Glottometrics 39
2017

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

External academic peers for Glottometrics

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 39 (2017), Lüdenscheid: RAM-Verlag, 2017. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 39 (2017) **ISSN 2625-8226**

Contents

Glottometrics 39, 2017, 1-37

A Bibliometric Analysis of *Glottometrics*

Yanni Lin[1], *Haitao Liu*[1,2]

Abstract. *Glottometrics*, one of the most authoritative journals in quantitative linguistics, has celebrated its 17th anniversary in 2017. In this paper, we conduct a bibliometric study of this journal. By statistical analysis of the basic data in all the 37 volumes published so far (2001-2017), we explore the publication profile, contributors, research content, and citations based on the self-built library and corpora. Results provide a glimpse of development and research status of quantitative linguistics. Suggestions of further improvements for this journal are also proposed.

Keywords: *Glottometrics*; bibliometrics; quantitative linguistics

1. Introduction

As a sub-discipline of linguistics, Quantitative Linguistics (or QL) studies linguistic phenomena (properties, structures, processes) and their interrelations, whose methodology is characterized by quantitative methods and instruments ranging from mathematical tools to simulation and modeling (Best, 2006; Köhler, Altmann, & Piotrowski, 2005). The International Quantitative Linguistics Association (IQLA) and the International Conference on Quantitative Linguistics (QUALICO) are two most important international forums for quantitative linguists. With special focalization and profession, *Journal of Quantitative Linguistics* and *Glottometrics* are deemed as the most authoritative journals in QL.

Capturing the research status of an area, as is known, is the starting point of forming a strategic visions and conducting scientific research. In library and information science, bibliometrics is used to analyze academic literature and evaluate research performance quantitatively, especially for universities, policy makers, research directors, librarians and researchers themselves. Nowadays in the Information Age, we have easy access to the research status and trends via content analysis and citation analysis. Databases (e.g. Web of Science, Scopus) and software (e.g. RefViz, CiteSpace, and Quosa) provide a more efficient way to detect burst terms, identify research fronts and visualize patterns and trends in scientific research.

As the names of *Glottometrics* and "bibliometrics*"* imply, the shared suffix *-metrics* suggests a methodological similarity between them: measuring textual objects. In quantitative sense, it is natural to see that bibliometric method is employed in analyzing the literature in QL. Through quantitative analysis of 66 issues in *Journal of Quantitative Linguistics*, Chen and Liu (2014) investigated the objects, aims, methodologies as well as focuses, shifts and representative achievements of QL.

In this study, a bibliometric study of *Glottometrics* is conducted. The research questions of our study are: (1) What is the publication profile of the journal? (2) Which authors,

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

countries and regions, and affiliations contribute most to the journal? (3) What themes do these articles focus on? Are there any shifts throughout the years? (4) Which of the source articles are cited most? What kinds of articles cite the journal? Which references occur most frequently in the bibliographies? We expect to provide a better overview of QL and suggestions for improving the academic impact of this journal.

The rest of this paper is organized as follows: Section 2 introduces the material and method used in this study; in Section 3, the results of bibliometric analysis are illustrated and discussed; the concluding remarks come in the final section.

2. Material and Method

Glottometrics (ISSN 1617-8351) is a scientific journal for the quantitative research of language and text published 2-3 times a year by RAM-Verlag in Germany. It has been indexed in Emerging Sources Citation Index (ESCI) since 2015, and then accepted for inclusion in Scopus since 2017. All issues are available as printed and electronic editions (pdf-files free download from its official homepage[2]). As for its aim and scope:

"The aim of *Glottometrics* is quantification, measurement and mathematical modeling of any kind of language phenomena. We invite contributions on probabilistic or other mathematical models (e.g. graph theoretic or optimization approaches) which enable to establish language laws that can be validated by testing statistical hypotheses."[3]

The editorial board of *Glottometrics* consists of the following members: G. Altmann (Univ. Bochum, Germany), K.-H. Best (Univ. Göttingen, Germany), R. Čech (Univ. Ostrava, Czech Republic), F. Fan (Univ. Dalian, China), P. Grzybek (Univ. Graz, Austria), E. Kelih (Univ. Vienna, Austria), R. Köhler (Univ. Trier, Germany), H. Liu (Univ. Zhejiang, China), J. Mačutek (Univ. Bratislava, Slovakia), G. Wimmer (Univ. Bratislava, Slovakia), and P. Zörnig (Univ. Brasilia, Brasilia). The majority of the editorial board are from the European countries except for two Chinese linguists Liu and Fan from Asia.

