Glottometrics 33

RAM-Verlag
2016

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

Prof. Dr. Haruko Sanada
Rissho University,Tokyo, Japan (http://www.ris.ac.jp/en/);
Link to Prof. Dr. Sanada: http://researchmap.jp/read0128740/?lang=english;
mailto:hsanada@ris.ac.jp
Prof. Dr.Thorsten Roelcke
TU Berlin, Berlin, Germany (http://www.tu-berlin.de/)
Link to Prof. Dr.Roelcke: http://www.daf.tu-berlin.de/menue/deutsch_als_fremd-_und_fachsprache/personal/professoren_und_pds/prof_dr_thorsten_roelcke/
mailto:Thosten Roellcke (roelcke@tu-berlin.de)

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: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. –33 (2016). – Lüdenscheid: RAM-Verlag, 2016
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.
Bibliographische Deskription nach 33 (2016)
ISSN 1617-8351

Contents

History

Book reviews

Glottometrics 33, 2015, 1-7

A Quantitative Analysis of English Compounds in Scientific Texts

Hanna Gnatchuk (University of Trier)[1]

Abstract. The given research focuses on a statistical analysis of English compounds in the scientific texts with a special emphasis on the parts of speech and the cohesion of the constituents. In order to conduct the research in question, we have analysed the books "The Power of Management Capital" (2008) which belongs to the Exact Science and "Tort Law" (2008) which concerns the Humanities. We have analysed each tenth page of the above-mentioned books. The treatment of the data was done with the help of statistical methods.

Key words: *English, scientific prose style, compounds, cohesion, statistical methods*

1. Introduction

According to Chris Baldick (1996), "stylistics can be defined as a branch of modern linguistics devoted to the detailed analysis of literary style or of the linguistic choices made by speakers or writers in non-literary contexts". I. Galperin (1981) outlines two tasks of stylistics. The first task consists in studying stylistic devices or expressive means (*phonetic stylistic devices*: assonance, onomatopoeia, alliteration; *lexical stylistic devices*: metaphors, personifications, metonymies, ironies, zeugmas, etc; *syntactic stylistic devices*: represented speeches, rhetorical questions, elliptical sentences, etc). The second task concerns the types of texts which are distinguished by the pragmatic side of the communication. These types are called functional styles. In such a way, Galperin (1981) distinguishes 5 functional styles of the English language: belles-lettres, publicistic, newspaper, scientific styles and the style of official documents. The focus of our attention is on the scientific prose style.

Scientific style originated from the essay. Gradually this style began getting rid of the apriority which was available in the essay by acquiring more logical organization of the information. The most characteristic features of scientific prose style are a syntactic structure of sentences and the choice of the lexemes. As far as the selection of the words is concerned, the scientific style takes into account its main task: to present the analysed phenomenon adequately and more precisely. Therefore, the words here have only one meaning. It is difficult to find the lexemes with metaphorical or other contextual meanings. Metaphors, metonymies, hyperboles, comparisons and other means are hardly to be found. In this case, terminology makes up the basis of the scientific style. Sometimes the most frequent words may become the terms due to their peculiar usage in the scientific work.

Another feature of the scientific prose style is the coinage of neologisms. It is the only style which provides the favourable conditions for the neologisms. The new notions require new words in order to designate themselves. It is possible to find here the frequent cases of affixations and conversions with the purpose of building new words. In such a way, the scientific style remains the main source for new words, word combinations and new meanings of the existing words.

[1] Address correspondence to: Dept. of English and American Studies, Alpen-Adria Universität, 65-67, 9020 Klagenfurt, Austria. Email address: agnatchuk@gmail.com

Dealing with a syntactic structure of a sentence, it is possible to reveal a system of conjunctions. Their usage is aimed at transferring a logical sequence of the information. Moreover, our attention should be drawn to such a process as de-semantization of such words as *consequence, connection, results, in connection with, in consequence, as a result* which took place at the earlier stages of the development of the analysed style. But the system of the conjunctions is not the only means of expressing a logical connection of separate parts. In this case, participial and infinitive constructions play a significant role here.

The division of speech into the paragraphs is very strict. The logical organization of paragraphs finds its reflection in this style to the higher extent. Each paragraph is intended to prolong the idea of the previous one. It is possible to separate the basic idea in each paragraph. In such a way, the completion of the idea can always be found here.

Distinguishing the main point out of a mass of facts is characteristic of this style. It can be achieved by means of syntactic as well as logical principles: the main idea is to be found in the principal clause, the subordinate – in the subordinate one. The additional information is separated by a dash.

It is worth mentioning that the system of the usage of the conjunctions was used in the earlier periods of this style quite differently. In particular, the authors of the scientific treatises were intended to reveal the interconnection, interdependence of the observed facts. That led to the unprofessional usage of the conjunctions which gave a rise to long paragraphs. In the process of the development and mastering the norms of the English language, the scientific style started deviating from the norms of the earlier established periods. In such a way, the scientific prose style reacted to the change of the literal norms which could also be found at lexical and phraseological levels. A considerable number of terms and their combinations were de-etymologized by enriching the literary language.

The abstract nature is appropriate to the word-stock of the scientific style. It is clear that this style is aimed at treating the environment facts. Therefore, it is necessary for the words to express the general features of the subjects in question. Furthermore, it is relevant to mention that less exact notions can be found in the Humanities in comparison with technical and natural sciences which deal with special formulas. Finally it is worth saying that the bookish words are quite prolific in the scientific style. The reason for the usage of these lexemes is that one searches for an adequate expression of a new idea in the process of exploring the facts.

Moreover, it is necessary to have a brisk look at phonetic, morphological, lexical and syntactic peculiarities of the scientific language:

Phonetic level

On the whole, phonetic level does not play a key role in the scientific texts. Nevertheless, it is impossible to neglect such phonetic features available in the scientific style as the gradual slowness of the tempo of words or the prolonged pauses on the notional word combinations (aimed at giving a better understanding of the content). These features make up the basis of phonetic aspect of the scientific text. In such a way, it is relevant to summarize all phonetic features in the following way:

a) The subordination of intonation according to the syntactic structure of a scientific language;
b) Standard character of intonation;
c) Stable character of rhythm;
d) Gradual slowness of the tempo.

Lexical level

The most peculiar feature of the given style is the abundance of the terms. The number of terms used in the scientific texts is not the same in the other styles. Moreover, the correct and logical definition of the notions is the necessary condition for the scientific language. Otherwise, the incorrect usage of the term is capable of misunderstanding the reader.

Morphological level

Abstract character of a scientific style can be found at the grammatical level, namely in the choice of a word's form and the structure of word combinations and sentences.

Syntactic level

The accurate structure of sentences is appropriate to the syntax of the scientific style due to the logical organization of the information. The most important feature is the predominance of complex sentences with different extended subordinations. Moreover, special attention is paid to the number of impersonal sentences. The experiments are usually described with the help of participles. The action of mechanisms (in the technical texts) is explained by means of passive constructions. The usage of such syntactical constructions is aimed at concentrating the reader's attention on the action or process.

Koyalan and Mumford (2011) emphasized the fact that writing the articles in English for non-English native speakers remains quite problematic. In this case they often ask for the help. D. Biber, B.Grey (2011) and other linguists pay attention to the differences between the colloquial and writing styles. They showed the advantage of the usage of the compressed constructions (i.e. nominal phrases). I. Martinez (2011) deals with the impersonal sentences in the scientific articles whereas B. Grey, V. Cortes (2011), M. Halliday (1976), are engaged with the ways of the cohesion of a text. Nevertheless, it would be relevant to enumerate the integral peculiarities of the scientific style:

- Logical sequence of the information
- Coherence
- Cohesion
- Abstractness
- Accuracy
- Objectivity
- Formality
- Information saturation

Taking into account the number of the peculiar features for the texts of a scientific style, we are intended to conduct the analysis of English compounds in the style under consideration.

2. A statistical analysis of English compounds according to the parts of speech

The purpose of the research consists in detecting the most frequent patterns of parts of speech for English compounds in the texts of a scientific style.

The data for analysis consist of two books: *Tort Law* (2008) which belongs to the Humanities and *The Power of Management Capital* (2008) which concerns the Exact Science.

The procedure of the analysis consists in counting the parts of speech within the text of a scientific style. Each tenth page was under analysis. In such a way, all patterns of English compounds have been written out. As a result, we have received the following 24 types of English compounds:

1) Noun + Noun: *business growth, cycle time, labour law, property rights, blood transfusion, stock prices, property damage, pregnancy certificate, consultation paper, flight controllers, law-duty, community law, management capital, market leadership, product development, etc*;
2) Noun + Noun + Noun: *charter flight business, balance sheet items, football league, business information management, delivery cycle time, supply chain networks, customer delivery times*, etc;
3) Adjective + Noun: *mandatory regulations, public authority, high-term, civil liability, monetary value, monetary compensation, supervisory authority, blue-chip, statutory powers, real estate, etc*;
4) Noun + Verb (ing): *cost-accounting, decision-making, danger-increasing, life-threatening, deposit-taking*;
5) Verb + Preposition: *breakthrough, return-on, carryovers, take-over, break-up*;
6) Noun + Participle 2: *staff-led, cost-driven, time-integrated, oil-fired*;
7) Verb (ing) + Noun: *dwelling-convection, trading-practices, breeding-value, starting-point*;
8) Phrases: *Court of Appeal Judgement, steam of leadership, quality-of-management discipline, brick-and-mortar*;
9) Adverb + Participle 2: *well-known, well-established, well-entrenched*;
10) Numeral + Noun + Noun*: twenty-first-century management, twenty-first-century business, twenty-first-century operation*;
11) Noun + Adjective: *businesswide, organizationwide, sundry;*
12) Preposition + Verb (ing): *overriding, ongoing, overarching*;
13) Numeral + Noun: *first-mover, first-leg, second-class;*
14) Noun + Noun + Noun + Noun: *management capital framework, information technology performance areas;*
15) Verb (ing) + Preposition: *passing-off, running-down;*
16) Preposition + Participle 2: *above-mentioned, out-moded;*
17) Noun + Participle 2: *fact-based, business-related*;
18) Noun + Preposition + Noun: *case-by-case, situation-by-situation*;
19) Preposition + Noun: *aftereffects, underfoot*;
20) Noun + Preposition: *start-up*
21) Adjective + Participle 2: *heavy-handed, deep-seated*;
22) Adverb + Preposition: *thereon*;
23) Adjective + Verb (ing): *wide-ranging*;
24) Preposition + Preposition*: throughout*

At this stage of the research we present the results in Table 1 where rank-frequency distribution, the general number of English compounds and their patterns are given. Though the number of classes is very large and the tail of the distribution is too long, one may capture the trend using the Zipfian power function with an additive constant 1, i.e. $y = 1 + ax^{-b}$. The computed values $y = 1 + 170.7951x^{-2.4537}$ yielding $R^2 = 0.9985$ are presented in the last column of Table 1.

Table 1
Rank-frequency distribution and the general frequency of English compounds in the texts of scientific style

Rank	Pattern	Number	Computed
1	Noun + Noun	172	171.79
2	Noun+Noun+Noun	30	32.17
3	Adjective + Noun	15	12.52
4	Noun + Verb (ing)	5	6.69
5	Verb + Preposition	5	4.29
6	Noun + Participle 2	4	3.10
7	Verb (ing) + Noun	4	2.44
8	Phrases	4	2.03
9	Adverb + Participle 2	4	1.77
10	Numeral + Noun + Noun	3	1.60
11	Noun + Adjective	3	1.47
12	Preposition + Verb (ing)	3	1.38
13	Numeral + Noun	3	1.31
14	Noun + Noun + Noun +Noun	2	1.26
15	Verb (ing) + Preposition	2	1.22
16	Preposition + Participle 2	2	1.18
17	Noun + Participle 2	2	1.16
18	Noun + Preposition + Noun	2	1.14
19	Preposition + Noun	2	1.12
20	Noun + Preposition	2	1.10
21	Adjective + Participle 2	2	1.09
22	Adverb + Preposition	1	1.08
23	Adjective + Verb (ing)	1	1.07
24	Preposition + Preposition	1	1.07
	Total	274	

In such a way, it is possible to summarize the following results in two points:

- We have detected 24 types of English compounds in the text of a scientific style. The analysis of the prose texts (which was earlier undertaken by the author) has shown that 18 types of the compounds are available in the novels. Here it is possible to suppose that the structure of English compounds is more prolific within the scientific texts. It can be explained by the fact that one needs specifications of meaning. This is considered to be one of the factors influencing the language evolution.
- Judging from the structure of English compounds, it is possible to detect the highest frequency of Noun + Noun pattern in the scientific style.

3. Statistical investigation of the types of cohesion in the scientific texts

According to Fan/Altmann (2007: 190), “cohesion is a property present at all language levels”. In order to analyse the cohesion, scientific texts are selected as a domain where

"microscopic observation of the English compounds" was conducted and then "establish a scale for their cohesion" (Fan/Altmann 2007 : 190).

Therefore, *the aim of the given analysis* is to scale the cohesive types of English compounds in the scientific texts: "The power of Management Capital" (2008) and "Tort Law".

The procedure of the research. All the English compounds have been written out in the books in question. Then we have classified the compounds according to the types of cohesion. We have received the following results:

Blank type (separate writing of the compounds): *consultation paper, times supply chain networks, business information management*, etc;

Within the blank type, we have distinguished the following subtypes:

a) Blank with a preposition: *Quality of Management Discipline, Court of Appeal Judgement, Steam of Leadership*;
b) Blank with a joining element: *communications products, sales growth, operations effectiveness, operations relationship*.

Hyphenized compounds (the hyphen unites the units): *return-on, take-over, high-risk, long-term, blue-chip, fact-based, decision-making, well-informed, above-mentioned, hands-on, air-traffic, wide-ranging.*

Hyphenized compounds with a preposition: *case-by-case, situation-by-situation;*

Joining (the units have a joining element): *throughout, aftereffects, underfoot, marshland, airspace, airport, framework, carryovers, ongoing, overarching.*

Joining with an inserting element: *groundswells*

In such a way, it would be relevant to give the results in the form of the table with the rank-frequency distribution, the cohesion types, and the computed numbers. Instead of applying a distribution, we simply use a function (i.e. a not normalized model) and can state that the usual Zipfian rank-frequency function is quite adequate.

Table 2
The rank-frequency distribution of the cohesion of the compounds
in the texts of scientific style

Rank	**Name**	**Frequency**	**Computed**
1	Blank	159	161.59
2	Hyphenization	65	53.90
3	Joining	39	28.36
4	Blank with a joining element	4	17.98
5	Blank with a preposition	3	12.63
6	Hyphenization with a preposition	3	9.46
7	Joining with an inserting element	1	7.41

Here, again, the rank-frequency sequence can be captured by the power function: $y = 161.5888x^{-1.5839}$ with $R^2 = 0.9798$, yielding a very satisfactory result.

Table 2 has shown that 7 cohesive types of English compounds are available in the scientific texts: blank (58.0 %), hyphenization (23.7 %), joining (14.2 %), blank with a joining element (1.5 %), blank with a preposition (1.1 %), hyphenization with a preposition (1.1 %), joining with an inserting element (0.4 %). The blank type of cohesion is highly frequent in the scientific style. Here the comparison should be drawn between the scientific

and prose texts (the last has been undertaken by the author in a separate article). In contrast to the scientific style, the joining type is quite prolific in the prose texts.

On the whole, it is relevant to make the following conclusions on the basis of the above-mentioned analysis:

- the cohesion of English compounds differs in two styles under analysis (prose and scientific), namely, blank cohesion is observed in scientific style according to the highest frequency (though this type is quite rare in the text of prose style); hyphenized and joining types predominate in the prose style;
- scientific style contains blank, hyphenized and joining compounds with an inserting element and preposition. On the contrary, we have found only joining with an inserted element in the prose style. It shows that scientific texts possess all the types of compounds with various inserting elements.

Leaning against the results we may conjecture that in English scientific texts there is an expressed tendency to apply a specific kind of compounding. In order to corroborate the results, the study must be continued and extended not only to other scientific texts but also to other text sorts. Finally, in order to generalize the result, the same aspects must be scrutinized in other languages. The Zipfian function may remain as it is but there will be differences in the parameters; further, deviations may be discovered in strongly synthetic or extremely analytic languages hence the results could be useful also for typological research.

References

Biber, D., Gray, B. (2011). Grammatical change in the noun phrase: the influence of the written language use. *English language and linguistics 15 (2),* 223-250.

Baldick, C. (2008). *Oxford Concise Dictionary of Literary Terms*. OUP Oxford, 384.

Fan, F., Altmann. G. (2007). Measuring the cohesion of compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of Typological and Quantitative Lexicology:* 190-209. Černovcy: RUTA.

Feigenbaum, Armand (2008). *The Power of Management Capital*. McGraw-Hill Professional, 218.

Galperin, I. (1981). *Stylistics.* Moscow: Vyshhaja shkola, 316.

Gray, B. Cortes, V. (2011). Perception vs. evidence. An analysis of this and these in academic prose. *English for specific purposes 30,* 31-43.

Halliday, M. A. K., Hasan, V. (1976). *Cohesion in English*. London: Longman.

Koyalon, A, Mumford, S. (2011). Changes to English as an Additional Language writers' research articles: from spoken to written register. *English for Specific Purposes 30,* 113-123.

Martinez, Illiana A. (2011). Impersonality in the research article as revealed by analysis of the transitivity structure. *English for Specific Purposes 20,* 227-247.

Glottometrics 33, 2015, 8-24

Quantitative Aspects of RST Rhetorical Relations across Individual Levels*

Hongxin Zhang[1], Haitao Liu[1, 2]

1. Department of Linguistics, Zhejiang University, Hangzhou, China.
2. Ningbo Institute of Technology, Zhejiang University, Ningbo, China.

Abstract. This study converts each tree in the RST (Rhetorical Structure Theory) Discourse Treebank into three trees with mere ultimate nodes of clauses, sentences and paragraphs respectively, examines the rank-frequency distribution of rhetorical relations along three taxonomies at the three granularity levels and finds they all abide by a right truncated modified Zipf-Alekseev distribution. It justifies considering rhetorical relations as a result of a diversification process and verifies the taxonomies in the corpus.

Keywords*:* *Rhetorical Structure Theory (RST), discourse treebank, rhetorical relations, distribution, diversification process*

1. Introduction

Among the various approaches to discourse analysis (Moore & Wiemer-Hastings, 2003; Taboada & Mann, 2006a, 2006b), Rhetorical Structure Theory (RST) (Mann & Thompson, 1987, 1988) is among the very few methods addressing both hierarchical and relational aspects of text structures and is recognized as the most employed discourse-structural analysis. Despite its name, it is not a theory but a method and a notational convention. This language-independent formalism is functional, addressing text organization through distinctively labelled rhetorical relations (also known as coherence relations and discourse relations) holding between text components, and explicating coherence by postulating a text as a hierarchically connected structure (Mann & Thompson 1988; Taboada & Mann, 2006a). Taboada and Mann (2006a, 2006b) provide overviews of RST.

Figure 1 presents a typical RST tree, clearly illustrating both hierarchical and relational dimensions of RST. The leaves of the tree, or elementary discourse units (EDUs) are minimal spans (here in this tree, clauses or equivalent units), which can be aggregated into bigger spans (e.g. 1-7, 2-3) through the joint of rhetorical relations. Most RST relations are asymmetrical, pointing from satellites (additional information dependent on the nucleus/nuclei) to nuclei (the salient part). In symmetrical relations (e.g. *Sequence*, *Contrast*), multi-nuclear spans are regarded as having an equal status.

* Address correspondence to: Haitao Liu, Department of Linguistics, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com

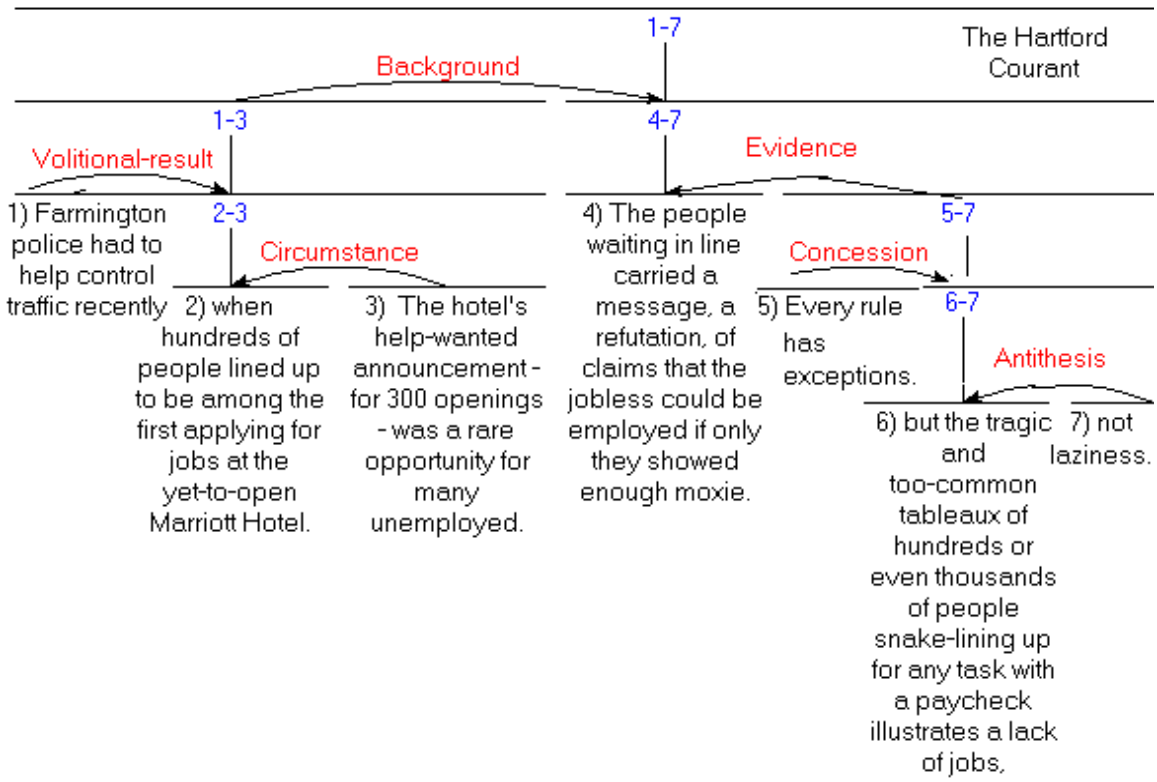


Figure 1. Diagram of an RST analysis: excerpt from *The Hartford Courant* (*source:* http://www.sfu.ca/rst/images/notlaziness.gif)

Quite a number of studies contribute to the quantitative investigation on the distribution patterns of RST relations.

Both Williams and Reiter (2003) and Carlson and Marcu (2001) notice the unequal distributions of RST relations: some are more frequent than others; some are more likely to occur at lower layers while others tend to be present at higher layers in rhetorical structure trees.

Motivated by the wish to investigate the probability distribution of discourse relations, Yue and Liu (2011) randomly choose 20 texts from a Chinese RST-annotated corpus (Yue & Feng, 2005), and find that the relations in them all abide by a right truncated modified Zipf-Alekseev distribution pattern. Their study justifies considering rhetorical relations as a result of a diversification process (Altmann, 1991; Altmann, 2005).

With the aim to examine the syntagmatic dimension of argumentation elements, Beliankou et al. (Beliankou, Köhler & Naumann, 2012) choose the Postdam Commentary Corpus (Stede, 2004), and look into the quantitative properties of motifs of RST relations in the corpus and the lengths of these motifs. They examine both R-motifs (uninterrupted sequences of unrepeated elements) and D-motifs (differing from R-motifs in that they follow a depth-first path and end with the end of a path). These motifs follow the hyperbinomial distribution and the mixed negative binomial distribution, respectively, which are linguistically interpreted as consequences of a diversification process, and a combination of two diversification processes, respectively.

These studies examine rhetorical relations between both EDUs and non-elementary expanded spans. Li et al. (Li, Wang, Cao, & Li, 2014) examine the distribution of RST re-

lations between mere terminal clausal nodes of EDUs (not necessarily independent clauses) in trees converted from the RST Discourse Treebank (RST-DT) (Carlson, Marcu, & Okurowski, 2002, 2003) which comprises of RST-annotated texts from *the Wall Street Journal*. They employ graph-based dependency parsing techniques along with two algorithms to convert the graphs into new dependency graphs with mere ultimate nodes of clauses.

Amid competing hypotheses about what constitutes EDUs, researchers agree that they shall be "non-overlapping spans of text" (Carlson & Marcu, 2001:2). Mann and Thompson (1988) argue that the unit size for RST analysis can be arbitrary, including paragraphs or even chapters. Taboada and Mann (2006a) also suggest granularity levels in accordance with the aims of the analyses. Despite the previous facts, attempts with large units were not so successful (Marcu, Carlson, & Watanabe, 2000).

Zhang and Liu (forthcoming) extend EDUs beyond sentences. Drawing on an analogy between syntactic and discourse trees and operating in line with the hierarchy principle of RST and its compositionality criterion, they convert each tree in the RST-DT into three trees with mere ultimate nodes being clauses, sentences and paragraphs, respectively. They examine the motifs of rhetorical relations along three taxonomies at the three granularity levels and also lengths of these motifs, and find these properties abide by the negative binomial distribution and positive negative binomial distribution, respectively. Their study demonstrates the applicability of RST analysis between same-level terminal units, including beyond-sentence level units. The newly-constructed discourse dependency trees boast unique analytical advantages and provide new research prospects.

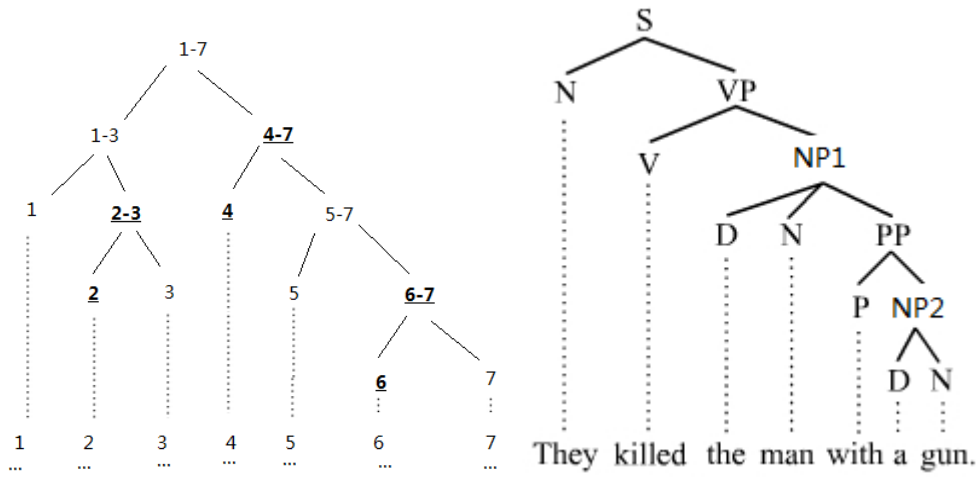
This work is a follow-up study of the previous one. It aims to examine the rank-frequency distribution of rhetorical relations *per se* along the same three taxonomies at the same three granularity levels, to check whether the rhetorical relations in the newly constructed discourse trees are a result of a diversification process. We posit our data also fit the right truncated modified Zipf-Alekseev distribution pattern in Yue and Liu (2011).

> *Hypothesis 1: RST relations at various levels abide by a common right truncated modified Zipf-Alekseev distribution.*

In the remainder of this paper, Section 2 presents methods of examining relations between mere terminal discourse units at distinct levels, detailing the tree conversion, taxonomies and granularity levels. Section 3 discusses the research findings and the last section addresses conclusions and proposals for future work.

2. Materials and method

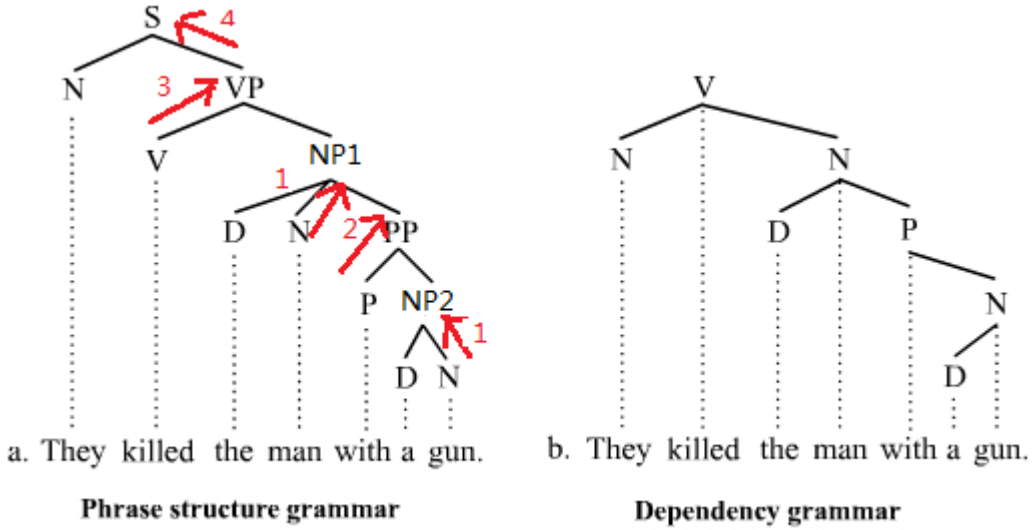
Figure 2(a) is an equivalent presentation of Figure 1, resembling in quite a number of ways a constituent syntactic tree like Figure 2(b). This analogy is the starting point for the tree conversion.



a. An equivalent representation of Figure 1 b. A constituent syntactic tree

Figure 2. An analogy between RST trees and phrase structure syntactic trees
(Source of *b*: *https://en.wikipedia.org/wiki/Constituent_(linguistics)*)

Figure 3(a) illustrates how to convert a constituent syntactic tree into a dependency one like Figure 3(b) (Liu, 2009a, 2009b).



(Source: *https://en.wikipedia.org/wiki/Constituent_(linguistics)*)
Figure 3. Transforming a constituent structure into a dependency structure

Borrowing this practice of promoting the more salient node to the top of the sub-tree along the vein, we follow the steps in Figure 4 and convert the sample discourse tree into Figure 5.

