the late Copper Age. Tischler (2002) supports a relationship with UR speakers. Kortlandt (2009) references Gimbutas' theory that the IEs moved from a primary homeland north of the Caspian Sea to a secondary

[27] Not adjusted to the 2013 revision.

homeland north of the Black Sea. Haarmann (2010) accepts lexical concordances as well as the established grammatical congruencies. Beekes (2011) writes "Uralic […] shows similarities to Indo-European with respect to essential aspects of the language system, such as the ending of the accusative, verbal endings and personal and demonstrative pronouns." Schalin (2015) presents a long comparison between Finnish words and their assumed UR and IE relationships. Häkkinen (2015) concludes, "So much we get from the Uralic anchor: the Kurgan theory seems to be the only credible one."

The numerous connections between IE and UR, as loans in either direction or as sharing a common ancestor, corroborate a prehistoric neighborhood somewhere on the border between Europe and Asia. By contrast, no trace of PIE languages other than historical Phrygian, Armenian, Iranian or Greek migrants has ever been found in Anatolia either before, during or after the presence of IE–Anatolians. In addition, according to all the rules of historical linguistics, both the known Hattian substrate and the Akkadian and Hurrian adstrates provide hard evidence for the migration of the Hittites into Anatolia.

The Dutch linguist Beekes (2011) sums this up writing "Extremely improbable is the theory of the British archeologist, Colin Renfrew, in his book Archaeology and Language (1987)."

4.3. Genetics

Based on the DNA markers of R1a1a-M17 in 26 specimens in the Krasnoyarsk region, Keyser et al. (2009) conclude that "[o]ur results corroborate the 'steppe hypothesis'." The basis for this claim is the lack of physical traits (blue-eyed, fair-haired, etc.) detected by the team, which undermines the Anatolia hypothesis of eastward Indo-Iranian migration. Recent aDNA research (Haak et al. 2015) has revealed that "Corded Ware people from Germany traced ~3/4 of their ancestry to the Yamnaya, documenting a massive migration into the heartland of Europe from its eastern periphery. […] These results provide support for the theory of a Steppe origin for at least some of the Indo-European languages of Europe." There has been a recent explosion of published results including one recent study (Callaway 2015) which concludes that "[t]he findings echo those of a team that sequenced 69 ancient Europeans […]. Both groups speculate that the Yamnaya migration was at least partly responsible for the spread of the Indo-European languages into Western Europe." This line of argument has been expanded by Allentoft (2015) who, in addition, confirmed a considerable North-Eurasian admixture which may be assumed to represent the Uralic substrate.

4.4. Cultural concepts and archeology

Most linguistics find support for their argument in the evidence provided by goods traceable as common in both the IE languages and datable find of the same goods in archeological excavations ("paleo linguistics")[28] through the Eurasian Steppe. The following briefly reviews only the most impressive and recently confirmed examples.

[28] This view had become discredited by outdated studies that applied cognates of IE salmon or oak words to geographical habitats without taking in account even small changes in meaning or species.

4.4.1. Metallurgy

Apart from a word for gold, only one common word for a metal, *$h_2$éyos* “copper” later partly extended to “bronze” and even “iron” (see Buck 1949, Beekes 2011; Huld 2012 characterized it as the “generic metal”) has been convincingly proven.[29]

Natural copper has been worked since the eighth millennium in Anatolia-Persia. Evidence for the first copper extraction (smelting) in Serbian Belovode between c. 5000-4600 BC has been (indirectly) confirmed. Copper tools have been discovered in Serbian Pločnik soon spreading into all directions, perhaps slowly establishing a network of ‘copper kings.’ Circumpontic metal craft becomes visible in the archeological record in the fourth millennium BC where a new weapon, the shaft-holed copper axe, dominates the finds between the Balkans and the Caspian Sea, throwing light upon new social conditions (Hansen 2009). All this may well have played a central role in the spread of PIE. Historically, trading networks have often established the use of a lingua franca.[30] A knowledge of **tin** required to produce tin bronze appears after 3200 ± 200 BC. The word for tin differs in the PIE subfamilies and thus represents a terminus ante quem for the split of PIE.[31]

4.4.2. Wheeled transport

Archaeological confirmation for wheels used for transport is currently dated from c. 3500 BC onward (Mischka 2011). The terminology for wheeled transport is clearly labeled by IE words in all IE languages, which is a very strong indicator that PIE was still closely associated at this time. The smaller representation of PIE wheeled-transport vocabulary in Hittite can easily be explained by migration into an area of more highly developed cultures with advanced knowledge of wheeled transport, which has been confirmed for the period after c. 3400 BC.

The phonologist Heggarty (2006) may be right to criticize linguists as sometimes being careless in concluding from attested meanings to PIE meanings, but he doubtless goes too far in his claim that IEs could have named their transport technology individually with their own words after its repeated invention, a theory that is rejected by the majority of Indo-Europeanists. He further speculates that the terminology could have been borrowed along with the technology from elsewhere. While it is likely that foreign goods and ideas would bring their “label” with them, it is equally likely that this label would undergo subsequent changes according to localized sound law in the receiving language and thus remain distinguishable to a historical linguist from originally inherited words. Stifter (2008) supports this widespread view stating that “[i]f transport terminology had spread across the IE world after the breakup of the proto-language, this would be recognizable by deviant sound correspondences, the unmistakable diagnostic tool of loan relationships as opposed to genetic inheritance.” One example is the meaning of Albanian word *rrotë*, clearly a loan from Latin (Holm 2011), which is typically limited to the meaning “wheel” as opposed to the inherited word *rreth*, which typically has a broad spectrum of meanings including “hoop”, “circle”, “around,” etc.

[29] In both western and eastern IE sub-families, here, Germanic, Italic, and Indo-Iranian.

[30] For example, the Hanseatic League brought Middle Low German to Scandinavia as a lingua franca.

[31] Many (if not all) specialists in Indo-European languages would agree that cognate terms in widely dispersed IE subfamilies are a strong indication of the knowledge of cultural goods and vice versa. By contrast, non-cognate terms (e.g. the worldwide distributed term “computer”) would suggest later acquirement.

Thus, Indo-Europeanists overwhelmingly maintain that PIEs knew the wheel. However, this is little more than a platitude. Huld's (2000) sophisticated approach which focuses on linguistic forms fails to clearly state that different branches of IE have different terms for the concept "wheel." These differences are not random. The use of wheels for transport spread over a time span so short that the scatter of datings does not reveal a clear source in space or time (Burmeister 2011). Archeologists have found quite different techniques of combining the wheel with its axle in the second half of the fourth millennium BC which reveal a striking spatial similarity with the distribution of their labels (cf. Fig. 3, blue circles). PIE languages presumably spoken in the central and eastern European plains and ridges from the area covered by the Corded Ware culture in the west to the Poltavka culture in the east share the term **kwekwlo-s* for "wheel." Two outliers from surrounding highly mountainous areas are not included in this communicational network. We firstly find fixed wheel–axle constructions with a square-cut fit for the hole and shaft exclusively around the Alps (the oldest confirmed combination was discovered in Stare Gmajne near Ljubjana c. 3328–3116 BC (Mischka 2011)). All the languages in this area use the term **roth₂-* for "wheel." Note that some of Old German-speaking regions later borrowed this word from the Celts along with their superior techniques before usage expanded into Latvian and Lithuanian (and even Estonian and Finnic in secondary cases) as well as Albanian via Latin. The third and less used term **h₂wrg(h)-* is represented only in Hittite and Tocharian in a form that suggests a common origin presumably north or south of the Caucasus, irrespective of whether it was the result of a possible linguistic change from **kwekw(lo-s)* or a reinterpreted PIE root.

To summarize, wheel (and wheeled-transport) terminology displays three already divergent yet definite IE sources all of which can be traced to the second half of the fourth millennium BC: The first source is predominant, the second indicates a different technique and the third indicates an IE coining by an early wheel region along the Circum-Caucasian trade routes for Hittite and Tocharian (with different endings). This third source indicates a separate area, presumably south of the Caucasus, suggesting a common and not too early separation of these two languages.

4.4.3. Burial rituals

A chain of graves sharing typical traits and dating from the North Pontic Eneolithic period between 4600 and 4300 BC (Govedarica 2004) was discovered in a wide area from Transylvania in the west to the Caucasian foothills in the east. The bodies were uniformly interred in flexed supine positions on an ochre base and equipped with zoomorphic scepters suggesting widely-dispersed elite of copper traders. These finds confirm the times suggested for the first split of PIE by Gimbutas (1994). The tradition spills over into similar practices in the subsequent Pit Grave/Yama32 horizon in which particular graves of higher-ranking individuals were furnished with goods needed in the afterlife and often sacrificed animals or wheeled vehicles (Anthony 2007, Fortson 2010: 11).

[32] Traditionally called Pit Grave culture. Now often referred to using the Russian adjective Yamnaya, part of the Kurgan culture.

4.4.4. Economy

A detailed discussion of the abundant archeological finds supporting a homeland in the far east of Europe can be found in Anthony (2007). A possible Aryan homeland in the upper Volga–Kama region and eastward is proposed by Carpelan (2001). However, the absence of common PIE words for any grain type originating in the Fertile Crescent, including southeastern Anatolia, provides strong evidence against an Anatolian homeland because, in this case, an IE name for any food plant not known to the original inhabitants should have survived (Diamond 1992).

4.4.5. Horse culture

Continuations in Old Lithuanian, Gothic, Old Irish and Latin in the west, and Mycenaean, Ancient Armenian, Anatolian, and Indo-Iranian in the east confirm the existence of a PIE root _*(h$_1$)ek|u/w_-o for “horse” and thus PIE knowledge of the horse in either its domesticated or wild form. The linguistic evidence coincides with the archaeological record, which describes horses in nearly all later IE cultures with evidence for them as animals of prey (see Fig. 3 below) as well as their representation in human culture. This is not insignificant because “[t]he horse is often thought of as the IE animal *par excellence*; it was important in PIE myth and ritual […].” Fortson (2010), and Beekes (2011) assert that “[t]he horse was certainly the animal which more than any other characterized the Indo-Europeans.” Common rituals of horse sacrifice have been confirmed in the Indic, Roman and Irish traditions (Fortson 2010: [2.26]).

This alone might appear insignificant because horses were found throughout almost all the Eurasian steppe zones during the Holocene (and before). However, given the important role of horses in PIE, it is indicative that between the fifth to fourth millennia horses were not found in Italy and Greece (Vila 2005) and were very rare in Anatolia. Horses are absent from human culture in pre–Bronze Age Anatolia between the fifth and third millennia. The equids depicted in the "hunt painting” from Çatal Höyük East roughly dated to c. 7000 BC were described as “wild donkeys” by Ankara Museum as of November 2014. Arbuckle et al. (2014) found no domestic horses in Anatolia. A PIE home in Anatolia and expansion along a southern route as calculated by Bou12 would therefore suggest a PIE term for “donkey,” which does not exist, as nearly all European terms for donkey go back to the Latin word *asinus*, itself a late loan. Horses were also absent from the Neolithic economy. This excludes a PIE origin in Anatolia, particularly for the era calculated by Bouckaert (2012/2013).

By contrast, horses constituted a considerable proportion of the prey animals in the Eurasian Steppe and are also represented in artifacts (see Anthony 2007), rituals and myths (Gaitzsch 2011) adding convincing weight to the argument for the Steppe as the original home of the PIE community.

Many may wonder why there is no common PIE term for “riding (on horseback);” however, this may be explained by the dozens of terms for everyday activities in any linguistic dialect map, and riding is likely to have been an everyday activity for peoples as closely familiar with horses as the PIEs. A linguistic map of modern-day German lists over a dozen words for “to speak,” and it would be ludicrous to conclude that Germans could not speak.

5. Conclusion

The claim of a new scientific discovery on the question of the Indo-European homeland following the publication of Bouckaert et al. (2012) was enthusiastically taken up by the media. The revised result of 2013 putting the date of the first split at around 5580 BC, however, did not fit in with any of the alternative hypotheses (Eurasian Steppe vs. Anatolia).

In trying to find the reason for that result, an initial analysis of the Bou12 database revealed 283 traits containing exclusively {0} and {?} codes (corrected in Bou13). Removing these traits alone resulted in an age reduction of 1000 years (Table 2a, test series 1 and 2). No calibrations and parameters had been changed and therefore cannot be the reason for the tremendous difference. Further assuming that the zeroes were not the reason for the reduction, but rather the mass of included {?} codes,33 we removed languages with many ({?}-coded) gaps. The same suspicion led Cha15 to omit many gap-affected languages and meanings. As described in chapter 2.2.3 (Table 2a), every step of reducing gap-affected languages similarly and significantly reduced the root age. Suspecting the poor dataset previously employed to be at fault, we created our own dataset “H17” (Holm 2016) based on the meanings already reduced from 207 to 100 by Morris Swadesh (1955, 1971) to improve its quality. In addition, all gap-affected languages were removed except the essential Hittite, Tocharian B and Avestan.

Bou13 not only removed the empty traits of the database but also switched from the Dollo model, originally favored for good reasons, to the covarion model because of its slightly better Bayes factors. Table 3 shows that a further reduction of {?}-codes yields much better Bayes factors than changing the model.