Up to June 30th, 2017, the journal has published altogether 37 volumes (330 articles), covering a time span from the year 2001 to 2017, which is divided into four time slices of five years for better discussion: Period I (2001~2005), Period II (2006~2010), Period III (2011~2015) and Period IV (2016~2017).

A lack of complete citation data of *Glottometrics* (2001~2017) in databases even including Scopus and Web of Science causes difficulties in bibliometric analysis. Thus lots of efforts are made to fulfil the fields of the Endnote library manually based on the information collected in the downloaded full texts. For the same reason, it is also difficult to visualize the patterns and trends in bibliometric instruments such as Web of Science and CiteSpace. Without the aid of these tools of high efficiency, items are counted in Microsoft Excel instead in our study.

After downloading all the articles as the source material from the homepage of *Glottometrics*, we first build an Endnote[4] library of metadata manually. Each record has 11 regular fields (namely, *type of work*, *author*, *year*, *title*, *volume*, *pages*, *keywords*, *abstract*, *country*, *affiliation*, *language*). Two additional fields, *viz.*, *research theme* and *research object* of a research article are also marked. Besides, the corpus of *keywords* and the corpus of *abstracts* are built respectively, each with four sub-corpora for different periods. Then, based on the

[2] URL: http://www.ram-verlag.eu/journals-e-journals/glottometrics/
[3] URL: http://www.ram-verlag.eu/wp-content/uploads/2012/09/Aims-and-Scope-Editorial-Board.pdf
[4] Endnote is a commercial reference management software package developed by Clarivate Analytics (URL: http://endnote.com/).

counts of the fields above, we give a statistical analysis of the journal profile (publication frequency, type of work, length of article, and language) and contributors (authors, countries and regions, and affiliations). Additionally, research content, to be more specific, research themes and their diachronic changes are tracked by using AntConc to generate the wordlists and N-Gram lists for the corpora of keywords and abstracts. Next, the frequently occurring cited references and the most cited source references are counted and described statistically; a bibliometric profile for citing articles is given with the help of citation data from Web of Science and Google Scholar.

3. Results and Discussion

3.1. Publication Profile

3.1.1. Publication frequency

The first volume of *Glottometrics* was issued in 2001. Over the past 17 years, 37 volumes (330 articles) have been published so far (up to June, 2017). Its publication frequency over the years is shown in Figure 1.

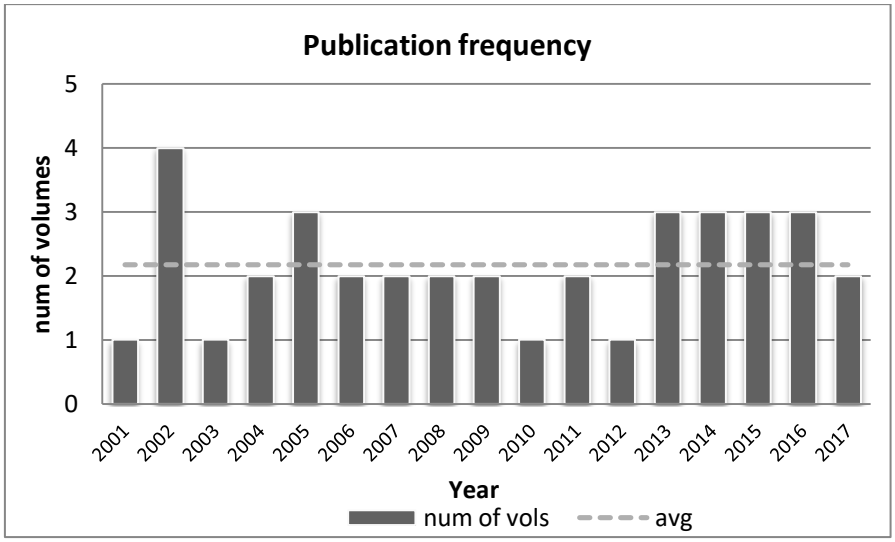

Figure 1. Volumes by year

Overall, the journal has kept its stated publication frequency of 2~3 times a year, except in 2001 (once), 2002 (4 times), 2010 (once) and 2012 (once).

The counts of articles in each volume are provided in Figure 2.