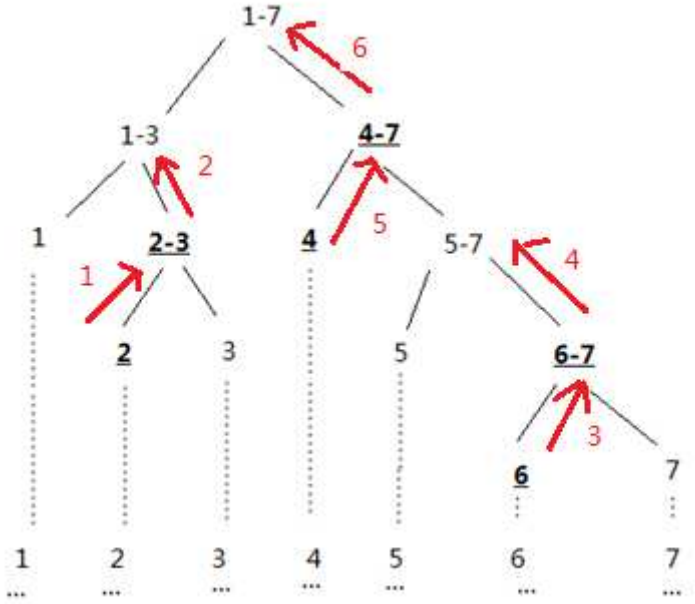


Figure 4. Steps of converting an RST tree

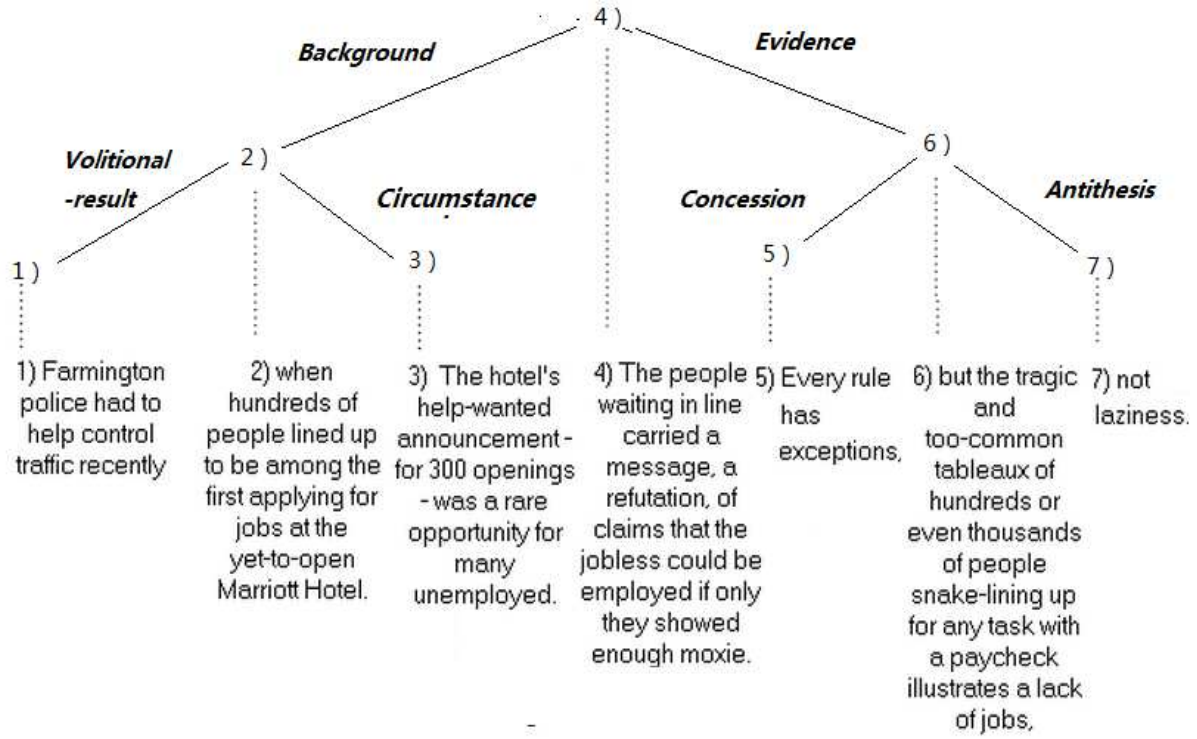


Figure 5. Reframing Figure 1 into relations between terminal clause nodes only

The conversion agrees with both the compositionality criterion of RST (Marcu, 2000) and the hierarchy principle. The former goes that for a rhetorical relation *R* holding between two textual spans, it also holds between their most significant textual units. The compositionality criterion is also inversely applicable. As an essential principle in RST (Taboada & Mann 2006a), the hierarchy principle boasts four constraints (completeness, connectedness, uniqueness and adjacency). In the newly-built dependency trees, each span is a unique node and each

tree, connecting all spans, constitutes a contiguous discourse.

Guided by these guidelines, we further build trees in the RST-DT into new ones with mere ultimate nodes of sentences like Figure 6 and finally trees with mere paragraph nodes.

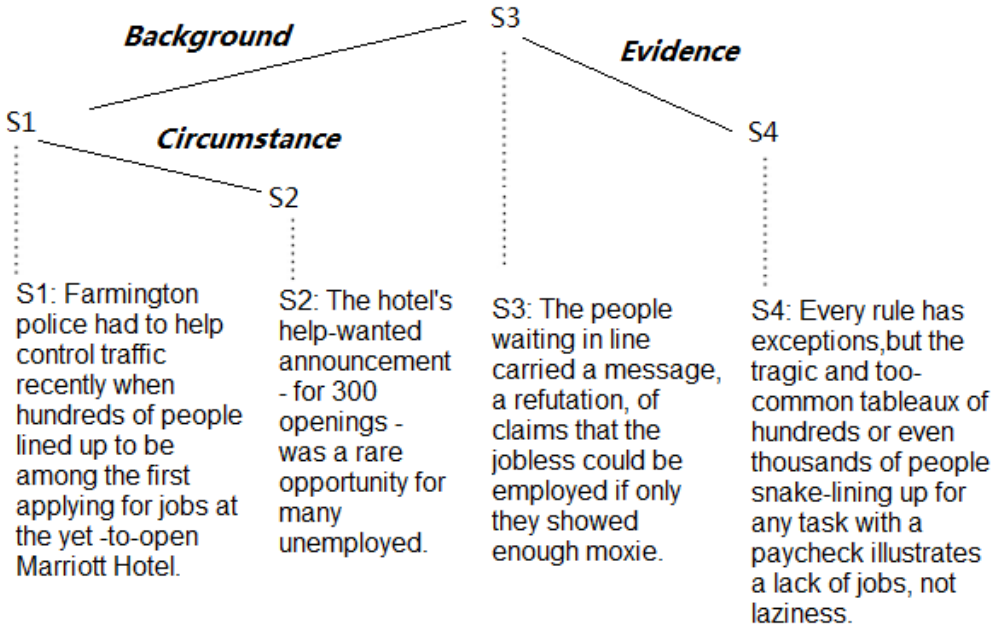


Figure 6. Relations between sentence constituents for the sample text

In this study, the three levels of discourse processes refer to a) building clauses into sentences, b) building sentences into paragraphs and ultimately c) building paragraphs into complete discourses. These processes are presented through the lens of RST relations, as RST relations play unique roles in the dynamic process of meaning construction and facilitate the comprehension of the discourse as an integrated whole.

In terms of multi-nuclear cases, we examine each nuclear-satellite pair, like the three relations (quite likely the same) in Figure 7.

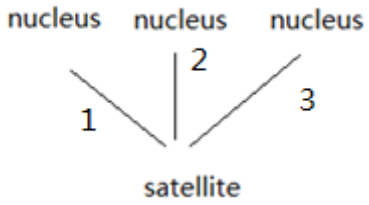


Figure 7. Parallel nuclei

Also germane to this study is the taxonomy of rhetorical relations. In RST, there are several different ways of deciding the granularity of relations *per se*; for instance, *Elaboration-object-attribute-e*, *Elaboration-object-attribute*, and *Elaboration* might be considered 1 or 2 or 3 types of relations. In the RST-DT, these three all appear. We are going to look at the distribution patterns from three taxonomies.

Taxonomy 1:

This is the most elaborate classification (e.g. regarding the previous 3 relations as 3 types). Carlson and Marcu (2001) point out that the inventory of rhetorical relations in the RST-DT is 78, but our study finds it otherwise: it is 86, including those ending

with *–e* (indicating embedding).

Taxonomy 2:

This classification is more general, grouping those with the same initial parts before the dash into the same type. For the above 3 *Elaboration*-related relations, we regard them as belonging to the same type — *Elaboration*. There are 37 such types in the RST-DT.

Taxonomy 3:

Carlson and Marcu (2001) detail the 78 (actually 86) types of relations in the RST-DT and partition them into 16 classes sharing some type of rhetorical meaning. Take for instance, *Comparison*. It can be the umbrella term for *Comparison, Preference, Analogy* and *Proportion*.

In particular, we are interested in three classes: *Elaboration, Topic-Comment* and *Temporal* to see how their representative members (Table 1) are quantitatively distributed. For the rest 13 classes, as they embrace only 2 to 4 representative members, with too few data, way too many types of distribution patterns will be possible and are thus less linguistically revealing. To avoid the sparse data problem, we exclude them from our discussion.

Table 1
Rhetorical relation classes with at least 5 representative members

Class	Representative members
Elaboration	*Elaboration-additional, Elaboration-general-specific, Elaboration-part-whole, Elaboration-process-step, Elaboration-object-attribute, Elaboration-set-member, Example, Definition*
Topic-comment	*Problem-solution, Question-answer, Statement-response, Topic-comment, Comment-topic, Rhetorical-question*
Temporal	*Temporal-before, Temporal-after, Temporal-same-time, Sequence, Inverted sequence*

We are addressing these taxonomies as we deem them justifiable if their distributions observe a regularity which is linguistically interpretable.

Hypothesis 2: The taxonomies of rhetorical relations used in the RST Discourse Treebank are results of diversification processes.

3. Results and discussion

In this part, each sub-section will address one hypothesis. Initially, we examine whether RST relations abide by the chosen distribution across levels. Following that, we investigate the ways to validate the taxonomies of rhetorical relations in RST analysis.

3.1 Distribution pattern of RST relations

Initially, we get the data of relations among constituents and rank (R.) them in a descending frequency (Freq.) so that the highest frequency has Rank 1 (Appendices 1 & 2). Table 2

details the rhetorical relations with Taxonomy 3, the 16-class classification at three levels.

Table 2
Rhetorical relations across levels (Taxonomy 3)

R.	Between clauses within sentences			Between sentences within paragraphs			Between paragraphs		
	Relation	**Freq.**	**%**	**Relation**	**Freq.**	**%**	**Relation**	**Freq.**	**%**
1	*Elaboration*	4476	38. 8	*Elaboration*	2004	56.2	*Elaboration*	1940	63. 3
2	*Attribution*	3639	31. 6	*Explanation*	547	15.3	*Explanation*	263	8. 6
3	*Background*	618	5. 4	*Evaluation*	255	7.2	*Evaluation*	243	7. 9
4	*Enablement*	567	4. 9	*Background*	194	5.4	*Background*	197	6. 4
5	*Contrast*	412	3. 6	*Contrast*	190	5.3	*Summary*	115	3. 8
6	*Cause*	391	3. 4	*Cause*	163	4.6	*Contrast*	113	3. 7
7	*Explanation*	321	2. 8	*Attribution*	42	1.2	*Cause*	87	2. 8
8	*Condition*	283	2. 5	*Summary*	41	1.2	*Comparison*	21	0. 7
9	*Manner-Means*	242	2. 1	*Comparison*	37	1.0	*Topic Change*	21	0. 7
10	*Temporal*	220	1. 9	*Condition*	22	0.6	*Topic-Comment*	21	0. 7
11	*Comparison*	130	1. 1	*Topic-Comment*	19	0.5	*Condition*	16	0. 5
12	*Evaluation*	120	1. 0	*Temporal*	18	0.5	*Enablement*	11	0. 4
13	*Summary*	99	0. 9	*Enablement*	16	0.4	*Manner-Means*	10	0. 3
14	*Topic-Comment*	11	0. 1	*Manner-Means*	16	0.4	*Temporal*	3	0. 1
15							*Attribution*	2	0. 1

In the process of fitting the data to the right truncated modified Zipf-Alekseev distribution, Altmann-Fitter 3.3 is employed to calculate R^2 and parameter values. Table 3 presents all the fitting results.

Table 3
Fitting the right truncated modified Zipf-Alekseev distribution to data
(T = Taxonomy)

	nodes	**relations**	R^2	**a**	**b**	**n**	**α**
T1:86 types	clauses	all	0.9764	1.38	0.04	78	0.30
	sentences	all	0.9964	0.12	0.28	58	0.43
	paragraphs	all	0.9971	0.02	0.30	54	0.47
T2:37 types	clauses	all	0.9443	1.50	0.13	33	0.38
	sentences	all	0.9969	0.13	0.30	36	0.49
	paragraphs	all	0.9982	0.20	0.27	37	0.59
T3:16	clauses	all	0.951	1.69	0.08	14	0.39

	nodes	**relations**	R^2	**a**	**b**	**n**	α
classes	sentences	all	0.9963	0.35	0.46	14	0.56
	paragraphs	all	0.9971	0.05	0.50	15	0.63
3 of the classes	clauses	elaboration	0.9993	1.16	0.36	15	0.59
	paragraphs	elaboration	0.9991	0.44	0.74	14	0.76
	sentences	elaboration	0.9973	0.52	0.66	12	0.74
	clauses	temporal	0.9801	0.26	1.18	6	0.39
	paragraphs	topic-comment	0.9968	0.00	0.66	6	0.37

The observations from the three different ways of classifying the rhetorical relations within the RST-DT all show striking agreement with the distribution. Table 4 is a typical example, presenting the approximation of between-sentence rhetorical relation data to the distribution. Figure 8 graphically illustrates the fitting.

Table 4
Fitting the right truncated modified Zipf-Alekseev distribution
to the data of RST relations between sentences within paragraphs
(f[i]:empirical frequency, NP[i]:theoretical frequency)

x[i]	***f[i]***	***NP[i]***	***x[i]***	***f[i]***	***NP[i]***
1	2004	2004.00	8	41	54.54
2	547	512.76	9	37	41.57
3	255	319.45	10	22	32.26
4	194	208.41	11	19	25.42
5	190	142.03	12	18	20.30
6	163	100.36	13	16	16.41
7	42	73.07	14	16	13.40

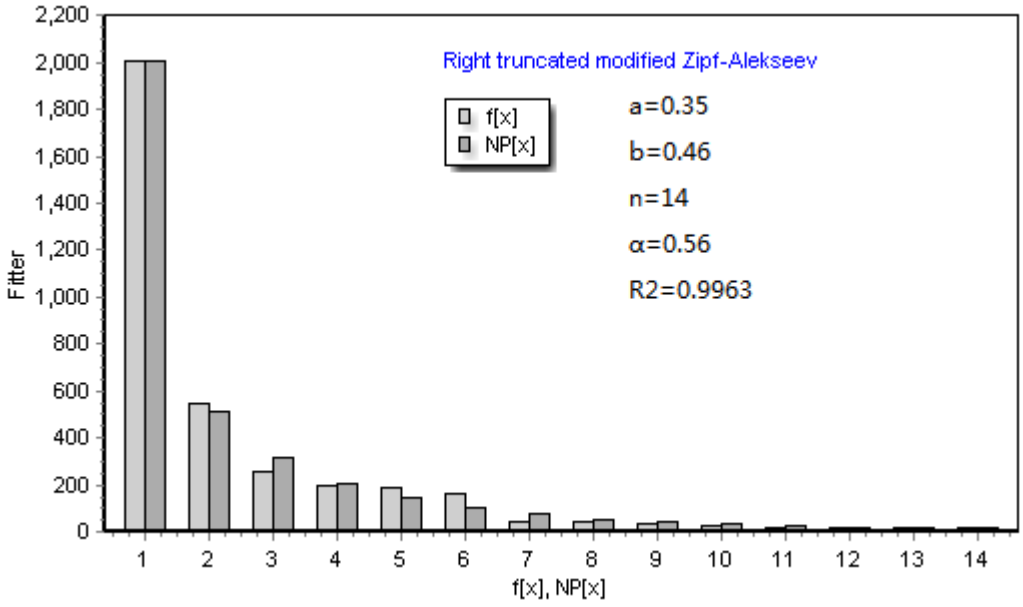


Figure 8. Graphic representation of Table 4

Even rhetorical relation classes with at least 5 representative members (Table 1, Appendix 2) display the same regularity (Table 3).

Elaboration is the class that covers the most types of relations, including *Example, Definition* and all the *Elaboration*-initial relations. It has been a particular problem area for the definition of rhetorical relations, particularly when the identification of its sub-types is not very clear (Taboada & Mann, 2006a). Interestingly, even this class of relations across levels fits perfectly with the same distribution (with R^2 ranging from 0.9973 to 0.9993).

Similarly, we examine the classes of *Temporal* and *Topic-Comment* at three levels in case there are more than 5 members at a certain level in the corpus. The relations of *Temporal* between clauses (R^2 = 0.9801) and *Topic-Comment* between paragraphs (R^2 = 0.9968) are both found to share the same distribution pattern.

These findings validate the first hypothesis: rhetorical relations (decided on three taxonomies) at various levels (including sub-classes) are all regularly distributed, following the same right truncated modified Zipf-Alekseev distribution. This is in agreement with the finding in Yue and Liu (2011). What this model means will be elaborated in the next sub-section.

The common distribution pattern gives a clear indication of a certain common mechanism of discourse processes through the joint of RST relations at all three levels.

3.2 Taxonomies in RST and the diversification process

This part canvasses the taxonomies of relations used in the RST Discourse Treebank.

3.2.1 Inventory of relations

A fixed inventory of relations are not required in many areas of linguistics (Taboada & Mann, 2006a), but the taxonomies of rhetorical relations, somehow subjective and intuitive, are among the most debated issues of RST. Without a uniform standard for annotation in play, the

agreement on certain values can sometimes be problematic (Scholman & Sanders, 2014). Generally, the taxonomy in RST is a set of relatively stable but nonetheless open relations.

Since the initial proposal of 24 in Mann and Thompson (1988), recurrent categories have been proposed and discussed (e.g. Sanders, Spooren, & Noordman, 1992; Louwerse, 2001; Carlson & Marcu, 2001; Taboada & Mann, 2006a). For instance, these 24 relations are extended to 30 on the RST website (http://www.sfu.ca/rst/). In the RST-DT, there are 86 relations. So long as they are not beyond "observability", innovation shall be encouraged as the unit division method won't necessarily be suitable for everyone (Taboada & Mann, 2006a: 437).

Also, there are some other alternative collections of rhetorical relations based on some alternate basis with the number ranging from 2 to 350 (Taboada & Mann, 2006a)!

In addition, the attempts to define the number of rhetorical relations are paired with those to compartmentalize them. To illustrate, Taboada and Mann (2006a) propose grouping them into 12 classes. And Carlson and Marcu (2001) identify 16 classes.

3.2.2 Empirical validations and justifications to rhetorical relations

Empirical validations and justifications to rhetorical relations proper or their taxonomies have been carried out. Most of them are basically experimental, involving either subjects' experiences with rhetorical relations or a comparison between two ways of representing the targeted texts. From the very outset of RST, descriptive adequacy and cognitive plausibility have been proposed as two main features (Sanders et al., 1992). Meyer, Brandt, and Bluth (1980) suggest that coherence relations, particularly when explicitly marked, help ninth-grade students in their discourse organization. It is also proved that marked coherence relations facilitate discourse segment processing (Haberlandt, 1982). Through psycholinguistic experiments, Sanders et al. (1992) prove that subjects are sensitive to different relations, which can be understood as a psychological salience of their taxonomy and as evidence to the understanding of coherence relations. Spooren (1997) focuses on underspecified coherent relations, proving that both speakers and hearers tend to use those relations cooperatively. Similarly, Sanders and Noordman (2000) find that relations explicitly marked result in faster processing. Den Ouden, Noordman, and Terken (2009) carry out an interesting investigation on the prosodic realization (segments, pitch range and articulation rate) of organizational features in 20 RST-annotated news reports. Through a comparison with the read-aloud version, the RST-annotated version is found to reflect organizational features of the texts, which in turn, correspond to prosodic characteristics. By means of an eye-tracking study, Rohde and Horton (2014) argue that anticipatory looks of the comprehenders reveal expectations about inter-sentential coherence relations.

A more direct proof is the practical applications of rhetorical relations, the most important of which include theoretical linguistics, discourse analysis, psycholinguistics and computational linguistics, way beyond its original objective of text generation (Taboada & Mann, 2006a).

The third dimension of validation comes from quantitative studies fitting RST data to certain distributions. Studies by Yue and Liu (2011) and Beliankou et al. (2012) are typical examples.

3.2.3 The diversification process

The discussion on the diversification process shall start from the Zipf's law (Zipf, 1935, 1949). This empirical law, named after Zipf due to his great contribution to it, was originally a law on word frequencies in natural language speech and texts. It states that only a few words are used very frequently, while many or most are used rarely and that the frequency of a word decays as a power law of its rank.

"The least effort" principle (Zipf, 1949) is the theoretical explanation for Zipf's law, which is a consequence of two competitive economic principles. For instance, the speaker tends to reduce the number of words for his least effort in production; consequently, many words will carry more than one meaning and in an extreme case, a word means everything. This speaker economic force gives rise to the unification force. But the listener, in his least effort in utterance comprehension, prefers each word to carry only one exact meaning. This listener economic principle thus constitutes the diversification force. The Zipf's law is actually an indication that the balance (or simultaneous minimization in the effort of both parties) is reached between these two opposing forces. To put in other words, this way, the cost of communicative transactions between speakers and listeners is optimized, hence an actualizeation of the least effort principle.

Language evolves with the two forces of unification and diversification in play. But Zipf's law is not restricted to language laws. It's a general law governing various fields of human behaviors.

Coming back to our case, for the attempts of enlarging, adapting, and categorizing rhetorical relations, we regard them as a diversification and unification process. As a process of enlarging the number of forms or meanings of any linguistic entity, the diversification process (also quite known in biology) occurs at various levels of language and covers an enormous scope of phenomena (Altmann, 1991). For example, words can enlarge their class membership without any formal change (e.g. the *head*, to *head*) or through derivation (e.g. *compose*, *composition*). A word can acquire different meanings, giving rise to polysemy (e.g. *polish*, *fine*) and every word can be associated with other words, acquiring connotations. Diversification and its opposite, unification, are also called Zipfian processes.

The starting points for the study of diversification are three general assumptions serving as the foundation of modeling (Altmann, 2005):

Firstly, the classes of the diversified entity from diversification form a decreasing rank-frequency distribution in case the classes represent a nominal variable or another discrete distribution where the classes represent a numerical variable.

Here in our case the entity is a nominal one. It means that an entity diversifies in one direction, and as a result, the frequencies of the diversified entities won't be equal, but rather, they can be ordered according to a decreasing frequency. If this very assumption goes right, it can constitute a criterion distinguishing the various taxonomies, claiming each to be "good", "useful" or "theoretically prolific", depending on the fitting of rank-frequency distributions (Altmann, 2005:647).

Secondly, the resulting classes do not stand alone, but rather, they are linked by mutual influence. And thirdly, the emerging dimension (or the diversified property) is linked with at least one other property of the same entity.

Concerning the case of nominal rhetorical relations in this study, they follow a decreasing

rank-frequency distribution and accord with the other two assumptions.

Following the three relevant assumptions, some models for these processes are suggested in Altmann (1991). Among them is the right truncated modified Zipf-Alekseev pattern, a known Zipf's law-related distribution to model the ranking law of diversified entities (Altmann, 1991; Köhler, 2012). For details about the linguistic interpretation of this distribution pattern, refer to Yue and Liu (2011). In our study, the robust fitting to this linguistically interpretable distribution suggests that the taxonomies (including the taxonomy of sub-types) in the RST-DT are justifiable and theoretically prolific. Therefore, the taxonomies can be regarded as results of a diversification process, which constitutes a validation of Hypothesis 2.

This research actually backs up Yue and Liu (2011), Beliankou et al. (2012) and Zhang and Liu (forthcoming). The data in all these three studies are found to abide by a rank-frequency distribution, collectively corroborating the idea that rhetorical relations and relation-induced properties are results of diversification processes.

4. Concluding remarks

This study converts each tree in the RST-DT, where there are both elementary clause nodes and expanded spans, into three new dependency trees where there are only terminal nodes of clauses, sentences and paragraphs, respectively. It examines discourse processes through the lens of RST rhetorical relations at three levels of organizing one level of units into the next immediate level. It yields the following research findings:

- The rank-frequency distribution of all the rhetorical relations at all levels is regular, regardless of the granularity of nodes (clauses, sentences or paragraphs) or the granularity of RST relations. They all follow the same Zipf's law-related distribution (right truncated modified Zipf-Alekseev distribution). Three classes with over 5 representative rhetorical relations are also found to behave in this uniform manner.
- The robust fitting of all sets of data by the same distribution pattern justifies three taxonomies (including the taxonomy of sub-types) in the RST-DT. We thus claim that the taxonomies are theoretically prolific. The fitting also corroborates the idea that rhetorical relations are a result of a diversification process.

To apply the findings to language in general, studies from more languages and genres are called for since this study only examines texts from *the Wall Street Journal*. We expect that investigations of other corpora and of languages other than English will yield comparable results.

In previous studies, we have examined the dependency distance and length-frequency relationship (Liu 2007, 2008; Jiang & Liu 2015). Further research efforts shall also cover more properties in the reframed trees, like complexity, number of layers, and various inventories, among many others. When these are done, we might expect to construct a synergetic discourse model.

Acknowledgments:

Both Department of Education of Zhejiang Province, China (Grant No. Y201223584) and the National Social Science Foundation of China (Grant No. 11&ZD188) supported this work. We are deeply indebted to Reinhard Köhler and Timothy Osborne for their helpful discussions. And we sincerely appreciate Chunshan Xu's help in proofreading the paper. Thanks also go to Haiqi Wu, who helped with data for the research.

REFERENCES

Altmann, G. (1991). Modeling diversification phenomena in language. In: Rothe, U. (Ed.), *Diversification Processes in Language*: *Grammar:* 33-46. Hagen: Rottmann.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., & Piotrowski, R.G. (Eds.), *Quantitative Linguistics. An International Handbook:* 648-659. Berlin: de Gruyter.

Beliankou, A., Köhler, R., & Naumann, S. (2012). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., & Köhler, R. (Eds.), *Methods and Applications of Quantitative Linguistics, selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)* (pp. 35-43). Belgrade, Serbia, April 26-29, 2012.

Carlson L., & Marcu D. (2001). *Discourse Tagging Reference Manual*, http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf, accessed at 08:50, Feb. 9, 2015

Carlson, L., Marcu D., & Okurowski M.E. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus]*. Philadelphia, PA: Linguistic Data Consortium.

Carlson, L., Marcu D., & Okurowski M.E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt, J., & Smith, R.W. (Eds.), *Current Directions in Discourse and Dialogue:* 85–112. Dordrecht: Kluwer Academic Publishers.

Den Ouden, H., Noordman, L., & Terken, J. (2009). Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication* 51(2), 116-129.

Haberlandt, K. (1982). Reader expectations in text comprehension. *Advances in Psychology* 9, 239-249.

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.

Li, S.,Wang, L., Cao, Z., & Li, W. (2014). Text-level discourse dependency parsing. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: 25–35. Baltimore, Maryland, USA, June 23–25 2014.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics* 15, 1-12.

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

Liu, H. (2009a). *Dependency Grammar: from theory to practice*. Beijing: Science Press.

Liu, H. (2009b). Probability distribution of dependencies based on Chinese Dependency Treebank. *Journal of Quantitative Linguistics* 16 (3), 256–273.

Louwerse, M.M. (2001). An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics* 12(3), 291-315.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin, New York: de Gruyter (= Quantitative Linguistics; 65).

Mann, W.C., & Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87–190). Marina del Rey, CA: Information Sciences Institute.

Mann, W.C., & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), 243-281.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization.* Cambridge, Massachusetts: The MIT Press.

Marcu, D., Carlson, L., & Watanabe, M. (2000). *The Automatic Translation of Discourse Structures*. Presented at the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00) (pp. 9–17). Seattle, WA.

Meyer, B.J.F., Brandt, D.M., & Bluth, G.J. (1980). Use of Top-level Structure in Text: Key for Reading Comprehension in Ninth-grade Students. *Reading Research Quarterly* 16(1), 72–103.

Moore, J., & Wiemer-Hastings, P. (2003). Discourse in computational linguistics and artificial intelligence. In: Graesser, A., Gernsbacher, M., & Goldman, S. (Eds.), *Handbook of Discourse Processes* (pp. 439–486). Mahwah, NJ: Erlbaum.

Rohde, H., & Horton, W.S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition,*133(3), 667–691.

Sanders, T., Spooren, W., & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes* 15(1), 1-35.

Sanders, T., & Noordman, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* 29(1), 37-60.

Scholman, M., & Sanders, T. (2014). Annotating coherence relations in corpora of language use. In: *Proceedings of the CLARIN Annual Conference*. Soesterberg, The Netherlands.

Spooren, W. (1997). The processing of underspecified coherence relations. *Discourse Processes* 24(1), 149-168.

Stede, M. (2004). The Potsdam commentary corpus. In: *Proceedings of the ACL 2004 Workshop on "Discourse Annotation"* (pp. 96-102). Barcelona, Spain.

Taboada, M., & Mann, W. (2006a). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3), 423-459.

Taboada, M., & Mann, W. (2006b). Applications of Rhetorical Structure Theory. *Discourse Studies* 8(4), 567-588.

Williams, S., & Reiter, E. (2003). A corpus analysis of discourse relations for natural language generation. In: *Proceedings of Corpus Linguistics* 2003, 899–908. Lancaster University.

Yue, M., & Liu, H. (2011). Probability distribution of discourse relations based on a Chinese RST-annotated corpus. *Journal of Quantitative Linguistics* 18(2), 107-121.

Yue, M., & Feng, Z. (2005). *Findings in a preliminary study on the rhetorical structure of Chinese TV news reports*. Paper presented at the First Computational Systemic Func-

tional Grammar Conference, Sydney, Australia.

Zhang, H., & Liu, H. (forthcoming). Motifs in Reconstructed RST Discourse Trees. *Journal of Quantitative Linguistics*.