The basal topologies sometimes differed considerably over the course of several hundreds of tests. These variations seemed to indicate what may be termed a “western versus eastern” dichotomy. Precisely this first-order dichotomy also resulted from a previous, lexicostatistical (=no chronology) calculation (Holm 2008) based on the best available IE dataset of around 1,140 verbal roots. The 95% highest probability density interval between c. 5190 to 3110 BC and a $\log_n$ (∝ or shape) posterior probability of −3,314 resulting in a final date of c. 4100 BC (see Table 5; more details in App. 1, tests 14 and 15) is much better than the approx. −50 000 obtained in previous approaches. The drop of 500 years obtained with the covarion model resulted again in worse posteriors and Bayes factors (Table 5, test series 15) with self-contradictory and illogical rate distributions, and can thus not be recommended.

This paper’s multidisciplinary discussion shows that the date of split achieved in the analysis corresponds to the Steppe hypothesis supported by major linguistic, archeological and recent genetic research. The further dispersal of the western and eastern limb around 3400 BC in particular corresponds to the three types of wheel–axle combinations and their different designations (chapter 4.4.2).

Finally it must be kept in mind that the handling of loans remains unsolved, and all results must be regard in relation to their whole probability density, as noted in Appendix 1 and visualized visually in Figure 3.

[33] Coding gaps in originally contained, but then omitted languages.

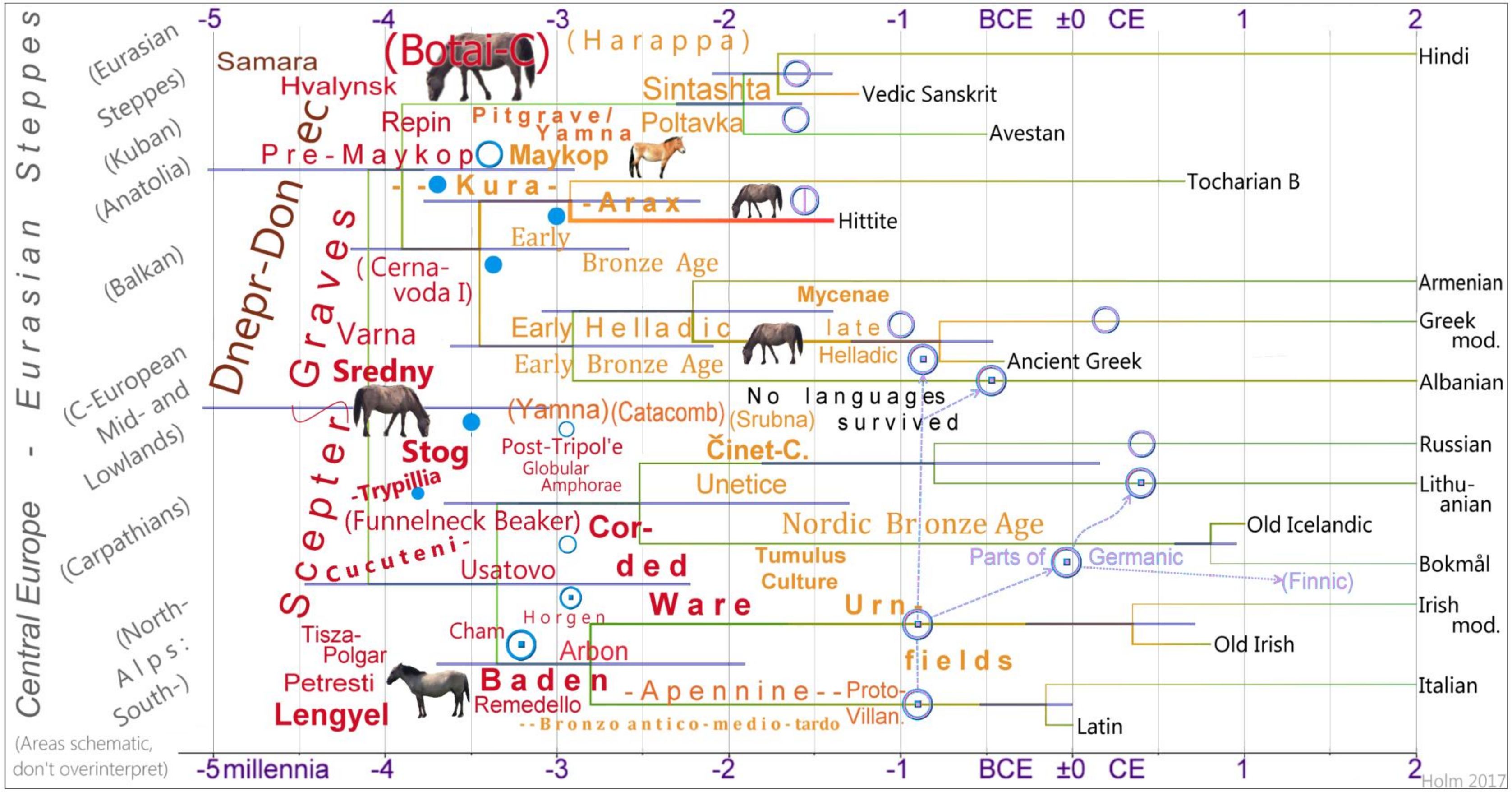

Fig. 3. Archeological features accompanying IE dispersal: The tree shows the main IE branches. The rates of linguistic change are indicated by color and thickness, from thin green = low rates to thick red =high rates; the violet bars indicate the 95% confidence intervals. The overlay roughly indicates in gray the probably passed geographical areas and archeological complexes at the correct times, not implying to be IE: dark red: Copper Age, orange: Bronze Age. Confirmed wheel types in blue color: smaller full circles: toys only; empty circles: **kwekwlo-s* (north and east); circles with square axle holes for *$roth_2$- in the west; and circles with a vertical bar for *$h_2wrg(h)$- in Anatolian and Tocharian. The same symbols after 2000 BCE in violet refer to the terms only! Horses are depicted where they have been proven to appear first in different cultures.

References

Alekseyenko, Alexander V, Christopher, J. Lee, and Marc, A. Suchard (2008). Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Systematic Biology* 57(5), 772-784.

Allentoft Morten E, Martin Sikora, Karl-Göran Sjögren, et al. (2015). Population genomics of Bronze Age Eurasia. *Nature 522(7555), 167-172.*

Anthony, David W. (2007). *The Horse, the Wheel, and Language: how Bronze-Age Riders from the Eurasian Steppes shaped the Modern World*. Princeton NY: Princeton Univ. Press.

Anttila, Raimo (1989). *Historical and Comparative Linguistics*. 2nd. rev. ed. Amsterdam: Benjamins.

Arbuckle, Benjamin S., S. W. Kansa, E. Kansa, D. Orton, C. Çakırlars, L. Gourichon, et al. (2014). Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey. *PLoS ONE* 9:e99845. doi:10.1371/ journal.pone. 0099845.

Beekes, Robert S.P., Michiel de Vaan (2011). *Comparative Indo-European Linguistics: An Introduction.* 2nd rev. ed. Amsterdam: Benjamins.

Baele, Guy, Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., and Alekseyenko, A.V. (2012). 'Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty' *Molecular Biology and Evolution 29(9), 2157-2167.*

Baele, Guy, Li, W.L.S., Drummond, A.J., Suchard, M.A., and Lemey, P. (2013). 'Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics' *Molecular Biology and Evolution 30(2), 239-243.*

Bergsland, Knut, and Hans Vogt (1962). On the validity of glottochronology. *Current Anthropology 3(2),115-153.*

Bouckaert, Remco, Ph. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, et al. (2012). Mapping the origins and expansion of the Indo-European language family, referring to supplementary material (SM), Additional supplementary material (AD). *Science* 337: 957-960; retrieved from http://www.sciencemag.org/content /337/6097/957/suppl/DC1; Corrected and revised in the following (2013). Mapping the origins and expansion of the Indo-European language family. Revision. *Science:* 342 (AAAS); retrieved from www.sciencemag.org.

Buck, Carl D. (1949). *A Dictionary of Selected Synonyms in the Principal Indo-European Languages.*Chicago: Univ. Chicago Press.

Burmeister, Stefan (2011). Innovationswege - Wege der Kommunikation; Erkenntnisprobleme am Beispiel des Wagens im 4. Jt. v.Chr. [Ways of Innovation-Ways of Communication; Problems in Recognition on the Example of the Wagon 4th Century BC].In: Hansen and Müller (editors), *Sozialarchäologische Perspektiven: Gesellschaftlicher Wandel 5000-1500 v. Chr. zwischen Atlantik und Kaukasus*: 211-240. Darmstadt: Zabern.

Callaway, Ewen (2015). DNA data explosion lights up the Bronze Age: Population-scale studies suggest that migrants spread steppe language and technology. *Nature 522, 140–141* (11 June 2015). doi:10.1038/522140a.

Campbell, Lyle (1990). Indo-European and Uralic tree names. *Diachronica 7(2),149-140.*

Carpelan, Christian, A. Parpola, P. Koskikallio (Eds) (2001). *Early contacts between Uralic and Indo-European: Linguistic and archeological considerations*. Helsinki: Suomalais-Ugrilaisen Seura.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European Steppe hypothesis. *Language 91(1), 194-244.*

Drummond, Alexei J., Simon Y.W. Ho, Matthew J. Phillips, Andrew Rambaut (2006). Relaxed Phylogenetics and Dating with Confidence, ***PLoS Biology*** *4(5), e88*.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8):1969-1973.

Dunn, Michael (2011), passim. Full dataset (word lists), referred to in the Bouckaert et al. Additional Supplementary Material. Retrieved from http://ielex.mpi.nl. 2015; available via http://corpus1.mpi.nl/qfs1/media-archive/eplc_data/dunn/Annotations/IE2012-lexical-data.txt.

Dyen, Isidor, J. Kruskal, & P. Black (1997). Comparative Indo-European database. Collected by Isidore Dyen. File IE-Rate 1; available http://www.wordgumbo.com/ie/cmp/iedata.txt. 7 February 2015.

Embleton, Sheila (1986). *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Fortson, Benjamin W. (2010). *Indo-European Language and Culture: An Introduction*. 2nd ed. Malden MA: Wiley-Blackwell.

Gaitzsch, Torsten (2011). *Das Pferd bei den Indogermanen. Sprachliche, kulturelle und archäologische Aspekte*. [The horse at the Indo-Germanics. Linguistic, cultural, and archeological aspects]. Berlin: LIT. German.

Gimbutas, Marija (1994). *Das Ende Alteuropas: Der Einfall von Steppennomaden aus Südrußland und die Indogermanisierung Mitteleuropas*. (W. Meid, Ed.). Revised translation of "The End of Old Europe: Intrusion of Steppe pastoralists from South Russia and the Transformation of Europe." In: *The Civilization of the Goddess: The World of Old Europe. Chapter 10.* San Francisco: Harper; 1991. Innsbruck: Inst. für Sprachwissenschaft. German.

Govedarica, Blagoje (2004). *Zepterträger–Herrscher der Steppen; Die frühen Ockergräber des älteren Äneolitikums im karpatenbalkanischen Gebiet und im Steppenraum Südost- u Osteuropas*.[Sceptre bearer –Rulers of the Steppes; the Early Ochre Graves of the Oldest Eneolithic in the Carpatho-Balkan Area and in the Steppes Area of Southeastern and Eastern Europe]. Darmstadt: Zabern.

Gray, Russell D., and Quentin D. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature. 426(6965), 435-438.*

Haak, Wolfgang, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, et al. (2015). Massive migration from the steppe is a source for Indo-European languages in Europe. Preprint arXiv:1502.02783.

Haarmann, Harald (2010). *Die Indoeuropäer; Herkunft, Sprachen, Kulturen* [The Indo-Europeans: Origin, Languages, Cultures] München: Beck.

Hansen, Sven (2009). Kupfer, Gold und Silber im Schwarzmeerraum während des 5. und 4. Jahrtausends v. Chr. In: Joni Apakidze, Blagoje Govedarica, and Bernhard Hänsel (eds.) *Der Schwarzmeerraum vom Äneolitikum bis in die Früheisenzeit(5000–500 V. CHR.). Kommunikationsebenen zwischen Kaukasus und Karpathen*. [PRÄHISTORISCHE ARCHÄOLOGIE IN SÜDOSTEUROPA 25. Intern. Fachtgg Tiflis / Georgien (17.-20. Mai 2007). Rahden/Westf.: Leidorf.

Häkkinen, Jaakko (2015). *Uralic evidence for the Indo-European homeland*. Available www.elisanet.fi/alkupera/UralicEvidence.pdf. Accessed 5 February 2015.

Heggarty, Paul (2006). Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data–and to Dating Language? In: Forster P, Renfrew C, editors. *Phylogenetic Methods and the Prehistory of Languages*. Cambridge UK: McDonald IAR.: 143-194.

Helimski, Eugen A. (2001). Early Indo-Uralic linguistic relationships: Real kinship and imagined contacts. In: Carpelan C, Parpola A, Koskikallio P. (Eds.), *Early Contacts* *between Uralic and Indo-European: Linguistic and Archaeological Considerations*. Helsinki: Suomalais-Ugrilaisen Seura: 147-206.