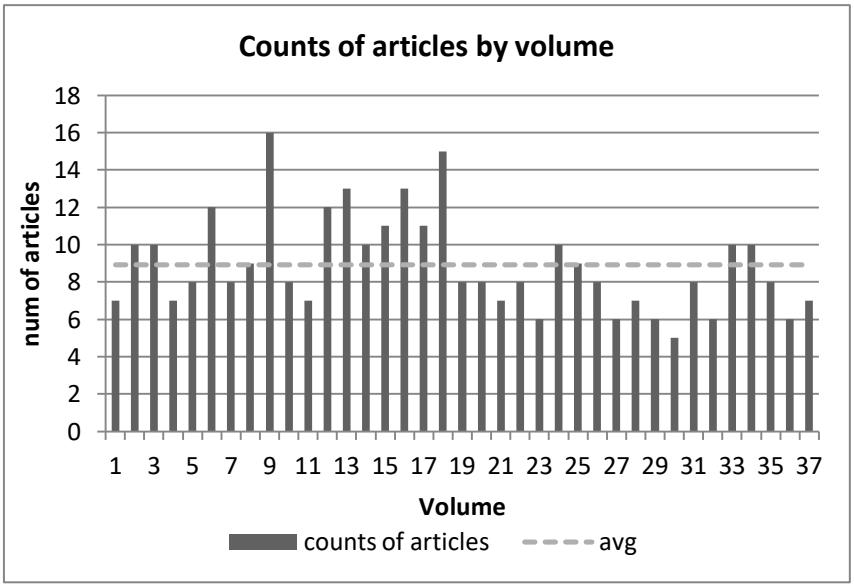

Figure 2. Articles by volume

Figure 2 displays the number of articles in a volume varies from 7 to 16 with an average of 9 over the years.

3.1.2. Types of Work

The articles of *Glottometrics* fall into six types: "general article", "book review", "history", "bibliography", "discussion" and "miscellanea". Among them, "history" is a featured type of work in the journal which introduces important linguists and their achievements in the history of QL. Figure 3 and Figure 4 show the number and proportion of each type of work as well as their diachronic changes in number.

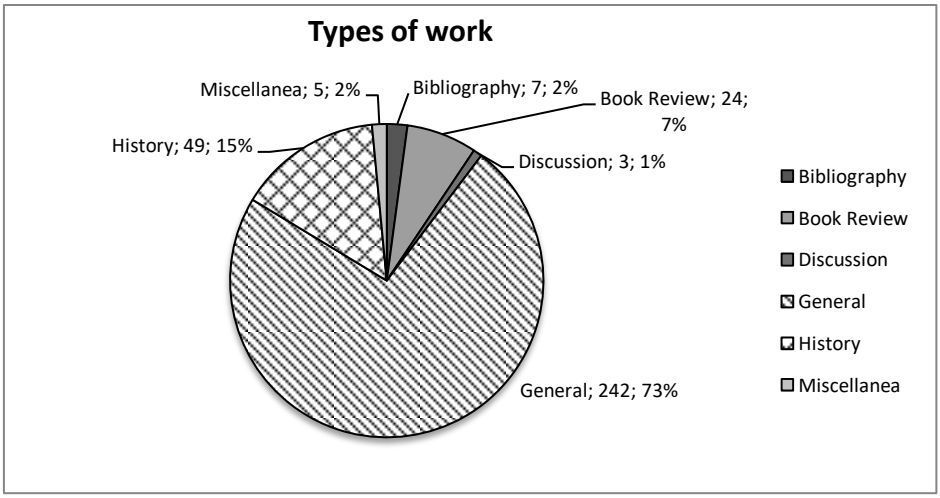

Figure 3. Type of work

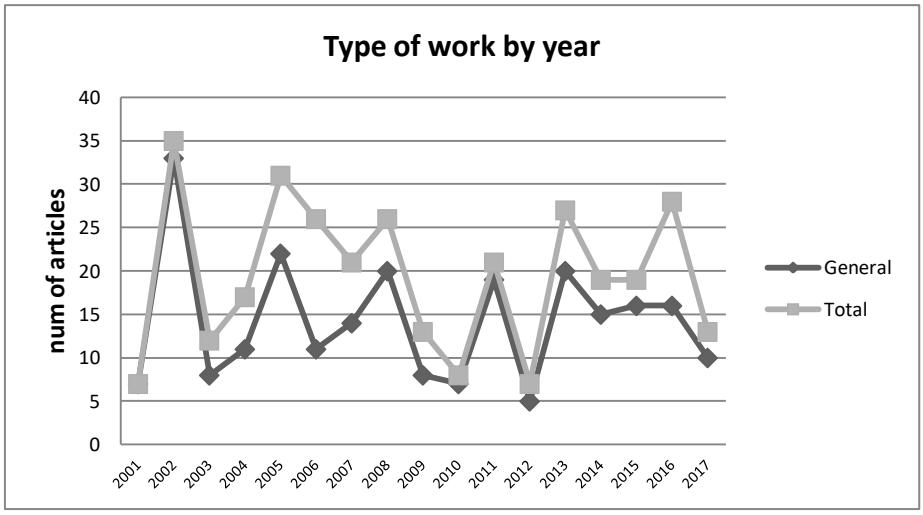

Figure 4. Type of work by year

As is shown in Figure 3 and Figure 4, general articles have kept the highest proportion in all types of articles over the years. Articles about history also take an important part especially from 2003 to 2010. The frequency of book reviews ever reached its peak in 2006 and 2013. Like other types of work, it appears unregularly in the timeline, accounting for just a small proportion.