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

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

Appendix 1

Rank-frequency data of Taxonomies 1 and 2

(R = Rank, 1 = Taxonomy 1, 2 = Taxonomy 2, C = Clauses as nodes, S = Sentences as nodes, P = Paragraphs as nodes)

R.	C2	S2	P2	R.	C1	S1	P1	R.	C1	S1	P1
1	4352	1753	1802	1	3455	1531	1427	40	18	4	3
2	3639	363	189	2	2644	363	204	41	18	3	2
3	605	207	135	3	840	205	189	42	17	3	2
4	553	139	118	4	549	147	135	43	16	3	2
5	251	124	112	5	511	139	120	44	14	2	2
6	239	108	90	6	398	123	118	45	14	2	2
7	220	103	85	7	233	103	87	46	13	2	1
8	212	96	79	8	219	91	86	47	13	2	1
9	200	91	77	9	199	91	79	48	13	2	1
10	162	82	65	10	183	80	78	49	13	1	1
11	148	75	60	11	158	75	77	50	12	1	1
12	139	66	58	12	141	75	65	51	11	1	1
13	102	46	36	13	135	66	60	52	9	1	1
14	99	45	22	14	129	46	50	53	9	1	1
15	94	44	20	15	123	45	36	54	8	1	1
16	94	42	15	16	107	43	34	55	7	1	
17	93	28	14	17	95	39	26	56	7	1	
18	50	27	12	18	91	32	22	57	7	1	
19	46	18	10	19	86	28	20	58	7	1	
20	43	17	8	20	85	28	15	59	7		
21	31	14	7	21	80	27	14	60	6		
22	30	12	7	22	80	17	10	61	5		
23	26	12	6	23	61	14	8	62	5		
24	20	9	4	24	61	12	8	63	5		
25	18	9	4	25	56	12	8	64	5		
26	14	7	3	26	54	12	7	65	5		

27	13	6	3	27	53	10	7	66	4		
28	13	4	3	28	46	9	7	67	4		
29	10	4	3	29	45	9	7	68	3		
30	7	3	3	30	42	9	6	69	3		
31	4	3	3	31	41	9	5	70	3		
32	1	2	3	32	40	7	4	71	2		
33	1	2	2	33	38	6	4	72	1		
34		1	2	34	34	6	3	73	1		
35		1	1	35	32	5	3	74	1		
36		1	1	36	30	5	3	75	1		
37			1	37	25	4	3	76	1		
38				38	20	4	3	77	1		
39				39	20	4	3	78	1		

Appendix 2
Rank-frequency data of chosen RST relations

Relations	***Elaboration***			***Temporal***	***Topic-comment***
Rank	**Between clauses**	**Between sentences**	**Between paragraphs**	**Between clauses**	**Between paragraphs**
1	2644	1531	1427	86	7
2	840	205	204	80	5
3	398	147	135	41	3
4	233	43	120	5	2
5	80	32	34	5	1
6	54	14	8	3	1
7	53	10	4		
8	45	9	3		
9	40	4	2		
10	25	3	1		
11	20	2	1		
12	18	2	1		
13	14	1			
14	11	1			
15	1				

Glottometrics 33, 2016, 25 - 31

Verbal vs. Adjectival Styles in Long Poems by A.S. Pushkin

Sergey Andreev

Abstract. This study is based on the methodology, suggested in the research by G. Altmann, S. Naumann, I.-I. Popescu (Naumann et al. 2012). It is used in our research for the analysis of the proportions and distribution of three parts of speech (nouns, adjectives and verbs) in the data-base, consisting of ten long poems by A.S. Pushkin, the great Russian poet.

The proportions of adjectives and verbs against nouns show the type of the author's poetic visualization of the world (static or dynamic) and the intensity of such description in general.

Keywords: *Russian, Pushkin, poetry, style*

Among language units highly relevant for textometric analysis are parts of speech (PoS) whose counts create a basis for solving a large number of important problems in various spheres of linguistic research such as authorship detection, automatic classification of texts and/or individual styles of the authors, discovering rhythmic peculiarities (in verses), finding out main features of text structure in various genres, etc. (Altmann 2014; Gasparov 2012a; Mikros 2009; Popescu et al. 2007; Tuzzi et al. 2009).

An important problem of how a poet visualizes the world can be resolved by finding out the type of relationship which exists between nouns, verbs and adjectives in his poetic texts. If themes in poetry are mostly expressed by nouns, verbs and adjectives give additional information about themes by specifying images (adjectives) and actions or states (verbs).

Studies of types of poetic description in verses have established proportions between adjectives and verbs, verbs and nouns (Popescu et al. 2013, Gasparov 2012a, Naumann et al. 2012), revealed important interrelations between parts-of-speech, and have brought about important results about the peculiarities of individual styles of poets as well as general tendencies in dynamic representation of the world in epic and lyrical poems.

In this study the type of relationship of static vs. dynamic representation of themes in verse texts in Russian is carried out by finding out proportions of verbs and adjectives against nouns. Speaking of the opposition of static and dynamic description as an important feature of style it should be mentioned that there are at least two more interpretations of style dynamics.

Firstly, it can refer to the changes that occur over years in the creative manner of an author, the study of alternations of text features, observed at different stages of creative activities of an author. This approach also includes investigations which are devoted to the chronology and dating of the texts (Brandwood 2009; Temple 1996).

Secondly, this term can be applied to the description of text elements of a poetic work from its beginning to end (Martynenko 2004), and sometimes in the opposite direction (Köhler et al. 2012: 82). In this case the fact that the texts have been written during a certain period of time is usually ignored.

Counts of PoS in Russian necessitate certain specifications.

Proper nouns were counted together with common nouns without reduction which was used by, e.g. in q-sum approach to authorship detection (Farringdon 1996) in which several proper nouns were taken as one word ('name').

Verbs include all their forms (personal forms, infinitives, participles, gerunds). By gerund in Russian a verbal form is understood which denotes an additional action to the main one ('while doing smth' or 'having done smth').

Adjectivized participles, which are included into the class of adjectives, are differentiated from the verbal participial forms on the basis of syntactic and semantic criteria (such as metaphorical meaning, constant quality, etc. for the adjectivized forms).

The data-base includes 10 long poems by the great Russian poet Alexander S. Pushkin. The long poems are: "Ten' Fonvizina" (*Fonvizin's Shade*); "Kavkazskiy plennik" (*Prisoner of the Caucasus*); *Vadim*; "Bratya razboyniki" (*The Robber Brothers*); "Bakchisarayskiy fontan" (*The Fountain of Bakchisarai*); "Tsigani" (*The Gypsies*); "Graf Nulin" (*Count Nulin*); *Tasit*; *Jezierski*; "Mednyj vsadnik" (*The Bronze Horseman*).

All the poems were written in iambic tetrameter (with few exceptions in a very small number of lines), thus forming a homogeneous basis. The total number of lines is a little less than 4000.

The first stage of analysis consisted in finding out proportions of verbs and adjectives in the texts of all poems (Table 1).

Table 1
Proportions of verbs and adjectives against nouns
in long poems by Pushkin

	Poem	Verb-noun	Adjective-noun
1	*Fonvizin's Shade*	0.359	0.222
2	*Prisoner of the Caucasus*	0.326	0.298
3	*Vadim*	0.327	0.288
4	*The Robber Brothers*	0.384	0.239
5	*The Fountain of Bakchisarai*	0.316	0.306
6	*The Gypsies*	0.573	0.275
7	*Count Nulin*	0.394	0.208
8	*Tasit*	0.376	0.258
9	*Jezierski*	0.283	0.216
10	*The Bronze Horseman*	0.366	0.286

The proportion of adjectives and nouns is to some extent more stable than that of verbs and nouns: the coefficient of variance for the former is 13.93% and 21.38% for the latter.

To test the deviation of the observed proportion from its expectation the following criterion was suggested (Naumann et al. 2012: 29):

$$u = \frac{p(S) - E(p)}{\sqrt{p(1-p)/(S+N)}},$$

where p is the proportion of the given part of speech S, $E(p)$ – expected proportion, $p(S)$ is the observed proportion in the text of the given part of speech S, N is the number of observed nouns (Naumann et al. 2012).

The expected proportion of verbs and adjectives against nouns in Russian can be obtained from the data of M.L. Gasparov (Gasparov 2012b: 327). According to his data at the beginning of the 19th century in Russian poetry nouns equalled 43% of all the words in the

poems, verbs – 22% and adjectives – 15%. Judging by this per cent representation of the occurrence of these three PoS, the proportion of verbs against nouns, calculated by dividing the number of verbs by the sum of the verbs and nouns is $E = 43/(22+43) = 0.33846$, and the proportion of adjectives against nouns equals 0.25862. These values are taken in this study as expected proportions. It should be noted that the verbal expected value in Russian poetry is very close to the value, used in the analysis of different Indo-European languages in the above-mentioned article, which is $E = 0.3333$ (Naumann et al. 2012: 29).

Taking the critical level as $p = 0.1$ for $df = \infty$, $|u| \leq 1.64$ will not be considered as relevant deviation from the expected level. In such case we shall speak about verb-balanced or adjective-balanced style. In case $u < -1.64$, the number of verbs or adjectives is significantly less than expected, if $u > 1.64$ the empirical number of these PoS exceeds the expectation level.

Table 2 contains the observed values of verbs, nouns and the *u*-criterion for 10 poems, Table 3 – the same data about adjectives. Statistically relevant deviations from the expected level are marked in bold type.

Table 2
Expected and observed proportion of verbs against nouns
in long poems by Pushkin

	Poem	Date of writing the poem	Verbs	Nouns	U-criterion
1	*Fonvizin's Shade*	1815	251	449	1.12
2	*Prisoner of the Caucasus*	1820	515	1064	-1.03
3	*Vadim*	1822	145	299	-0.53
4	*The Robber Brothers*	1821-1822	197	319	**2.17**
5	*The Fountain of Bakchisarai*	1821-1823	365	789	-1.59
6	*The Gypsies*	1824	943	702	**20.13**
7	*Count Nulin*	1825	305	469	**3.27**
8	*Tasit*	1829-1830	205	340	**1.86**
9	*Jezierski*	1832-1833	119	301	**-2.39**
10	*The Bronze Horseman*	1833	358	619	**1.85**

The most important conclusion which can be drawn from the data of Table 2 is that lyrical poems by Pushkin are characterized by clearly marked verbal style which corresponds to the observation of M. Gasparov that Pushkin's verse is even more dynamic than prose of Tolstoy and Chekhov (Gasparov 2012a). One exception from this verbally intensified style is his unfinished poem Jezierski, displaying "deficiency" of verbs. A certain tendency to nominal style is also seen in *The Fountain of Bakchisarai*, and in three other poems the verbal style may be considered as balanced.

The other conclusion concerns the correlation of the period of creative activity of the poet when he wrote the poems and the style. Pushkin's earlier poems seem to be more balanced in style, less markedly verbal than those written later.

Table 3
Expected and observed proportion of adjectives against nouns
in long poems by Pushkin

	Poem	Date of writing the poem	Adjectives	Nouns	U-criterion
1	*Fonvizin's Shade*	1815	128	449	**-2.018**
2	*Prisoner of the Caucasus*	1820	452	1064	**3.516**
3	*Vadim*	1822	121	299	1.380
4	*The Robber Brothers*	1821-1822	100	319	-0.905
5	*The Fountain of Bakchisarai*	1821-1823	348	789	**3.654**
6	*The Gypsies*	1824	266	702	1.149
7	*Count Nulin*	1825	123	469	**-2.826**
8	*Tasit*	1829-1830	118	340	-0.048
9	*Jezierski*	1832-1833	83	301	**-1.901**
10	*The Bronze Horseman*	1833	248	619	**1.844**

Speaking of adjectives it can be noted that the picture is less uniform. In three poems adjectives are obviously ‘deficient’ in number and in other three cases they exceed the expected level. It should be noted that the group of poems with pronounced adjectival style includes highly romantic *Prisoner of the Caucasus*, *The Fountain of Bakchisarai* as well as realistic style poem *The Bronze Horseman*.

Contrary to what was observed for verbs, in case of adjectives the period of poetic activities and the age of the poet had no influence on the choice between adjectival and not-adjectival types of style, cf. *Fonvizin's Shade vs. Prisoner of the Caucasus, Jezierski vs. The Bronze Horseman*.

Comparing the data of Tables 2 and 3 several oppositions between verbal and adjectival types of style can be discovered. The strongest opposition is observed in *Count Nulin* in which verbal style is expressed rather vividly whereas adjectival description is below the expected level. This phenomenon can be called compensation – the decrease of static description of themes is compensated for by dynamic representation. Very close to it is *The Fountain of Bakchisarai* in which the same tendency of contrast of styles can be supposed. In this case the adjectival style prevails and verbality has a tendency to be deficient.

In three poems *The Robber Brothers*, *The Gypsies* and *Tasit* verbal style exists at the background of adjective-balanced style (the expected level of the number of adjectives is observed), forming thus a binary opposition in which adjectival style is a zero member. In the poem *Prisoner of the Caucasus* vice versa the adjectival style is observed at the background of a verb-balanced style. *The Gypsies* is also characterized by verb-balanced style, on the one hand, and adjective number deviation from the expected value, on the other. But in this case the deviation consists not in an increase, but in a drop of adjective descriptiveness.

In two cases (*The Bronze Horseman* and *Jezierski*) the description tendency is the same for both PoS. *The Bronze Horseman* is characterized at the same time by verbal and adjectival styles and in case of *Jezierski* the number of both verbs and adjectives is below the expected level. This tendency can be called the tendency of intensification. In the first case it means the intensification of descriptive aspect, in the second case – the intensification of a fall in

detalization ("zero-description"). In the poem *Vadim* the style is highly balanced as both with adjectives and verbs the observed values correspond to the expected level.

At the next stage of analysis the possibility of the style alternations over the texts was investigated. The increase or decrease of the verbal or adjectival styles in the texts from their beginning to end was studied according to the method of dynamic view of text suggested in the above-mentioned article (Naumann et al. 2012: 24–26). In our case we counted the number of verbs and adjectives before every noun obtaining cumulative sums and thus forming a sequence of values. In graphic representation nouns are reflected on the x-axis and cumulative sums of verbs or adjectives on the y-axis. To capture the trend of such sequences the authors of the article propose to use the power function $y = ax^b$ in which the parameter b reflects changes in style.

Using this method we found out the parameters a and b and the determination coefficient r^2 for 10 poems. The results are given in Tables 4 and 5. For $b > 1$ verbal or adjectival style is increasing, if $b < 1$ it is falling. If $b = 1$ verbs and adjectives in regard to nouns are distributed uniformly through the text.

Table 4
Changes of verbal style

Poem	*A*	*b*	r^2
Fonvizin's Shade	0.62	0.974	0.997
Prisoner of the Caucasus	0.14	1.178	0.994
Vadim	0.31	1.071	0.982
The Robber Brothers	0.14	1.267	0.992
The Fountain of Bakchisarai	0.20	1.132	0.996
The Gypsies	0.07	1.333	0.995
Count Nulin	0.13	1.260	0.998
Tasit	0.26	1.144	0.996
Jezierski	0.14	1.183	0.994
The Bronze Horseman	0.16	1.204	0.997

The coefficient of determination is high, showing that the power function captures well the trend.

In all cases the intensity of style changes is not very large. Parameter *b* demonstrates a very moderate growth of verbality in 8 texts out of 10. Two texts (*Fonvizin's Shade* and *Vadim*) demonstrate practically uniform, smooth distribution of verbs over the whole text. Slight intensification of the verbal style in poems seems to be their characteristic feature. It is observed regardless of whether the style in the poem is verbal as in *The Gypsies,* etc., or verbally neutral as in *Prisoner of the Caucasus,* or even verbally-deficient as in *The Fountain of Bakchisarai* and *Jezierski.*

Table 4
Changes of adjectival style

Poem	*a*	*b*	r^2
Fonvizin's Shade	0.34	0.973	0.996
Prisoner of the Caucasus	0.63	0.947	0.999
Vadim	0.25	1.085	0.998
The Robber Brothers	0.31	0.997	0.994
The Fountain of Bakchisarai	0.65	0.944	0.998
The Gypsies	0.24	1.071	0.996
Count Nulin	0.19	1.062	0.991
Tasit	0.26	1.144	0.996
Jezierski	0.09	1.202	0.990
The Bronze Horseman	0.72	0.905	0.995

The distribution of adjectives against nouns displays even stronger evenness than that of verbs. Only in two cases it is possible to speak of some relevant changes. These are *Tasit* and *Jezierski* which display certain growth of adjectival style towards the end and, to a certain extent, *The Bronze Horseman* in which a slight decrease in the number of adjectives through the text is observed.

On the whole it is possible to say that though these 10 long poems by Pushkin are lyrical, they nevertheless are characterized by verbal style, which is more expected in prose or, if verses are considered – in epic poems. The verbal style in lyrical poems does not change considerably from beginning to end.

In most cases the intensification of one style (usually verbal) is observed at the background of the balanced style of the other one (usually adjectival). In other words the tendency of compensation when deficiency of one feature is compensated for by the intensification of the other is not very marked here. It is possible to say that the competition of verbal and adjectival styles is realized not as an equipollent but as a privative opposition with one zero member.

The other tendency – the tendency of intensification of description when a simultaneous rise or drop in both types of description occurs (so that both verbal and adjectival styles should become prominent or weakened at the same time) is displayed rather weakly.

And lastly, the general trend (romantic or realistic) to which each of the poems belongs does not influence the ratio of verbal and adjectival styles.

References

Altmann, G. (2014). Supra-sentence levels. *Glottotheory* 5(1), 25–39.

Brandwood, L. (2009).*The chronology of Plato's dialogues*. Cambridge: Cambridge University Press.

Farringdon, **J. M.** (1996). *Analyzing for Authorship. A Guide to the Cusum Technique*. Cardiff: University of Wales Press.

Gasparov, M.L. (2012a). Tochniye metody analyza grammatiki v stihe [Exact methods of verse analysis]. In: *V.L. Gasparov. Izbranniye trudy. Lingvistika stiha. Analyzy I interpretatsyy*. Vol. 4: 23-25. Moscow: Yaziky Slavanskoy kul'tury.

Gasparov, M.L. (2012b). Fonetica, morphologiya i syntaksis v borbe za styh [Phonetics, morphology and syntax in struggle for verse]. In: *V.L. Gasparov. Izbranniye trudy. Lingvistika stiha. Analyzy I interpretatsyy.* Vol. 4: 23-25. Moscow: Yaziky Slavanskoy kul'tury.

Martynenko, G.Y. (2004). *Ritmico-smyslovaya dynamica russkogo klassicheskogo soneta* [Rhythmic and sense dynamics of the Russian sonnet]. Saint-Petersburg: Saint Petersburg state university.

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

Mikros, G.K. (2009). Content words in authorship attribution: An evaluation of stylometric features in a literary corpus. In: Reinhard Köhler (ed.), *Studies in Quantitative Linguistics 5: Issues in Quantitative Linguistics:* 61–75. Lüdenscheid: RAM-Verlag.

Naumann, S., Popescu, I.-I., Altmann, G. (2012). Aspects of nominal style. *Glottometrics* 23, 23–55.

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

Popescu, I.-I., Čech, R., **Best, K.-H., Altmann G. (**2013). Descriptivity in Slovak lyrics. *Glottotheory* 4 (1), 92–104.

Temple, J.T. (1996). A Multivariate Synthesis of Published Platonic Stylometric Data. *Literary and Linguistic Computing* XI (2), 67–75.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics* 19, 42–48.

Glottometrics 33, 2016, 32

Introduction

Discussion is the „daily bread“ of science. Some problems are discussed for decades in separate articles and journals, other ones are the objects of conferences or omnibus volumes. Often, the critical point disappears because a new “school” does not consider it worth of discussion. Mostly, the criticized author does not even learn that his approach had some weak points because one cannot read everything and the critics do not send him their discovery. Even if today it is simple to convey news, it is easier to wait until the concerned author himself reacts – it can take several years and usually, in some years, the problem is not topical any more.

Glottometrics ventures to accelerate this historically well-known procedure and opens a rubric in which a certain topical problem can be directly discussed. The criticized authors can take part in the discussion but need not. The rubric is dedicated to articles reviewing a problem or even a world view but it must be connected in some way with quantitative linguistics.

We begin with the discussion concerning the problem of dependence in text. The problem is not new but there are that many aspects propagated by various linguistic schools that even a survey of views would fill several books. One should not forget that in science, we construct views, not truths, and try to corroborate plausible hypotheses as well as possible. Unfortunately, there are as many languages as there are Men, because everybody uses even the same language differently. There is no final corroboration because there are always some boundary conditions which cannot be captured by a quantitative linguist and do not interest a qualitative linguist. Qualitative linguistics searches for rules and enumerates the exceptions, quantitative linguistics searches for models of phenomena and tests the models, just as in physics. In quantitative linguistics one gets deeper and is ready to abandon a falsified hypothesis; in qualitative linguistics one adheres to a “school” and follows the prescriptions just as in a religion. Projects are accepted only if they are in line with the dominant school – represented by the members of the commissions.

Here we want to open the door for direct criticism and direct response. This is the way of connecting colleagues living in different continents and present their opinions to those living in other continents.

The Editorial Board

Glottometrics 33, 2016, 33-34

Liberating Language Research from Dogmas of the 20th Century

Ramon Ferrer-i-Cancho[1] & Carlos Gómez-Rodríguez[2]

Abstract. A commentary on the article "Large-scale evidence of dependency length minimization in 37 languages" by Futrell, Mahowald & Gibson (PNAS 2015 112 (33) 10336-10341).

Keywords: dependency length minimization, syntactic dependencies, linguistic theory

Central to the inspiring contributions of E. Gibson and collaborators to language research is the idea that a wide range of phenomena, e.g., ambiguity resolution, parsing difficulties or even our notion of sentence "grammaticality", could be manifestations of a principle of dependency length minimization (e.g., references to Gibson's work of Futrell, Mahowald, & Gibson, 2015), in stark contrast to the view of generative linguistics at least.

In a recent study of impressive breadth, Futrell, Mahowald and Gibson (2015) have provided evidence of dependency length minimization across languages by means of various baselines. Paradoxically, the random baselines incorporate constraints on word order that are likely to be consequences of the very principle of dependency length minimization. Futrell et al. argue that their "Free Word Order Baseline" does not obey any particular word order rule, however, it is not actually free because crossing dependencies are not allowed. A truly free word order baseline, and indeed a fully null hypothesis, is one where the $n!$ possible linearizations of the n units (words) of a sentence are allowed *a priori*, as in the pioneering research on dependency length minimization by Ferrer-i-Cancho that Futrell et al. (2015) cite. Furthermore, a large body of theoretical and empirical research strongly suggests that non-crossing dependencies arise as a side effect of pressure to reduce dependency lengths (see Gómez-Rodríguez & Ferrer-i-Cancho 2016, Ferrer-i-Cancho & Gómez-Rodríguez, 2015 and references therein).

Therefore, investigating dependency length minimization with random baselines or an "optimal baseline" where crossings are not allowed is not only theoretically superficial, but also unnecessarily complicated and most worryingly, indicates subordination to the division between competence and performance, a dogma of generative linguistics that Gibson and collaborators have challenged in the past. Futrell et al.'s "Consistent Head Direction Baseline" is another example of baseline that is likely to incorporate dependency length minimization in its very definition: consistent head direction might be a consequence of dependency length minimization (Ferrer-i-Cancho, 2015a, 2015b). For instance, once the

[1] Complexity & Qualitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega. Jordi Girona Salgado 1-3. 08034 Barcelona, Catalonia, Spain.
Address correspondence to: rferrericancho@cs.upc.edu.
[2] LyS Research Group, Departamento de Computación, Facultade de Informática, Universidade da Coruña, Campus de A Coruña, 15071 A Coruña, Spain.

verb is placed last (as in SOV order), dependency length minimization predicts that, consistently, the dependents of the nominal heads of S and O should precede their heads. Similar arguments can be made for the ”Fixed word order baseline”: dependency length minimization predicts the relative placements for certain dependencies, e.g. adjectives with respect to their nominal heads, verbal auxiliaries with respect to their verbal heads, and so on (Ferrer-i-Cancho, 2015a).

Surprisingly, Futrell et al. take for granted dogmas behind principles and parameters theory, where the consistent branching is assumed (not explained) and its direction is determined by a parameter. In contrast, tendencies for consistent branching and its direction are less parameter-consuming predictions of a mathematical theory of dependency length minimization (Ferrer-i-Cancho, 2015a, 2015b).

In sum, Futrell et al.’s research on dependency length minimization is an example of radical empirical research that attempts to remain theoretically agnostic but, paradoxically, turns out to gullibly accept tenets of theoretical linguistics of the past century. Those tenets can be summarized as a belief in the existence of word order constraints that cannot be explained by evolutionary processes or requirements of performance or learning, and instead require either (a) heavy assumptions that compromise the parsimony of linguistic theory as a whole or (b) explanations based on internal constraints of obscure nature.

Our commentary has focused on the problems of Futrell et al.’s analysis for the construction of a general theory of language that is both highly predictive and parsimonius. Other issues have been reviewed by Liu, Xu, and Liang (201).

Acknowledgments

This commentary is a slightly extended version of the letter that we submitted to PNAS and was rejected. R.F.C is funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and the grant TIN201457226-P from MINECO (Ministerio de Economia y Competitividad). C.G.R is partially funded by the MINECO grant FFI2014-51978-C2-2-R and Xunta de Galicia (grant R2014/034 and an Oportunius program grant).

References

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*.

Ferrer-i-Cancho, R. (2015a). The placement of the head that minimizes online memory. A complex systems approach. *Language Dynamics and Change* 5, 141-164.

Ferrer-i-Cancho, R. (2015b). Reply to the commentary “Be careful when assuming the obvious”, by P. Alday. *Language Dynamics and Change 5, 147-155.* doi: 10.1163/22105832-00501009

Ferrer-i-Cancho, R., & Gómez-Rodríguez, C. (2015). Crossings as a side effect of dependency lengths. *http://arxiv.org/abs/1508.06451*.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences 112*(33), *10336-10341*. doi: 10.1073/pnas.1502134112

Liu, H., Xu, C., & Liang, J. (2016). Dependency length minimization: puzzles and promises. Glottometrics (this issue).

Glottometrics 33, 2016, 35-38

Dependency Length Minimization: Puzzles and Promises

Haitao Liu[1a,c], *Chunshan Xu*[a, b] *and Junying Liang*[a,]

Abstract. In the recent issue of PNAS, Futrell et al. claim that their study of 37 languages gives the ***first*** large scale cross-language evidence for Dependency Length Minimization, which is an overstatement that ignores similar previous researches. In addition, this study seems to pay no attention to factors like the uniformity of genres, which weakens the validity of the argument that DLM is universal. Another problem is that this study sets the baseline random language as projective, which fails to truly uncover the difference between natural language and random language, since projectivity is an important feature of many natural languages. Finally, the paper contends an "apparent relationship between head finality and dependency length" despite the lack of an explicit statistical comparison, which renders this conclusion rather hasty and improper.

Key words: dependency length minimization, cross-language, projectivity,

For decades, dependency length (distance) minimization has been pursued as a universal underlying force shaping human languages. In a recent issue of PNAS, Futrell, et al. (2015) suggest that dependency length minimization (DLM) is a universal property of human languages and hence supports explanations of linguistic variation in terms of general properties of human information processing. However, this statement is much exaggerated and far-fetched.

First of all, it is claimed in the paper that this is the ***first*** large scale cross-language evidence for DLM, since "previous comprehensive corpus-based studies of DLM cover seven languages in total". However, this is absolutely **NOT** true. In fact, there have been some large scale cross-language studies of DLM. For example, Liu (2008) has compared dependency distance of 20 natural languages with that of two different random languages, and pointed out that dependency distance minimization is probably universal in human languages. Evidently, the two articles share the same research objective, the same research findings, and similar research methodologies.

There are some minor differences in the specific methods used in these two works. For example, Furtell et al (2015) hold dependency relations constant and draw random word order, while Liu (2008) held word order constant and drew random dependency relations. But such minor differences cannot deny the fact that the two works adopt similar research methods: both

[1] [a]Department of Linguistics, Zhejiang University, Hangzhou, CN-310058, China;
[b]School of Foreign Languages, Anhui Jianzhu University, Hefei, CN-230601, China;
[c]Ningbo Institute of Technology, Zhejiang University, Ningbo, CN-315100, China.
Address correspondence to: *jyleung@iip.zju.edu.cn*

are based on the comparison between the dependency length (distance) of natural languages and that of corresponding artificial random languages. This method has also been used in an earlier study of two languages (Ferrer-i-Cancho 2004). The above difference in methods has no significant influence on the results of research, since it merely reflects the different ways to construct random languages in which the distribution of dependency length is randomized. Of course, it is perfectly acceptable and even encouraging for researchers to test previous findings with somewhat different methods. Anyway, any scientific finding must be subject to repeated tests. However, as far as this PNAS paper is concerned, we are much curious and puzzled why and how the authors could cite the work of Ferrer-i-Cancho and Liu (2014), which clearly introduces and largely dwells on previous DLM study based on 20 languages, but still claim that their PNAS paper is the *first* large scale cross-language evidence for DLM, and that "previous comprehensive corpus-based studies of DLM cover *seven* languages in total".

What is more, dependency length is sensitive to many factors. Linguistic properties, say DLM, may feature in one genre of language, but become vague and weak in another. Therefore, it is more desirable, especially in cross-language studies, to use a parallel corpus, or at least, corpora with the same genres, annotated with similar syntactic annotation schemes or drawn from native dependency treebanks (Jiang and Liu 2015). In the present study, however, it is not clear whether these conditions are satisfied by judging from the materials and methods, and hence there is some doubt in the validity of the argument that DLM is universal in all these languages.

As recently suggested, DLM bears closely on the rarity of crossing dependencies (Ferrer-i-Cancho 2013), and the authors also mention projectivity as one pervasive property of word order that can explain (or be explained by) DLM. What puzzles is that the baseline word order is set as projective. If projectivity is one feature of human language that contributes to DLM, it is desirable for a study of DLM to set baseline word order as non-projective so as to reveal the influence of projectivity on human languages in general. Projective baseline word order in this article fails to reveal the role of projectivity in DLM. In comparison, two baseline word orders respectively set as non-projective and projective may well throw much more light on DLM in natural languages, which has been adopted in previous works (Liu 2007, 2008). Also directly related to DLM is the distribution of dependency distance or the proportion of adjacent dependencies (AD) in natural languages. Previous studies have indicated that AD accounts for at least nearly half of all dependencies in any language investigated so far(Liu 2008), that the frequency of dependency drops dramatically with the increase of length (distance)(Liu 2007), and that a distribution of dependency distance is not influenced by variation in sentence length(Jiang and Liu 2015). These findings explain why DLM is persistently found in human languages and hence should have been mentioned in this article.