Holm, Hans J. (2005). Genealogische Verwandtschaft. In: Köhler R, Altmann G, Piotrowski RJ, editors.*Quantitative Linguistics: An International Handbook.* [HSK-Series 27, Chapter 45]. Berlin: De Gruyter.

Holm, Hans J. (2008). The Distribution of Data in Word Lists and its Impact on the Subgrouping of Languages. In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications*. Proc. of the 31th Annual Conference of the German Classification Society (GfKl), University of Freiburg; 2007. Heidelberg-Berlin: Springer.

Holm, Hans J. (2011). “Swadesh lists” of Albanian revisited and consequences for its position in the Indo-European languages. (Translated and updated from: Albanische Basiswortlisten und die Stellung des Albanischen in den indogermanischen Sprachen. Z. Balk; 2009; 45-2). *Journal of Indo-European Studies 39(1-2), 45-99.*

Holm, Hans (2016). H17= Glottochronological database for Morris Swadesh’s final 100 concepts and 17 representative Indo-European languages. Work in progress, available online via www.hjjholm.de; References.

Huld, Martin E. (2000). Reinventing the Wheel: The Technology of Transport and Indo-European Expansion. In: Jones-Bley K, Huld ME, Della Volpe A. (Eds.), *Proceedings of the 11th Annual UCLA Indo-European Conference, LA, June 4-5, 1999*.

Huld, Martin E. (2012). Some Observations on the Development of Indo-European Metallurgy. In: *Archaeology and Language: Indo-European Studies Presented to James P. Mallory*, JIES Monograph 60 : 281-356.

Kassian, Alexei, and Ilya Yakubovich (2011). Annotated Swadesh wordlists for the Hittite (Old Hittite) language (Anatolian group, Indo-European family). [Text version of database, created 14/10/2011]. Available http://starling.rinet.ru/new100/ana.pdf.

Keyser, Christine, C. Bouakaze, E. Crubézy, V. G. Nikolaev, D. Montagnon, T. Reis, B. Ludes et al. (2009). Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Human Genetics* 126(3): 395-410. DOI: 10.1007/s00439-009-0683-0.

Kloekhorst, Alwin (2008). *Etymological Dictionary of the Hittite Inherited Lexicon*. Leiden–Boston: Brill.

Koivulehto, Jorma (2001). “The earliest contacts between Indo-European and Uralic speakers in the light of lexical loans.” In: C. Carpelan, A. Parpola, P. Koskikallio (eds). *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations 242: 235-263.* Helsinki: Mémoires de la societé Finno-Ougrienne.

Kortlandt, Frederik C.C. (2009). Uhlenbeck on Indo-European, Uralic and Caucasian. *Historical Linguistics 122(1), 39-47.*

Lemey, Philippe, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard (2009). Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biology* 5:e1000520. doi:10.1371/journal.pcbi.1000520.

Mallory, James P., and Douglas. Q. Adams (2006). *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford: Oxford Univ. Press.

Manning, Katie, A. Timpson, S. Colledge, E. Crema, K. Edinborough, T. Kerig, and St. Shennan (2014). The chronology of culture: a comparative assessment of European Neolithic dating approaches. *Antiquity 88(342), 1065-1080.*

Meier-Brügger, Michael (2010). *Indogermanische Sprachwissenschaft* [Indo-European Linguistics]. 10th ed. Berlin: De Gruyter. (9th ed. available in English).

Mischka, Doris (2011). The Neolithic burial sequence at Flintbek LA 3, north Germany, and its cart tracks: a precise chronology. *Antiquity 1;85(329), 742-758.*

Pereltsvaig,Asya, and Martin Lewis (2015). *The Indo-European Controversy; Facts and Fallacies in Historical Linguistics.* Cambridge University Press.

Rambaut, Andrew (2014. *FigTree*: Tree Figure Drawing Tool, V1.4.2. Edinburgh: Institute of Evolutionary Biology Available http://tree.bio.ed.ac.uk/

Rambaut, Andrew (2014). http://beast.bio.ed.ac.uk/Model-selection, Retrieved oct, 2016.

Rambaut, Anrew et al. (2003-2013). *Tracer*. *MCMC Trace Analysis Tool*, version 1.6. http://beast.bio.ed.ac.uk/.

Rédei, Károly (1986). *Zu den indogermanisch-uralischen Sprachkontakten* [Upon the Indogermanic-Uralic Language Contacts]. Wien: SBÖAW 6;468.

Renfrew, Andrew C. (1987). *Archaeology and Language: The Puzzle of Indo-European Origins*. London: Pimlico.

Ringe, Don, Tandy Warnow, Ann Taylor (2002). Indo-European and computational cladistics. *Transactions of the Philological Society 100(1), 59-129.*

Rix, Helmut, Martin Kümmel, Thomas Zehnder, Reiner Lipp, Brigitte Schirmer (2001). *Lexikon der Indogermanischen Verben* [Lexicon of the Indogermanic Verbs]. 2. ed. Wiesbaden: Reichert.

Ryder, Robin J. (2010). *Phylogenetic Models of Language Diversification*. PhD. Oxford, UK: The Queen's College.

Schalin, Johan (2015). *Lexicon of Early Indo-European Loanwords Preserved in Finnish*. Available http://tcoimom.suntuubi.com/?cat=10, and [13].

Seebold, Elmar (1970). Versuch über die Herkunft der indogermanischen Personal-en_dungssysteme [Trial upon the Origin of the Indogermanic Personal Ending Systems]. *Zeitschrift für vergleichende Sprachforschung 85(2), 145-211.*

Swadesh, Morris (1855). Toward greater accuracy in lexicostatistic dating. *International Journal of American Linguistics 21(2), 121-137.*

Swadesh, Morris (1971). *The origin and diversification of language*. Sherzer J, editor. Chicago: Aldine (posthumous).

Starostin, Sergej, translated by N. Evans, and I. Peiros. (2000). Comparative-historical linguistics and lexicostatistics In: Renfrew, Colin, A. McMahon, and L. Trask (eds.) *Time Depth in Historical Linguistics*: 223-266.Cambridge UK: McDonald IAR.

Starostin, Sergej. A. (2005). The most recent result of Sergei Starostin (Workshop on the chronology in linguistics, Santa Fe 2004). Cited from Blažek V. From august schleicher to sergei starostin; on the development of the tree-diagram models of the Indo-European languages. *Journal of Indo-European Studies 35(1-2), 82-109.*

Stifter, David (2008). Review of Heggarty 2006. *LINGUIST List*. 6 October 2008.

Tischler, Johann (2002). Bemerkungen zur Urheimatfrage [Remarks upon the Urheimat question]. In: Fritz M, Zeilfelder S (Hrsg): *Novalis Indogermanica, FS Günter Neumann zum 80. Geburtstag*. Graz: Leykam.

Uhlenbeck, Christianus C. (1937). The Indogermanic mother language and mother tribes complex. *American Anthropologist* 39(3). Available: http://onlinelibrary.wiley.com/doi/10.1525/aa.1937.39.3.02a00020/pdf. Accessed 28 October 2009.

Vila, Emmanuelle (2005). Data on Equids from late fourth and third millennium sites in Northern Syria. In: Mashkour, Marjan. (ed.)*Equids in time and space; papers in honor of Véra Eisenmann; Procs 9th ICAZ Conference* 2002; Durham. Oxford: Oxbow Books.

7. Appendices

Appendix 1. Detailed results of cited test series examples

Appendix 1. Detailed results of cited test series examples

1	2a	2b	2c	2d	2e	3	4	6	7a	7b	7c	7d	7e
Test Ser. #	Data Source		Test Type - Exam-ples	Cod-ing of gaps	Traits	Model, Parameter	BS PS	Chain length [M]	Results				
	Name in text; Properties	Abbr. Taxa							Root Date BC (!)	Root mean BC	First split; topology	95% HPD Range "ago" (!)	-ln Poste-rior
1	Bou12	B 103	A-a l-12	{?}	6280	A= Dollo as in Bou12	Infy.	50	(12 runs)	**6500** ±80	Ana-Toc	Typically, 10232-7057	Typically, 52 230
2	Bou13 = revised align-ment		A-a	{?}	5997		Infy.	50	5531	**5508** ±104	Ana-Toc	8888-6015	51 588
			A-b						5622			9348-6409	51 591
			A-c					100	**5507**			9224-6382	51 590
			A-d						5371			8973-6223	51 583
3	Bou13, minus three gapped langu-ages (Luv, Lyc, TocA)	B 100	**A-a**	{?}	5866		Infy.	50	**5046**	**5048** ±62	Hit-ToB	8164-6101	50 540
			A-b				25k		4990		Balkan,IndIra	8288-5967	50 544
			A-c				50k		5113		Ind-Ira	8264-6139	50 537
			A-d				Infy.		4983			8172-6040	50 559
			A-e				Infy.		5107			8300-6115	50 562
4	Cha15, no 6 gapped langs. (Luc, Lyc, Osc, Umb, oPer, Kur)	C 97	A-a	{?}	5755		Infy.	60	4825	**4835** ±15	Ind-Ira	8037-5786	48 755
			A-b					50	4852		Balkan,IndIra	8297-5843	48 752
			A-c					54	**4828**		Hit-Toc	8012-5769	48 755
5	Bou13, no 52 gapped languages (Hit, ToB, Ave kept)	B 51	A-a	{?}	3981		Infy.	50	4771	**4722** ±57	Hit-ToB, Balkan unforced	8335-5565	27 206
			A-b						4641			8155-5533	27 205
			A-c						**4723**			8392-5559	27 206
			A-d						4752			8381-5624	27 206
6	Bou12	B	B-a	{?}	6280	Covarion	Infy.	30	8189	**8381**	Ana-Toc	13520-7314	51 996

		103	B-b			as in Bou13 (Science)		60	8572	±192		14965-7293	51 997
			B-c					30	**8382**			13767-7639	51 994
7	Bou13 (revd.)	B 103	C-r 1-10	{?}	5997			50 100	(10 re-runs)	**7870** ±1612		c. 15 500 to 3 960	c. 52 400
-	Bou13 (revd.)	B 103	=Publ.	{?}	5996	Co-varion Modified by “allow Identical” Parameter	Infy.	50	“5579”	**5579**	Ana-Toc	9351-5972	47 769
8	Bou13 (revd.)	B 103	D-a D-b	{?}					5533 5643	**5588** ±78		9251-6031 9487-5729	49 006 48 994
-	Cha15-B1Repl. (with changes to B&a)	C 103	=Publ.	{?}	5992		25k	20 only	“5750”	**5750**		9720-6180	~48 170
-	Cha15-B2	C	=Publ.	{?}	5754				“4810”	**4810**		8780-5400	~46 220?
9		97	E	{?}	5755		100	50	4898	**4898**		9145-5396	46 256
10	17 languages (9 mod., 8 extinct)	H 17	F-a **F-b** F-c	{?}	760	Dollo	100	75	5047 **5055** 5068	**5056** ±9	Hittite	8976-5395 8956-5405 8897-5401	3924 **3906** 3906
11		H 17	G-a **G-b** G-c	omitted	658	Dollo			4783 **4794** 4803	**4793** ±10	Hittite-Tocharian-B	8537-5391 8490-5340 8520-5420	3584 **3559** 3584
12		H 17	H-a **H-b** H-c	{?}	760	Cov., allow Identical		200	4175 **4243** 4269	**4229** ± 48.5	Hittite, self-contradicting branch rates	8216-4635 **8232-4682** 8264-4716	4196 **4194** 4303
13		H 17	I-a **I-b** I-c	omitted	658	Cov., allow Identical		100	4120 **4178** 4246	**4181** ±63	Hittite	8276-4651 **8512-4638** 8276-4651	3827 **3854** 3827
14		H 17	GW-a **GW-b** GW-c	omitted		Dollo, allow Identical		75	4099 **4100** 4107	**4102** ±4.36	West : East dichotomy	7212-5038 **7234-5109** 7188-5054	3556 **3556** 3556

15		H 17	**<u>IW-a1</u>** IW-a2 IW-a3	o-mit-ted		Cov., allow Identical			3515 3525 **<u>3531</u>**	**3524** ±8	West : East dichotomy, Rates confused	6701-4595 6746-4594 **<u>6711-4608</u>**	3852 3851 **<u>3852</u>**

Legend: Test series #; 2a:Data source: Name in text; 2b: Abbr. in file with number of languages; 2c: Test Type, underlined: File attached as example; 2d: Handling of gaps; 2e: Number of traits; 3. Model details; 4: Population Size; 6: Million mcmc runs; 7: Results; 7a: Root date BC; 7b: Test type mean ± adjusted standard deviation; 7c: Primary split: Hit(tite), Toc(harian)B); 7d: TRACER: 95 % HPD ago; 7e:TRACER: negative log_n Median Posterior.

Ap**pendix 2:** Date priors for extinct languages (Leaf heights L), means of node (N) heights.