3.1.3. Lengths of Article

The length of an article is also calculated as displayed in Table 1.

Table 1
Pages per article of each type of work

Type of work	Min (pages)	Max (pages)	Avg (pages/article)
Bibliography	2	28	9
Book Review	1	10	4
Discussion	2	6	4
General	4	46	14
History	2	33	6
Miscellanea	1	13	6
Total	1	46	12

The lengths of an article vary greatly both within and across different types of work: an average length for all the articles is 12 pages; a book review or a miscellanea can be as short as only one page, while a general article can reach as long as 46 pages.

3.1.4. Languages

All the submissions to *Glottometrics* are written in either English or German. Chronological changes in proportions of the two languages with and without the consideration of type of work are given in Figure 5 and Figure 6 respectively.

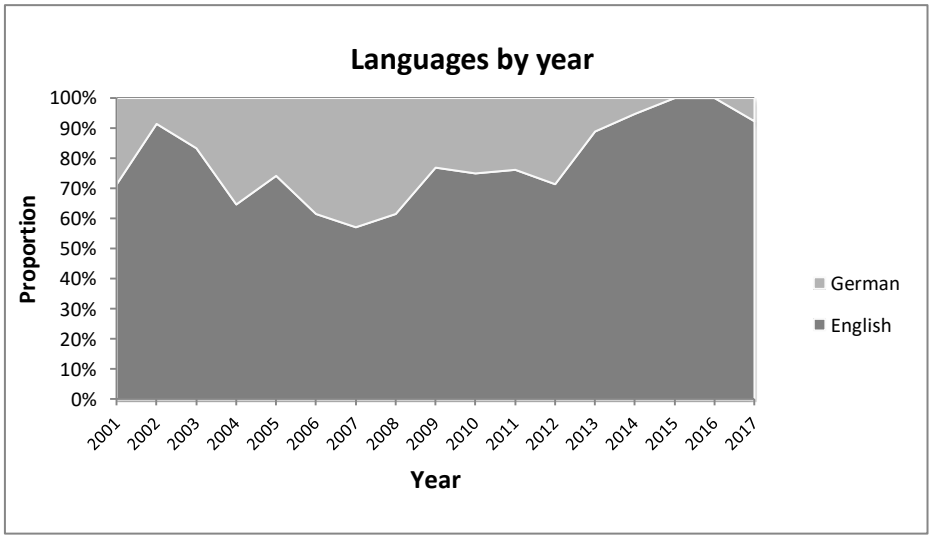

Figure 5. Languages by year

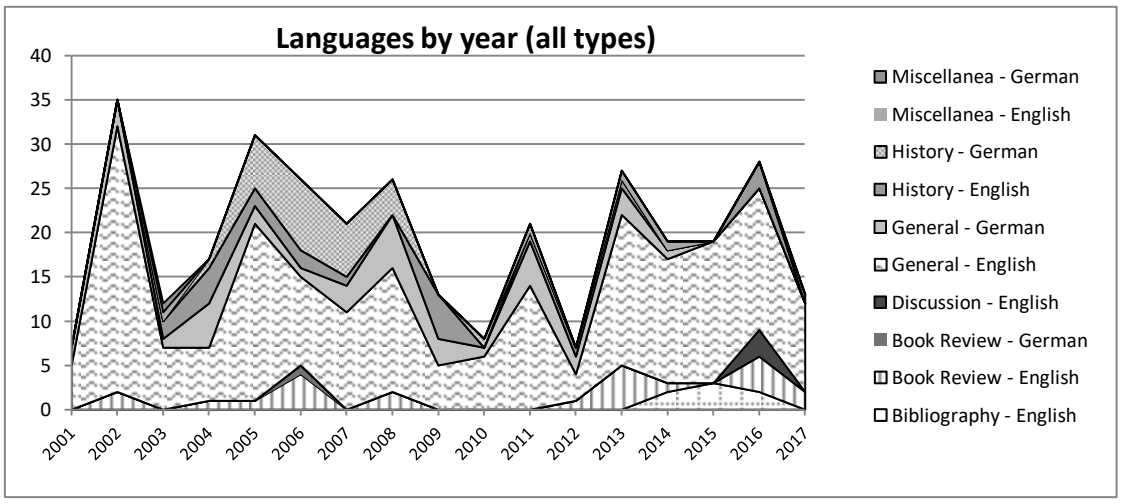

Figure 6. Languages by year and type