In the concluding part, the authors contend that an "apparent relationship between head finality and dependency length is a new and unexpected discovery". Nevertheless, it seems not apparent enough that dependency length is directly related to head-dependent order: no explicit statistical comparison is made in the present paper. Hence, the conclusion seems rather hasty, lacking solid supporting data. Theoretically, SVO order is in favor of DLM, as has been mathematically proven by Ferrer-i-Cancho (2015). But language is complex, constrained by multiple factors whose interactions may lead to no significant distance difference between VO and OV languages. In fact, existent corpus-based researches point to no definite relations between head placement and dependency distance. Gildea and Temperley (2010) find that

German, as an OV language, has longer dependency distance than English, a VO language, but Hiranuma (1999) finds no difference between English and Japanese, which is an OV language, while Liu (2008) finds that Chinese, which is a VO language, has the longest mean dependency distance in all the languages that have been investigated. More importantly, another study (Liu and Xu 2012) that has quantitatively investigated 15 different languages clearly suggests no correlation between dependency distance and head placement. These findings indicate that, for complex systems like language, it is too casual to draw a relation between them based on one single study.

Taken together, Futrell et al. intend to address the dependency length minimization as a universal quantitative property of human languages. However they do overstate the significance of their study: it is definitely not the **first** large scale evidence of DLM, but a repetition of some previous works, though with slightly different methods. Further, they do not include adequate non-cognitive factors in mind. Finally, this paper is impaired by a lack of systematic review and references to related studies mentioned above in particular and dependency grammar in general (Hudson 2010), and due to this lack, it is legitimate to question the originality of this study because it is largely dissociated and disconnected from previous findings.

Futrell et al. have potentially displayed an intriguing domain for large-scale cross-linguistic research on dependency distance. However, the methodology itself is basically a repetitive effort of previous studies, and the data presented are not sufficient enough to support the conclusions made in this paper. This work uses more languages than previous studies - probably thanks to the fact that much more dependency treebanks are available today than in the past. However, simply using more languages in the study is insufficient to amend the drawbacks mentioned above.

References

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

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

Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change 5(1), 114-137.*

Ferrer-i-Cancho, R., Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory 5(2), 143–155.*

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *PNAS 112(33), 10336-10341*. www.pnas.org/cgi/doi/10.1073/pnas.1502134112

Gildea, D., Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science 34, 286–310.*

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

Hudson, R. (2010) .*An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Jiang. JY., Liu, HT. (2015). The effects of sentence length on dependency distance, depend-

ency direction and the implications – based on a parallel English–Chinese dependency Treebank. *Lang. Sci. 50, 93–104.*

Liu, HT. (2007). Probability distribution of dependency distance. *Glottometrics 15,1–12.*

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

Liu HT, Xu, CS. (2012). Quantitative typological analysis of Romance languages. *Poznan Stud. Contemp. Linguist*. *48(4), 597–625.*

Glottometrics 33, 2016, 39-44

Response to Liu, Xu, and Liang (2015) and Ferrer-i-Cancho and Gómez-Rodríguez (2015) on Dependency Length Minimization

Richard Futrell, Kyle Mahowald, and Edward Gibson[1]

Abstract. We address recent criticisms (Liu et al., 2015; Ferrer-i-Cancho and Gómez-Rodríguez, 2015) of our work on empirical evidence of dependency length minimization across languages (Futrell et al., 2015). First, we acknowledge error in failing to acknowledge Liu (2008)'s previous work on corpora of 20 languages with similar aims. A correction will appear in PNAS. Nevertheless, we argue that our work provides novel, strong evidence for dependency length minimization as a universal quantitative property of languages, beyond this previous work, because it provides baselines which focus on word order preferences. Second, we argue that our choices of baselines were appropriate because they control for alternative theories.

Introduction

In recent work, we addressed the question of whether dependency length---the distance between syntactically related words in natural language sentences---is shorter than one would expect under random baselines (Futrell et al., 2015). This idea has linguistic relevance because if one hypothesizes a universal pressure to minimize dependency length, one can explain a variety of universal properties of languages, including many of the word-order universals noted by Greenberg (1963). Evidence that language users perfer word orders with shorter dependency length than chance supports this hypothesis, known as the dependency length minimization (DLM) hypothesis. The DLM hypothesis is theoretically attractive because it is motivated by general human information processing constraints: minimizing dependency length minimizes the online memory load for human sentence parsing and generation.

Two recent articles have raised important criticisms of our work (Liu et al., 2015; Ferrer-i-Cancho & Gómez-Rodríguez, 2015).

Random Trees and Random Word Orders

First, Liu et al. (2015) note correctly that we failed to cite a previous large-scale empirical study with similar aims. In particular, Liu (2008) compares average dependency length in attested sentences of 20 languages to dependency length in random trees. Not acknowledging this important prior work was an error on our part. The reason for this omission is that, in all honesty, we did not fully understand this paper and its relationship to ours until conversations with Liu and colleagues after publication. But these are not good reasons: we acknowledge that we should have made more of an effort to understand and acknowledge prior similar work. Consequently, we apologize and we urge anyone pursuing research relating to our

[1] Address correspondence to: {futrell, kylemaho, egibson}@mit.edu
Department of Brain and Cognitive Sciences Massachusetts Institute of Technology.

paper to also study Liu (2008). This prior work will be acknowledged in a correction to the PNAS article.

Nevertheless, we believe the difference between the Liu (2008) baselines and ours is non-trivial, such that our work represents new large-scale evidence for the DLM hypothesis. Liu (2008) uses a "random tree" baseline, comparing dependency length in attested dependency trees to dependency length in random ordered trees with the same numbers of nodes. For example, the dependency length of a sentence with a tree such as in Figure 1 is compared to the dependency length induced by random ordered trees as in Figure 2. The baseline trees do not share any syntactic structure with the attested trees they are compared to, beyond their length. In contrast, Gildea & Temperley (2010) and Futrell et al. (2015) use "random word order" baselines, keeping the syntactic dependency structure of attested sentences constant and investigating random word orders given that syntactic structure, subject to a number of linguistic constraints. For example, dependency length for a sentence such as in Figure 1 is compared to dependency length in a sentence with different word order but the same (unordered) dependency tree structure, as in Figure 3. Attested dependency length is shorter than both the random tree and random word order baselines.

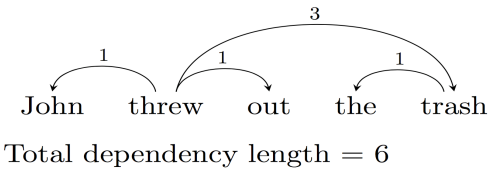


Figure 1. A possible sentence with its dependency tree and sum dependency length

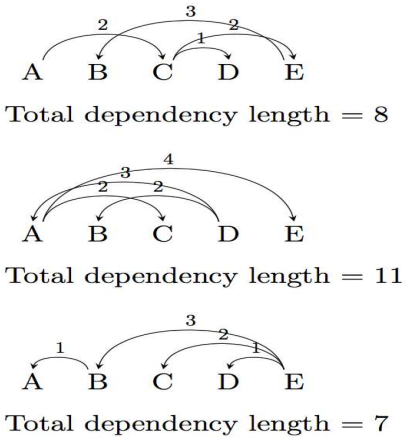


Figure 2. Some random trees based on the sentence in Figure 1 according to the Liu (2008) random tree baseline.

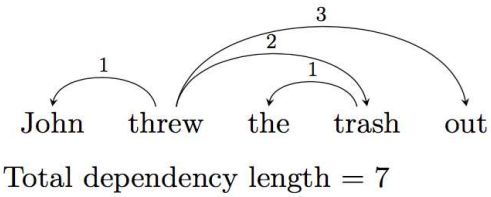


Figure 3. A random permutation of the sentence in Figure 1 according to a random word order baseline, specifically the head-fixed projective baseline in Futrell et al. (2015). This particular baseline permutes sister nodes while maintaining head direction.

Our finding that attested dependency length is shorter than random word order baselines shows that, *given* a syntactic structure, language users and language grammars tend to prefer the word order that minimizes dependency length. This finding supports the DLM hypothesis and provides direct evidence for a specific mechanism (word order preferences) by which dependency length minimization is accomplished.

On the other hand, the finding that attested dependency length is shorter than the random *tree* baselines supports the DLM hypothesis in a more general form and is consistent with many possible mechanisms that shorten dependency length, including non-syntactic mechanisms. For example, it is consistent with the idea that languages might disprefer structures which inevitably create long dependencies, such as high arity trees. It is also consistent with the hypothesis that language users prefer sentences with structures that create long dependencies, and might structure discourse to avoid such sentences. For example, the sentence (1) “A man who was wearing a hat arrived” has a long dependency between the subject “man” and the verb “arrived” because the relative clause “who was wearing a hat” intervenes between them. Language users might prefer to instead say (2) “A man arrived”, avoiding the relative clause between the subject and the verb, and perhaps mentioning the information about the hat in another sentence later in discourse, or perhaps dropping it altogether. Though language users are ultimately achieving the same or similar communicative goals in saying sentence (1) and sentence (2), they are doing so by expressing different propositional content in each sentence. The mechanisms by which dependency length minimization is accomplished in comparison to a random tree baseline are thus highly general: in addition to word order preferences, languages might have tree structure preferences; and language users might strategically choose *what content to express*, in addition to what word order to use, in order to avoid long dependencies.

In summary, comparing to random tree baselines can show DLM as a result of many mechanisms, including the content that people choose to express and/or the word orders they use in sentences. So the finding that attested dependency length is shorter than this baseline supports an influence of DLM on discourse structure or syntactic structure or both. Comparing to the random word order baseline, on the other hand, shows specifically that the word orders that people prefer, *given* the content they choose to express, are those that minimize dependency length. That is, it shows unambiguously that DLM as a pressure affects syntactic structure and word order in particular. Because our findings are *only* compatible with dependency-length-minimizing preferences in word order, we believe they provide novel, strong evidence for the DLM hypothesis as it pertains to syntax. Our claim is that, all else being equal, language users prefer linearizations with short dependency length. Only the comparison to a random word order baseline supports this claim unambiguously. So we see this work as a complement of Liu (2008) and related work, strengthening the body of evidence for the DLM hypothesis, rather than a repetition.

The difference between random tree baselines and random word order baselines can also explain some discrepancies between our work and previous findings. For example, we find relatively long dependency lengths for head-final languages such as Japanese and Turkish, whereas Hiranuma (1999) finds that dependency length in Japanese is highly optimized. Hiranuma (1999)'s finding is specifically that Japanese speakers drop verbal arguments to achieve dependency length minimization, trusting that the language comprehender will be able to infer the missing arguments from discourse context. Our finding is that, given the set of words and the dependency tree that Japanese speakers want to express, they choose orders with longer dependency length than, say, English speakers.(This finding remains unexplained.)

Projective Baselines

The second major issue raised in both Liu et al. (2015) and Ferrer-i-Cancho & Gómez-Rodríguez (2015) is our choice of baselines for comparison. We use projective linearizations, meaning that when a dependency tree is drawn over a linearized sentence, none of the arcs of the tree cross. We also use linearizations incorporating other factors that might conceivably influence word order: a pressure for fixed word order, and a pressure for consistency in head direction. These three factors---projectivity, head direction consistency, and fixed word order---all have the effect of reducing dependency length, and so it has been argued for the first two that they need not be considered separate factors, but rather the result of DLM. Ferrer-i-Cancho & Gómez-Rodríguez (2015) argue that our use of these baselines is redundant for this reason.

We believe comparison to these baselines provides stronger evidence for DLM than comparison only to a fully nonprojective baseline, because it shows that the phenomenon of short dependencies must be explained *even if* independent factors affecting word order are assumed. Since DLM can explain the phenomena attributed to these other factors, the most parsimonious theory seems to be that DLM is the only factor influencing word order. But we can only make this argument after showing that the shortness of dependencies persists as a phenomenon even after controlling for these other hypothetical factors. For example, suppose we had found that attested dependency length was *not* shorter than the projective random baselines[2]. One would be left with the question of why, if DLM is the main factor influencing language structure, German speakers pass up opportunities to minimize dependency length. Then one could argue that DLM is not a good explanation for projectivity, since word orders are not minimized for dependency length beyond what is needed to establish projectivity, which itself might have independent motivations (such as enabling polynomial-time parsing). Since we found that dependency length *is* shorter than this baseline in many languages, this line of argumentation is no longer available.

For the sake of completeness, we provide a comparison of attested dependency lengths with dependency lengths in random nonprojective linearizations in Figure 4. For this baseline, the dependency tree is linearized by shuffling nodes at random. The baselines from Futrell et al. (2015) are also shown. The figure shows that dependency length is much shorter than the nonprojective baseline, and that the projective baselines are much more conservative than the nonprojective baseline. We felt that including the nonprojective baselines in the original paper would be redundant, since Ferrer-i-Cancho (2006) showed that projective trees on average

[2] Which would not have been surprising given previous work: Gildea & Temperley (2010) found much weaker minimization in German than in English.

have shorter dependency length than nonprojective trees, and Kuhlmann (2013) (among others) showed that natural language dependency trees are overwhelmingly projective.

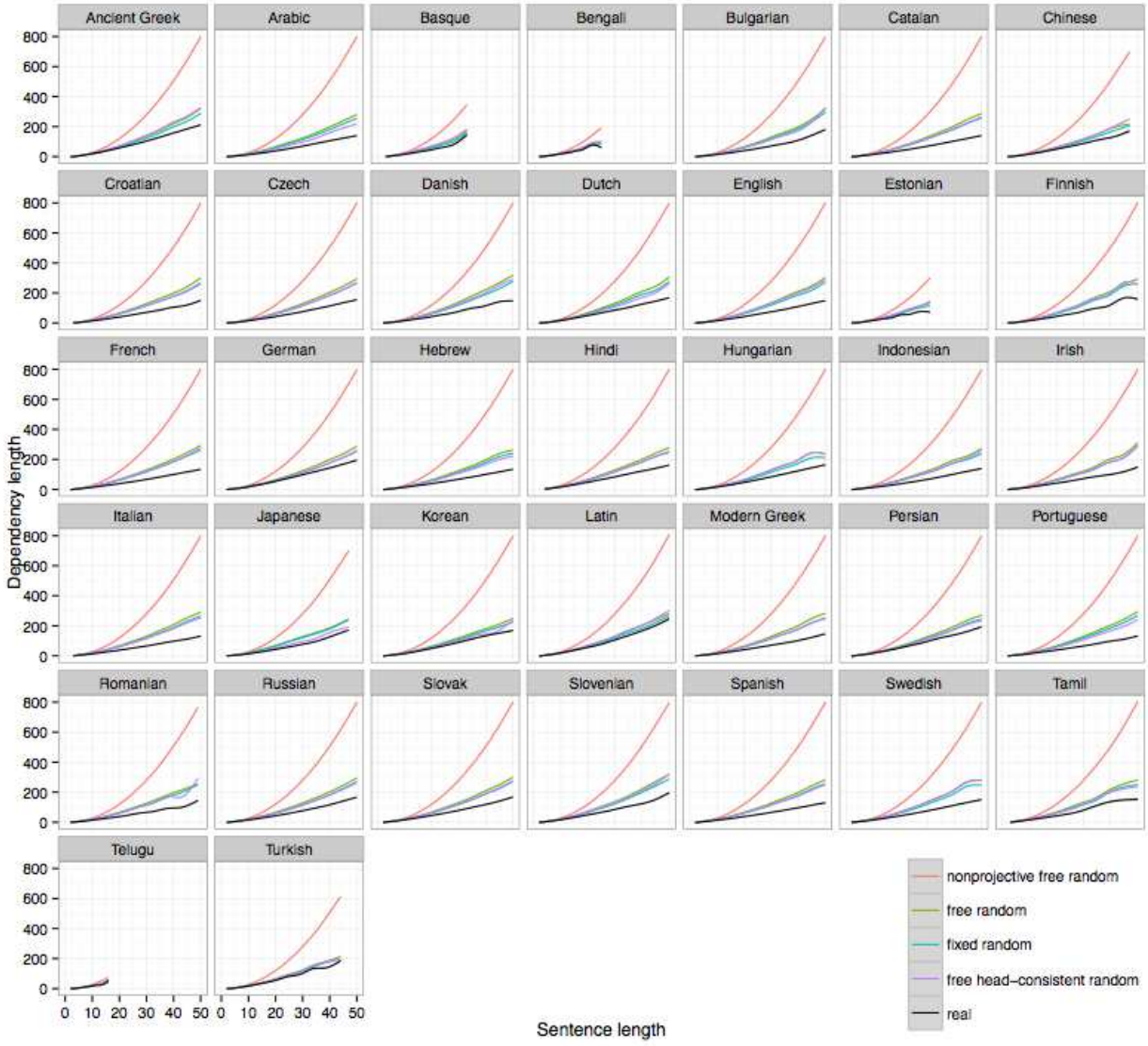


Figure 4. Dependency length as a function of sentence length, for real sentences (black), the free nonprojective baseline (red), and several baselines from the paper. All data except for the free nonprojective baseline were present in the original paper.

We also want to stress that, contra Ferrer-i-Cancho & Gómez-Rodríguez (2015), controlling for these possible alternative factors affecting word order does not imply that we are accepting traditional nativist or Universal Grammar-based hypotheses. These factors have possible functional explanations, just as DLM does. Fixed word order can be motivated by efficient communication of relation types; consistent head direcion can be motivated by compression of grammars; and projectivity can be motivated by the time complexity of parsing, where parsing to projective trees is cubic-time but parsing to fully nonprojective trees is NP-hard. In general, we aimed to include the most conservative reasonable baselines.

Other Issues

Liu et al. (2015) also raise a number of more specific criticisms. They claim that the uniformity of genres of the text in our corpora could be a confounding factor. The criticism is valid: It is true that our corpora were primarily (but not entirely) written text from newspapers and novels. Nevertheless, we would find it surprising if DLM universally influenced novels and newspapers but not language use in general. We welcome any work which controls for this possible issue.

Finally, Liu et al. (2015) also note that in our original paper we state that head-final languages appear to have longer dependencies than more head-initial or head-medial languages, but we do not provide statistical tests of this claim. We intended this remark not as a main claim of the paper, but as a conjecture intended to draw attention to the wide variation between languages in their dependency length, and possible typological implications of that variation. Working out the correct statistical methodology and gathering the right data to make this a strong empirical claim would require another whole paper. The question of variation in dependency length has also been a major focus of Liu's research. We feel that explaining this variation is the most interesting direction for future dependency length research, and we hope to join our present critics in future investigations of this phenomenon.

References

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters 76(6), 1228.*

Ferrer-i-Cancho, R., & Gómez-Rodríguez, C. (2015). Liberating language research from dogmas of the 20th century. *arXiv, 1509.03295.*

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), *10336–10341.*

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science 34(2), 286–310.*

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In: J. Greenberg (ed.), *Universals of Language:* 73–113. Cambridge, MA: MIT Press.

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

Kuhlmann, M. (2013). Mildly non-projective dependency grammar. *Computational Linguistics* 39(2), 355–387.

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

Liu, H., Xu, C., & Liang, J. (2015). Dependency length minimization: Puzzles and Promises. *arXiv, 1509.04393.*

Glottometrics 33, 2016, 45-55

How Statistics Entered Linguistics: Pierre Guiraud at Work. The Scientific Career of an Outsider

Gabriel Bergounioux

University of Orléans

1. Introduction

Looking back, Pierre Guiraud (1912-1983) stands out conspicuously from the rest of the French academic world. His career, his work and his chosen topics pioneered a novel conception of how computation could be applied to linguistics. This approach was not understood in his time by French academics, perhaps due to the fact that he was the only humanities scholar to venture into a field that had been largely pre-empted by mathematicians (see Hérault & Moreau 1967), even though, motivated by natural language processing, mathematicians focused on parsing rather than on statistics, as did Maurice Gross for example in the same issue (Gross 1967). Of course, one has to take into account both the internal hierarchy in mathematics, where statistics were ranked low on the scale amid Bourbaki's logicist conceptions, and the desire to differentiate computer science in its early stages from electronics. As a matter of fact, despite Guiraud's copious production (eighteen books) in the famous paperback encyclopaedia collection "Que sais-je?", he never wrote one on the topic he knew so well, quantitative linguistics.

1. A short biography

Pierre Guiraud was born in Sfax (Tunisia) on September 26th 1912 and died on February 2nd 1983. His mother quickly divorced and when she died in Paris, a few years later, the young orphan was raised by two aunts in Genolhac, a small village located in Gard (south of France). He moved to secondary school in Alès and was awarded his undergraduate degree (*licence de lettres*) in Montpellier in 1934. He held a position as a teacher in Aubusson (Creuse) and Chatellerault (Vienne). Lacking the requisite qualifications (*agrégation*) to be a secondary school teacher in France, he accepted a position abroad as French language assistant in Chisinau (Romania) in 1939. Meanwhile, he joined the British Intelligence Service where he was promoted, at the end of the war, to the rank of colonel and received the D.S.O. for his action. When Chisinau and all the territory east of the River Prut (eastern Moldova) were occupied by the Soviet Union in June 1940, in accordance with the German-Soviet Pact signed in August 1939, Guiraud was repatriated to Bucharest (Romania) where the Vichy government had set up a secondary school. The "lycée français" was closed in June 1941 when Romania entered the war on the side of the Axis powers. From 1943 to April 1944, Guiraud was employed as a French language teacher in Hungary where he acted as a spy for the United Kingdom. Back in Bucharest, he was immediately arrested by Antonescu's police. In August 1944, Marshal Antonescu was toppled and, as the country joined the Allies, Guiraud was released and returned to France.

Since his initial academic studies did not allow him to obtain a position in higher education in France, he took up a position as a lecturer at the University of Swansea at the end of the 40s where he prepared his doctoral thesis (a Higher Doctorate, or "doctorat d'état", involving much more extensive research than a current PhD) to apply for a position as professor. He became a professor at Groningen (the Netherlands) and, following a reform of the legal framework in France, at Nice (1964) and also taught as a visiting professor at Bloomington in the same years. He spent the remainder of his academic career at the University of Vancouver until his retirement (August 1978) in France.

As neither a former student of the École Normale Supérieure, nor an "agrégé", Guiraud was considered an outsider as were, in those days, A. J. Greimas or Roland Barthes (on this topic, see the interview with Greimas in Chevalier & Encrevé 2006), and he failed when he applied for a chair at the Sorbonne, despite an attempt to portray himself as a follower of Charles Bruneau by dedicating his thesis to him. At that time, Bruneau held the only French Language chair, established for Ferdinand Brunot (Bruneau's former teacher) at the beginning of the 20th century. But Bruneau had not kept pace with new trends in linguistics and Guiraud's remoteness was not on his side, despite the encouragement of Robert-Léon Wagner (1905-1982), an acknowledged grammarian of the Sorbonne and the École Pratique des Hautes Études.

2. Linguistics in France: a policy of containment towards statistics

For a long time, the French syllabus in the universities was dominated by literary studies but nonetheless made a B.A. dependent on acquiring specialized knowledge. Undergraduate studies were divided in four parts: the least important one (aka "the 4th certificate") because it was a technical one and not an aesthetic one, was the "certificate of grammar and philology (= Old French)", of which a small part was devoted to stylistics. This organization had been decided during the 1870s, the starting point of an academic structure designed on the German model, a process completed in 1896 and retained until it was updated in the 1960s (Bergounioux 1998).

At first glance, it seems that Guiraud missed his aim three times in his career:

(i) When he tried to renew the stylistics studies of his time by means of statistics, during the 50s and 60s, an approach that was deemed unacceptable before the major reforms of higher education;

(ii) When he proposed a new deal in linguistics where stylistics and semantics would play a leading role. Despite the special orientations given by Benveniste and Martinet in general linguistics, by Ducrot in semantics and by Jakobson, Mounin, and Ruwet in poetics, he remained outside the scope of the new trends, however;

(iii) Lastly, when he studied etymology in connection with semiology. As he was more interested by Lazare Sainéan's, Lucien Tesnière's, or even Gustave Guillaume's working hypotheses, he remained isolated, far from the functionalist and generativist schools then prevailing.

Nevertheless, when defending his doctoral thesis, Guiraud had the opportunity to adopt a stance on stylistic questions, in particular on an internationally renowned poet, Paul Valéry (1871-1945). But his method was original. Although he had signed a contract with the Éditions du Seuil to write an academic literary study entitled *Valéry par lui-même*, in his thesis *Langage et versification d'après l'œuvre de Paul Valéry* (1953) [Language and versification based on Paul Valéry's work] he did not deal with any biographical topics but

devoted himself entirely to formal questions of literary work, in particular metrics and sound symbolism.

The positioning of this research differed from the approach of mathematicians who favoured logical formalisms in which poetic and lexical studies were discarded in favour of syntax and phonology. Nor did Guiraud's recourse to the enumeration of tokens tally with the survey conducted during this period for the definition of "Français Fondamental" [Basic French] (Gougenheim *et al.* 1956/1964). Although both these initiatives appeared to converge, almost to the year, in introducing word counts into language sciences, the differences between them are very great. First, Basic French concerned non-literary language. Based on an oral survey, it focused exclusively on spoken, even colloquial French. Second, as its objective was the teaching of French, especially French as a foreign language, this led to the preparation of dictionaries and textbooks published by an educational publisher (Didier). Guiraud, in contrast, undertook a very ambitious analysis of an author who is notoriously difficult to understand. His study was published in the highly ranked collection "Linguistique" of the Société de Linguistique de Paris. A significant fact, pointed out by the lexicographer Alain Rey, was that:

> From his beginnings, by his very conception of syntax and stylistics, and his constant interest in quantifiable formal features – Guiraud was one of the main introducers of language statistics in France – he sought to reconcile and articulate the essential forces that are at work in language and more broadly in semiosis (Rey 1985: 48).

In the introduction to his doctoral thesis, Guiraud justified his approach as follows:

> I must now say a word about the method. I had always thought it would be interesting to count all the components of a text until all the possible combinations had been exhausted (...). As I progressed in this direction I rapidly acquired the certainty of being on the right track. It seemed to me more and more that every style corresponds not to a purely quantitative definition but rather to a standard deviation from a norm (...). In summary, three guides should help the reader to navigate through this essay: (...).
> 3° a statistical analysis of these problems; and the claim that literary expression and style are "standard deviations" which justify our analytical method. (Guiraud 1953: 15-17)

3. The use of statistics: seeking scientific certainty in the humanities

While the end of the introduction to the thesis was addressed to all the lovers of pure literature who would not appreciate the book, Guiraud first explained how he was led to use statistics:

> The analysis of my predecessors' innumerable studies, however, suggested some doubt about the value of my original project, as most of these studies seemed to me very fragile. The analysis of a standard deviation presupposes the establishment of a standard and a measurement system. Soon I felt lost in the complexity and mystery of numbers and turned for a time to mathematics. This research resulted in two studies currently in press: one is a bibliography of statistical linguistics that contains an analysis of nearly two thousand books and papers on the topic and a discussion of the applications of statistics to problems of language; the other is an attempt to analyze the statistical characteristics of vocabulary. I tried to address the issue with as much mathematical

> rigor as I could. I provide – from a theoretical viewpoint in the first study, and a pragmatic one in the second – the qualitative value, the limits and the conditions of application of statistics in the analysis of language (Guiraud 1953: 16).

It is no small paradox that counting was required by the analysis of poetry, not in terms of the number of syllables as usual but in terms of words or of phonemes. In the summary of the book statistics are mentioned, apart from the introductory and the concluding parts, in the following chapters:

Ch. II "Rhythm"
> Statistical study of the frequency of mute *e* which proves that this rate is abnormally high for some poets (...) Valéry has the highest frequency of mute *e* among all our poets (58 sq.)

Ch. IV "Rhyme"
> Statistical analysis of rhyming dictionary (108-109)
> Identical rhymes. Statistical review (115-117)
> Frequency of isometric rhyming words (124)

Ch. VII Extension of meaning
> Valéry's high frequency of derived words (179-180)

While the importance accorded to statistics may seem slight, there are numerous other accounts in percentages and a roster of statistical tables enumerates 17 frequency distributions.

Although the main innovation of Guiraud's doctorate was the use of statistics, among the two hundred items listed in the bibliography there are only two explicit references, both of them to Zipf's work. One is under the heading "Phonetic and phonological system of the French language" where the book *The Psycho-Biology of Language* (Zipf: 1935) is incorrectly cited as "*Psychology of Language*", the other under the heading "Vocabulary and syntax: parts of speech" in which "Human behavior and the principle of least effort" (1949) is mentioned.

4. An example of literary study in the light of statistics: Apollinaire (1953)

In 1953, Guiraud published (in French) his *Index of the Vocabulary of Symbolisme I. Index of the Words of Alcohols by Guillaume Apollinaire*. In his foreword, Wagner draws attention to the difficulties which had arisen with phonetics and semantics and he emphasizes the results obtained by linguistic statistics applied to literature, and which complement the survey conducted by Gougenheim *et al.* on spoken French. Quoting Eluard, Wagner highlights the specificity of stylistic devices in modern poetry, even if he regrets the lack of a table of rhymes.

As Wagner points out, the original idea behind this program was shared by a few linguists:

> Fortuitously and independently, without knowing each other, Mr Pierre Guiraud and I were following the same path. A chance encounter led us to work together; first, to correct our mutual prejudices. Statistics can be off-putting and it took me some time to convince Mr P. Guiraud that his tables and his calculations could find, so to speak, a literary application. After discussing matters on an equal footing, I can say – I believe in

> both our names –, that as long as there are more indexes, they will from now on more conveniently meet the needs of readers for whom they have been written. (p. III-IV)

The book is a short, 29-page monograph, with half a page to explain how the lemmatisation had been done, one page for theme-words and one more for key-words, and one and a half pages for POS distribution. The remainder of the book is an alphabetical list of words with an asterisk preceding words which are not on Van der Beke's list (1929). Unsurprisingly, these words are proper names, poetic words (Apollinaire had a special liking for them, some of which are unknown even to French readers, such as *dulie* or *sistre*), non lemmatised words and compound expressions. Nevertheless, one can note that Van der Beke had omitted *ciseaux* (scissors), *médicament* (drug), *voisin* (neighbour) and… *vocabulaire* (vocabulary).