Dates of N(ode), L(anguage)		Calendar dates	Calibration b2k[34] / ago	Reasons and sources
N1	Rus-Lit:	1100 BCE	3100 ±600	The P-Baltic Bronze Age differs from the presumably P-Slavic Černoles Culture (Marshall Cavendish 2010:1030). Previous glottochronological studies gave c. 1100 BCE (Bou12/13), or 1210 BCE (Burlak/Starostin 2001) for the split.
N2	N-Germanic	900 CE	1100±200	Settlement of Iceland with HPD 850-922 CE (Sveinbjörnsdóttir 2016), which from 1050 onward considerably split from “Old Norwegian”, however, for the final split, different sources give dates between 1200 to 1500 CE (Torp 2004: 56). Deciding for the computations is the time of attestation, which, for the Old Norse literary works, mainly based on Old Icelandic, lies between the tenth through thirteenth centuries, or 1100±100 CE.
L1	Old Icelandic (B&A “Old Norse”):	900 to 1300 CE, with mean around 1100 CE	900±100	
N2	Irish-Italic (Celtic-Romanic)	after 2800 BCE	4240 ±600	David Anthony (2007:367) assumes that “thousands of Yamnaya kurgans in Eastern Hungary suggest a more continuous occupation … by a larger population of immigrants … could have spawned both pre-Italic and Pre-Celtic.” Such expansions from the area are attested for the Baden Culture (3500 – 2800 BCE), at the end outreaching to the north and south of the Alps. Tribe of Latins assumed to live near Rome since c. 1000 BCE. Our calibration equals the result of B&a (2012), and tests with reduced datasets based upon Bou12/13 and Chy15 data with results between 6500 and 5500 BCE, and can thus not be responsible for a lower root age.
L2	Old Irish	8th to 9th CE	1200 ±75	Bible glosses preserved on the Continent 8th to 9th century CE (Lucht 2007: 6).

[34] All employed as “normal priors”, because BEAST too often fails to accept uniform priors of the same extension.

L3	Classical Latin	75 BCE - 75 CE	2000 ±75	Meier-Brügger, E427
N3	Balkan Branch	3200 to 1700 BCE	4450 ±600	The later Balkan languages split from south-eastern groups after the end of the Cernavoda I Culture c. 3200 BCE, which itself had come “from the east” (Mallory 1997; Anthony 2007:260, “Cernavoda after 4000 BCE”). Since c.-1650 the Mykenes were already Pre-Greek.
L4	Ancient Greek	400 to 700 BCE	uniform 2700 to 2400	Meier-Brügger [E418]: “Anfang des 7. Jh”; Beekes (2011:24) “end of the 8th century. with Homer.” Thus probably -700 earliest date of origin of Homeric epics (Ilias, Odyssey) with editorial changes to -300. Bou12/13 use the relatively late date of “Classical Attic” 2400±50 b2k.
L5	Hittite	1650 to 1200 BCE	3400±250	Bou12/13 insert 3450±125 b2k, Cha15 3400±100 b2k. Kassian/ Starostin (2011) claim many of the words in our list to be attested for Old Hittite, for which the Russian Wikipedia (with newer sources) gives 1650 to 150 BCE. However, Meier-Brügger [E410] holds “Old Hittite attestations since 1570”, and Beekes (2001:20) writes “Bulk of attestations from 13th century.”
L6	Tocharian B	650 CE	1350 ±75	From sixth to eighth (12th) centuries., thus 500-800, with the bulk probably 650 CE.
N4	Indo-Iranian	1800 to 1000 BCE	3400±300	The Andronovo Culture, flourishing between the 18th to the 14th -10th centuries from the Ural river in the West to the Altai Mountains in the East is widely assumed as “Aryan” cradle. (Anthony 2007: 18th to 12th century; Kuz'mina 2007).
L7	Avestan	600 to 400 BCE	2500±75	With Bou12/13; Chy15: 550-450 BCE. Avestan attestations are overwhelmingly Young-Avestan (Meier-Brügger: E406: 6.-5. Jh. v. C.).
L7	Vedic (Sanskrit)	1500 to 1200 BCE	3250 ±250	With Chy15, 3250±250 b2k. The composition of the Rigveda is dated to roughly between c. 1500–1200 BCE. (Flood 1996: 37; Witzel 1995: 4; Anthony 2007: 454); thus older than Bou’s12/13 3000±100 b2k.

Remarks: 1. Note that the here given standard deviation σ comprises c. 68.3 % of the data, and 2σ would comprise 95.4 %. 2. A Greek split (assumed to have happened shortly before departure of Mycenaean, attested in Linear B texts from the end of the 15th century BCE) is not used, because it should not be the time of split between the here only employed Homerian vs. recent Greek.

References:

Anthony, D. (2007). The Horse, the Wheel, and Language: how Bronze-Age Riders from the Eurasian Steppes shaped the Modern World. Princeton Univ. Press, Princeton, NY.

Beekes, Robert S.P., Michiel de Vaan (2011). Comparative Indo-European Linguistics: An Introduction. 2nd rev. ed. Amsterdam: Benjamins.

Bouckaert et al. (2012). Mapping the origins and expansion of the Indo-European language family, in SM, Science 337 (6097): 957-960 (24 August, 2012). Not updated in 2013.

Burlak, S.A. & Starostin, S.A. (2001). Vvdenie v lingvističeskuju komparatistiku. Mockva: MGU-RGGU. Pp 82-105.

Flood, G. D. (1996). An Introduction to Hinduism, Cambridge University Press.

Holm, H. J. (2008). The Distribution of Data in Word Lists and its Impact on the Subgrouping of Languages. In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds): Data Analysis, Machine Learning, and Applications. Proc. of the 31th Annual Conference of the German Classification Society (GfKl), Univ. Freiburg, 2007. Heidelberg-Berlin: Springer.

Holst, J. H. (2009): Armenische Studien, Wiesbaden, Harrassowitz.

Kassian, A., G. Starostin (2011). Annotated Swadesh wordlists for the Hittite (Old Hittite) language (Anatolian group, Indo-European family). [Text version of database, created 14/10/2011]. From http://starling.rinet.ru/new100/ana.pdf.

Kulke, H., Rothermund, D. (2004): A history of India. London: Routledge. 4th ed.

Kuz'mina, E. E. (2007), in: Mallory, J. P., ed. *The Origin of the Indo-Iranians*. Leiden: Brill.

Lucht, Martina (2007): Der Grundwortschatz des Altirischen. Doctoral dissertation. Bonn: RFW-University.

Marshall Cavendish (2010). World and its Peoples: Europe. Marshall Cavendish Reference, New York.

Mallory JP, & DQ Adams (1997). *Encyclopedia of Indo-European Culture*. London: Fitzroy Dearborn Publishers.

Meier-Brügger, M. 2010. Indogermanische Sprachwissenschaft. De Gruyter. 9th edition.

Ringe D, T Warnow, A Taylor (2002). Indo-European and computational cladistics. Trans. Phil. Soc. 100-1:59-129.

Sveinbjörnsdóttir ÁE, Ch Bronk Ramsey, J Heinemeier (2016): The Settlement Date of Iceland Revisited: Evaluation of 14C Dates from Sites of Early Settlers in Iceland by Bayesian Statistics. Radiocarbon 58-02: 235 – 245.

Torp, Arne (2004): Nordens sprog med rødder og fødder. Nord 2004:10. København: Nordisk Ministerråd.

Witzel, M. (2005): Indocentrism: autochthonos visions of Ancient India. In Edwin F. Bryant, and Laurie L. Patton (eds.), The Indo-Aryan controversy. Evidence and interference in Indian history. Routledge, London, New York, pp341-404.

Witzel, M. (1995). "Early Sanskritization: Origin and Development of the Kuru state". Electronic Journal of Vedic Studies.

Zimmermann, Th. (2008). Steinerne Rundgräber der inneranatolischen Frühbronzezeit. In Arch. KorrespondenzBl. 38: 191-200.

App. 3. Examples of cited Input Files, one of each series Available from the author on personal request

T 2. B103 x5997;Ac=5507.xml	T 6. B103 x6280;Bc=8382.xml	T10. H17 x760;Fb=5055.xml	T13. H17 x658;Ib=4169.xml
T 3. B100 x5866;Aa=5046.xml	T 7. B103 x5997;Cc=7917.xml	T11. H17 x658;Gb=4794.xml	T14. H17 x658;GWb=4100.xml
T 4. C 97 x5755;Ac=4828.xml	T 8. B103 x5996;D=5533.xml	T12. H17 x760;Hb=4173.xml	T15. H17 x658:IWa1=3524.xml.
T 5. B 51 x3981;Ac=4723.xml	T 9. C 97 x5755;Ea=4898.xml		

Mastering the measurement of text's frequency structure: an investigation on Lambda's reliability

Rafaël Poiret[1], *Haitao Liu* [1,2]

Abstract. Lambda is a measure of frequency structure that has been presented to be independent of text size (Popescu, Čech & Altmann, 2011). We demonstrate in this study that Lambda is obviously dependent on text size, confirming the findings of Čech (2015). Based on the assumption that Lambda was independent of text size, Popescu, Čech & Altmann (2011) investigated into its capacity to detect text genre. We find that Lambda is still able to distinguish genres, but only very different ones. We also propose an experimental method based on Chinese to observe if Lambda is really able to measure the degree of analytism/synthetism of a text (Popescu, Čech & Altmann, 2011). We find that this method is promising. Moreover, our results corroborate with the assumption that Lambda has this property.

Keywords *: Lambda, Chinese, text genre, French, text size*

1. Introduction

The seeking of the formula able to measure the vocabulary richness of a text has attracted many intrepid statisticians. Vocabulary richness, in quantitative linguistics, is the proportion of different words in a text.

A well-known measure used to calculate the vocabulary richness is TTR (type-token ratio). TTR is the number of types divided by the number of tokens in a text sample. The problem with this measure is its dependence on the text size. Indeed, it is not reliable to compare the vocabulary richness of two samples of different sizes.

Popescu, Čech & Altmann (2011) proposed Lambda which measures the frequency structure of a text and is able to detect its vocabulary richness. They insisted mostly on the independence of Lambda on text length. But when they verified this assumption, Popescu, Čech & Altmann (2011) did not pay attention to the fact that Lambda may be influenced by other factors. According to their view, Lambda is sensitive to authorship, to genre, and to degree of analytism/synthetism of a given text. However, the corpus they used to analyze the relation between Lambda and text length consists of texts of different genres in different

[1] Department of Linguistics, Zhejiang University, China ; [2] Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com

languages, which may lead to unreliable results. This has been partly corrected by Čech (2015), who re-examined the dependence of Lambda on text size. In order to get rid of the influence of different languages, he focused on Czech and English separately. An obvious dependence of Lambda on text size has been discovered for both languages. The author found, for each of them, the intervals in which there is no influence of text size on Lambda. This study is a first counter-expertise concerning this property of Lambda, but it did not remove the possible effect of other factors on Lambda. Thus, this attempt is not totally satisfying.

Based on the assumption that Lambda was not influenced by text size, Popescu, Čech & Altmann (2011) tried to demonstrate that this measure was able to detect text genre. They worked on 16 different genres in 15 languages. They established a ranking of genres expressed by Lambda. This ranking is shared by different languages. However, because of the dependence of Lambda on text length, this genre defferentiating capacity of Lambda should be re-investigated.

As to Lambda's sensibility to the degree of analytism/synthetism of a text, Popescu, Čech & Altmann (2011) analyzed the Lambda of prose texts from 25 languages and found that Lambda was able to measure the degree of analytism/synthetism of texts and groups them according to their source language. What we propose here is to use one language, Chinese to verify this property. There is no blank between characters in Chinese written system. Thus, segmentation tools are employed to tokenize Chinese texts. Linguistic strategy may vary from different tools. Some may have strategy tending toward analytism, others toward synthetism. If Lambda is really sensitive to the morphological properties of texts, this measure should vary according to the strategy employed by the segmentation tool.

In this study we will work on the following research questions :

- Is Lambda dependent on text size?
- Is Lambda able to detect text genre?
- We propose a method to investigate on the property of Lambda to detect the degree of analytism/synthetism of one text. Is this method promising? Does it corroborate with the assumption that Lambda has this property?

2 Material and Methods

Lambda is not only based on words but also on the rank of their frequencies, because it includes the arc length *L*. The arc length is defined as the sum of Euclidean distances between neighboring distances. Below, f_r is the frequency of $r -$ the most frequent token, *V* is the total number of tokens.

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

Popescu, Čech & Altmann (2011) transformed *L* so that it was not dependent on the text size, and proposed Lambda. *N* is the total number of tokens in the text.

$$\Lambda = \frac{L\left(Log_{10}N\right)}{N} \qquad (2)$$

In the following part, we focus on the corpus specifically constructed for the present study and the corresponding method we employed ; the first research questions using French texts and the other two with Chinese. Lambda is computed with the software QUITA[2], the $U-$ test and the Brown-Forsythe test are performed with the Python-based ecosystem of open-source software Scipy[3] and SPSS[4].