As is shown in Figure 5 and Figure 6, the majority of articles are written in English. During the first decade, German articles, especially book reviews and introductions to QL history were commonly seen. Since the journal was indexed by ESCI, all the articles have been written in English.

3.2. Contributors

3.2.1. Authors

A rank of contributing authors is given in Figure 7 (among all the 201 authors, those who contribute less than three articles are not shown in this shortlist).

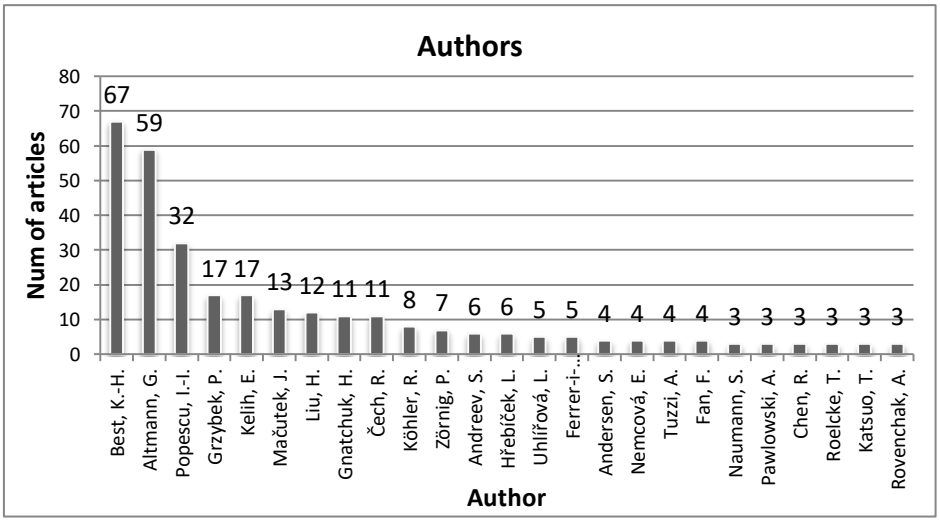

Figure 7. Authors (all types of work, freq. >= 3)

Figure 7 clearly shows that Best and Altman are leading scholars, contributing more than 50 articles to *Glottometrics*. Other authors like Popescu, Grzybek, Kelih, Mačutek, Liu and Gnatchuk are quite productive as well.

When type of work is taken into consideration, results of counts of authors are shown in Figure 8 (for general articles), Figure 9 (for introductions to QL history) and Figure 10 (for book reviews) respectively.

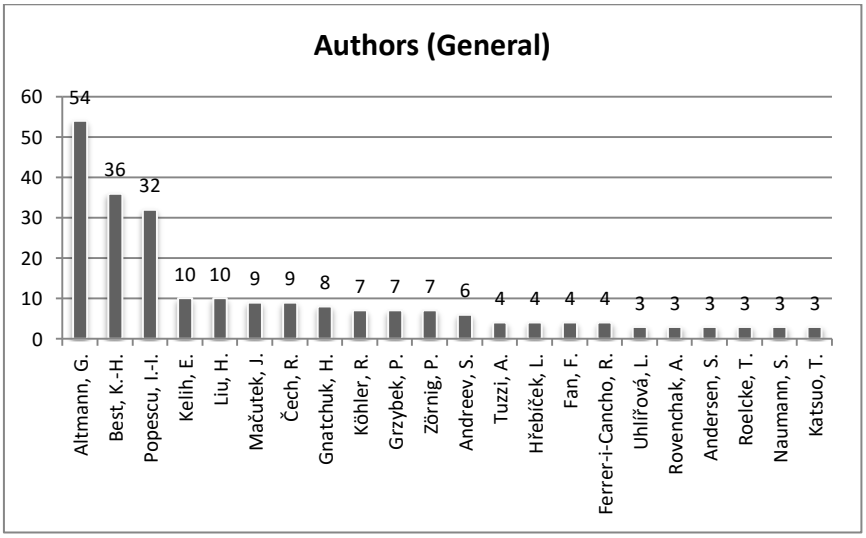

Figure 8. Authors (general articles, freq. >= 3)