5. Guiraud as a reference in statistical linguistics: counting and techniques

So, forsaking literary studies as they had been practised previously, Guiraud adopted a quantitative approach. During this period, he published a series of indexes to prepare the ground for an inventory of the vocabulary of the Symbolist poets (1953-1954 and 1960a) and, with the assistance of Robert W. Hartle, of Jean Racine's tragedies (the general title of the series was "Great seventeenth-century French dramatists", but in fact only Racine was analyzed), with the support of R.-L. Wagner. The data obtained by such painstaking and tedious compilations did not result in a lot of papers. A compilation of nine of them (Guiraud 1969) out of a total of thirty gives a single reference in the table of contents to "statistics", in the chapter: "Language and style: form".

One year later, in an anthology co-authored with P. Kuentz, statistics was again mentioned only in passing. Guiraud just quoted a text by Dolezel when presenting the statistical theory of poetic language (1970: 62-4) before introducing his own work (1954a) on the opposition between *theme-words* (the words most frequently used by an author) and *keywords* (the words whose frequency deviates from the normal range in an author) (1970: 222-4).

At the same time, he conducted a comprehensive and up-to-date database of bibliographical references (1954b) as a result of the decision taken at the sixth Congrès International des Linguistes [International Congress of Linguists] in Paris, to establish a committee for linguistic statistics to investigate what has been published. For this second title in the series, Guiraud supervised a team comprising Joshua Whatmough, Thomas D. Houchin, Jean Puhvel, and Calvert W. Watkins, all from the Department of Comparative Linguistics at Harvard University. While it is strange that one of the most inventive and creative linguists of his generation spent ten years as a researcher compiling a bibliography and counting tokens in literary texts (even if some of these tasks were done by his wife), we can consider that it is the price he had to pay to compensate for his lack of academic qualifications.

Meanwhile, Guiraud wrote a short methodological essay of 116 pages, entitled "The statistical characteristics of vocabulary", dedicated to R.-L. Wagner and published in 1954. Two thirds of the book are devoted to "The distribution of words", the last third to "The lexicon of poetry". The last part applies the theoretical principles outlined in the early chapters and it is exemplified by the Symbolist poets' vocabulary. Quoting Henmon (1924) at the very beginning, Guiraud followed in the footsteps of pioneering studies and complied with the guidelines of the "Français Fondamental" program, with which he was never associated: in the bibliography of Gougenheim *et al.* (1964), for example, Guiraud is referred to only once, versus ten references to René Michéa on related topics.

Now let's look at the first paragraph of the foreword, entitled "Language and numbers":

> Any language event can be defined by its frequency in discourse; between this frequency and all its psycho-physical characteristics, constant and strict relationships are established. Linguistics, which studies the elements of sounds and their mutations, the structures of grammatical forms, the meanings of words and the mechanism of changes which transform them, generally ignores one of their most important and most significant features: frequency. (Guiraud 1954a: 1)

If we have a closer look at this excerpt, we can see that there are two differences with the philological tradition and also with Saussure's theory embraced by Wagner and also by Guiraud. Instead of the *langue/parole* (language/speech) distinction, Guiraud employed the word *discours* (discourse) which was not commonly used in French linguistics at the time (it was to become widespread in the 1960s). Admittedly, he was influenced by English terminology. Moreover, he did not confine himself to lists of words but he included in his work the three main linguistic domains (phonology, morpho-syntax and semantics) and the two approaches, synchronic and diachronic. The use of statistics was therefore both an improvement in the definition of the scientific object of study (*discours* instead of *parole*) and an advancement of the method.

The book was primarily intended for linguists even if it established a link between lexicography and stylistics. Thus after a presentation of Zipf – reiterated in a short paper to the *BSL* (Guiraud 1955b) – he devoted a few pages to Yule (1944) in order to preserve the relationship to literary studies, but apparently this attempt at conciliation convinced neither linguists nor professors of literature. A conclusion to this research resulted in (Guiraud 1960b) where he tried to go beyond the aims of a method, by taking into account the difficulties entailed by using statistics.

Problems and Methods of Statistics in Linguistics (1960)

Except for three subsequent papers, this book was Guiraud's last contribution to the topic. A brief foreword outlines the plan, divided in two parts: five chapters deal with “method”, and seven chapters with “problems”, most of which are reprinted or revised articles. Chapter one lists ten areas to which linguistic statistics can be applied: (i) methodology; (ii) phonetics (= phonology); (iii) metrics and versification; (iv) indexes and concordances; (v) lexical distribution and frequencies; (vi) semantics; (vii) morphology; (viii) syntax; (ix) child language; and (x) philology. This broad coverage makes it clear that the implementation of statistics can reorganize linguistics at large.

A wide variety of areas are itemized and the key authors are mentioned. In methodology, following Herdan (1956) and Miller (1951), Guiraud enumerates the following authors:

> While our field may claim the patronage of the greatest names in linguistics, Whitney, Reinach, Riemann, Gaston Paris, Saussure, Troubetzkoy, it was not before the 40s that it became aware, thanks to Zipf, Yule, and Ross, of the possibilities of an analysis based on a rigorous methodology. Until then we had quantitative linguistics but which could not be called statistical linguistics (Guiraud 1960b: 6).

In chapter 2, “Postulates and limits of the method”, Guiraud characterizes linguistics as an observational science grounded on statistics, like sociology or economics:

> Linguistics is the typical statistical science; while statisticians are well aware of that, most linguists are still unaware of that fact. This is because the separation between

> literary and scientific disciplines limits the number of researchers who can address aesthetic issues using fairly complex mathematics (...). (*Ibid.*: 15)

He further assumes that there is a cognitive substructure underpinning this phenomenon:

> [These facts] allow us to imagine language as a sum of the mental images that exist objectively in the speaker's brain in the form of marks or engrams in memory. What is more, it can be plausibly argued that each sign is present together with its frequency. In this way, there are as many engrams as the number of times that the word has been received and the frequency of the sign, far from being an accident of speech, is an objective attribute of the language that is just as important as its form or its meaning. Under this assumption – which is confirmed more strongly every day – any speech or text can be considered to be a sample of a linguistic state that reflects its numerical structure as well as the possibilities of its semantic performances. (*Ibid.*: 17-18)

In the original text, there are two occurrences of “ingramme” instead of “engramme”, a word coined by the German psychologist Richard Semon in 1904, and translated into English and French (Larousse dictionary, 1932). This probably means that this odd spelling is patterned after the American one, perhaps after Miller’s books. Then, Guiraud says, five difficulties are encountered: (i) the qualitative dimension of language; (ii) the distortions of measurements performed on speech, not on language; (iii) the heterogeneity of data; (iv) the complexity of language, and (v) the size of the problem, which is an obstacle to data processing. On the last point, Guiraud predicts an increasing use of electronic machines and he mentions, as an example, what was being carried out at MIT.

Chapter three is a re-issue of (Guiraud & Wagner 1959) with an unexpected psychological incursion into characterology (probably inspired by McCormick (1920) more than by Le Senne (1945)):

> The real problem is the characterology of the language. That is to say, we must begin by defining a method similar to the method of anthropology or of graphology, a kind of linguistic bertillonnage [from the Bertillon system]. It is questionable whether this is possible. (*Ibid.*: 27)

Three core issues are discussed: the genealogical relationship of languages (without considering linguistic typology), linguistic chronology and, with respect to literature, authorship attribution. Even though the aim assigned to statistics is to take linguistic tasks beyond description and classification to a science of causes, the paper concludes with a definition of the general principles of quantitative stylistics.

Chapter four, “Statistical analysis (how to describe)”, is a presentation for dummies (i.e. linguists) of statistical method, especially the use of tables. Chapter five, “Statistical analysis (how to interpret results)”, is a continuation of the previous chapter, distinguishing between quantitative linguistics and statistics in linguistics, in contrast to Grammont’s claims (1923). Chapter six, “Language and information” is a re-issue of an article first published in the *Journal de Psychologie* (1958). The seventh chapter, “Estoup-Zipf Equation and information substrate in verbalisation” links statistics and information and quotes the stenographer Jean-Baptiste Estoup’s proposal (1912), as a precursor to Zipf, and Mandelbrot (1961) on the statistical interpretation of data.

Chapter eight, “Estoup Zipf Equation and statistical characteristics of vocabulary” begins with two considerations regarding word status (“word definition is not relevant in practice”) and mental projection (“the vocabulary of a text reflects the mental lexicon from

which it has been drawn"). On the second point Guiraud expresses a difference of opinion with Mandelbrot:

> Mr Mandelbrot thinks that distribution is a characteristic of the vocabulary of the text and has a constant slope for this text. I think that the distribution is a characteristic of the lexicon of the text, that is to say a characteristic of all the words from the memory storage of which the words of the text are derived. (*Ibid.*: 87)

Sampling requires particular attention to the number of words, especially for pedagogical purposes (there are recurrent references to Gougenheim *et al*. 1956), since there is an inverse relationship between the frequency of a word and the quantity of information that may be deduced from it.

Chapter nine, "Distinctiveness structure and statistical distributions of phonological systems", correlates the distinctive features with the frequency of phonemes, in an attempt to compare the viewpoints of Zipf and Martinet or Haudricourt. The linguistic changes that have taken place from Latin to modern French are scrutinized, a reflection pursued in chapter ten, on the effects of loanwords: "Loanwords and phonological balance", written as a tribute to Walther von Wartburg and first published in the *Zeitschrift für Romanische Philologie* (1958). Foreign words, by introducing distortions in the phonotactic and statistical distributions, allow the assignment of semantic values to a certain number of sound concatenations, for example, says Guiraud, "KA- has a negative connotation in many words; B- contributes to creating many onomatopoeic words" (*Ibid*.: 123). This suggestion will guide his further enquiries into phonosemantism.

Chapters eleven and twelve conclude this book, dealing with "The evolution of Rimbaud's style and the chronology of *Illuminations*" and "The phonetic structure of verse", i.e. stylistics and metrics. There is neither a conclusion, nor a bibliography.

This book is in some respects the acme of Guiraud's work on statistical linguistics. Compared with the ten subdivisions of the initial enumeration (see above), we can note that methodology takes the lion's share (chapters I to V). Overall, phonetics is covered in chapters IX to X, metrics and versification in chapters XI to XII (placed at the end of the book, in spite of the fact that they were at the beginning of the list), indexes and lexical distribution in chapter VIII, semantics in chapters VI to VII, basically grounded on information theory. There is no part devoted specifically to the other topics (morphosyntax, child language, philology) and no special discussion of language training or didactics which pioneered the work in this field.

In assessing the points at issue, besides an open-mindedness with respect to new trends in psychology (characterology) and mathematics, Guiraud returned to his initial subject of interest: literature; but he pointed in the direction of two new topics, word characterization and semantics.

6. How to quantify what is uncountable? From disambiguation to metaphor

A recurring problem in the field of linguistic statistics could be worded as follows: how can one count lexical units or tokens which are identical in appearance (the same character strings) but that fall into different categories? For example, rather than lemmatisation, which requires additional processing, Guiraud dealt with the question of French *locutions* (fixed expressions or chunks) in his book on the theme (1960c). In a phrase, each word, defined as a

cluster of letters between two blanks, should not be counted separately but as a whole, as a macro-unit. So the same token can be classified in two different ways. The same problem occurs with homonyms, especially homographs, which must be distributed under different headwords.

This question was first approached by Guiraud through the example of slang (1956a) and the concept of "morpho-semantic field", coupled with etymology (*BSL*, 1956b), a path undertaken much earlier in *Valéry* (1953) about sound symbolism (131-150). This transfer of an infra-lexical semantic level is developed, for the first time, in a systematic and comprehensive way, in "The morpho-semantic field of the root T.K." (*BSL*, 1963c) and later in *Le Français Moderne* (1966). In 1967, in his masterpiece *Structures étymologiques du lexique français*, Guiraud synthesizes the findings and deals with issues relevant to the etymological structure of the French lexicon. The observed regularities induced a form of statistical determinism and thereby, the idea that it was possible to predict meaning on the basis of a purely phonetic assessment. Some basic combinations of phonemes (consonants mainly) in specific fields based the principles of etymology on particular sound sequences, by means of a consonant frame. Unlike conceptual metaphor, the sounds organize the content. So, he shares the views of other authors, ranging from Le Senne's and Berger's characterology to Lacan's conceptions, on psychoanalytic issues.

Over the years, Guiraud's thinking on the role of statistics in linguistics had evolved. By the late 1960s, he no longer envisioned the statistical approach as a merely quantitative computation but as an intuitive recognition of the link between the distribution of the letters in a text, or in a list of words, and its global signification. To a certain extent, it was still a matter of quantitative linguistics but it was no longer a matter of statistical linguistics. And even in stylistics, when Guiraud attempted to follow in the footsteps of his predecessors and continued to build on the heritage left by Bruneau, after having distanced himself from Marouzeau or Cressot because he was a lot more interested in Bally's and Spitzer's work, the time had now come for analysts such as Barthes, Kristeva, Todorov or Genette to prevail.

Conclusion

Despite his position as leader in the field of statistical linguistics, and his pioneering work, Guiraud never received the recognition he deserved. Working far from Paris, even outside France until 1964, without the academic qualifications expected of a professor at the Sorbonne, he was trapped by his inability to respond to changing circumstances. Linguistics and literature, that he had always attempted to reconcile, had become two distinct and quite antagonistic domains in the universities and his broad professional network seemed to be outdated at a time when new linguistic schools sprang up. His sole contribution to Martinet's guidebook "Language" in the famous "Encyclopédie de la Pléiade" is truly symbolic: "The secondary functions of language".

There was no room for him in French linguistics in this period. Neither before the 50s for academic reasons, nor during the 50s and 60s, when the confrontation between Benveniste and Martinet had split the field into two factions, nor since the 60s when the generativists (Ruwett), the harrissians (Dubois), the "énonciativistes" (Culioli) and the semanticians (Ducrot) discussed guidelines for phonology, syntax and semantics, not for lexicology or statistics. Even poetics was, at the time, controlled by Jakobson, Ruwet, and Mounin and prosody by Meschonnic or Roubaud. Although Guiraud dedicated his 1967 book to "Hjelmslev, Guillaume, Jakobson, Benveniste, and Martinet", he remained alone, without any successors. Through this position of outsider, however, his professional career sheds light on the conditions in which French quantitative linguistics emerged.

References

Guiraud, P. (selected bibliography)

(1953). *Langage et versification d'après l'œuvre de Paul Valéry*. Paris: Klincksieck.

(1953-1954). *Index du vocabulaire du symbolisme*, avant-propos de R.-L. Wagner. Paris: Klincksieck.

[1. Index des mots d'« Alcools » de G. Apollinaire; 2. Index des mots des poésies de P. Valéry; 3. Index des mots des poésies de S. Mallarmé; 4. Index des mots des « Illuminations » d'A. Rimbaud; 5. Index des mots des « Cinq grandes odes » de P. Claudel; 6. Index des mots des « Fêtes galantes », de « La Bonne chanson » et des « Romances sans paroles » de P. Verlaine]

(1954a). *Les Caractères statistiques du vocabulaire*. Paris: PUF.

(1954b). *Bibliographie de la statistique linguistique*. Utrecht-Anvers: Spectrum.

(1954c). L'évolution statistique du style de Rimbaud et le problème des *Illuminations*. *Mercure de France 322, 201-234*.

(1954d). *La Stylistique*. Paris: PUF.

(1955a). *La Sémantique*. Paris: PUF.

(1955-1964). *Index du vocabulaire de la tragédie classique*. Paris: Klincksieck.

(1955b). A propos des caractères statistiques du vocabulaire et de l'équation de Zipf. *Bulletin de la Société de Linguistique de Paris LI(1), 236-239*.

(1956a). *L'Argot*. Paris: PUF.

(1956b). Les champs morpho-sémantiques. Critères externes et critères internes en étymologie. *Bulletin de la Société de Linguistique de Paris LII(1), 265-288*.

(1958). Langage, connaissance et information, *Journal de Psychologie Normale et Pathologique, juillet-septembre, 302-318*.

(1958). Emprunts et équilibre phonologique. *Zeitschrift für romanische Philologie, 74(1-2), 78-88*.

(1960a). *Index du vocabulaire du symbolisme*. Paris: Klincksieck. [7. Index des mots d'« Une saison en enfer » de Rimbaud]

(1960b). *Problèmes et méthodes de la statistique linguistique*. Paris: PUF.

(1960c). *Les locutions françaises*. Paris: PUF.

(1963a). La mécanisation de l'analyse quantitative en lexicologie. *Etudes de Linguistique Appliquée 2, 33-46*.

(1963b). Structure des répertoires et répartitions fréquentielles des éléments de la statistique du vocabulaire écrit. *Communications et Langage, 37-48*.

(1963c). Le champ morpho-sémantique de la racine T.K. *Bulletin de la Société de Linguistique de Paris LIX(1), 135-155*.

(1965). Diacritical and statistical models for languages in relation to the computer. In: D. Hymes (ed.), *The Use of Computers in Anthropology [1962] 235-254*. La Haye: Mouton.

(1966). De la grive au maquereau : le champ sémantique des noms de l'animal tacheté, *Le Français Moderne (octobre), 280-308*.

(1967). *Structures étymologiques du lexique français*. Paris: Larousse.

(1969). *Essais de stylistique*. Paris: Klincksieck.

Secondary sources

Augé, P. (1932). *Larousse du XX^e^ siècle*, Paris: Larousse.

Bergounioux, G. (ed.) (1998). Un siècle de linguistique en France. 1. Institutions et savoirs. *Modèles linguistiques XIX(2)*.

Bouton, Ch. (ed.) (1985). Hommage à Pierre Guiraud. *Annales de la Faculté des Lettres et Sciences Humaines de Nice 52*. Paris: Les Belles Lettres.

Chevalier, J.-C. ; Encrevé, P. (2006). Combats pour la linguistique, de Martinet à Kristeva: essai de dramaturgie épistémologique. Lyon: ENS Editions.

Estoup, J.-B. (1912). Gammes sténographiques. Paris: Institut sténographique.

Gougenheim, G.; Michéa, R.; Rivenc, P.; Sauvageot, A. (1964). L'Élaboration du français fondamental. Paris: Didier. [First printed as L'Élaboration du français élémentaire, 1956]

Grammont, M. (1923/1936). Le vers français. Ses moyens d'expression, son harmonie. Paris: Champion.

Gross, M. (1967). Linguistique et documentation automatique. *Revue de l'enseignement supérieur 1-2, 47-55*.

Henmon, V.; Allen, Ch. (1924). *A French Word Book based on the Count of 400,000 Running Words*. Madison: Bureau of Educational Research, University of Wisconsin.

Hérault, **D.; Moreau, R.** (1967). La linguistique quantitative. *Revue de l'enseignement supérieur 1-2, 113-127*.

Herdan, G. (1956). *Language as choice and chance*. Groningen: P. Noordhoff N. V.

Kuentz, P. (1970). *La Stylistique : Lectures*. Paris: Klincksieck.

Le Senne, R. (1945). *Traité de caractérologie*. Paris: PUF.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In: *Structure of Language and its mathematical aspects*. Providence: American Mathematical Society.

McCormick, L. H. (1920). *Characterology: An Exact Science Embracing Physiognomy, Phrenology and Pathognomy, Reconstructed, Amplified and Amalgamated, and Including Views Concerning Memory and Reason and the Location of These Faculties Within the Brain, Likewise Facial and Cranial Indications of Longevity*. New York - Chicago: Rand McNally.

Miller, G. A. (1951). *Language and Communication*. New York - London: McGraw Hill.

Rey, A. (1985). [Obituary]. In: Ch. Bouton (ed.) *Annales de la Faculté des Lettres et Sciences Humaines de Nice 52: 47-49*. Paris: Les Belles Lettres.

Ross, A. S. C. (1950). Philological Probability Problems. *Journal of the Royal statistical Society 12 (B): 19-59*.

Semon, R. (1904) *Die Mneme als erhaltendes Prinzip im Wechsel des organischen Geschehens*. Leipzig: Engelmann.

Vander Beke, G. E. (1929). *French Word Book*. New York: Macmillan.

Wagner, R.-L. (1959). La méthode statistique en lexicologie. *Revue de l'Enseignement Supérieur 1, 154-159*.

Yule, G. U. (1944). *On the Statistical study of Literary Vocabulary*. Cambridge: CUP.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Cambridge (Mass.): Harvard University Press.

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

Glottometrics 33, 2016, 56-72

Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis

Valérie Beaudouin
Télécom ParisTech, I3 (UMR 9217)
valerie.beaudouin@telecom-paristech.fr

1. Introduction

While the dream of artificial intelligence (AI), of a machine capable of dialoguing in a natural language, of understanding texts and so of generating them, or even of translating them, has run up against a wall, inductive approaches for the exploration of texts have been developed, with lower theoretical ambitions but greater efficacy. The purpose of such approaches is to identify phenomena and regularities in a corpus of texts and to infer laws from them.

A discourse, or text, being the raw material of numerous human and social sciences, this current has not been restricted to a particular discipline, such as linguistics. These methods have been, and still are, widely used in many different disciplines.

From the 1960s to the 1990s, long before "text mining" became fashionable, France witnessed an exceptionally active period in the field of automated text analysis, exploiting the new affordances provided by IT: digital corpora, statistical algorithms and computing power.

A research field in this territory has grown up, with its laboratories, academic journals, reference books, symposiums, internal controversies, and currents… It brings together researchers coming from different disciplines (literature, linguistics, politics, sociology…). Its multidisciplinary aspect, and the diversity of the objects of research that its methods have been used on, comes from the very ubiquity of human language as a tool. Beyond their different goals and disciplines, the actors of this field are motivated by the common need to mine the text that is the material of their research.

The diffusion of these methods within the social sciences has been associated with the commitment of researchers who have devoted a large part of their activities to developing and diffusing the tools and software that put these methods into practice. The French school of Data Analysis was a major actor in this development, and at its core were Jean-Paul Benzécri and his colleagues; the influence of these founders is still vivid in the practice of text mining, because the algorithms and software carry their philosophy, as we will show below.

In this article, we have attempted to trace the history of the statistical analysis of textual data, focusing on the influence of Benzécri's work and school, and to make explicit their theoretical positions, clearly opposed to AI and to Chomskyan linguistics. After a presentation of the intellectual project, as an inductive approach to language based on the exploration of corpora, we present the principles of correspondence analysis, which is the main method developed in the Data Analysis School, used for corpus analysis but also for many other types of datasets. Then, we will focus on textual data analysis, a set of methods to analyse a corpus of texts (answers to open-ended questions, set of newspapers articles, corpus of literary works…). Based on the fact that software programmes have played a major role in the use of these statistical techniques, we shall examine a selection of these, display their specificities and their underlying theoretical bases.

In the process, we had to face the question of how to name this field, which has evolved considerably. For purposes of clarity, we shall use as the generic term 'textual data analysis', as used during the emblematic colloquium of this community, the JADT (*Journées Internationales d'Analyse des Données Textuelles – Textual Data Statistical Analysis*), even if the most currently used term today is text mining. This JADT conference was founded in 1990 (in Barcelona), with a scientific committee head by Ludovic Lebart. Since then, this international conference takes place every second year in a different European country.

1 The origins of textual data analysis

From the middle of the 1960's, Jean-Paul Benzécri, his colleagues and students introduced and developed a series of methods, which is commonly designated as "Analyse des Données" (Data Analysis) and that we can consider as the precursor of data mining and "big data". The methods could be applied to all kinds of data, textual data being a particular kind. .

Jean-Paul Benzécri, born in 1932, alumnus of the Ecole Normale Supérieure, obtained his Ph.D. in mathematics (topology) in 1955 at Princeton University under the direction of mathematician Henri Cartan. He started his career at the University of Rennes as an assistant professor in 1960. In 1965, he was promoted as a professor at ISUP, the Statistical Institute of the University of Paris, where he spent the rest of his career (Armatte, 2008). He is a mathematician, mainly interested in linguistics. When he was in Rennes, he introduced a mathematical linguistics course that revealed his turn to linguistics and the beginning of data analysis.

Benzécri is unanimously considered the father of the French School of Data Analysis.

In a nutshell, the principle of correspondence analysis consists in setting the data in rectangular "tables", in the form of matrices, in order to be able to apply data analysis methods to these tables. The tables were initially contingency tables (or cross tables that represent the frequency distribution of two qualitative variables). Correspondence analysis, initially adapted to contingency or cross tables, was extended to other kinds of tables, as disjunctive tables (Multiple Correspondence Analysis) and can be used on all kinds of tables with positive numbers. The idea is to identify the pattern of the relation between two sets of elements put into the table. In the case of a text corpus, the tables contain texts in their rows and words in their columns; at the intersection of a row and a column, there is an indicator of the presence or frequency of the word in the text.

Data analysis algorithms allow the information contained in the matrices to be synthesised. Factor analysis attempts to reorganise the matrices so that the first dimensions contain the maximum amount of information; classification methods allow for the identification of homogenous subgroups of texts and words. The School of Data Analysis often combines factor analysis and classification.

1.1 The origin of data analysis

In *A History and prehistory of data analysis* written in 1975 and published in 1982, Benzécri traces the origins of data analysis, explains correspondence analysis and put it in relation to current related works (Benzécri, 1982). As he explains in his introduction, after a chapter on "chance science" ("science du hasard"), he distinguishes three steps for the improvement of multidimensional statistics (or multivariate data analysis): biometry from Quetelet to Pearson, the works of Sir Ronald Fisher and psychometrics (from Spearman to Guttman). By these means, he draws a personal history of the origins of correspondence analysis (Armatte, 2008)

to which he dedicates the last part of the book. Although he underlines the originality and homogeneity introduced by his method, he also presents related works.

The origins of data analysis go back to the beginning of the century. Psychologists were the pioneers in the exploration of multidimensional data and factorial analysis, as analysed by Olivier Martin (Martin, 1997). Spearman, the British psychologist, by analysing the links between students' academic results and their mental aptitudes (Spearman, 1904), believed that he had shown the existence of a general aptitude or intelligence *factor*, which was later given the letter G. Subsequently, not just one, but several factors were sought from increasingly numerous data. Here lie the origins of *factor* analysis.

Correspondence analysis, a branch of factor analysis, started with Fisher, during the 1940s (Fisher, 1940). For Benzécri, by exploring discriminant analysis, Fisher developed the basic equation of correspondence analysis. Then, in 1961, Kendall and Stuart elaborated the canonical methods for the analysis of contingency tables (Kendall and Stuart, 1961). This allowed them to calculate the parameters used to test the hypothesis of independence between rows and columns.

Benzécri explains that he used the name of correspondence analysis for the first time in 1962 and presented the method in 1963 at the College de France (Benzécri, 1982, p. 101). Correspondence analysis is a generic term used as an umbrella.

He was aware of the work by psychometrists and was in contact with Shepard at Bell Labs who had introduced "multidimensional scaling" (Rouannet, 2008). His mathematical linguistics course at the University of Rennes lays the foundation of data analysis as it will be developed by the school.

1.2 The main contribution of Benzécri

Correspondence Analysis is often presented as an adaptation to categorical (or discrete) data of Principal Components Analysis (Greenacre and Blasius, 2006; Hill, 1974; Murtagh, 2005) or very close to muldimensional scaling (Hill, 1974). How can we specify the originality of the Benzécri's contribution to multidimensional analysis?

His main contribution was to show the full algebraic properties of the method and to display its interest: the testing of the independence of rows and columns, but above all the description of how data diverge from this hypothesis, by representing "proximities", the associations that exist between rows and columns, on factorial maps (Diday and Lebart, 1977). The map, a data visualisation of the proximities between individuals and between variables, is the central output for the interpretation. The accent on visualization methods is a key to understanding the success of the Data Analysis School. What was a complex set of data was organized as a "space" for the benefit of the analyst, and suddenly the cloud of data became accessible to interpretation as a whole, with a structure that could be explored, discovered, commented on and displayed. This approach differs from the more classic (and widespread in English literature at the time) approach of testing hypotheses on data sets.

Benzécri was not only interested in algorithms: data analysis constitutes for him a *global framework*, and this is his second main contribution. It first includes data preparation: how to transform any kind of data into a rectangular table with positive numbers that can be analysed. Correspondence analysis can be applied to almost all kinds of tables after suitable data transformation. It also includes a global set of aids to interpretation: the computation of contributions allows for measuring the quality of the representation on the map and the projection of supplementary variables gives to the practitioner complementary elements for interpretation. The association of correspondence analysis with clustering methods (in part-

icular with ascending hierarchical classification) allows a deeper understanding of data, and a simpler interpretation.

Finally, the framework gives a unique method (correspondence analysis and classification) instead of a profusion of algorithms, hard to understand for non-statisticians.

The framework is clearly oriented for users and practitioners by offering a methodological frame, with a particular attention to the display of results.

Benzécri devised and authorised the diffusion of a global framework for analysing "large tables", but he was above all guided by a theoretical and philosophical ambition, which directly interests us here.

1.3 The philosophy of Benzécri

As a mathematician turning towards linguistics, Benzécri became interested in data analysis methods not as psychological tools (a discipline which has been at the origin of a very large number of developments), but instead as a research tool for linguistics: “Correspondence analysis was initially proposed as an inductive method for linguistic data analysis” (Benzécri, 1982, p.102), “It was mainly with a view to studying languages that we became involved in the factorial analysis of correspondences” (Benzécri, 1981, p. X). His theoretical ambition was to open the doors to a new linguistics, in an era that was dominated by generative linguistics. He was opposed to the idealistic thesis of Chomsky who, in the 1960s, considered that only an abstract modelling could reveal linguistic structures. Against this thesis, Benzécri proposed an inductive method of linguistic data analysis "with, on the horizon, an ambitious tiering of successive researches, leaving nothing about form, meaning or style in darkness" (Benzécri, 1981, p. X). In this sense, he was quite close to the objectives of Bloomfield and Harris, who aimed at constructing the laws of grammar from a corpus of statements, with a distributionalist approach. The methods Benzécri developed were from his point of view more efficient for an in-depth understanding of language than the works on statistical linguistics carried out by Guiraud or Muller (Guiraud, 1954; Muller, 1977) which he found interesting but too exclusively focused on vocabulary (Benzécri, 1981, p. 3).