2.1 Lambda and text size

Lambda is a measure that react to different aspects of texts, i.e. genre, authorship, vocabulary richness and morphology. However, the effect of text size must first be considered, which can be best done with texts from a single language. That is the reason why we chose to work on texts of one language, French, and of one genre, law text. The law genre suffers very slightly from the influence of authorship. We selected texts belonging to one specific subgenre « Decisions » of the « Constitutional Council ». In this way, we also control the degree of vocabulary richness of the texts. Texts have been written between 2011 and 2016. We obtained them from the official website of the French Constitutional Council[5] using the Python module Beautiful Soup[6]. We got 1092 texts which lie in the interval $N \in$ <114, 17717>. After having calculated Lambda for all the texts, we computed four different regressions : power, logarithmic, exponential and linear, with their corresponding R^2 value.

In order to continue investigating in the way Lambda evolves among different sizes of texts, we computed the mean Lambda for different intervals. Because we wanted to make easier the comparison between our results, we used the same intervals as Čech (2015) did. We computed the regressions and the corresponding R^2 to see which one had the best fit to the results. We tried to find if there were any significant change between the mean Lambda of two subsequent intervals. In order to do that, we followed the method of Čech (2015). We calculated the $U-$ test and the p-value for each couple of two subsequent intervals.

[2] https://code.google.com/archive/p/oltk/

[3] https://www.scipy.org/

[4] http://www.ibm.com/analytics/fr/fr/technology/spss/

[5] http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/les-decisions/acces-par-date/decisions-depuis-1959/les-decisions-par-date.4614.html

[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

2.2 Lambda and genres

Our corpus is composed of 8 genres (Table 1). The language is Chinese. If one can easily find large quantities of official texts available on different government websites, it is another story for genres like scientific texts or novels, which may be under copyright regulations, or simply not available in digital version. Thus, it is not easy to get exactly the same amount of data for each genre. We present below the genres and the corresponding number of texts *n* we worked on in this study. We tokenized these texts with PyNLP[7], a Python wrapper around ICTCLAS2015[8].

Table 1
Corpus for each genre

Genre	*n*	Comment & precision
Child	148	The texts have all been scraped from http://story.beva.com
Media	106	Articles from the website of the People's Daily. We scraped articles of three sub-genres : culture, legal and society.
Official	859	Law texts, of the subgenre « Procedural law » (诉讼法; *sùsòng fǎ*) scraped from http://www.law-lib.com. They have been written between 2011 and 2016.
Sanwen	64	A Chinese subgenre of prose literature. The author exposes its feelings and opinions. We chose texts from 1980 until 2007.
Scientific	30	Scientific essays belonging to economic field.
Translation	20	Chinese translation of prose literature texts in French, German and Russian. For this genre, we did not pay attention to the date of writing. The texts come from http://www.xieguofang.cn/index.htm
Xiaopin	34	Xiaopin is a kind of Chinese comedy. We found the texts on Baidu.
Xiaoshuo	177	Xiaoshuo is a genre of Chinese prose literature, similar to the genre of novel. They have been written between 1980 and 2015.

When working on the capacity of Lambda to detect genres, Popescu, Čech & Altmann (2011) did not pay attention to the size of the texts they used. However, this capacity may not have anything to do with the size of the texts. We built two corpura. In one, the size of the texts has

[7] https://github.com/tsroten/pynlpir

[8] http://ictclas.nlpir.org/

not been controlled. In the other one, texts size belong to a fixed interval, $N \in < 600, 2600 >$.

We have computed the mean Lambda for each genre of these two corpura. We will see if the genres ranking they express is the same or not. If the capacity of Lambda to detect genre does not have anything to do with text size, we should expect that the two rankings are the same. If the two rankings are different, this should justify investigation on how the mean Lambda of each genre evolve along different text size.

Before comparing the mean Lambda of each genre for the two sets of data, we must verify if, for each set, Lambda is sensitive to the genre difference. In order to do so, we decided to apply One-Way Anova. We checked if it met the two conditions: whether all the Lambda values have normal distribution, and whether they satisfy the homogeneity of variances.

Table 2

Test of normality for the $N \in$ <600, 2600> corpus

	Shapiro-Wilk		
	Statistic	**df**	**Sig.**
Child	0.941	20	0.245
Medias	0.985	60	0.661
Official	0.999	663	0.952
Sanwen	0.949	29	0.176
Scientific	0.937	25	0.128
Translation	0.843	7	0.105
Xiaopin	0.975	27	0.736
Xiaoshuo	0.926	19	0.147

Table 3

Test of normality for the size non-controlled corpus

	Shapiro-Wilk		
	Statistic	**df**	**Sig.**
Child	0.987	148	0.170

Medias	0.984	106	0.237
Official	0.999	859	0.980
Sanwen	0.964	64	0.061
Scientific	0.939	30	0.083
Translation	0.925	20	0.125
Xiaopin	0.973	34	0.545
Xiaoshuo	0.986	177	0.070

From Table 2 and 3, we can see that all the data for both corpus obey the normal distribution condition, as $p > 0.05$. However, none of them statisfy the homogeneity of variances. We used the non-parametric test Brown-Forsythe.

2.3 Lambda and the degree of analytism/synthetism

In this section, we will present our method based on Chinese to verify the sensitivity of Lambda to the degree of analytism/synthetism of one text. The text unit on which Lambda is based is token, i.e. a character string delimited by blanks, punctuation, beginning and end of a text. However, there is no blank between Chinese characters. That is the reason why we use segmentation tools, to add blanks between words before processing Chinese texts. Since the notion of word itself is not well defined, different segmentation tool may give very different results. Some may tend toward analytism, some may tend toward synthetism. We say that a language is more analytic when the grammatical links are conveyed by distinct words (Le Trésor de la Langue Française Informatisé). It is traditionally opposed to synthetic language which tends « to gather many morphemes in one unique word »[9] (Dubois et al., 1973). Most of the words in Chinese are composed by only one morpheme, but we can find some exceptions. The morpheme 过 (*guò*) indicates that a situation, expressed by the verb it follows and on which it depends, has been experienced. In the sentence:

他	去	过	五	次	北京
tā	*qù*	*guò*	*wǔ*	*cì*	*běijīng*
he	go	GUO	five	time	Beijing

'He went to Beijing five times'

[9] Translation from French made by the authors of this paper

过 (*guò*) informs us that ‘go to Beijing’ has been experienced in the past. There could be two ways to tokenize the segment « 去 过 » : the first one separates the grammatical morpheme 过 (*guò*) from the verb 去 (*qù*), the second one adjoins them together. The first choice will make this segment more analytical, the second one more synthetical.

Each segmentation tool may differ in its morphological approach to Chinese. If Lambda has the property to detect the degree of analytism/synthetism of a text, it should be sensitive to this sort of difference. The approach chose by the segmentation tool can be defined by observing how some more or less grammaticalized elements depending on verbs, like 过 (*guò*) we presented above are treated. We need to build a closed list of elements and analyze how they have been tokenized by each segmentation tool. They are presented in Table 4. As the category to which these elements belong is still object of debates in Chinese linguistics, we employed the ones proposed by a reference book (Li & Thompson, 1989). Numerous studies agree on the grammaticalization of these elements (Huang, Ching & Yu, 2008; Li, 2001; Peyraube, 2006). This is the main point here.

Table 4
Grammaticalized elements

Category		**Element**	**Literal meaning**	**Semantic feature**	**Example & translation**	
Resultative	Phase	到 (*dào*)	‘To have reached’	Phase	我闻到 *wǒ + wén + dào* I + smell + DAO	‘I smelt’
		完 (*wán*)	‘To have finished’	Phase	我吃完 *wǒ + chī + wán* I + eat + WAN	‘I finished to eat’
	Directional	下去 (*xiàqù*)	‘Going down’	Continuation	我活下去 *wǒ + huó + xiàqù* I + leave + XIAQU	‘I’m still alive’
		起来 (*qǐlái*)	‘Rising up’	Inchoation	我笑起来 *wǒ + xiào + qǐlái* I + laugh + QILAI	‘I laughed’
Aspect marker		过 (*guò*)	‘Pass’	Experiential	我去过 *wǒ + qù + guò* I + go+ GUO	‘I went’

	着 (*zhe*)	'A move in'	Durative	我活着 *wǒ + huó + zhe* I + live + ZHE	'I live'

The resultatives 到 (*dào*) and 完 (*wán*), presented in Table 4 indicate an aspectual meaning (Sun, 2013). They express « something more like the *type* of action described by the first verb or the degree to which it carried out than its result » (Li & Thompson, 1989). The elements 起来 (*qǐlái*) and 下去 (*xiàqù*) have meanings of inchoation and continuation (Che, 2014). Chang (1993) notes that both of them have lost their spatial meaning, and became grammaticalized. Chao (1968) designated 起来 (*qǐlái*) and 下去 (*xiàqù*) as verbal suffixes, putting them in the same category as 过 (*guò*) and 着 (*zhe*). Li & Thompson (1989) consider 过 (*guò*) and 着 (*zhe*) as experiential and durative markers respectively. We chose the novel of Su Tong, *My Life As Emperor* (我的帝王生涯; *wǒ de dìwáng shēngyá*) published in 1992, and tokenized it with 5 different segmentation tools. They are all well-known to provide high accuracy. We obtained 5 different tokenized files. We extracted from them the segments containing the elements of Table 4. We checked this extraction manually. We calculated, for each file, the percentage of times that the grammatical elements were split from the main verb they follow. The more these elements are split, the more the given text tends toward analytism.

3 Results and Discussion

3.1 Lambda and text size

We computed four different regressions : power, logarithmic, exponential and linear, with their corresponding R^2 value.

Table 5
The R^2 value for each trend line

Regression	R^2
Power	**0.5922**
Logarithmic	0.5852
Exponential	0.4329
Linear	0.395

Table 5 indicates that power provides the best fit, with R^2 of 0.5922. We show below the graph of the distribution of Lambda along text of different size, from the shortest one, to the longest one. The dashed line represents the power fit.

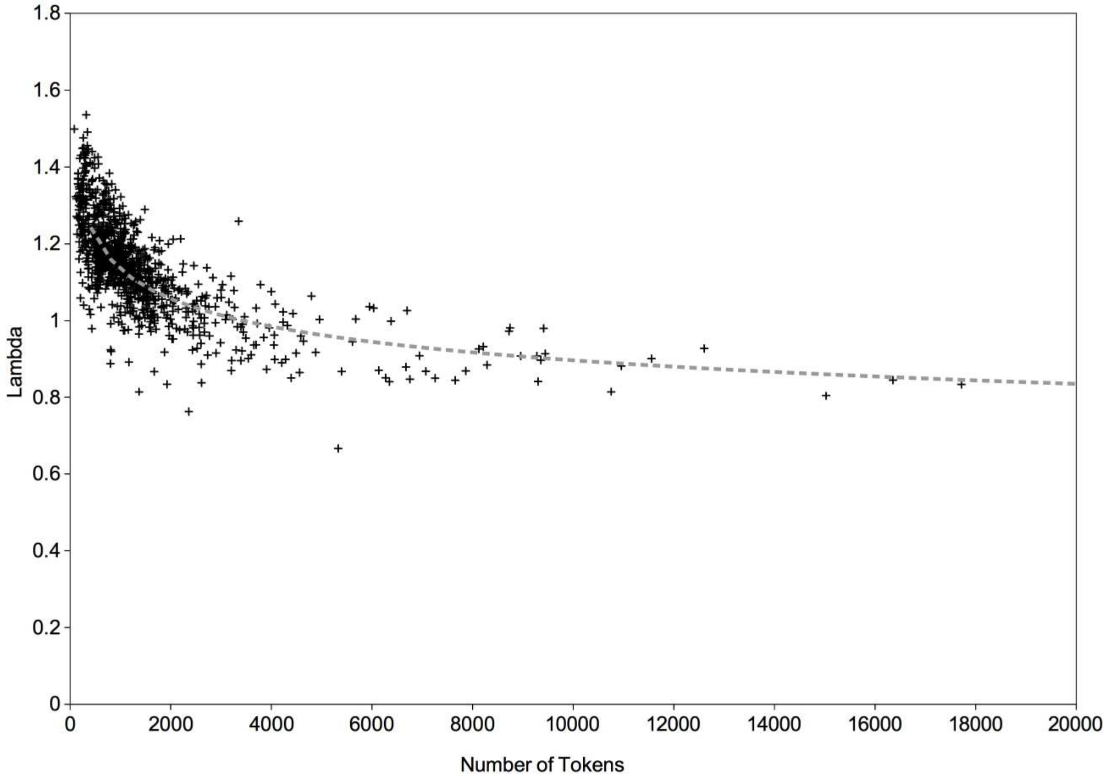

Figure 1. The distribution of Lambda along text size $N \in$ < 114, 17717 >

Figure 1 permits us to confirm a certain relationship between Lambda and text size. Čech (2015) got an inverted bell-shape with his two monolingual corpura containing different genres. Here the Lambda decreases in the interval $N \in$ <100, 2000>, and then continues descending in a low slope until the endpoint. The difference of Lambda for one short interval of N, $N \in$ <100, 600> (with n = 314) is important. It goes from 0.98 to 1.54. This disparity of Lambda observed in a factor-controlled corpus (language, genre, authorship, size) means that the results given by this measure should be interpreted with cautiousness. We computed the mean Lambda for different intervals (Table 6).