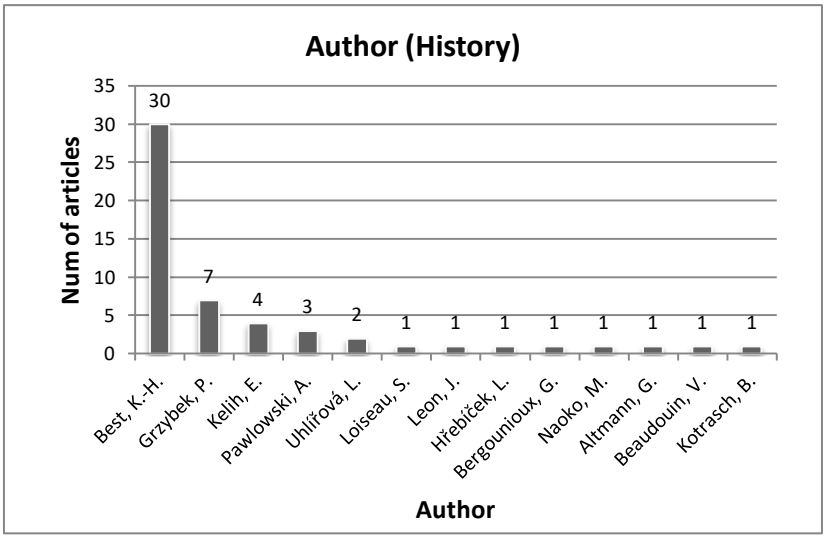

Figure 9. Authors (history)

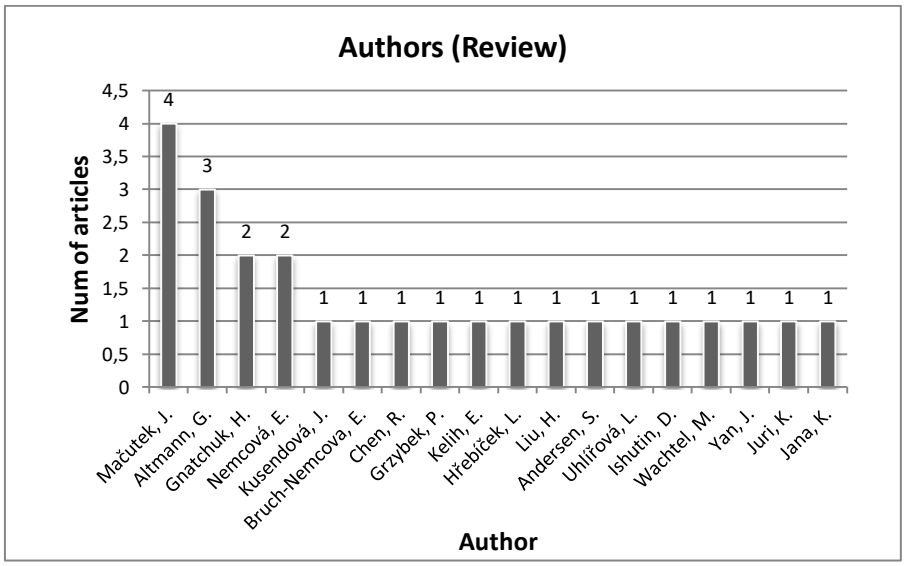

Figure 10. Authors (book reviews)

As for general articles, as is seen in Figure 8, Altmann and Best switch to the other's position, while the ranking order is more or less the same as that in Figure 7. Figure 9 and Figure 10 show that Best is the leading scholar in contributing introductions to history and Mačutek in book reviews.

Co-authors are commonly seen within the network of research community. For example, Popescu and Altmann, Grzybek and Kelih, Best and Altmann, have kept their long-time cooperation.

3.2.2. Countries and Regions

All the published 330 articles are written by 201 authors from 25 countries and regions. A pie chart of these countries and regions' contributions is given in Figure 11.