> We propose a method aimed at the fundamental problems that interest linguists. And this method (…) will consist in a quantitative abstraction, in the sense of starting from tables of the most varied data, it will construct, through calculation, quantities that could measure new entities, situated at a higher level of abstraction than that of the facts that were initially collected. (Benzécri, 1981, p. 4)

By identifying factors, there can be doubt that an operation of *abstraction* has indeed been carried out. The computer gives neither any names nor meanings to the entities that it has extracted; it is up to specialists to provide their interpretations.

Benzécri’s philosophical ambition was to reassign value to the inductive approach, and thus to oppose idealism:

> For we condemn the idea that, from principles lightly received, idealism can through a dialectic, even if it is suborned to mathematics, derive certain conclusions; then, to such a priori deductions, we oppose induction which, a posteriori, from the basis of observed facts attempts to rise up to what orders them. (Benzécri, 1968, p. 11)

He criticised idealistic theories that suppose the existence of a model and check its relevance approximately through observation. He doubted that it was possible to reduce a complex object into a combination of elementary objects, "for the order of the composite is worth more than the elementary properties of its components" (Benzécri, 1968, p. 16).

The objective that he thought to be attainable through data analysis was being able to be extract "from the mush of data the pure diamond of true nature". The passage from data to

abstract entities, from darkness to light, was made possible in his eyes thanks to data analysis and the "novius organum" of the computer: “The new means of calculation allow us to confront complex descriptions of a large number of individuals, and so place them on flat or spatial maps, in reliable images that are accessible to intuitions from the nebular of initial data” (Benzécri, 1968, p. 21). As an auxiliary for synthesis, the computer is a mental tool: after Aristotole’s *organum* and the *Novum Organum* conceived by Bacon, is not this *Novius Organum* "the newest tool"? (Benzécri, 1968, p. 24).

> After all, it can be seen just how much analysis is free from a priori ideas. From data to results, a computer, insensitive both to expectations and to the researcher’s prejudices, proceeds on the large and solid basis of facts that have previously been defined and accepted as a whole, then counted and ordered according to a programme which, given that it is incapable of understanding, is also incapable of lying. (Benzécri, 1968, p. 24)

> Finally, among all the, often contradictory, a priori ideas that each problem inspires in profusion, a fitting choice is made: even more, some ideas which, a posteriori, and after a statistical examination of the data, seem to have been quite natural a priori, would not always have occurred to the mind. (Benzécri, 1968, p. 24)

1.4 Influence

The contribution of Benzécri (a unified frame for data analysis oriented to users) greatly contributed to the diffusion of correspondence analyses in France in all the physical, social, human, and biological sciences: they were, and still are, extremely successful as a display of results. Pierre Bourdieu played an important role in the diffusion of the method as his influence in social sciences increased. Bourdieu’ theory was profoundly inspired by correspondence analysis when he analysed the social space as a field of tensions for example in *Distinction* (Bourdieu, 1984). Rouanet explains that “For Bourdieu, MCA provides a representation of the two complementary faces of social space, namely the space of categories - in Bourdieu’s words, the space of properties - and the space of individuals. Representing the two spaces has become a tradition in Bourdieu’s sociology” (Greenacre and Blasius, 2006, p. 167).

The Data Analysis School has been, and still is, widely present in the field of social sciences, and its approach continues to be used very regularly. Publication of such research, however, runs up against the fact that English-speaking publications favour hypothetic-deductive approaches. The purely exploratory dimension, aimed at bringing out forms and models from data, does not have the same legitimacy as other approaches; they are too descriptive, instead of being explicative. Yet, it is well known that hypothetic-deductive methods are fragile, because of the order of causality which is pre-established at the moment when a hypothesis is determined. Consequently, the data analysis school had a wider diffusion in France than in other countries.

In Paris, Benzécri put together a large team of data analysis researchers, as can be seen in their numerous collective publications under his direction. The main publications of Benzécri consist of treaties, handbooks and a history.

The treaty on Data Analysis is constituted of two volumes: the first (Benzécri, 1973a) is dedicated to taxonomy and reviews all the classification and clustering methods, the second (Benzécri, 1973b) to correspondence analysis.

A *History and prehistory of data analysis*, redacted by Benzécri in 1975 and published in 1982 (Benzécri, 1982), constitute a state of the art of correspondence analysis and situates the originality of his approach.

For Benzécri this book is an introduction to the series of handbooks *Pratiques de l'analyse des données* published at the beginning of the 1980's: the first volume is dedicated to correspondence analysis (Benzécri, 1980), but in the 1984 edition, an added chapter concerns classification. The second is more theoretical and the third is dedicated to linguistics: *Pratique de l'analyse des données. 3 Linguistique et lexicologie* (Benzécri, 1981).

Each of his volumes involved a large number of contributors, 30 for example for *Linguistique et lexicologie*.

The Journal of data analysis (*Cahiers d'Analyse des Données*) based on an idea of Michel Jambu (Armatte, 2008) stands as the main outlet for articles in the field of data analysis, extended to textual data analysis. This journal was published from 1976 to 1997.

An element that distinguishes Benzécri's work is the organisation of his collective books that all propose: theory, examples of applications from very large fields (natural and human sciences) and programs to be reused in different computers. This structure is an element that explains the important diffusion of methods. The statistical procedures were explicit and shared (an open source approach before its time). At the end of the 1980', several correspondence analysis procedures were included in the leading statistical software packages of the time, notably SPSS, BMDP, and SAS (Greenacre and Blasius, 2006). Nowdays they are implemented in "R", the open source package for statistical computing (Husson et al., 2009).

At ISUP, Benzécri along his co-workers had an important flow of students, estimated at 180 master students per year and 40 Ph.D. (Armatte, 2008) who contributed to the diffusion of methods.

Although cluster analysis is also an important part of Data Analysis School, we will focus on Correspondence Analysis, which can be considered as the core of Benzécri's innovation.

2 Correspondence Analysis

The presentation of correspondence analysis in this section is based on the chapter dedicated to this topic in *Histoire et préhistoire de l'analyse des données* (Benzécri, 1982, p. 101-131), on the introduction in the volume dedicated to linguistics and lexicology (Benzécri, 1981, p. 73-135) and on the *Handbook* (Benzécri, 1992).

Correspondence analysis is a method that gives a geometrical representation of the associations between two sets of elements in correspondence as they appear in a table. It is applied to a specific kind of data: a table of correspondence between the two sets of elements (correspondence or concordance table). Statistical tests are usually used to reject the idea of independence of variables or attributes. The Benzécri's approach is exploratory and descriptive. The main originality of correspondence analysis is to represent, in a geometrical way, the extent to which the independence of observations and attributes is *not verified*. For Benzécri, independence between rows and columns lacks scientific interest; what is interesting is precisely the detail of *how* they interact.

2.1 From a correspondence table to profiles

Correspondence analysis firstly requires one to transform raw data, for example a corpus, into a contingency table, that crosses two sets of elements, a set I (individuals or observations) and a set J (variables or attributes). At the crossing point of a row and a column, we get the number of occurrences of the attribute j in the observation i, k(i,j). Two examples will clarify.

Suppose we are interested in analysing theatre plays. We can build a table, I representing the set of plays, and J the vocabulary that we can find in the plays. In this case, k(i,j) will represent the number of occurrences of the word j in the play i. In the table, there are as many rows as elements in the set I (plays), m, and as many columns as there are in the set J (words), n. Rows are individuals and columns are properties. Let's take another example from (Benzécri, 1982, p. 103). In order to analyse the distribution of nouns and verbs in a corpus, we can build a table where rows are nouns and columns are verbs and at the intersection of a row and a column, we have the number of sentences where the noun is the subject of the verb.

In order to compare the distribution of the two sets of elements, row and column profiles are calculated: f^i_j is *k(i,j)/ki.* (where $k_{i.} = \sum_{j=1}^{n} k(i,j)$, ie the sum of frequencies on the line i). The profile of i will be f^i_J, a vector made of the sequence of f^i_j $(f^i_J = \{f^i_j \mid j \in J\})$

Symmetrically, the profile of an element j will be $f^j_I = \{f^j_i \mid i \in I\}$.

2.2 Representing the distance between profiles

How do we compare the profiles of different elements (rows or columns of the table)? We need a space and a distance. Correspondence analysis uses a Euclidean space and a distributional distance, or the chi-square distance, which is a distinctive feature of correspondence analysis. The distance between i and i' will be defined as follows:

$$d^2(i, i') = \sum \{(f^i_j - f^{i'}_j)^2 / f_j \mid j \in J\}$$

Each element i (resp j) of set I is represented by its profile and is assigned a mass proportional to the total of the row. The set of the profiles fiJ constitutes a cloud N(I) in a multidimensional space. Respectively, a cloud N(J) is defined for the profiles fjI.

The main idea is to reduce the complexity of the cloud and to find a way to represent most of the information in a lower dimension space. For this, the center of gravity of the cloud is calculated and the dispersion of the cloud around its center of gravity is measured (inertia). Then the factor axes, or principal axes of dispersion, are constructed. Points are projected on those axes, and their coordinates on these axes are called factors. In the plan defined by the first two axes we can have the best projection of the cloud (which minimizes the loss of information).

A distinguishing feature of correspondence analysis is the perfect symmetry of the roles assigned to the two sets I and J in correspondence. This permits the simultaneous representation of the two clouds on the same axes.

The main objective is to visualize the distance between observations or attributes, i.e. the distance from a random distribution. The algorithm produces a set of 'aids to interpretation' that allows the researcher to interpret the results properly.

Often correspondence analysis is combined with hierarchical clustering: the classification is based on the coordinates of the elements on the factor axes.

3 Instruments at the service of the humanities and social sciences

Innovations rarely come from isolated individuals. They emerge and are diffused through networks, collectives and institutions, in which individuals meet and exchange, in which innovations circulate, are discussed, improved and criticised. The diffusion of textual data analysis is no exception to this rule.

Laboratories, journals and lectures have progressively contributed, thus stimulating exchanges and debates. But in this specific field of research, IT tools have become the major players in the diffusion of methods and the organisation of this network. On the one hand, they crystallised the theoretical debates within the community and, on the other, raised the question of economic, or more modestly commercial, factors linked to these methods.

For the diffusion of these methods has been supported for economic reasons: in the sector of surveys and marketing, the possibility of conducting quantitative research on qualitative data, in other words to introduce measurement into the analysis of discourse, provides an interesting opportunity.

After quickly examining the institutions that have contributed to bring to life this scientific speciality of textual data analysis, we will then focus on a few emblematic textual statistics programmes, while showing how each tool bears the marks of the environment in which it was developed (the discipline, type of corpus and the questions raised by researchers) and how this milieu interacts with the researchers' own objectives.

3.1 Places

After Rennes, ISUP, in Paris, became the centre of elaboration and diffusion of data analysis. Benzécri's seminar at ISUP was attended by most prominent statisticians and researchers in this area. This field was far broader than just textual data analysis as we have seen, but the audience included key figures such as Ludovic Lebart, who also paid particular attention to texts.

Crédoc (*Centre de recherche pour l'observation des conditions de vie*) was for a long time a powerhouse in the field of textual statistics. Ludovic Lebart worked there for many years (1971-1988), setting up and directing the survey *Aspiration et Conditions de vie des Français*. With André Morineau, he was behind the development of Spad (*Système portable d'analyse de données*) (Lebart and Morineau, 1982) and its extension devoted to texts 'Spad.T' (Lebart et al., 1989) which was also based on the work and findings of Eric Brian (Brian, 1986). The Lebart & Morineau's programmes were, up to the year 1987, distributed by a non-profit organization, Cesia in a freeware context and served many researchers or data analysts in the pioneer era of what was to become text mining. Spad had been designed to analyse quantitative surveys and Spad T for the analysis of answers to open-ended questions. The implementation of the algorithms was guided by the framework of surveys with open-ended questions. A data centre in the basement of Crédoc, shared with the Cepremap, another research centre on economics, and connected to Circé (a regional computing centre in Orsay, *Centre Inter Régional de Calcul Électronique*) provided the possibility to develop and test these tools on data and was the meeting point of a community also involving statisticians such as Jean-Pierre Fénelon (Fénelon, 1981) or Nicole Tabard (pioneer of geographic information systems) (Lebart et al., 1977). A few years later, in the "Prospective de la Consommation" department, Saadi Lahlou developed a research axis based on the applications of lexical analysis in the social sciences (Yvon, 1990; Beaudouin and Lahlou, 1993; Lahlou, 1992;). He contributed to the diffusion of these methods in the field of social psychology.

At Crédoc, Spad was used, but also Alceste, which had been developed by Max Reinert (Reinert, 1990, 1987), and could analyse sets of texts other than open-ended questions. Lexical statistics became a tool for the study of social representations (Lahlou, 1998) and led to a reflexion about the interpretation processes (Lahlou, 1995). Lahlou started a collaboration with M. Reinert to develop tools on the Unix platform and to process greater volumes of text. The large number of *Cahiers de recherche* from Credoc published on these subjects, and the contracts using these methods, bear witness to the dynamism of this centre at the time.

Portability on Mac, Unix and Windows ensured an enduring success of Alceste software in the social sciences in France, and as the software's dictionaries extended to other languages, to further countries.

The laboratory "*Lexicologie et textes politiques*" was set up in 1967 at the Ecole Normale Supérieure in St-Cloud. It has been attached to various different bodies over time, and some of its activities are now located in the Icare laboratory of the ENS in Lyon, while others are at Paris III. The analysis of political discourses stands as the backbone of the unit, with a methodological reflexion branch that explores the place occupied by machines in lexicometry, for the analysis of texts. Pierre Lafon (Lafon, 1984) and André Salem (Salem, 1987) undertook more specifically the setting-up of statistical analysis tools: "these two linguist-mathematicians […] were advised in their methods by the masters of 'data analysis' (Jean-Paul Benzécri) and of probability theory (Georges-Théodule Guilbaud)" (Tournier, 2010). It was in this laboratory that reflexions about corpus linguistics started in France (Habert et al., 1997) and more exactly reflexions regarding annotation systems and the enrichment of texts. André Salem's Lexico programme is one of the tools created in this context. It includes correspondence analysis. It can be distinguished from other software on two points: the identification and processing of repeated segments (sequences of words allowing for the introduction of a notion of syntax) (Salem, 1987) and a detailed processing that measures the chronological evolution in the corpus (Salem, 1995). Correspondence analysis allows to show the distances between sub-parts of a text corpus and to visualise, if relevant, the chronological evolution of texts. An attachment to political and trade-union discourses was specialty of this laboratory.

In the South of France, at the University of Nice, another laboratory was founded in 1980, which accorded a significant role to machines. Etienne Brunet, a literary scholar who had been a computer amateur since the end of the 1960s, set up an active research pole at the university, based in the laboratory *Bases, Corpus, Langage*. Brunet designed a tool, Hyperbase, which was particularly suited to the analysis of very large volumes of literary texts (Brunet, 1988), but also political texts (Mayaffre, 2000), which opened up bridges with the laboratory in St Cloud. The software includes a correspondence factor analysis from the programs developed by J-P Fénelon and his colleagues. It gives a visualisation of distances between words and sub-parts of texts projected on the map. For example, figure 1 represents the result of the correspondence analysis applied to a table containing in rows the different works of Rabelais (capital letters, PANT for Pantagruel) and in columns the personal pronouns.

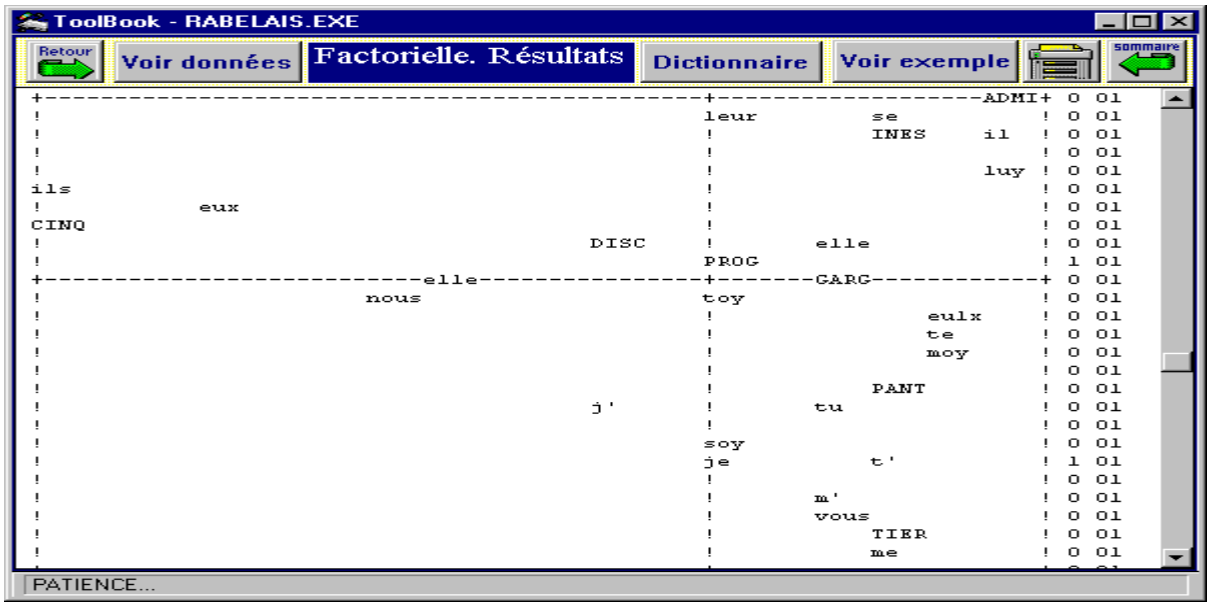


Figure 1. Hyperbase Factorial Analysis
(http://ancilla.unice.fr/~brunet/PUB/hyperwin/analyse.html)

This tool was distributed in the community of humanities researchers. This laboratory explored large corpora from the Frantext database, an exceptional collection of digitized literary works. Since 2001, it has had its own journal, *Corpus*, whose current editor-in-chief is Sylvie Mellet. Two volumes (Brunet, 2009, 2011) collected the main papers published by Etienne Brunet .

Other sites have also played an important role: the IBM scientific centre led by François Marcotorchino, the team headed by Dominique Labbé in Grenoble and other sites abroad, such as Sergio Bolasco's team at the Sapienza in Rome…

The *Journées internationales d'Analyse des Données Textuelles*, which have been organised every second year since 1991, stand as a point for rallying, but also enlarging, the community of researchers in this field. Mostly French-speaking, it also welcomes Italian and Spanish researchers from the same field. The systematic publication of the papers and the availability online from André Salem and Serge Fleury's journal *Lexicometrica* (http://lexicometrica.univ-paris3.fr/jadt/) thanks to Paris III, constitute a corpus of experiences.

Lebart and Salem's book, *Analyse statistique des données textuelles*, published by Dunod in 1988 (Lebart and Salem, 1988) and republished in 1994 (Lebart and Salem, 1994), then translated into English as *Exploring Textual Data* (Lebart et al., 1998), has become the reference manual in this field .

3.2 Programmes

Publications played a decisive role in the diffusion of methods of textual analysis, explaining the algorithms, displaying possible usages on corpora, and multiplying examples of application. But the diffusion of usages has mainly taken place through the tools themselves, which have been major vectors in the appropriation of methods that are sometimes viewed with mistrust by the world of the social sciences and the humanities. In each case, we shall underline the particularities of the programme: preparation of corpora (selection of texts and variables), processing algorithms and interpretation. We will focus on two software pro-

grammes that where the most innovative for text analysis in Benzécri's tradition: Spad T and Alceste.

3.2.1 Spad T

As we have seen, Spad T is an extension of Spad (*Système portable pour l'analyse des données)* which allows for the analysis of answers to open questions in surveys. Spad and Spad T were both designed and coded by Ludovic Lebart and André Morineau at the data centre of Crédoc and Cepremap (see above).

The unit of analysis (each row of the table) is the individual in the survey, characterised by their answers to open and closed questions. But it can also correspond to a group of individuals, according to variables such as age, or level of education, with all the individuals having the same variable value constituting *one* text (a row in the table). For example, figure 2 is the result of the correspondence analysis of a cross tabulation between words (from answers to an open question[1]) and individuals grouped by educational level.

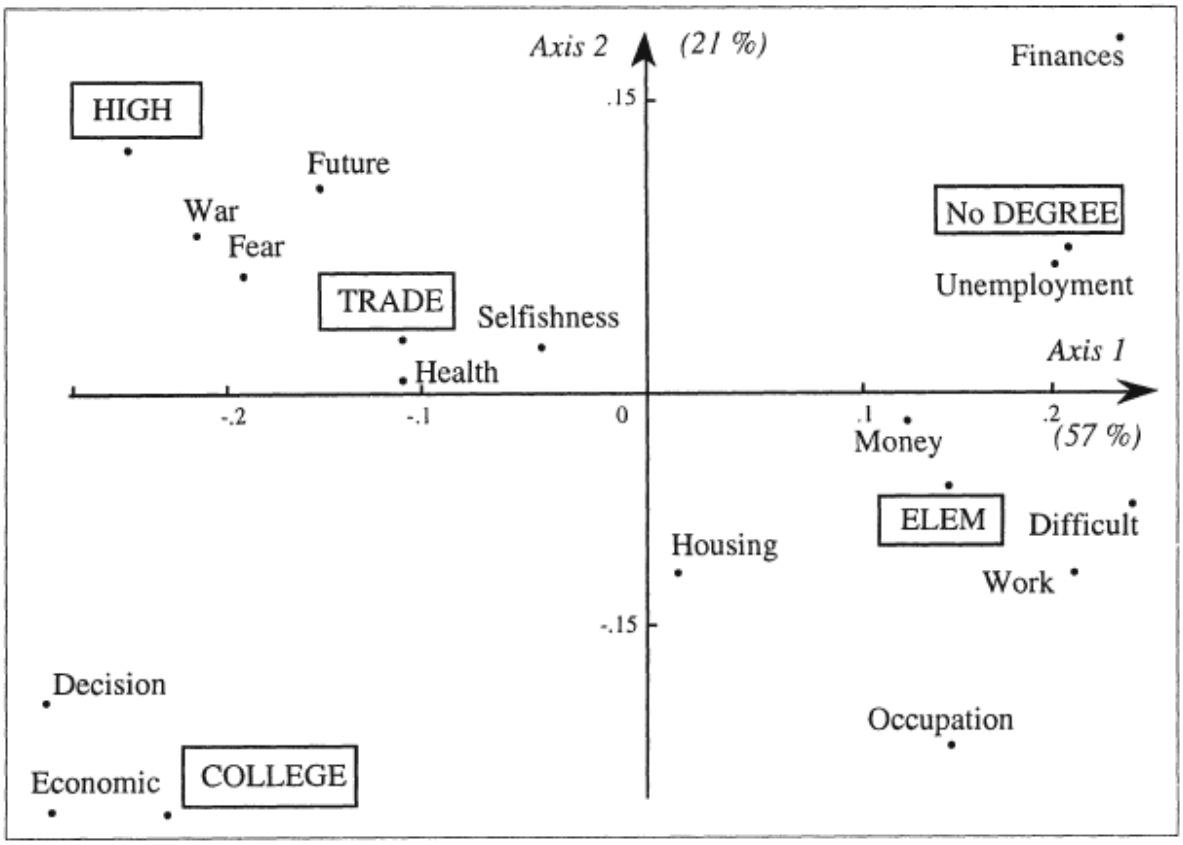


Figure 2. Proximities among words and among educational level
(Lebart et al., 1998, p. 52)

For the words entered in the tables (i.e. making up the columns of the table), Spad T proceeds as follows: it keeps the graphic forms and the words, as they appear in the text, and uses no form of lemmatisation (that is taking graphic forms back to their roots, or dictionary entries); with a frequency threshold, it eliminates rare and very short words (under 3 letters, for example), which is a way to exclude grammatical words (articles, pronouns…). As the answers

[1] The question was "What are the reasons that might cause a couple or a woman to hesitate having children?" (Lebart et al., 1998).

are reduced throughout the chain leading from the survey to the processing (investigators tend to keep only the main points when noting down answers, the entry clerks often also simplify anyway), and the corpus in question is full of redundancies, this rather brutal "cleansing" has in practice little impact on the results.

Spad T offers a full palette of data analysis procedures. The most classic approach is to carry out a correspondence analysis in a table crossing the answers in the rows with the words used in the columns. Then, based on factor coordinates, an ascending hierarchical classification (clustering) is carried out. The principle consists of bringing together in pairs the answers that are most alike in terms of the vocabulary used, and to advance progressively so as to arrive at a predefined number of classes.

To assist interpretation, it is possible to obtain for each class its specific vocabulary (the words that are significantly more present in this class than in the others), and the most characteristic answers. As Spad T is consistent with Spad, it is possible also to add the values of other variables to the survey, which are over- or under-represented in the class. Spad T includes a most useful "Tamis" (sieve) procedure which systematically tests the interaction of a given modality with every other modality of every other variable in the survey, and orders them by decreasing degree of significance. This enables profiling a class and orienting interpretation and testing without any preconception, in the very explorative spirit of the Data Analysis School.

To sum up, Spad T[2] is particularly well suited to a specific usage context (quantitative surveys) and well-defined types of corpora (answers to open questions). The data analysis and interpretation assistance algorithms are extremely robust, and the usage context means that the simplistic vocabulary reduction creates no problems. The originality of the approach is the possibility to incorporate metadata (*i.e.* information on individuals who produced the text), and then to situate the texts regarding the characteristics of the speaker or writer.

It should be noted here that one of the flaming debates that animated the community was precisely on this issue of lemmatisation; some defended the idea of working on "raw" graphic forms (Lafon, 1984), while others considered that lemmatisation (the reduction of forms to their lemma) was an indispensable prerequisite to any processing, as can be seen in the defence mounted by Muller in his introduction to Lafon's book. The pros considered it was a necessary step to avoid ambiguity of forms (homonymy) while the cons thought it leads to a loss of information: plural/singular, masculine/feminine, person, time being meaningful. This debate provoked heated discussions at almost every JADT conference until the possibility of keeping at the time the raw and the lemmatized form was provided.

3.2.2 Alceste

The methodology of ALCESTE (*Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte*) was designed by Max Reinert (1993, 1983); it was inspired by the field of data analysis, Reinert being also a participant of Benzécri's seminar. However, Reinert's preoccupations took a particular orientation. He considered a corpus as a sequence of statements produced by a subject-utterer. Thus, the text is modelled in a table containing statements in rows, bearing the mark of the subject-utterer, and words or lexemes in columns, referring to objects in the world (without any preconceptions about the "reality" of these objects). The objective is then to bring out "lexical worlds".

[2] Ludovic Lebart has made available to the public a software programme, DTM-VIC (http://www.dtmvic.com/), which shares the same properties as Spad for analyzing both numerical and textual data.

> A lexical world is thus at once the trace of a referential site, and the index of a form of coherency linked to the specific activity of the subject-utterer, which we shall call a local logic. (Reinert, 1993, p. 9)

Thanks to statistical procedures, which associate statements using the same type of vocabulary, the method is able to identify different lexical worlds, which could be interpreted as "visions of the world". For example, in his study of *Aurélia* by Nerval, Reinert (Reinert, 1990) identified three types of world by classifying the statements: the imaginary world, the real world and the symbolic world, each of which bears the mark of a certain relationship with the narrator.

Let's describe Alceste in a nutshell. The input is a text or a set of texts, described by some extra textual variables, which describe the communication situation. The output is a typology of the statements that constitute the corpus. A statement is defined as a point of view from a subject about the world. The clustering process is based on the similarity/dissimilarity of words inside the statements. Each cluster of statements is interpreted as a lexical world, which reflects a world view.

This theoretical orientation has consequences on the way analysis is carried out. Let us start with textual units. Reinert attempted to identify the notion of a statement: a point of view about the world that bears the trace of a subject. But how to define automatically the notion of an statement given that it does not necessarily coincide with the notion of a sentence, and no punctuation marks allow it to be identified clearly? As there is no satisfactory solution to this problem, Reinert offered a heuristic: make two possible segmentations of the corpus into textual units while varying the length of the units. Thus, one table would contain in its lines the textual units from the first segmentation, and a second those from the alternative one.

What vocabulary elements are kept in the table's columns? As with Spad T, a frequency threshold allows rare words to be eliminated (this has virtually no impact on the final result since calculation is done on co-occurrences). A lemmatisation process reduces the words to their roots and above all provides an identification of the elements of speech (nouns, verbs, pronouns…). Given the perspective adopted by Reinert, only "full" words, with reference points, are kept for the analysis, and not grammatical words (articles, etc.), which form the text's cement.

On these matrices, which cross textual segments and lemmatised words, Alceste carries out a descending hierarchical classification, using an original algorithm devised in 1983 (Reinert, 1983) which is particularly suited to sparse matrices (with over 90% "0's"). The idea is to take all of the textual segments and to divide them into two groups, in such a way as the groups will be as homogenous as possible in terms of the vocabulary used, while also being as distant as possible from each other. The procedure is then reiterated on the larger remaining group until the requested number of classes has been obtained. This classification process is iterative and leads to a typology. Technically, the descending hierarchical classification uses factor analysis. Once the first axis is calculated, a hyperplane is slid along the axis to split the cloud into two sub-clouds until it maximises the inertia between both while minimizing the intra-class inertia. This defines the first two groups, and the process is reiterated (Reinert, 1983).

This is where the heuristic proposed by Reinert comes into play again: on each of the tables that have been made, a descending hierarchical classification is carried out, then the two analyses are compared, so that only the most stable typological classes in both analyses will be conserved. What is more, this provides a procedure which can optimise the number of end classes. For example, the figure 3 shows the result of the double classification on Aurélia (Reinert, 1990). At the end, three classes will be kept : 8 <->9, 10 <->11 and 11<->10.