Table 6
The mean Lambda for each interval

Interval	***n***	**Mean lambda**	**Variance**
101-200	25	1.3116	0.0038
201-300	55	1.2974	0.001
301-400	50	1.2679	0.0139
401-500	40	1.2414	0.0112
501-1000	430	1.1839	0.0045
1001-1500	230	1.12766	0.0054
1501-2000	99	1.07511	0.0046
2001-2500	45	1.0393	0.0058
2501-3000	29	1.0051	0.0051
3001-4000	32	0.9941	0.0069
4001-6500	28	0.9392	0.0073
6501-9000	14	0.9073	0.0031
9001-20000	12	0.8793	0.0027

We calculated the four regressions, and the corresponding R^2 to see which one had the best fit to the results.

Table 7
The R^2 value for each regression

Regression	R^2
Exponential	**0.9903**
Linear	0.989
Logarithmic	0.8744
Power	0.849

Table 7 indicates that exponential regression has the best fit to the distribution of our data, with a R^2 equal to 0.9903. Apart from this regression, all the three others fit the distribution well.

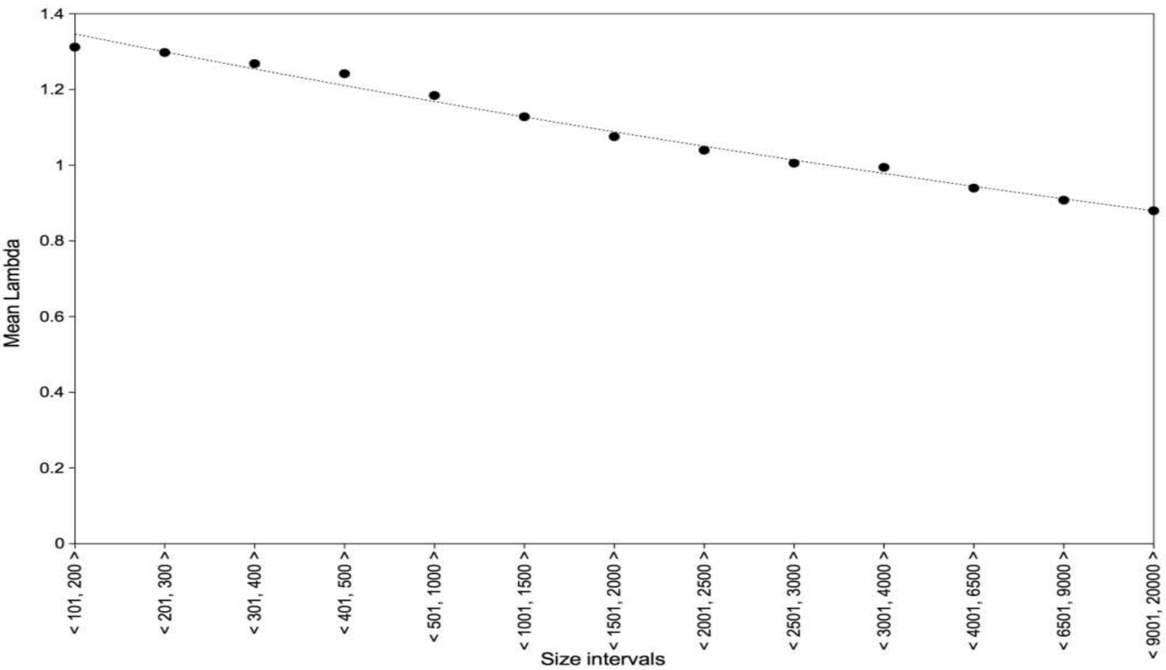

Figure 2. The distribution of mean Lambda along different intervals

Figure 2 is drawn from the data of Table 6 and the dashed line represents the exponential regression. The first point is 1.311 and the last one is 0.879. The line descends straightly from the first point to the endpoint. This shows and confirms the evident dependence of Lambda on text size. We tried to find if there were any significant change between the mean Lambda of two subsequent intervals. We calculated the $U-$ test and the p-value for each couple of two following intervals. The results are presented in Table 8 below.

Table 8
The $U-$ test and p-value for each couple of following intervals

Interval	**Mean Lambda**	$U-$ test	**p-value**	**Interval**	**Mean Lambda**
< 101, 200 >	1.311	0.78	0.435	< 201, 300 >	1.297
< 201, 300 >	1.297	1.38	0.1685	< 301, 400 >	1.268
< 301, 400 >	1.268	1.12	0.2617	< 401, 500 >	1.241
< 401, 500 >	1.241	**3.38**	**0.0007**	< 501, 1000 >	1.184
< 501, 1000 >	1.184	**9.64**	**0.0**	< 1001, 1500 >	1.128
< 1001, 1500 >	1.128	**6.27**	**0.0**	< 1501, 2000 >	1.075
< 1501, 2000 >	1.075	2.7	0.0069	< 2001, 2500 >	1.039
< 2001, 2500 >	1.039	1.96	**0.0494**	< 2501, 3000 >	1.005
< 2501, 3000 >	1.005	0.56	0.5779	< 3001, 4000 >	0.994

< 3001, 4000 >	0.994	2.51	**0.012**	< 4001, 6500 >	0.939
< 4001, 6500 >	0.939	1.45	0.1469	< 6501, 9000 >	0.907
< 6501, 9000 >	0.907	1.33	0.1828	< 9001, 20000 >	0.879

In Figure 3 and 4, the intervals where the change of mean Lambda is significant are marked with black lines.

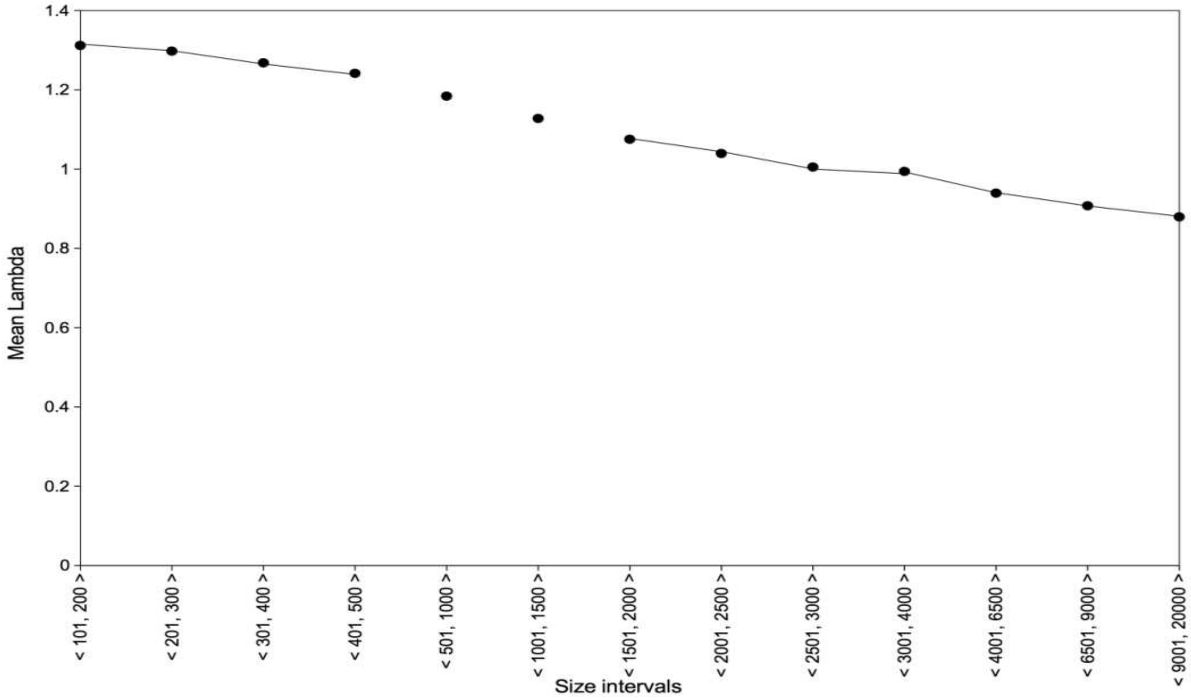

Figure 3. Significant difference ($U -$ test)

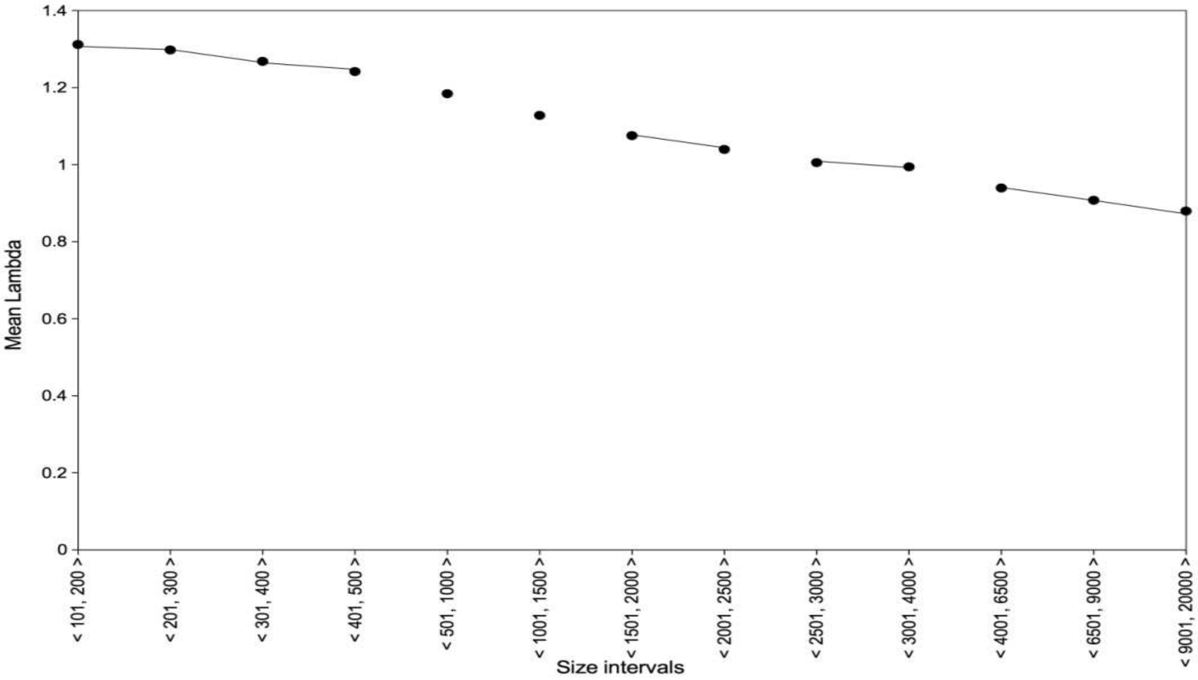

Figure 4. Significant difference (p-value)

The $U -$ test indicates significant differences between subsequent intervals in $N \in < 401, 2000 >$. The p-value gives the same intervals, but finds two more pairs of subsequent intervals $N \in < 2001, 2500 >$ and $N \in < 2501, 3000 >$; $N \in < 3001, 4000 >$ and $N \in < 4001, 6500 >$. These results are very different from what Čech (2015) got. In his study, the non-significant intervals are, for both languages, at the beginning and at the end of the distribution. This difference may be explained by genre, but we have no data to propose any further explanation.

3.2 Lambda and genres

As the conditions are not met to use the One-Way Anova, we computed the Brown-Forsythe test to verify whether Lambda was sensitive to different genres of text on two different sets of data. One is a corpus where the size of texts has been controlled, with $N \in < 600, 2600 >$. The other is a corpus constructed without control on the text size. According to the Brown-Forsythe test, the difference between the Lambda values of each genre for the two corpura is significant, as $p < 0.05$. Then we can compare the ranking given by the mean Lambda.

Table 9
Comparison of lambda means for both corpus

Ranking	**Size between 600 and 2600 tokens**			**No size limit**		
	Genre	***n***	**Mean Lambda**	**Genre**	***n***	**Mean Lambda**
1	Sanwen	29	1.608	Sanwen	64	1.542
2	Translation	7	1.515	Media	106	1.471
3	Xiaoshuo	19	1.467	Translation	20	1.388
4	Media	60	1.443	Scientific	30	1.366
5	Scientific	25	1.381	Child	148	1.218
6	Xiaopin	27	1.24	Xiaopin	34	1.212
7	Official	663	1.195	Xiaoshuo	177	1.211
8	Child	20	1.081	Official	859	1.196

Table 9 indicates that the rankings of genres given by Lambda based on the two corpura are different. In both, sanwen is placed on the first position, but translation goes from the second in *N* ∈ <600, 2600> corpus to the third in size non-controlled corpus, child from the last one to the fifth one, xiaoshuo from the third one to the last one. This proves that genre ranking given by Lambda is not absolute. Some genres' frequency structure may be more susceptible to variation than others, and the variation of text size may be a factor of this variation. The mean Lambda of official are quite the same : 1.196 and 1.195. On the other hand, the genre of xiaoshuo has a mean Lambda of 1.467 in one corpus, and 1.211 in the other. This finding shows how do the mean Lambda varies among different size intervals. Figure 5 represents the evolution of the mean lambda for each genre along different size intervals, from *N* = <0, 600> to *N* = <7601, 8600>.