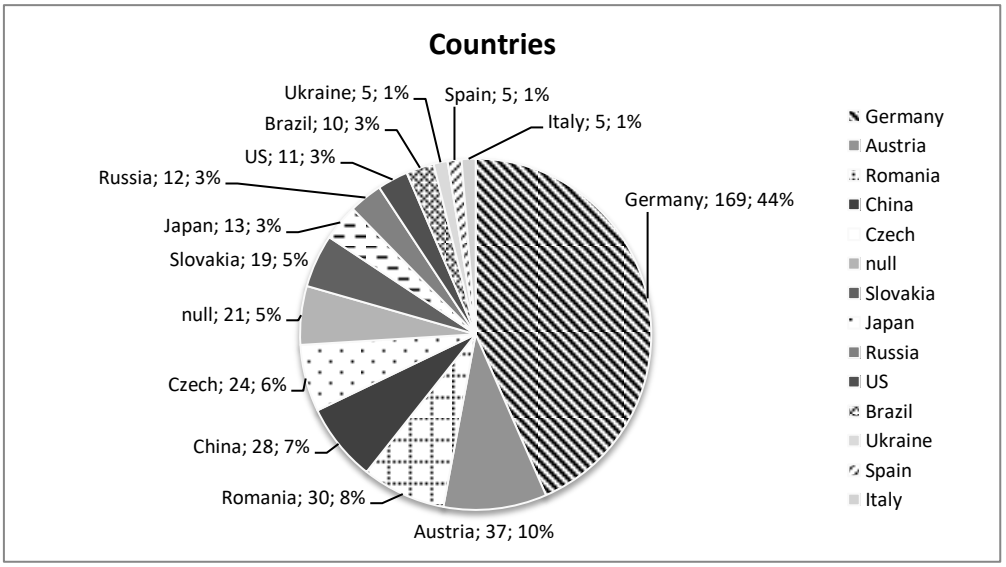

Figure 11. Countries and regions
(Note: "null" means information missing in this field.)

It is seen that the contributing countries and regions are mainly from Europe, US, China and Japan. Germany takes the champion position (44%), followed by Austria (10%). Romania (8%). China (7%), Czech Republic (6%) and Slovakia (5%) fall between the scope between 5% and 10%. The following countries and regions also have contributions to the journal (<= 4%, not displayed in Figure 11): Japan, Russia, US, Brazil, Ukraine, Spain, Italy, Canada, UK, Germany, India, Egypt, Belgium, South Korea, Argentina, Iran, Poland, France and Sweden. Of course, it should be noted that there is still 11% missing data.

For the general articles only, the chronological changes in the counts of countries and regions can be seen from Figure 12.

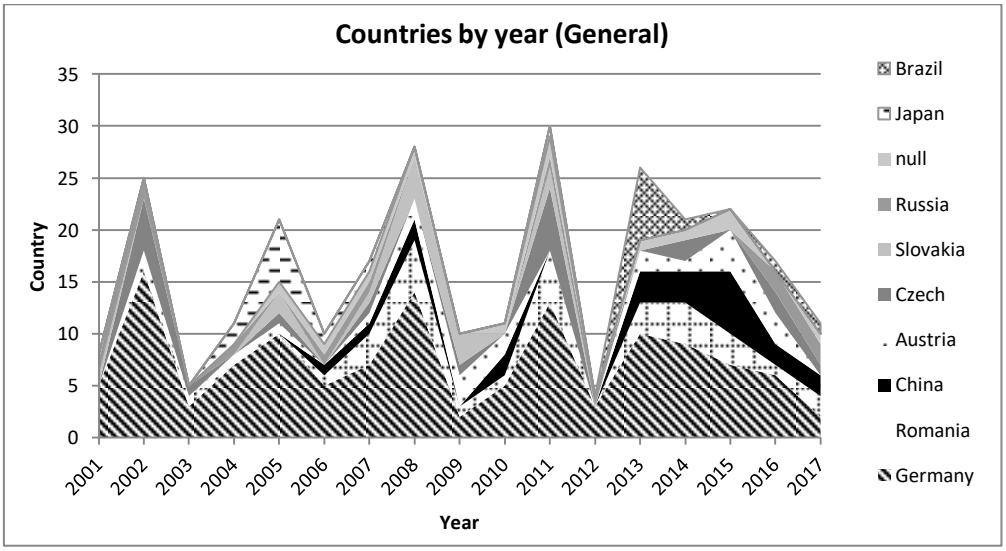

Figure 12. Countries and regions by year (general articles)

As is shown, Germany has maintained its overwhelming influence in QL research throughout the world. In recent years, the emergence of non QL-tradition countries and regions including China, Brazil and Russia is clearly seen. Japan reached a climax in 2005 and then underwent a decline afterwards.

3.2.3. Affiliations

As part of metadata of a citation, counts of affiliations are given in Figure 13.