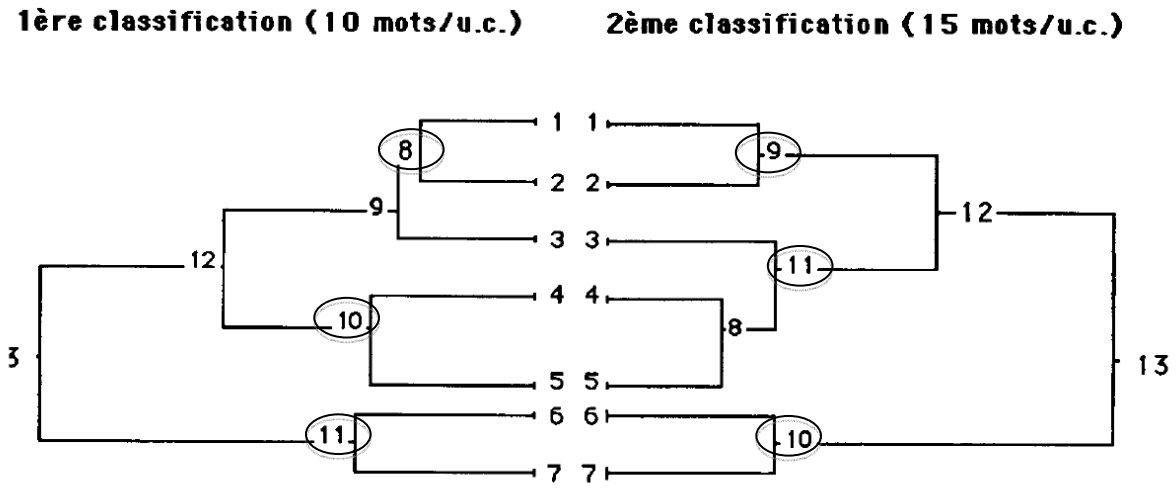


Figure 3. Descending hierarchical classification with Alceste (Reinert, 1990)

In this process, the new main axis is calculated separately for each successive sub-cloud and the result is amazingly robust, compared to other classification techniques which are based on a single factor analysis.

Each class of the typology is characterised by a list of words that make up the specific vocabulary of the class, in comparison with the entirety of the corpus, using the most characteristic textual segments of the class, and the most representative values of the illustrative variables. The whole can be visualised on a factor analysis plane. These interpretation aids allow for a characterisation of the lexical-semantic field appertaining to each class and give a picture of which external production factors best explain its particularities. (Schonhardt-Bailey et al., 2012) provide what is so far the sole detailed and illustrated description of the Alceste algorithm in English. Alceste has been used for analysing corpora of answers to open questions, literary works, newspaper articles, semi-directed interviews, forum interactions, film reviews, dictionary articles…

4 Conclusion and perspectives

Jean-Paul Benzécri and his colleagues developed a global framework for data analysis (correspondence analysis and clustering methods). Those inductive methods were defined for linguistic purposes, but were widely used in other disciplines, for text analysis but also for quantitative data. The efficiency of those approaches for exploring data and for building hypotheses of research has been widely proven by thousands of publications.

In linguistics, textual data analysis opened the path to a systematic study of language based on corpora, corpus linguistics, with the assumption that field-collected texts, in natural contexts, are the best way to infer sets of rules.

Although the research in statistics and computing sciences has much evolved, in particular with machine learning techniques, it is interesting to note that those “old” techniques are still used by researchers in the social sciences. To do so, the textual data analysis tools have been adapted to larger corpora. While a corpus containing 2,000 answers was considered to be a large one during the 1980s, we now process ones with tens of thousands, or even millions of texts. The textual statistic tools were developed with programming languages

which have sometimes since become obsolete, such as Fortran, and were often limited in their size when it came to processing. Updating them to make them appropriate to current volumes sometimes requires codes to be written anew. For example, Max Reinert's Alceste software was entirely reprogrammed by Pierre Ratinaud, and renamed Iramuteq (http://www.iramuteq.org/), with a more modern interface and the capacity to process far larger volumes. Such re-writing can raise problems of intellectual property rights, in that the approaches and the classification algorithms are virtually identical. In the same way, TXM developed for the Textométrie project (http://textometrie.ens-lyon.fr/), reuses and modernises old algorithms, while opening up an enrichment of the lexical data with morpho-syntactic, phonetic or other traits. In such cases, there have been no fundamental changes made to the algorithms of data analysis themselves which is a proof of their efficiency for social scientists.

The methods discussed above are based mainly on the analysis of the distribution of frequencies and co-occurrences of words in texts. The main unit of analysis is the word in its textual context. But, before long, the reduction of a text to a "bag of words" seemed too reductive and the introduction of finer descriptive traits of texts became necessary. Benzécri and his colleagues (Benzécri, 1981) already imagined the introduction of annotations although the technologies were not operational. The methods gradually improved thanks to natural language processing tools, which allowed syntactic, semantic and even prosodic aspects to be taken into account. A text could be associated with a series of descriptive characteristics, concerning different linguistic levels. In this perspective, influenced by (Biber, 1989) who aimed at inductively constructing textual typologies from descriptive traits, the field of corpus linguistics grew up (Habert et al., 1997). Let us take for examples of its application, the TypTex project (Habert et al., 2000), the characterisation of a corpus of texts according to morpho-syntactic traits by (Malrieu and Rastier, 2001; Rastier, 2011) or the attempt to articulate phonetic, morpho-syntactic, rhythmic and semantic characteristics by Beaudouin, (2002). To sum up, approaches that exploited the progress made in the natural language processing no longer limited themselves to words, but now included other levels of linguistic analysis (phonetics, syntax, semantics…). The principles of correspondence analysis and clustering are therefore now applied to much larger tables than they used to be.

The new frontier for textual data analysis is the analysis of web documents. Text was the first medium to enter into the digital world, before images, sounds or videos. It is thus quite natural that the statistical study of texts should have started long before other contents. In France, the digitization of large sections of literature on the Frantext database combined with mathematical and statistical progress in the area of data analysis have fostered the remarkable rise of the field of textual data analysis. Today, digitalisation has reached the entirety of cultural productions and, as a recent development; more and more production is "born digital". This has opened new research questions. It is no longer possible to reduce the Web to text only, so it will be necessary to enrich the current methods with resources that appertain to the Web's particularities (multimedia, hypertextual, imbricated in reception, dynamic) and develop approaches that combine different methods, textual statistics being just one among others.

5 Bibliography

Armatte, M., (2008). Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri: un cas de généalogie rétrospective. *Journl Electronique d'Histoire des Probabilités et de la Statistique 4, 1–22.*

Beaudouin, V. (2002). Mètre et rythmes du vers classique - Corneille et Racine. Champion, coll. Lettres numériques. Paris.

Beaudouin, V., Lahlou, S. (1993). L'analyse lexicale: outil d'exploration des représentations. CRÉDOC, Cahier de Recherche, n°48, Paris.

Benzécri, J.-P. (1968). La place de l'a priori, "Organum". In: *Encyclopedia Universalis. pp. 11–24.*

Benzécri, J.-P. (1980). *Pratique de l'analyse des données. Analyse des correspondances & classification. Exposé élémentaire*. Paris.

Benzécri, J.-P. (1982). *Histoire et préhistoire de l'analyse des données.* Paris: Dunod.

Benzécri, J.-P. (1992). *Correspondence Analysis Handbook.* New-York, Basel, Hong Kong: Marcel Dekker, Inc.

Benzécri, J.-P. et al. (1973a). *L'analyse des données. 1 La taxinomie*. Paris: Bordas.

Benzécri, J.-P. et al. (1973b). *L'analyse des données. 2 L'analyse des correspondances.* Paris: Bordas.

Benzécri, J.-P. et al. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie.* Paris: Dunod.

Biber, D. (1989). A typology of English texts. *Linguistics 27, 3–43.*

Bourdieu, P. (1984). *Distinction. A Social Critique of the Judgement of Taste*. Harvard University Press.

Brian, E. (1986). *Techniques d'estimation et méthodes factorielles, exposé formel et application aux traitements de données lexicométriques*. Ph.D., Orsay.

Brunet, E. (1988). *Le vocabulaire de Hugo.* Paris : Slatkine-Champion.

Brunet, E. (2009). *Comptes d'auteurs - Tome 1. Etudes statistiques, de Rabelais à Gracq.* Paris : Honoré Champion.

Brunet, E. (2011). *Ce qui compte. Ecrits choisis, tome II. Méthodes statistiques*. Paris: Honoré Champion.

Diday, E., Lebart, L. (1977). L'analyse des données. *La Recherche* p. 15–25.

Fénelon, J.-P. (1981). *Qu'est-ce que l'analyse des données?* Paris: Lefonen.

Fisher, R.A. (1940). The precision of discriminant function. *Annals of Eugenics 10, 422–429.*

Greenacre, M., Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire.* Paris: PUF.

Habert, B., Illouz, G., Lafon, P., Fleury, S., Folch, H., Heiden, S., Prevost, S. (2000). Profilage de textes: cadre de travail et expérience. In: *JADT'2000. 5èmes Journées Internationales d'Analyse Statistique Des Données Textuelles,* Lausanne, 9-11 Mars 2000.

Habert, B., Nazarenko, A., Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin/Masson.

Hill, M.O. (1974). Correspondence Analysis: A Neglected Multivariate Method. *Journal of the Royal Statistical Society 23, 340–354.*

Husson, F., Lê, S., Pagès, J. (2009). *Analyse des données avec R.* Rennes: Presses Universitaires de Rennes.

Kendall, M.G., Stuart, A. (1961). *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.* Hafner Publishing Company.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris : Slatkine-Champion

Lahlou, S. (1992). Sialors: "bien manger"? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus. *Cahiers de recherche*. Paris: CRÉDOC.

Lahlou, S. (1995). Vers une théorie de l'interprétation en analyse statistique des données textuelles. In: S. Bolasco, A. Salem (eds), L.L. (Ed.), *JADT 1995. III Giornate Internazionali Di Analisi Statistica Dei Dati Testuali.* CISU, Roma: p. 221–228.

Lahlou, S. (1998). *Penser manger. Alimentations et représentations sociales*. Paris: PUF.

Lebart, L., Morineau, A. (1982). *SPAD: Système Portable pour l'Analyse des Données*.

Lebart, L., Morineau, A., Bécue Bertaut, M. (1989). *Spad.T: Système portable pour l'analyse des données textuelles.*

Lebart, L., Morineau, A., Tabard, N. (1977). *Méthodes et logiciels pour l'analyse des grands tableaux.* Paris*:* Dunod.

Lebart, L., Salem, A. (1988). *Analyse statistique des données textuelles*. Paris: Dunod.

Lebart, L., Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data*. Dordrecht, Boston: Kluwer Academic Publisher.

Malrieu, D., Rastier, F. (2001). Genres et variations morphosyntaxiques. *TAL 42(2), 547-577.*

Martin, O. (1997). Aux origines des idées factorielles. *Histoire & Mesure 12(N), 197–249.*

Mayaffre, D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux guerre.* Paris-Genève: Slatkine-Champion,

Muller, C. (1992). *Principes et méthodes de statistique lexicale*. Paris: Larousse, 1977, réimpression Champion-Slatkine, 1992.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R.* Boca Raton: Chapman & Hall/CRC.

Rastier, F. (2011). *La mesure et le grain.* Paris: Honoré Champion.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données VIII(2), 187–198.*

Reinert, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud. *Bulletin de Méthodologie Sociologique 13(1), 53-90.*

Reinert, M. (1990). ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval. *Bulletin de Méthodologie Sociologique*, n°26, p. 24-54.

Reinert, M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société 66, 5–39.*

Salem, A. (1987). *Pratique des segments répétés*. Paris : Klincksieck.

Salem, A. (1995). La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert. In: *Langages de La Révolution (1770-1815)* (Actes Du 4ème Colloque International de Lexicologie Politique). Paris: Klincksieck.

Schonhardt-Bailey, C., Yager, E., Lahlou, S. (2012). Yes, Ronald Reagan's rhetoric was unique — but statistically, how unique? *Presidential Studies Quarterly*, vol. 42, n°3, p. 482–513.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology 15, 201–292*.

Tournier, M. (2010). Mots et politique, avant et autour de 1980. Entretien. Mots. *Les langages du politique 94, 211.*

Yvon, F. (1990). L'analyse lexicale appliquée à des données d'enquête: états des lieux, CRÉDOC, Cahier de Recherche, n°5.

Glottometcrics 33, 2016, 73-100

List of Journals Containing Contributions to Quantitative Linguistics

Tim Rostin, Trier

In this contribution, a survey of journals containing quantitative linguistic research articles is presented. The journals are ordered alphatetically and show the number of contributions to quantitative linguistics as far as they are already entered into the Bibliography of Quantitative Linguistics.

Hopefully this survey provides some help for researchers and students to keep track of the rapidly growing jungle of papers and to give awareness to lesser-known periodicals in which one might not expect related research.

The Bibliography of Quantitative Linguistics (BQL) is an ongoing project at Trier University under the direction of Reinhard Köhler, on which the author of this survey has been working as a student assistant. The resulting list is therefore by no means an exhaustive anthology since the data base is constantly updated by back issues of journals and by new publications.

However, as of February 2016 the project contains roughly 5000 relevant entries from 1097 different journals which cover many different aspects of quantitative linguistics, from corpus linguistics to computational linguistics to psychological experiments that quantitatively examined language in some way. Publications other than periodicals are excluded from this list.

Suggestions are welcome and should be sent to Tim Rostin, rost2701@uni-trier.de.

Journal	Number of Contributions
Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique / Classe des Sciences: Bulletin de la ... Bruxelles	2
ACLC Working Papers	3
ACM Computing Surveys	4
ACM Transactions on Information Systems	7
Acta Baltico-Slavica	1
Acta et Commentationes Universitatis Tartuensis = Tartu Riikliku Ülikooli Toimetised [Transactions of Tartu University]	4
Acta Linguistica	5
Acta linguistica Academiae Scientiarum Hungaricae	9
Acta psychologica	22
Acta Universitatis Carolinae. Philologica	1
Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica-Physica-Chemica	2
Acustica	1
Advances in Applied Probability	1
Advances in Complex Systems	5
Advances in Reading/Language Research	1
Aevum	1
African Language Review	1
African Language Studies	1
Afrika und Übersee	1
Akademie der Wissenschaften und der Literatur in Mainz / Abhandlungen der Geistes- und Sozialwissenschaftliche Klasse	1
Akademija nauk SSSR <Moskva>: Izvestija Akademija Nauk SSSR / Serija literatury i jazyka	1
Akademija nauk SSSR <Moskva>: Izvestija Akademija Nauk SSSR / Serija literatury i jazyka [before 1963: / Otdelenie...]	2
Akademija Nauk SSSR Moskva: Doklady ... = Comptes perdus de l'académie des Sciences de l'URSS	3
ALFA: Revista de Linguística	12
Alkalmazott matematikal lapok	2
ALLC Bulletin	39
ALLC Journal	20
Alta frequenza	1
AMANT (American anthropologist), New Series	1
American Anthropologist	2
American Antiquity	1
American Documentation	1
American Journal of Philology	10
American Journal of Psychology	23
American Mercury	1
American Sociological Review	4

American Speech	6
Angles on the English-Speaking World	2
Anglia - Zeitschrift für englische Philologie	6
Anglistik und Englischunterricht	1
Animal Behaviour	2
Annalele romîno-sovietice	1
Annales de l'Université de Paris	1
Annales des Télécommunications	1
Annals of Human Genetics	1
Annals of Otology, Rhinology and Laryngology	1
Annual Review of Psychology	2
Anthropological Linguistics	23
Anthropos	4
Antiquité classique	1
Anzeiger für Slavische Philologie	9
Applied Intelligence	1
Applied Linguistics	16
Applied Psycholinguistics	14
Arbeiten aus Anglistik und Amerikanistik	6
Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen	1
Arbeitspapiere der Universität Bern	1
Archiv für das Studium der neueren Sprachen und Literaturen	2
Archiv für die gesamte Psychologie	1
Archiv für vergleichende Phonetik	3
Archív orientální	16
Archive of Neurology and Psychiatry	1
Archives de psychologie	1
Archives néerlandaises de phonétique expérimentale	10
Archives of Acoustics	3
Archivum latinitatis Medii Aevi	1
Archivum linguisticum	1
Archiwum akustyki	3
Arkiv foer nordisk filologi. Lund	3
Asian and African Studies	9
ASSB (Annales de la société scientifique de Bruxelles), Series 1	1
Automatic Documentation and Mathematical Linguistics	14
Babel	2
Bălgarski ezik	11
Bastauyš mektep. Almaty	1
Behavior Research Methods	2
Behavior Research Methods & Instrumentation	2
Behavioral and Brain Sciences	1
Behavioral Science	4
Beijing hangkong xueyuan xuebao = Journal of Beijing Institute of	1

Aeronauties and Astronautics [北京航空学院学报]	
Beijing shifan daxue xuebao (ziran kexue ban) = Journal of Beijing Normal University (Natural Science) [北京示范大学学报 (自然科学版]	2
Beiträge zur Fachdidaktik	2
Beiträge zur Geschichte der deutschen Sprache und Literatur	2
Beiträge zur Phonetik und Linguistik	1
Beiträge zur romanischen Philologie	2
Beiträge zur Sprachkunde und Informationsverarbeitung	5
Beiträger zur Linguistik und Informationsverarbeitung	11
Bel'ckij gosudarstvennyj pedagogičeskij institut im. A. Russo: Učenye zapiski	2
Bell Labs Technical Journal	1
Bell System Technical Journal	8
Bell Telephone System Technical Publication	1
Berichte. Institut für Phonetik der Universität Köln	1
Berliner philologische Wochenschrift	3
Biblische Zeitschrift	2
Biological Cybernetics	2
Biological Psychology	1
Biology Letters	1
Biometrics	2
Biometrika	22
Biuletyn fonograficzny. Poznań	7
Biuletyn Polskiego Towarzystwa Językoznawczego	18
Biuletyn zarzadu głownego RSW "Prasa"	1
Brain	1
Brain and Language	4
British journal of psychology	6
British journal of psychology (London, England : 1953)	1
Brno Studies in English	1
Bulletin d’études orientales	1
Bulletin de la Faculté des Lettres de Strasbourg	1
Bulletin de la Société de Linguistique de Paris	19
Bulletin de la société polonaise de linguistique	1
Bulletin de psychologie	3
Bulletin des jeunes romanistes	3
Bulletin d'information du laboratoire d'analyse lexicologique	2
Bulletin of High Points (New York)	1
Bulletin of the Academy of Sciences, Georgia	4
Bulletin of the Central Institute of English	1
Bulletin of the Deccan College Research Institute	2
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering	1
Bulletin of the institute for research in English teaching	4

Bulletin of the international institute for linguistic sciences = Kokusai-gengo-kagaku-kenyshō	2
Bulletin of the Psychonomic Society	1
Bulletin of the Wisconsin Association of modern language teachers, Madison, WISC	1
Bulletin voor Taalwetenschap	1
Business Systems Research	1
Cahiers de lexicologie	18
Cahiers de linguistique théorique et appliquée. Bucarest	14
Cahiers de l'institut linguistique de Louvain	5
Cahiers du CERAT	1
Cahiers Ferdinand de Saussure	1
Cahiers Vilfredo Pareto	2
Canadian Journal of Linguistics	2
Canadian Slavonic Papers	1
Canadian Social Science	1
Časopis pro moderní filologii,Praha.	6
Centre de recherches et d'applications linguistiques <Nancy>: Cahiers du CRASP, Comptes perdus de l'académie des sciences de Paris	1
Cerebral Cortex	1
Česká literatura	5
Československá rusistika	5
Český jazyk a literatura. Praha	1
Cesty moderni jazykovedy	1
Changjiang xueshu = Yangtze River Academic [长江学术]	1
Chaos, Solitons & Fractals	1
Child Development	7
Child Language Teaching and Therapy	1
Childhood Education	1
Chinese Monthly	1
Chinese Science Bulletin	3
Chosonohak	1
CIIR Technical Report	1
Classica et Mediaevalia	2
Classical Journal	3
Classical Philology	8
Classical Review	3
Classical Weekly	2
Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology	1
Cognition	9
Cognition and Brain Theory	1
Cognitive Development	2
Cognitive Psychology	9
Cognitive Science	4

College English	7
Communications in Statistics, A: Theory and Methods	1
Communications of the ACM	5
Complex Systems	1
Complexity	6
Complexus	1
Comptes rendus des séances de l'Académie des Sciences	3
Comptes rendus hebdomaires des séances de l'Académie des sciences	1
Computational Linguistics	29
Computational Linguistics. Budapest	2
Computational Statistics & Data Analysis	1
Computational Statistics and Data Analysis	2
Computer	1
Computer Engineering and Applications	1
Computer Processing of Chinese and Oriental Languages	2
Computer Speech and Language	2
Computer Studies in the Humanities and Verbal behaviour	5
Computers and the Humanities	28
Computers in the Schools	1
Confinia Psychiatrica	2
Corpus	1
Corpus Linguistics and Linguistic Theory	62
Cortex	1
Current Anthropology	2
Current Directions in Psychological Science	1
Current Psychology Letters [Online]	1
Current Science	2
Cybernetica	1
Czech and Slovak Linguistic Review	1
Dacoromaia	1
Dalian haishi daxue xuebao (shehui kexue ban) = Journal of Dalian Maritime University (Social Sciences Edition) [大连海事大学学报(社会科学版)]	2
Denshi-gijutsu-sōgō-kenyūsho <Niihari>: Bulletin of the electrotechnical laboratory	2
Der mathematische und naturwissenschaftliche Unterricht	1
Deutsch als Fremdsprache	1
Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte	1
Deutschunterricht für Ausländer	1
Developmental Psychology	1
Diachronica	4
Dialektstudier	2
Dianzi xuebao = ACTA Electronica Sinica [电子学报]	1
Die lebenden Fremdsprachen	1

Die Musikforschung	1
Die Naturwissenschaften	3
Die neue Schulpraxis	1
Die neueren Sprachen	1
Die slawischen Sprachen	1
Die Sprache	1
Die Umschau in Wissenschaft und Technik	1
Die Unterrichtspraxis	3
Die Welt der Slaven	7
Discourse Processes	1
Dissertation abstracts international. Ann Arbor, Michigan	2
Dyslexia	3
Econometrica	1
Editorial office for contemporary foreign languages	1
Educational Research	1
Educational Research Bulletin	1
Educational Review	1
ELANS (Études de linguistique appliquée - Nouvelle série)	13
Electronic Journal of Human Sexuality	1
Electronic Journal of Vedic Studies	1
Elementary English	1
Engineering Cybernetics	1
Englische Studien	1
English and Germanic Studies	1
English for Specific Purposes	3
English Journal	1
English Language and Linguistics	3
English Language Teaching	1
English Studies	4
English Today	1
Entropy	3
Eos	1
EPL (Europhysics Letters)	1
Eranos	1
Essex Research Reports in Linguistics	1
Ethnomusicology	5
Études de linguistique appliquée	9
Études Finno-Ougriennes	1
Études germaniques	1
Euhemer	1
Euphorion	1
European Physical Journal B	1
Europhysics Letters	1
Europhysics Letters (EPL)	2

Evolutionary biology	1
Expert Systems with Applications	2
Ezik i literatura	1
Fernmeldetechnische Zeitschrift	1
Filologičeskij sbornik	1
Finnisch-Ugrische Forschungen	3
Finnisch-Ugrische Mitteilungen	11
Folia Linguistica	16
Folia linguistica historica	7
Folia Orientalia	3
Fonetică şi dialectologie	1
Foreign developments in machine translation and information processing	6
Foreign Language Teaching and Research	1
Forensic Linguistics	2
Forschungen und Fortschritte	5
Fortschritt der Psychologie	1
Forum der Letteren	2
Foundation and Trends in Information Retrieval	1
Foundations of language	2
Fractals	2
Frankfurter Phonetische Beiträge	2
Französische Studien	1
Fremdsprachenunterricht	1
French Review	6
French Studies	1
Fu Jen Studies	3
Functions of Language	1
Furman Studies, Furman University Bulletin	1
Fuza xitong yu fuzaxing kexue = Complex Systems and Complexity Science [复杂系统与复杂性科学]	1
Gdańskie Studia Językoznawcze	1
Gdańskie Zeszyty humanistyczne / Filologij rosyjska	1
Gdańskie Zeszyty humanistyczne / Prace jezykoznawcze	2
General Linguistics	3
Genetic Psychology Monographs	1
Genetics	1
Geographical Analysis	1
Geographical Review	1
Geographische Zeitschrift	1
Gercenovskie čtenija	2
German Quarterly	12
Germanistik. Tübingen	4
Germanistische Linguistik	10
Glossa	1

Glossologia	1
Glotta	3
Glottometrics	60
Glottometrika	60
Glottotheory	108
Goi Kenkyu = Studies on vocabulary	1
Göteborgs Högskolas Årsskrift	1
Göttinger Beiträge zur Sprachwissenschaft	7
Gramma	1
Gravesaner Blätter	1
Grazer linguistische Studien	1
Groninger Arbeiten zur Germanistischen Linguistik	1
Grundlagenstudien aus Kybernetik und Geisteswissenschaft. Tübingen	11
Gymnasium	1
Hamburger phonetische Beiträge	1
Handelingen van het Nederlands Filologencongress	1
Harvard Studies and Notes in Philology and Literature	1
Harvard Studies in Classical Philology	3
Hebrew computational linguistics	2
Hermathena	1
Hespéris	2
Hibbert Journal	1
Hiroshima daigaku bungakubu [Hiroshima University Studies]	1
Hispania	10
Homme	1
Human Brain Mapping	2
Human Physiology	1
IBM Journal of Research and Development	3
ICAME Journal	9
IEEE ASSP Magazine	1
IEEE Transactions of Automatic Control	1
IEEE Transactions on Computers	1
IEEE Transactions on Information Theory	12
IEEE Transactions on Knowledge and Data Engineering	2
IEEE Transactions on Pattern Analysis and Machine Intelligence	1
IEEE Transactions on Systems, man, and cybernetics	2
IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans	1
IETE Journal of Research: Journal of the Institution of Electronics and Telecommunication Engineers	1
Incontri linguistici	1
Indian Journal of Linguistics	1
Indian Linguistics	3
Indiana University Studies	1
INDJAL (Indian Journal of Applied Linguistics / Style, Structure and	1

Criticism)	
Indogermanische Forschung	4
Indogermanische Forschungen	21
Informační bulletin pro otázky jazykovědné	4
Information and Control	20
Information Processing & Management	8
Information Processing Society of Japan Journal	2
Information Retrieval	1
Information Sciences	7
Information Storage and Retrieval	3
Information Theory	1
Inostrannye jazyki v škole	1
Inostrannye jazyki v vyšej škole	5
Institut za Bălgarski Ezik <Sofija> : Izvestija	1
Institute of Electrical and Electronics Engineers Transactions ofnprofessional communication	1
Institute of the Classical Studies <London>: Bulletin of the ... of the University of London	1
Intermediair	1
International Economic Review	1
International Forum on Information and Documentation	1
International Journal of American Linguistics	20
International Journal of Applied Linguistics	6
International Journal of Bifurcation and Chaos	1
International Journal of Corpus Linguisitcs	2
International Journal of Corpus Linguistics	5
International Journal of Dravidian linguistics	3
International Journal of General Systems	1
International journal of Geographical Systems	1
International Journal of Man-machine studies	2
International Journal of Psycholinguistics	1
International Journal of Slavic Linguistics and Poetic	5
International Journal of the Sociology of Language	2
International Journal of Translation	1
International review of applied linguistics in language teaching	8
International review of the aesthetics and sociology of music	3
International Statistical Review	1
Internationales Archiv für Sozialgeschichte der deutschen Literatur	1
Interstate Bulletin	1
Intus News	1
IPO Annual Progress Report	1
IRE Professional Group on Information Theory, Transactions	1
IRE Transactions on information theory	3
Isis	1
Issledovanija po fonologii	1

Issledovanija po strukturnoj tipologii	2
Italian Journal of Applied Linguistics - Statistica applicata	1
Itogi nauki i techniki	1
Izvestija akademii nauk azerbajdžanskoj SSR <Baku> / Serija literatury i jazykov	1
Izvestija Akademii Nauk GSSR / Serija jazykov i literatury	1
Izvestija Akademii Nauk Kirgizskoj SSR, Frunze	1
Izvestija akademii nauk latvijskoj SSR	2
Izvestija Imperatorskoij Akademii Nauk / Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg	3
Izvestija sibirskogo otdelenija akademii nauk SSR	1
Jahrbuch des Committee on Modern Language Teaching	1
Jahrbuch für classische Philologie	1
Jazykovědné aktuality	2
Jazykovedný časopis	12
Jazykovedný sborník	1
Jazykovye universalii i lingvističeskaja tipologija	1
Język Polski	18
Języki obce w skole	1
Językoznastwo	1
Jiangnan daxue xuebao (ziran kexue ban) = Journal of Jiangnan University (Natural Science Edition) [江南大学学报 (自然科学版)]	1
Jisuanji gongcheng = Computer Engineering [计算机工程]	1
Jisuanji gongcheng yu yingyong = Computer Engineering and Applications [计算机工程与应用]	1
Joint Publications Research Service	1
Jornal of Applied Probability	11
Journal Belge de neurologie et psychiatrie	1
Journal de la société de statistique de Paris	2
Journal de psychologie normale et pathologique	2
Journal of abnormal and social psychology	4
Journal of Applied Psychology	4
Journal of Applied Statistics	1
Journal of Chemical Information and Computer Sciences	1
Journal of Child Language	3
Journal of Chinese Linguistics	2
Journal of Classification	1
Journal of Cognitive Neuroscience	3
Journal of Cognitive Science	1
Journal of Communication	2
Journal of Computational and Graphical Statistics	1
Journal of Computer and system sciences	1
Journal of Consulting and Clinical Psychology	1
Journal of cybernetics	1
Journal of Documentation	15

Journal of Econometrics	1
Journal of educational psychology	4
Journal of educational research	2
Journal of English linguistics	7
Journal of experimental child psychology	2
Journal of experimental pedagogy	3
Journal of experimental psychology	10
Journal of experimental psychology / General	3
Journal of experimental psychology / Learning, Memory, and Cognition	7
Journal of experimental psychology, human learning and memory	1
Journal of Experimental Psychology: Learning, Memory, and Cognition	1
Journal of General Psychology	8
Journal of genetic psychology	1
Journal of Information Processing	1
Journal of Informetrics	2
Journal of Intelligent Information Systems	1
Journal of juvenile research	1
Journal of Language and Linguistics	1
Journal of Law and Information Science	1
Journal of Learning Disabilities	6
Journal of Librarianship and Information Science	1
Journal of Linguistics	8
Journal of Machine Learning Research	1
Journal of Marketing	1
Journal of mathematical analysis and applications	2
Journal of Mathematical Physics	1
Journal of mathematical psychology	4
Journal of Memory and Language	46
Journal of Multivariate Behavior Research	1
Journal of music theory	9
Journal of Natural Language Processing	3
Journal of Neurolinguistics	1
Journal of Philology	1
Journal of philosophical logic	1
Journal of Phonetics	6
Journal of Physical Studies	1
Journal of Pragmatics	2
Journal of psycholinguistic research	15
Journal of Psychology	4
Journal of Quantitative Linguistics	441
Journal of Reading	2
Journal of Reading Disabilities	1
Journal of Scientific and Industrial Research	2
Journal of Semantics	1