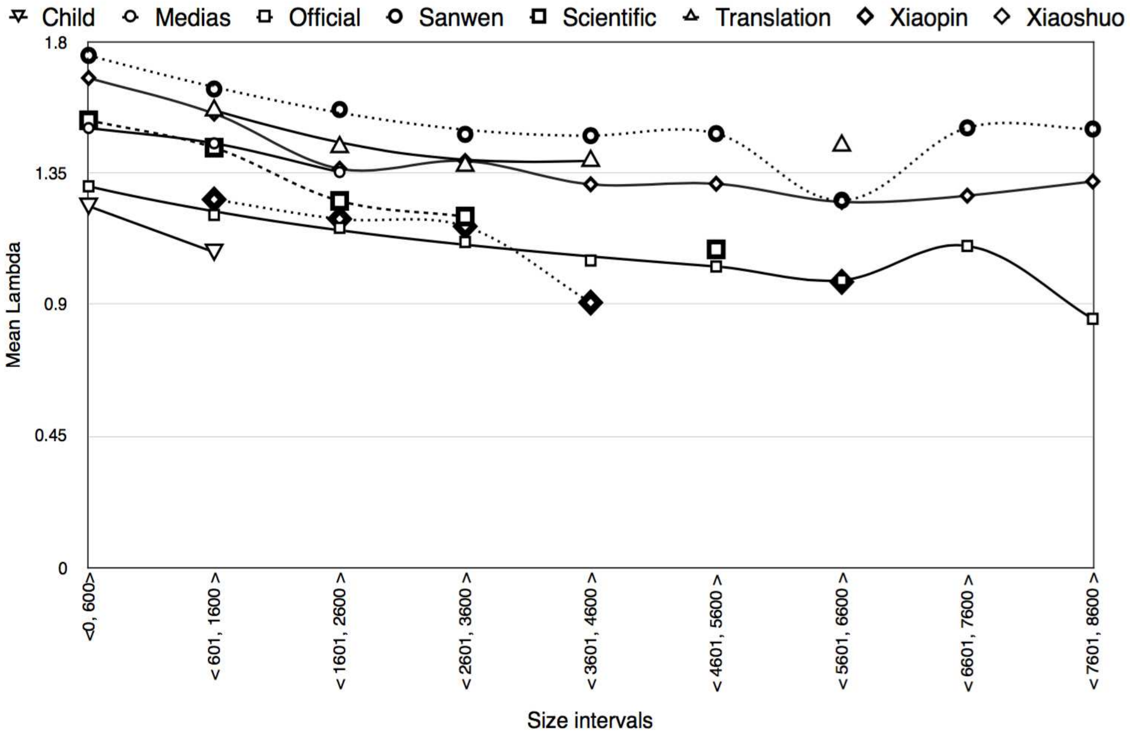

Figure 5. The evolution of mean lambda for each genre among different size intervals

The first remark we can make observing Figure 5 is that the variations of mean Lambda along different intervals do not take the same form for all the genres. Some genres are more regular than others. The line of official genre descends very regularly along seven subsequent intervals, where the line of xiaoshuo genre is obviously drawing a wave. However, the endpoint of the official genre line is still much lower than its starting one. This can be explained by the dependence of Lambda on text size demonstrated above.

The second remark is that Lambda may still be sensitive to the genres, but only for genres with less common characteristics, i.e. that are very different. We can observe that, in fact, the rankings of different genres overlap very often. In *N* ∈ <0, 600>, media ranking is under scientific, and then it goes over in the next interval. Xiaopin genre started over the official genre, and it finishes under. However, some genres have very different mean Lambda along different size intervals. There is no rank overlapping for child and sanwen, official and medias, official and scientific, official and xiaoshuo, official and translation, official and sanwen.

This finding may explain the results of Zhang & Liu (2015). The authors used Lambda to see whether the genre characteristics of modern Chinese novels since 1919 were significantly different, and reached the conclusion that they were not. It may be explained by the fact that Lambda is not sensitive to very slight variations of genres.

3.3 Lambda and the degree of analytism/synthetism

We tokenized the same text with 5 segmentation tools, and calculated their Lambda.

Table 10
The results of Lambda for the different text segmentation

Text	**Lambda**
Jieba	1.393
Stanford	1.297
FNLP	1.258
Segtag	1.025
PyNLP	1.015

Table 10 shows that Lambda results are really different, the lowest one is 1.015, given by PyNLP and the highest one, given by Jieba, is 1.393. A quantitative linguist using Lambda to work on Chinese text should be fully aware of how much the segmentation tool used may influence its results.

Table 11
Percentages of verbal complements and suffixes splitted from the main verb

Segmentation tool	**Lambda**	**Resultative**				**Aspect marker**		**Mean**
		Phase		**Directional**				
		到 **(*dào*)**	完 **(*wán*)**	下去 **(*xiàqù*)**	起来 **(*qǐlái*)**	过 **(*guò*)**	着 **(*zhe*)**	
Jieba	1.393	38.8%	18.2%	50.0%	95.8%	48.4%	72.2%	53.9%
Stanford	1.297	23.5%	9.1%	80.0%	95.8%	59.3%	65.7%	55.6%

FNLP	1.258	26.7%	54.6%	100%	100%	80.2%	96.8%	76.4%
Segtag	1.025	53.7%	100%	90%	94.4%	83.5%	92.6%	85.7%
PyNLP	1.015	58.4%	100%	90%	100%	84.6%	94.1%	87.9%

As one can see from Table 11, the ranking given by the percentages of mean follows exactly the one given by Lambda. In other words, the more the segmentation tool splits the grammaticalized elements from the main verb, the lower Lambda is. We can see that PyNLP seems to have an analytical approach to Chinese. It splits very often the resultatives from the main verb they follow. It also splits very often the verbal suffixes from the main verb. The file tokenized by PyNLP has the lowest Lambda, which indicates, according to Popescu, Čech & Altmann (2011), a tendency toward analytism. Jieba has a more synthetic approach. Only half of the 下去 (*xiàqù*) are split from the main verb they follow. The file tokenized by Jieba has the highest Lambda, indicating a tendency toward synthetism. These observations indicate that the method is promising to observe the capacity of Lambda to measure the degree of analytism/synthetism of a text. It also corroborates with the assumption that Lambda has this property, which is good news for Chinese linguistics that a quantitative formula could help to investigate into the inner workings of Chinese morphology. However, more work has still to be done in this direction. Firstly, our methodology could be improved. In particular in the way we have constructed our data. The element 起来 (*qǐlái*) can be overlapped :

他	喜欢	起	那	个	女孩	来	了。
tā	*xǐhuan*	*qǐ*	*nà*	*gè*	*nǚhái*	*lái*	*le.*
He	like	QI	this	GE	girl	LAI	LE

‘He started to like that girl’ (Chang, 1993)

In this example, 起 (*qǐ*) and 来 (*lái*) are separated by the object, of the love, 那个女孩 (*nà gè nǚhái*; ‘this girl’). But we only extracted 起来 (*qǐlái*) when the two characters 起 (*qǐ*) and 来 (*lái*) were adjacent. Another limitation of this experimental research is that we chose a small sample of six grammaticalized elements. Last but not least, we only worked on the verbal morphology, but the morphology of Chinese is not limited only to verbs. It can concern nouns, with for example the morpheme 们 (*men*), which indicates number :

朋友	们
péngyǒu	*men*
friend	MEN

‘Friends’

The Lambda of Chinese is very high for a so-called analytic language. According to our data, the Lambda of Chinese is often between 1.2 and 1.4, sometimes even approaching 1.6, rarely under 1. These results are not very far from the ones of synthetic language (Popescu, Čech & Altmann, 2011). However, Chinese is a language traditionally considered as analytic. Then, why ? a) It could be explained by the inability of Lambda to detect this kind of text property. But Popescu, Čech & Altmann (2011) demonstrated throughout a quite in-depth investigation that Lambda has this capacity, and our results corroborate with this assumption. b) It may be caused by the morphology strategy of the segmentation tool used. We found that, our tool seems to tend toward an analytic approach of Chinese. These two conclusions have still to be verified, but meanwhile new questions have to be raised about the reasons of this high Chinese Lambda:

1) Is it a matter of the definition of what we call « analytic » ?
2) Is it related to some properties of Lambda that have not been discovered yet ?
3) Or is it because Chinese is not that analytic ?

4. Conclusion

In this study, we investigated on the reliability of Lambda, and reached the following conclusions:

Lambda is dependent on text size. Čech (2015) already demonstrated this correlation, but the corpus he used was not totally satisfying. As Lambda is a measure of word frequency, it reacts to many different factors. In order to verify the independence/dependence of Lambda on text size, one should use a corpus that suffer as less as possible from variations in terms of authorship and vocabulary richnes. That is why we worked on a very specific subgenre and on one language. Our finding corroborates with the ones of Čezh (2015) : lambda is dependent on text size.

Since Lambda has been proposed (Popescu, Čech & Altmann, 2011), some studies used Lambda to work on the genre of texts. As Lambda is dependent on text size, we had to verify the reliability of this property again, paying attention to the size of texts. We found that Lambda was still able to differentiate the genre of texts, but only for genres that are obviously different, like child stories and medias. Lambda would not be sensitive enough to variations among subgenres.

The method we proposed to investigate on the property of Lambda to detect the degree of analytism/synthetism of a text is promising. We used one Chinese text, tokenized with different segmentation tools. It seems that the more the morphological strategy adopted tends toward an analytic approach of Chinese language, the lowest Lambda is. And a low Lambda indicates that the text tends toward analytism. This finding corroborates with the assumption that Lambda can detect the degree of analytism/synthetism of a text.

Acknowledgements

This study was partly supported by the Fundamental Research Funds for the Central Universities (Program of Big Data PLUS Language Universals and Cognition, Zhejiang University) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

Baayen, H., Van Halteren, H., Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, vol. 11, no 3, p. 121-132.*

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In: Baker, et al.(eds.) *Text and Technology*, Amsterdam/Philadelphia: John Benjamins, p. 233-250.

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator. Target, vol. 12, no 2, p. 241-266.

Chang, S. M. (1993). V+qi(-lai) compounds in mandarin Chinese. *Computational Linguistics Society of R.o.c.*

Che, D. (2014). *The syntax of particles in Mandarin Chinese*. (Doctoral dissertation, The University of Hong Kong (Pokfulam, Hong Kong)).

Čech, R. (2015). Text length and the lambda frequency structure of a text. In: Mikros, G. K., Mačutek, J. (eds.) *Sequences in Language and Text*. Berlin: De Gruyter, 71-88.

Chen, R., Liu, H. (2015). Ideologies of Supreme Court Justices: Quantitative Thematic Analysis of Multiple Opinions of “Bush v. Gore 2000”. *Glottotheory, vol. 6, no 2, p. 299-322*.

Dubois, J. et al. (1973), *Dictionnaire de Linguistique et des Sciences du Langage*, Paris: Larousse.

Évrard, É. (1964). Les mystères des vocables. *Bull. Amis de l'Université de Liège, vol. 36, p. 33-55.*

Fang, Y., Liu, H. (2015). Comparison of vocabulary richness in two translated hongloumeng. *Glottometrics 32, 54-75.*

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and linguistic computing, vol. 13, no 3, p. 111-117.*

Hoover, David L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities, vol. 37, no 2, p. 151-178.*

Huang, S. M., Ching, S., & Yu, H. (2008). Grammaticalization of directional complements in Mandarin Chinese. 語言暨語言學, 9.

Hubert, P., Labbé, D. (1988). Un modèle de partition du vocabulaire. *Études sur la richesse et la structures lexicales, p. 93-114.*

Jamak, A., Savatić, A., Can, M. (2012). Principal component analysis for authorship attribution. *Business Systems Research, vol. 3, no 2, p. 49-56.*

Kubát, M., Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics, vol. 20, no 4, p. 339-349.*

Li, C. N., Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Li, F. (2001) Origine et évolution du complément directionnel complexe en chinois. *Cahiers de linguistique-Asie orientale, vol. 30, no 2, p. 179-214.*

Peyraube, A. (2006). Motion events in Chinese. *Space in languages: Linguistic systems and cognitive categories, vol. 66, p. 121-135.*

Popescu, I-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text,* Berlin-New York: Mouton de Gruyter, p. 555-565.

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

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

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

Rousseau, A. (2001). Réflexions sur les catégories des unités linguistiques: comparaison de l’allemand et du chinois. *Linx. Revue des linguistes de l’université Paris X Nanterre, vol. 45, p. 59-70.*

Sun, C. (2013). Chinese Resultative Verb Compounds: Lexicalization and Grammaticalization. *Breaking Down the Barriers, p. 625-649*

Taine-Cheikh, C. (2002). À propos de l'opposition “type synthétique” vs “type analytique” en arabe. In: *4th Conference of the International Arabic Dialectology Association (AIDA) p. 234-243*

Thoiron, P. (1986). Diversity index and entropy as measures of lexical richness. *Computers and the Humanities, vol. 20, no 3, p. 197-202.*

Tweedie, F., Baayen, H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, vol. 32, no 5, p. 323-352*.