Journal of Social Psychology	1
Journal of Software	1
Journal of speech and hearing disorders	4
Journal of Speech and Hearing Research	1
Journal of Statistical Mechanics: Theory and Experiment	1
Journal of Statistical Physics	1
Journal of Statistical Planning and Inference	2
Journal of the ACM	5
Journal of the acoustical society of America	20
Journal of the American Oriental Society	5
Journal of the American Society for Information Science	8
Journal of the American Society for Information Science and Technology	2
Journal of the American Statistical Association	13
Journal of the Elisha Mitchell scientific society	1
Journal of the Institute of Actuaries	1
Journal of the international Folk Music Council	1
Journal of the international phonetic association	1
Journal of the Polynesian Society	1
Journal of the Royal Asiatic Society of Great Britain & Ireland	1
Journal of the Royal Statistical Society / B	6
Journal of the Royal Statistical Society. Series A (General)	1
Journal of the Royal Statistical Society. Series A (Statistics in Society)	1
Journal of the Royal Statistical Society. Series A: Statistics in Society	5
Journal of theoretical biology	1
Journal of Verbal Learning and Verbal Behavior	87
Journal of Zhejiang University Science C	1
Journalism Quarterly	1
Kalbotyra	5
Keiryo Kokugogaku = Mathematical Linguistics [計量国語学]	315
Kibernetika	1
Kielikello: kielenhuollon tiedotuslehti	1
Klagenfurter Beiträge zur Sprachwissenschaft	2
Klagenfurter geographische Schriften	1
Kōdias, Code	1
Kokugogaku	1
Königlich-Sächsische Gesellschaft / Philosophisch-historische Klasse: Berichte über die Verhandlungen ...	3
Kopenhagener Beiträge zur germanistischen Linguistik	2
Kratylos	3
Kul'tura i pis'mennost Vostoka	1
Kultura i Społeczeństwo	1
Kurier der Rumänischstudenten	1
Kvantitativnaja lingvistika i avtomaticeskij analiz tekstov [Quantitative linguistics and automatic text analysis]	1

Kwartalnik Neofilologiczny	9
KYB (Kybernetika)	3
Kybernetiky a jeji využiti	1
La parole	1
La Pathologie générale	1
Ladinia	5
Language	98
Language & Communication	1
Language and Cognitive Processes	3
Language and Communication: Special Issues	1
Language and Linguistics Compass	3
Language and Literature	1
Language and Speech	41
Language and style	1
Language Dynamics and Change	1
Language Learning	1
Language Learning & Technology	1
Language Problems and Language Planning	1
Language Research	1
Language Resources and Evaluation	1
Language Sciences	5
Language Teaching and Linguistic Studies	2
Language Testing	4
Language, Culture and Curriculum	1
Language: Supplements	1
Languages	1
Langue française	1
LDV-Forum	4
Le français moderne	11
Le Monde Oriental	1
Le Moyen Age	1
Learning Disabilities Research & Practice	2
Learning Disability Quarterly	1
Leeds Studies in English and kindred languages	1
Les Cahiers de l'analyse des données	1
Les langues modernes	8
Leuvense Bijdragen	1
Lexicographica	1
Lexis	1
Library Trends	1
Lietuvos matematikos rinkinys	1
Limba Română	3
LIMP	1
Lingua	27

Lingua e stile	2
Lingua Nostra	1
Lingua Posnaniensis	20
Linguistic Analysis	1
Linguistic Communications	1
Linguistic Discovery	1
Linguistic Inquiry	4
Linguistic Sciences	2
Linguistic Typology	6
Linguistica [= Lingvistika]	24
Linguistica Antverpiensia	2
Linguistica slovaca	2
Linguistica uralica	1
Linguistics	3
Linguistics - An Interdisciplinary Journal of the Language Sciences	45
Linguistics in Amsterdam	4
Linguistik und Datenverarbeitung	2
Linguistique	5
Linguistique et Mathématiques	1
Linguistische Arbeitsberichte	1
Linguistische Berichte	16
Lingvistyčni Studiji	1
LISTENER	1
Listy filologické	1
Literární noviny	1
Literary and Linguistic Computing	25
Literaturblatt für germanische und romanische Philologie	1
Litteraria	3
Lore and Language	1
Magazin für Stenographie	1
Magyar nylev. Budapest	11
Mašinnyj perevod i prikladnaja lingvistika	3
Matematičeskoe prosveščenie	1
Matematyka	2
Materiały stowarzyszenia stenografów i maszynistek polskich	1
Mathematical Programming	1
Mathematics Magazine	1
Mathématiques et sciences humaines	4
Mathematisch-Physikalische Semesterberichte	1
Mechanical translation	3
Mediaeval studies	1
Medizinische Psychologie	1
Melos	1
Mémoires de la Société de linguistique	1

Mémoires de la société finno-ougrienne	3
Memory & Cognition	17
Meta: Journal des traducteurs	1
Methodica	3
Metodološki zvezki	1
Metódy analýzy a interpretácie hudby z historického a systematického aspektu	1
Michigan Quarterly Review	1
Mind	2
Minsk state linguistic university bulletin	1
Miscellanea barcinonensia	1
Mitteilungen des Instituts für Orientforschung	1
Mnemosyne	1
Modern language forum	6
Modern language journal [= The Modern Language Journal]	25
Modern language notes	7
Modern language quarterly	1
Modern language review	1
Modern languages	5
Modern Philology	5
Moderna språk	1
Moldavskij jazyk i literatura	1
Molecular Biology and Evolution	1
Monatshefte für deutschen Unterricht	4
Monatsschrift für höhere Schulen	1
Mondo Ladino	1
Mots	1
Movoznavstvo	1
Münchener Studien zur Sprachwissenschaft	1
Musikometrika	18
Muttersprache	13
Nachrichtentechnische Forschungsberichte	2
Nachrichtentechnische Zeitschrift	1
Naroda prosveta	1
Narody Azii i Afriki	1
Naše řeč	7
Natural Language and Linguistic Theory	1
Natural Language Engineering	2
Nature	12
Naturwissenschaftliche Rundschau	1
Naučno-Techničeskaja Informacija	8
Naučno-techničeskaja informacija / Serija 2: Informacionnye processy i sistemy	5
Naučnye doklady vysšej školy / Filologičeskie nauki	4
Nauka. Moskva	1

Naukovy Visnyk Cerniveckoho Universitetu	9
Naukowyj wisnyk Czerniwezjkoho universytetu. Serija: Hermans'ka filolohija wypusk [Wissenschaftliche Beiträge der Universität Czernowitz. Germanische Philologie]	11
Neuphilologische Mitteilungen	1
Neuphilologische Zeitschrift	1
Neural Computation	1
New Shakespeare Society Transaction Series	1
New Testament Studies	2
Non-Linear Analysis: Real World Applications	1
Norsk tidsskrift for språkvidenskap	4
Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles	1
Novum Testamentum	1
Nowa Kultura	1
Nuovo cimenta a cura della societa italiana difisica	1
Nyelvtudományi közlemények	10
Obučenie čteniju v nejazykovom vuze	1
Occasional papers on linguistics	1
Oceania	2
Oklahoma Academy of Sciences: Proceedings	2
Onomastica	1
Onze Taal	1
Onze Taaltuin	2
Open Linguistics	4
Orbis	4
Oriens Extremus	1
Orientalia christiana periodica, commentarii de re orientali aetatis christianae sacra et profana	1
Ornicar	1
Osnabrücker Beiträge zur Sprachtheorie	1
OSU Working Papers in Linguistics	1
Oxford Slavonic papers	1
Pallas	1
Pamiętnik literacki. Warszawa	8
Papers from the 10th Regional Meeting. Chicago Linguistic Society	1
Papers of Regional Science	1
Papers of the Chicago Linguistic Society	2
Papiere zur Linguistik	4
Pattern Recognition	2
Pattern Recognition Letters	1
Pedagogical Seminary	16
Pedagogika	1
Perception and Psychophysics	2
Perceptual and Motor Skills	1

Philologica	1
Philologica Pragensia	5
Philological Quarterly	4
Philologische Wochenschrift	1
Philologus: Zeitschrift für das klassische Altertum und sein Nachleben	4
Philosophical Journal	1
Philosophical transactions of the Royal Society	1
Philosophy of Science	2
Phoenix	1
Phonetica	27
Phonetica Pragensia	2
Phonologica	1
Phonology	3
Physica A: Statistical Mechanics and its Applications	23
Physical Review A	1
Physical Review E	2
Physical Review Letters	3
Physicalia Magazine	1
Physics of Life Teviews	2
Physikalische Blätter	2
PLoS ONE	5
Podstawowe Problemy Wspólczesnej Techniki	1
Poetics = Poetyka = Poėtika	15
Pokroky matematiky, fyziky a astronomie	1
Polonica	1
Popular science monthly	1
Poradnik językowy	18
Portugiesische Forschungen der Görres-Gesellschaft	1
Poznań Studies in Contemporary Linguistics	2
Prace filologiczne	11
Prace Instytutu Podstawowych Problemów Techniki PAN	1
Prague Studies in English	2
Prague Studies in Mathematical Linguistics	65
Praxis	1
Prikladnaja lingvistika	1
Princeton Conferences on Information Science and Systems	1
Problems in Transmission of Information	1
Problemy kibernetiki	5
Problemy peredači informacii	1
Problemy shkolnogo uchebnika [Problems of School Textbooks]	1
Proceedings of the American Philosophical Society	2
Proceedings of the Annual Meeting of the Berkeley Linguistics Society	3
Proceedings of the Department of Education	1

Proceedings of the National Academy of Sciences of the United States of America	1
Proceedings of the National Academy of Sciences of the USA	2
Procesamiento del Lenguaje Natural	1
Progress Report on Speech Research '78, Report on Pattern Information Processing System (PIPS)	1
Przegląd Humanistyczny	1
Psichologičeskij žurnal	1
PSiCL (Poznań Studies in Contemporary Linguistics)	1
Psychological Bulletin	3
Psychological Monographs	6
Psychological Record	10
Psychological Reports	2
Psychological Research	3
Psychological Review	5
Psychological Studies	1
Psychologische Beitrage	1
Psychologische Forschung	2
Psychometrika	2
Psychonomic Bulletin & Review	2
Psychonomic monograph supplement	1
Psychonomic Science	3
PTL - A Journal for Descriptive Poetics and Theory of Literature	1
Publications de l'Institut de statistique de l'Université de Paris	1
Publications of the modern language association of America	17
Qingbao kexue = Information Science [情报科学]	3
Qingbao xuebao = Journal of the China Society for Scientific and Technical Information [情报学报]	1
Qiqihar daxue xuebao (zhexue shehui kexue ban) = Journal of Qiqihar University (Philosophy and Social Science) [齐齐哈尔大学学报 (哲学社会科学版)]	1
Quaderni di Semantica	1
Quality & Quantity	4
Quantitative linguistics	1
Quarterly Journal of Speech	2
Rapa Nui Journal	2
Rassegna di studi etiopici	1
Raster	1
Reading and Writing: An Interdisciplinary Journal	2
Reading in a Foreign Language	4
Reading Psychology	1
Reading Research Quarterly	4
Recherches et méthodes nouvelles au service de l'enseignement des langues vivantes	1
Recherches sémiotiques = Semiotic inquiry	1
Recueil linguistique de Bratislava	3

RELC Journal	1
Rendiconti	1
Reports on Progress in Physics	1
Research in Ancient Chinese	1
Research in the Teaching of English	1
Research on Language and Computation	1
Review of Educational Research	2
Revistă de lingvistică şi ştiinţă literară	2
Revista italiana di linguistica applicata	1
Revista nacional de cultura	1
Revue Belge de philologie et d'histoire	1
Revue de l'Ecole nationale des langues orientales	1
Revue de l'enseignement supérieur	2
Revue de l'enseignement Supplements	1
Revue de linguistique romane	12
Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur	72
Revue de mathématiques pures et appliquées	2
Revue de musicologie	1
Revue de philologie, de littérature et d'histoire anciennes	3
Revue de ponétique appliquée	2
Revue de statistique appliquée	2
Revue des études latines	3
Revue des études slaves	2
Revue des langues vivantes	3
Revue française de science politique	2
Revue française de sociologie	1
Revue générale des sciences pures et appliquées	3
Revue Informatique et Statistique dans les Sciences humaines	67
Revue internationale d'onomastique	1
Revue Romane	1
Revue Roumaine de Linguistique	19
Revue Thalès	1
Rivista di linguistica	2
Rocnik orientalistyczny	1
Rocznik Slawistyczny	2
Roczniki polskiego towarzystwa matematycznego. Seria 2: Wiadomości matematyczne	1
Romance notes	1
Romance Philology	4
Romania	5
Románica	1
Romanische Forschungen	1
Romanistisches Jahrbuch	2
Ruch Filozoficzny	1

Ruch literacki	1
Russian linguistics	3
Russkaja razgovornaja rec	1
Russkaja reč'	1
Russkij filologičeskij vestnik	1
Russkij Jazyk	1
Russkij jazyk v nacional'noj škole	10
Russkij jazyk zu rubežom	1
Sananjalka	1
Sankhyā: The Indian Journal of Statistics	1
Sankhyā: The Indian Journal of Statistics / B	3
Sapostavitelno Ezikoznanie	1
Saratovskij gosudarstvennyj universitet im. N. G.	1
Saturday Review	1
Sbornik naučnych trudov	1
Scandinavian Journal of Psychology	3
Scienca revuo	1
Science	8
Science China Information Sciences	1
Scientific American	4
Scientometrics	2
Sdělovací technika	1
Semiosis	1
Semiotic Inquiry	1
Semiotica	7
Semiotika i informatika	14
Semiotische Berichte	2
Shakespeare Quarterly	2
SIAM Review	1
Significance	1
SKASE Journal of Theoretical Linguistics	1
SKY Journal of Linguistics	1
Slavia	2
Slavia occidentalis	2
Slavia orientalis	1
Slavic and East European Journal	1
Slavic and East European Studies / Études Slaves et Est-Européennes	3
Slavica	1
Slavica Pragensia	1
Slavistična revija	1
Slavjanskaja filologija	2
Slavjanskoe jazykoznanie	1
Slovenská literatúra	4
Slovenská řeč	3

Slovo a slovesnost	27
SMIL Quarterly	1
Smith College Classical Studies	1
Social Science Information	1
Society for the study of the indigenous languages of the Americans	1
Sociological Methods & Research	1
Sociological Models and Research	1
Sociometry	1
Soobščenija Akademii nauk Gruzinskoj SSR	5
South Pacific Journal of Psychology	1
Sovetskaja pedagogika i škola	2
Sovetskaja tjurkologija	1
Sovetskoe Finno-ugrovedenie	4
Soviet Physics Doklady	1
Speech communication	4
Speech monographs	1
Speech, Music and Hearing: Quarterly Progress and Status Report	1
Spektator	1
Sprache im technischen Zeitalter	1
Sprache und Datenverarbeitung	2
Sprache und Kognition	1
Sprachforum	1
Sprachkunst	1
Sprachtypologie und Universalienforschung	11
Sprachwissenschaft	2
Språkvetenskapliga Sällskapets i Uppsala färhandlingar	1
Sprawozdania PAN	3
Sprawozdania z Posiedzeń Komisji Językowej Towarzystwa Naukowego Warszawskiego	2
Sprawozdania z posiedzeń komisji naukowych oddziału PAN w Krakowie	2
Sprawozdania z posiedzeń komisji orientalistycznej	2
Sprawozdania z posiedzeńkomisji oddziału krakowskiego polskiej	1
Statistica Sinica	1
Statistical methods in linguistics	22
Statistical Science	1
Statistics and Computing	1
Statistics and Probability Letters	1
Statistics in Medicine	2
Statistique et applications linguistiques	1
STEK	1
Stenograf polski	2
Stenograf. Ežemesjačnyj žurnal, posvjaščennyj voprosam naučnoj i praktičeskoj stenografii	1
Strukturnaja i matematičeskaja lingvistica	17

Strumenti critici	1
Studi i problemi di critica testuale	1
Studi italiani di linguistica teorica e applicata	1
Studia Anglistica posnaniensia	2
Studia filozoficzne	1
Studia Leibnitiana	1
Studia Linguistica	19
Studia neophilologica	2
Studia Polonistyczne, Uniwersytet im. A. Mickiewicze w Poznaniń	2
Studia religioznawcze	1
Studia Universitatis "Babes-Bolyai" Seria Informatica	1
Studia z filologii polskiej i słowiańskej	1
Studier i nordisk filologi	4
Studies in Language	1
Studies in linguistics	1
Studies in Philology	5
Studies in Second Language Acquisition	2
Studies in Slavic and General Linguistics	2
Studies in the linguistic sciences	1
Studies of Chinese Language	1
Studii şi cercetări lingvistice	5
Studium Generale	2
Style	3
Suomalais-ugrilaisen seuran aikakauskirja/ Journal de la Societé Finno-Ougrienne	1
Survey of English Usage, UCL	3
Svensk tidskrift för musikforsking	1
Symbolae Osloensis	1
Synthese	3
System	3
TAINF	1
Target	1
Teachers' College Record	2
Teorija peredači soobščenij	1
Teorija verojatnostej i ee primenenije	1
TESOL Quarterly	2
Texas studies in literature and language	1
The American mathematical monthly	1
The American Naturalist	1
The American Statistician	2
The Annals of Mathematical Statistics	3
The Annals of Statistics	1
The Annals of the Harvard Computation Laboratory	1
The Bible Translator	1

The Bulletin of the Phonetic Society of Japan	1
The Classical Quarterly	21
The Elementary English Review	1
The European Physical Journal Special Topics	1
The Higher Education	1
The Incorporated List	1
The Journal of Abnormal and Social Psychology	1
The Journal of Comparative Germanic Linguistics	1
The Journal of Experimental Education	1
The Mathematical Scientist	2
The Mental Lexicon	1
The New Scientist	1
The New Shakspere Society Transitions	1
The Philippine Journal of Linguistics	1
The Prague bulletin of mathematical linguistics	58
The Psychology of Learning and Motivation	1
The Reading Teacher	1
The Review of English studies	7
The Sciences	1
The Seventeenth Century	1
The Speller	1
The Statistician	1
The Theory of Science	1
The William and Mary Quarterly	2
Theoretical Linguistics	4
Theory and Practice in Language Studies	1
Theory of Probability and its Applications	2
Tianjin shifan daxue xuebao = Journal of Tianjin Normal University [天津师范大学学报]	1
TISUANG	1
TJDL	1
Töid keelestatistika alalt [Papers on linguostatistics]	3
Tōkyōgaikokugodaigaku ronshū = Area and Culture Studies [東京外国語大学論集]	58
Tongji daxue xuebao (shehui kexue ban) = Journal of Tongji University (Social Science Section) [同济大学学报 (社会科学版)]	1
Traduction automatique	1
Transactions and Proceedings of the American Philological Association	10
Transactions of the American Philosophical Society	1
Transactions of the Bibliographical Society, 4th Series	1
Transactions of the New York Academy of sciences	1
Transactions of the Philological Society	6
Transactions of the Society of Instrument and Control Engineers	1
Travaux de linguistique et littérature	9
Travaux de l'institut phonétique de Strasbourg	2

Travaux du cercle linguistique de Nice	1
Travaux du cercle linguistique de Prague	3
Travaux linguistiques de Prague	2
Trends in Genetics	1
Trends in Information Management	1
Trudy AN Litovskoj SSR, Serija obščestvennych nauk	1
Trudy instituta kibernetiki akademii nauk gruzinskoj SSR	2
Trudy Samarkandskogo gosudarstvennogo universiteta	1
Tushu qingbao gongzuo = Library and Information Service[图书情报工作]	1
Ucenye zapiski	1
Učenye zapiski kišinevskogo gosudarstvennogo universiteta im. V. I. Lenina	2
Učenye zapiski Mockovskogo oblastnogo pedagogičeskaja instituta im. N. K. Krupskoj	1
Učenye zapiski vuzov Litovskoj SSR	1
UCL Working Papers in Linguistics	1
Ungarische Jahrbücher	2
Universal Access in the Information Society	1
Universität Kiel / Seminar für Allgemeine und Indogermanische Sprachwissenschaft: SAIS Arbeitsberichte aus dem Seminar ...	1
University of California Publicatins in English	2
University of California Publications in American Archaeology and Ethnology	1
University of California publications in Linguistics	1
University of California publications in Statistics	1
University of Colorado studies	1
University of Iowa Studies: Child Welfare	1
University of Nebraska Studies	2
University of Virginia Abstracts of Dissertations	1
University of Wisconsin studies in language and literature	1
Universum	1
Uniwersytet Gdański / Wydział Humanistyczny: Zeszyty Naukowe Wydziału Humanistyczny / Filologia Polska / Prace językoznawcze	1
Uniwersytet Śląski: Prace naukowe Uniwersytetu Śląskiego w Katowicach. Prace językoznawcze	1
Uppsala Universitets Årsskrift = Acta Universitatis Upsaliensis	1
Ural-Altaische Jahrbücher	9
Uralische Philologie	1
Vestnik akademii nauk SSSR	1
Věstník AV åľR	1
Vestnik belaruskaha dzaržaunaha universitėta inja U. I. Lenina Serija 4	2
Věstnik královské české akademie věd a uměni	1
Věstník Královské české společnosti nauk	1
Vestnik Leningradskogo gosudarstvennogo universiteta	2
Vestnik Moskovskogo universiteta / Serija 1: Filologija	2

Vestnik obščestvennych nauk akademii nauk Armjanskoj SSR	2
Vestnik vyššej školy	1
Vigo International Journal of Applied Linguistics	1
Virittäjä	12
Vision Research	1
Voprosy informacionnoj teorii i praktiki	2
Voprosy jazykoznanija = Problems of Linguistics [Вопросы языкознания]	34
Voprosy kazachskoj fonetiki i fonologii	1
Voprosy Psichologii [Problems of Psychology]	2
Voprosy romano-germanskoj filologii i metodiki prepodavanija inostr. jazykov	1
Voprosy statistiki reči	5
Vox romanica	3
VPJ	1
Vyčislitel'nye sistemy. Sbornik trudov instituta matematiki SO AN SSSR	3
Waiguoyu (Shanghai waiguoyu daxue xuebao) = Journal of Foreign Languages [外国语 (上海外国语大学学报)]	1
Waiyu jiaoxue yu yanjiu = Foreign Language Teaching and Research [外语教学与研究]	2
Waiyu yu waiyu jiaoxue = Foreign Languages and Their Teaching [外语与外语教学]	2
Web Journal of Formal, Computational and Cognitive Linguistics	1
Wiener Archiv für Psychologie, Psychiatrie und Neurologie	1
Wiener linguistische Gazette	1
Wiener slavistischer Almanach	2
Wiener slavistisches Jahrbuch	1
Wiener Studien	1
Wiener Zeitschrift für die Kunde des Morgenlandes	1
Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik	1
Wirkendes Wort	2
Wissenschaftliche Zeitschrift der Humboldt-Universität Berlin / Gesellschafts- und sprachwissenschaftliche Reihe	1
Wissenschaftliche Zeitschrift der Universität Halle / Gesellschafts- und sprachwissenschaftliche Reihe	1
Wissenschaftliche Zeitschrift der Universität Jena / Gesellschafts- und sprachwissenschaftliche Reihe	3
WMU Journal of Maritime Affairs	1
Word	28
Word Ways - The Journal of Recreational Linguistics	2
Yale classical studies	1
Yearbook Modern Language Forum	1
Yearbook of the national society for the study of education	4
Yulin shizhuan xuebao = Journal of Yulin Teachers College	1

(Philosophy and Socical Science) [玉林师专学报]	
Yuyan Wenzi Yinyong = Applied Linguistics [语言文字应用]	3
Yuyan yanjiu = Language study [语言研究]	1
Z polskich studiów sławistycznych	3
ZADS (Zeitschrift des allgemeinen deutschen Sprachvereins), Wissenschaftliche Beihefte	1
Zastosowania Matematyki	8
Zbornik za filologiju i lingvistiku	1
ZDULB (Zeitschrift für Dialektologie und Linguistik) - Beihefte	4
Zeitschrift der deutschen morgenländischen Gesellschaft	2
Zeitschrift für alttestamentliche Wissenschaft	1
Zeitschrift für angewandte Mathematik und Physik	2
Zeitschrift für Anglistik und Amerikanistik	1
Zeitschrift für Antikes Christentum	1
Zeitschrift für Balkanologie	1
Zeitschrift für deutsche Philologie	3
Zeitschrift für deutsche Sprache	1
Zeitschrift für Dialektologie und Linguisitik	5
Zeitschrift für empirische Textforschung	6
Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie	1
Zeitschrift für experimentelle und angewandte Psychologie	16
Zeitschrift für französische Sprache und Literatur	2
Zeitschrift für germanistische Linguistik	1
Zeitschrift für Klinische Psychologie	1
Zeitschrift für Literaturwissenschaft und Linguistik	9
Zeitschrift für Mundartforschung	4
Zeitschrift für pädagogische Psychologie	2
Zeitschrift für Phonetik und allgemeine Sprachwissenschaft	3
Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung	25
Zeitschrift für Physik	1
Zeitschrift für Psychologie	3
Zeitschrift für romanische Philologie	8
Zeitschrift für Russisch-Unterricht	1
Zeitschrift für Semiotik	2
Zeitschrift für slavische Philologie	1
Zeitschrift für Slawistik	2
Zeitschrift für Sprachwissenschaft	6
Zeitschrift für vergleichende Sprachforschung [also: Kuhn'sche Zeitschrift - KZ]	14
Zeszyty językoznawcze	5
Zeszyty Naukowe Uniwersytetu Łodzkiego	1
Zeszyty Naukowe Wyższej Szkoły Pedagogicznej w Opolu	1
Zeszyty Naukowe Wyższej Szkoły Pedagogicznej w Szczecinie	1
Zeszyty Prasoznawcze	3

Zeszyty teoretyczne stowarzyszenia stenografów i maszyni stek polskich	1
Zhejiang daxue xuebao (renwen shehui kexue ban) = Journal of Zhejiang University (Humanities and Social Sciences) [浙江大学学报(人文社会科学版)]	2
Zhongguo yuwen = Chinese language] [中国语文]	1
Zhongwen xinxi xuebao = Journal of Chinese Information Processing [中文信息学报]	10
Zielsprache Deutsch	1
Znanie - sila	1
Zoological Studies	1
ZPHSK (Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 1961-1992)	1
Zprávy státního těsnopisného ústavu	1
Žurnal psichologii pedagogij i psikotechniki	1

Book review

Hanna Gnatchuk: *Sound Symbolism. A phonosemantic analysis of German and English consonants*. Saarbrücken: Akademiker Verlag, 2015, 96 pp.

Reviewed by Denys Ishutin (shutndenis@mail.ru)

The debatable problem of a direct linkage between sounds and meanings (sound symbolism) has been a matter of concern throughout centuries. The author of the book focuses on the detection of the meanings for English and German consonants in the human psyche and at a textual level statistically. In such a way, subjective (in the human psyche) and objective (at a textual level) types of sound symbolism have been analyzed in the present book. In this project, the author emphasizes the importance of following correct methodological demands in order to receive authentic and objective results.

On the whole, the book consists of three chapters. The first chapter "Theoretical fundamentals of phonosemantics as a linguistic discipline" deals with a thorough overview of theoretical problems in the given branch of linguistics. A careful attention is paid to the previous studies of the analyzed phenomenon as well as to its typology (classification). A critical consideration is given to the unresolved problems in phonosemantics – character (universal or national), methodology and nature of sound symbolism (which factors evoke the phenomenon). Moreover, the author sheds light upon the results of the experimental studies of sound symbolism (subjective sound symbolism) as well as the research at a lexical level. She outlines the methodological errors in the procedures of the previous research. Finally, she gives a thorough description of phonesthemes and the studies on this issue.

The second chapter "The investigation of subjective sound symbolism" deals with the detection of the meanings for German and English consonants in the human psyche. In this case, the author undertakes a psycholinguistic experiment among German and English native speakers by giving the questionnaires with the necessary instructions. The data have been statistically processed with the help of Osgood's Semantic Differential, chi-squared test and Chuprov's coefficient K. The following scales of the Osgood's Semantic differential have been used: (1) weak – strong, (2) pleasant – unpleasant, (3) slow-fast, (4) small – big, (5) cruel – kind, (6) rough – smooth. The aim of the chi-squared test is to corroborate or falsify statistically significant connections between sounds and meanings in the human psyche. The degree of this connection is determined with the help of Chuprov's coefficient K (the coefficient of contingency). Therefore, the results of the psycholinguistic experiment have shown that German and English consonants express certain meanings. The semantics of the sounds has turned out to depend upon the voice of the consonants. Moreover, the meanings have been systematized for each consonant statistically. The results can be useful for the creation of brand names.

The third chapter "The investigation of objective sound symbolism" is engaged with a quantitative analysis of English and German poems and prose excerpts. The objective of this chapter is to corroborate the connection between the frequencies of the consonants and the mood of the poems (prose extracts). Therefore, the author divides the poems and prose pieces into two groups – optimistic and pessimistic group of texts. The optimistic group includes the poems with a positive description of the events (marriage, joy, gratefulness, etc) and pessimistic description (sorrow, war, etc). In such a way, the frequencies of each consonant in each group have been found. The data have been statistically treated by means of the chi-squared test and Chuprov's coefficient K. As a result, the author has corroborated the connection between the mood of the texts and the usage of certain consonants. The outcomes can be of great use to the authors who are intended to make a certain impression on the audience by means of the usage of certain sounds.

As a matter of fact, the study is a good starting point for the investigation of the human view of the world in relation both to the human internal states and to the evolution of human abilities from the biological point of view. The words in which some phonosemantic relations are detected can be ordered chronologically according to our evolution from primitive life of organisms up to the human intellectual abilities. A cooperation with biologists is to be recommended. At the highest, human, level perhaps psychologists could intervene. Though linguists consider phonosemantics a matter of linguistics, an interdisciplinary approach would be of advantage for several humanistic disciplines.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

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