Zhang, C., Liu, Haitao. (2015). A quantitative investigation of the genre development of modern Chinese novels. *Glottometrics,* vol 32, p.9-20.

Book Reviews

Kelih, Emmerich, *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax.* Frankfurt am Main: Peter Lang Verlag 2016, 272 pages.
Reviewed by **Gabriel Altmann**

The book reviewed is, as a matter of fact, a good introduction into the methodological problems of quantitative linguistics. The author searches for the links between the size of phoneme inventory and other properties of language. Since the number of language properties is practically infinite, he restricts the questioning to phonological (also suprasegmental), syllabic, morphological, morpho-syntactic, lexical-semantic domains and typology. He correctly mentions that the definition and quantification of properties is not "given" but constructed by us. Every linguist has his own methods or adheres to a certain school.

The author's knowledge of literature, old and new, is enormous (the bibliography stretches across 35 pages). He connects classical linguistics with the new streams and focuses above all on synergetic linguistics, which represents the study of mutual dependencies in language. The book is illustrated with a number of links between phoneme inventory and other properties. Weak correlations are shown but mathematics is avoided in order to make the book readable to many linguists. The hypothesis that there is a link between size of the inventory and number of speakers is rejected.

The confirmation of any hypothesis in linguistics is a matter of degree. Whatever hypothesis is tested, one always finds exceptions which falsify the derivation. The author emphasizes the role of boundary conditions, which must – unfortunately – be searched for in every language. Hence any linguistic hypothesis presented in the book is a task for teams.

At the beginning of quantitative work one usually computes the correlation between two properties, but this is not the final aim. The author leans against the Köhlerian synergetics in which the requirements of speaker and hearer and their effects on some lawlike processes in language are taken into account. The book considers especially Slavic languages and examples from other ones taken from the abundant literature. Unfortunately, one can never say how many languages must be analyzed in order to obtain a valid law. The validation is not merely a problem of testing but also that of deriving the hypothesis from an existing background theory. However, mathematical models are no "truth" but only our trials to express a matter in a formal language which can further be processed and joined with other models.

In more progressed works one avoids linear relationships, whose existence in language is problematic, but especially the assumption that something in language is normally distributed. Since language is in eternal movement (development), the attractors hold the equilibrium which may be displayed by the difference in parameters of functions, but in every language a time comes when boundary conditions disturb something and the self-regulations re-establish the equilibrium without which no communication is possible.

The book highlights the fact that even the "lowest" level of language, namely the size of the phoneme inventory, is not isolated but must be inserted into the net of dependencies. A longer chapter is dedicated to the relation between inventory size and mean word length. Unfortunately, word length is measured in terms of phoneme numbers and not in that of its immediate phonetic components, namely syllables. Skipping a level leads rather to fractals, polynomials, Fourier series, etc. and is not in agreement with Menzerath's law. Nevertheless, this is just a kind of confirmation of Menzerath's law.

The book has many facets and is an excellent introduction to the problems of quantitative linguistics. It forces the linguist to take into account the methodology, which is a daily problem in natural sciences. It is a pity that it has been written in German but we hope that an English translation will appear soon.

Haitao Liu, Junying Liang (eds.) (2017). *Motifs in language and text.* Berlin/Boston: de Gruyter Mouton. 271 pp. (= Quantitative Linguistics vol. 71).
Reviewed by **Hanna Gnatchuk**

The reviewed book contains 13 articles describing a very abstract entity introduced into linguistics by R. Köhler inspired by the musicologist Boroda (1982). Only three articles are written by European authors, the other ones are written by Chinese linguists, a very important sign of the intensive development of quantitative linguistics in China. The articles are ordered alphabetically according to the family name of the first author. Today, the study of motifs of various kinds is a very promising object because it enables us not only to use models applied to other units but show a higher level of language.

The first article (**A.P. Beyer, Persistency of higher order motifs: 1-12)** performs a syllable count per word and evaluates DNA sequences up to 10-th order, that means, up to very abstract entities. The motifs are evaluated by Shannon's entropy and the Hurst indicator. Unfortunately there are no formulas, thus the reader cannot check everything. The qualitative motifs of DNA are transformed in quantitative ones. It is not clear whether the corpus and the database were taken as wholes, i.e. as mixed samples, or each text and species separately.

In the second article (**R. Čech, V. Vincze, G. Altmann, On motifs and verb valency: 13-36**) the authors study the full valency of verbs in Czech and Hungarian sentences, express them numerically and construct the motifs. For the rank-frequency relation of motifs the Zipf-Mandelbrot distribution is used; for the spectrum, the usual transformation of this distribution is used. The relation between length and frequency is slightly more complex, namely concave, hence the authors use the Lorentzian function. Nevertheless, the average length is monotone decreasing. In the Appendix one finds tables of all motifs.

The third article (**H. Chen, J. Liang, Chinese word length motif and its evolution: 37-64**) concerns word length in written and spoken Chinese, an extremely complex problem. They take into consideration 20 texts from talk shows and a journal (year 2013) respectively and measure the word length in terms of syllable numbers. They construct the word length distribution and fit to their rank-frequencies the power function. All results are displayed both in tables and in graphs. Then they construct length motifs and fit to their spectra the hyper-Pascal distribution. The article shows that both word lengths and their motifs change regularly in the history of Chinese. This fact is shown in the change of the parameters of the power function in 6 time periods. The time spans are shown in the Appendix. The authors show also the development of entropy of word length motifs. As far as known, any lengths in languages are captured today by means of the Zipf-Alekseev function (cf. Popescu, Best, Altmann 2014). Since the authors show all numbers, the reader can test the newer model.

In the article by **R. Chen (Quantitative text classification based on POS-motifs: 65-85)** the author analyses Chinese and English texts and computes for all the TTR, hapax proportions, Popescu's richness indicator, Entropy, Repeat rate and Gini's coefficient and tries to show that POS-motifs can be used for distinguishing text types (news, essays, official, academic and fiction texts). This is made by means of discriminant analysis whose results are presented in colored figures and tree forms separately for Chinese and English. The author expresses also warnings and shows which indicator may be used for the given classification. In any case, she is very critical and emphasizes the conventionality of definitions. This is especially important in qualitative motifs which may be constructed in different ways.

A further problem associated with motifs is the question whether they can be used for discriminating the authorship of texts. In: **L-motif TTR for authorship identification in Hongloumeng and its translation (87-108)** by **Yu Fang** the question is scrutinized whether the famous Chinese novel has been written by one or two authors, how a translator presents the two styles, how the parameters differ and measures the vocabulary richness. If one would analyze all old texts using motifs, evidently a new linguistic discipline would be created in which old qualitative questions would be tested quantitatively.

Perhaps the most complex problem is the definition of motives in the script. There are many possibilities out of which **Wei Huang** in **Length motives of words in traditional and simplified Chinese scripts (109-132**) shows one of them. The measurement of complexity is performed in the script, that is, in the secondary language. The author develops a method of counting the strokes and the components in a sign but not the kind of their mutual connection. That means, he has chosen one of the writing types which is an admitted method. If one would do the same for the dozens of Latin scripts used e.g. in WORD, one would obtain dozens of different results. Since the results of the author are clear, he constructs "length"-motif and evaluates 20 texts. He computes the types and tokens of motifs and applies to their ranks the power function. The spectra are captured by the Hyper-Poisson distribution. All numbers and figures are presented, there are even tables of the individual parameters. A method of measuring the simplification of Chinese sings into the Japanese katakana and hiragana can be found in Sanada, Altmann (2008).

A chapter on dependency grammar considering the dependency distance (DD) of individual words in the sentence is presented in **Yingqui Jing** and **Haitao Liu**: **Dependency distance motifs in 21 Indo-European languages (133-150).** Also here, the motifs are "higher" units taking into account the sentence construction. One prepares a graph of dependencies in the sentence and writes the complete DD-sequence from which the motifs can be stated mechanically. The authors analyze 21 languages and consider their article a further contribution to a possible typology of languages, this time from a syntactic point of view.

A further European cooperation of a Greek linguist and a Slovak mathematician can be found in **Mikros, G.K., Mačutek, J.: Word length distribution and text length: Two important factors influencing properties of word length motifs (151-163)** where the authors show how the kinds/types of motifs increase with increasing length of the text. They study an enormous number of Greek and Ukrainian texts and show figures displaying some type-token and type-text length relations. The formulas known from other domains could be applied in this domain, too. This is a further evidence of the fact that motifs are "legal" linguistic entities.

While in the previous article the relation of motifs to text length has been shown, in the next article, **Yaqin Wang: Quantitative genre analysis using linguistic motifs (165-180)** it is shown that L- and F-motifs can be used for distinguishing text types. The author analyzes texts concerning Applied science, Arts, Belief, Commerce, Imaginative texts, Leisure, Natural science, Social science and World affairs. He uses the Zipf-Mandelbrot law and compares the parameters in individual text types. He shows also that the parameter *b* depends on *a*. As is usual, one begins with English texts given by the BNC but for generalization one will be forced to study other languages, too. Besides, the continuation of this work requires a list of possible text types, a very complex task.

Motifs of parts of speech and syntactic dependencies are the object of the article **Jingqui Yan: The rank-frequency distribution of parts-of-speech motifs and dependency motifs in the deaf learners´ compositions (181-200).** It is very positive that somebody analyzes the texts written by deaf learners subdivided into three age-groups. The results seem to corroborate the "normal" results. However, there are some points that may be avoided in further processing this data. The author uses the Zipf-Mandelbrot function but one cannot find it in the article. Thus one does not know what are the parameters A, B (Table 4), b, a; why attributions and adjectives are abbreviated equally (like A); how the POS motifs have been won (there are several ways) and since the basic numbers are not presented, the results cannot be checked or further used by readers. It would be very useful if all numbers could be presented in a separate article. The aim of the author: to show the linguistic maturity of the writer is an important aspect of language both practically and theoretically.

Since motifs may be constructed at any level, **Jiang Yang** in **Quantitative properties of polysemy motifs in Chinese and English (201-216)** analyzes polysemy data in two languages. To each word the degree of its polysemy is ascribed and the sequence is rewritten in terms of motifs. The author states that the rank-frequency ordering follows the Zipf-Mandelbrot distribution. There are problems with the rank-frequency distribution of motif lengths which follows the mixed negative binomial distribution in both languages. The author explains the necessity of applying a mixed model by two-fold abstraction but there will be problems with the subsumption of such a model in Köhler's synergetic model. Further, he explains some differences by the dynamic character of English words and the conservative behavior of Chinese ones; and by the stronger context-dependence of Chinese words. This is a god beginning but these properties must be first quantified – a task for future research.

The only article concerning the frequency motifs is **The words and F-motifs in the modern Chinese version of the *Gospel of Mark* (217-229)** by **Cong Zhang.** Usually, quantitative linguists avoid the analysis of religious texts but in this case it was a correct decision because the author compared the six versions of the Gospel of Mark created between 1855 and 2010 showing thereby how "holy" texts change. A nice figure displaying the oscillation of frequencies shows why it is reasonable to consider F-motifs. The frequencies follow the power function with changing parameters. A special chapter is dedicated to the relation between word length and F-motifs and states that both length and the F-motifs change in the development of Chinese. Though there are some formulas and many numbers the authors seems to avoid "theorizing" which will be necessary in the future.

The last article by **Hongxin Zhang** and **Haitao Liu, Motifs of generalized valencies (231-260)** considers again valencies and the motifs constructed from the consecutive numbers. Again, Chinese and English are concerned, the authors used the Prague Czech-English Dependency Treebank and the Peking University Multi-view Chinese Treebank. They clearly formulate their three hypotheses, namely: 1. Are motifs of generalized valencies regularly distributed? 2. Are motif lengths regularly distributed? 3. What is the interrelation between motif length and length frequencies? For the first question the answer is the right truncated modified Zipf-Alekseev distribution. The same holds for the lengths. If one analyzes the frequency distribution whose independent variable is length, one obtains the Hyperpoisson distribution. Unfortunately, no formulas are presented, one already believes that all linguists know the formulas by heart. In a monumental appendix, the readers can see all necessary numbers and check the results.

The volume as a whole is an excellent survey of the problems concerning the motif, this modern entity existing in all domains of language. In the future, it would be good to perform complete projects with a more extensive view of motifs, i.e. comprising all levels of language, all known units, many of their quantified and measured properties, and above all, to extend the investigation to languages for which there are also old texts, in order to study the development. The present volume shows that linguistics can have other aspects than those developed by structuralists and generativists.

References

Boroda, M. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J., Boroda, M., Nadarejšvili, Š.I. (eds.), *Sprache, Texte, Kunst. Quantitative Analysen: 231-262*. Bochum: Brockmeyer.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language*. Lüdenscheid: RAM-Verlag.

Sanada, H., Altmann, G. (2008). On two simplifications of the Japanese writing systems. In: Altmann, G., Zadorozhna, I., Matskulyak, Y. (eds.), *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th Anniversary of Professor V. Levickij: 472-480*. Chernivcy: Books.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

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

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4.* 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.