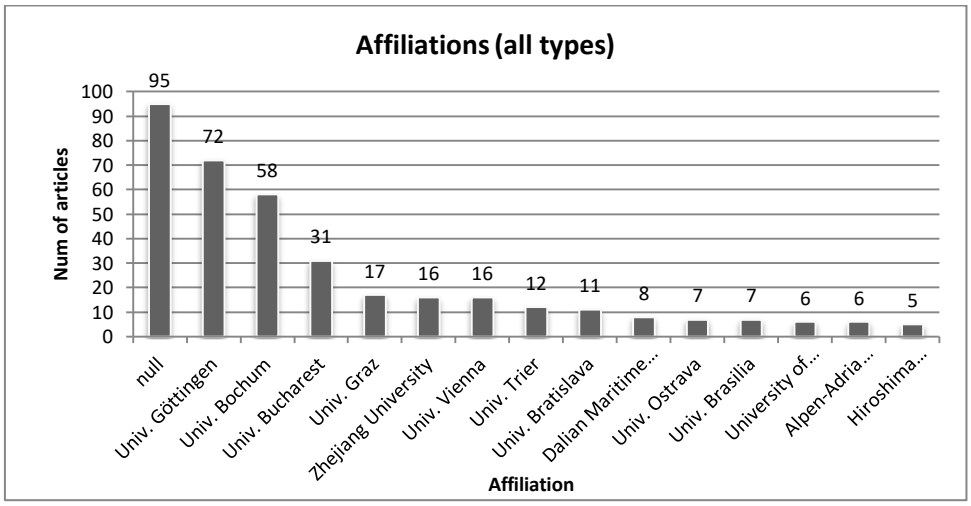

Figure 13. Affiliations (freq. >= 5)

In Figure 13, the information of affiliations is missing in quite a number of articles in *Glottometrics*. According to our limited statistics about the rest, Univ. Göttingen contributes most to the journal, followed by Univ. Bochum and Univ. Bucharest. The results are directly related to the authors. For instance, Univ. Göttingen, ranking first on the list, is the institution to which productive authors like K.-H. Best are affiliated.

Changes of affiliations for general articles over the years are shown in Figure 14.

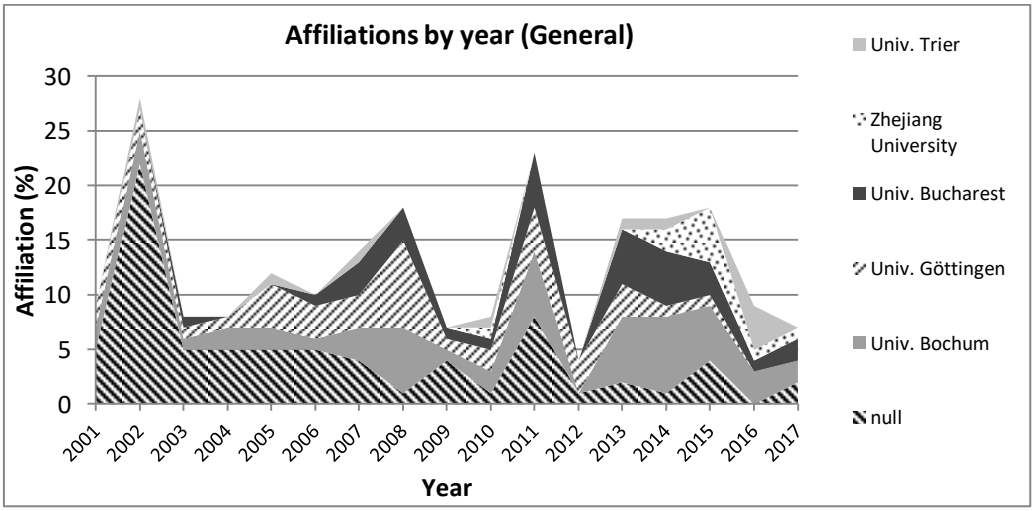

Figure 14. Affiliations by year (general articles)

It is shown in Figure 14 that Univ. Göttingen saw an obvious rise and fall at the turning point of the year 2008. Contributions of Univ. Bochum and Univ. Bucharest fluctuate greatly over the years, and those of Zhejiang University and Univ. Trier have dramatic increase in recent years.

3.2.4. Funding

Funding for research projects maintains and develops vigorous research activities by providing material foundation. In an article, funding acknowledgement provides a better context and confirmation of significance of research. Of all the 330 articles, there are 33 specifying their funding acknowledgements. The articles with funding acknowledgements are displayed in Table 2.

Table 2
Articles with funding acknowledgements

No.	Title of article	Fund	Country
1	An Optimization Model of Global Language Complexity	the Research Fund of CEMA University	Argentina
2	Entropy of a Zipfian Distributed Lexicon	the Brazilian agencies CNPq and FAPEMIG	Brazil
3	The Impact of Code-switching on the Menzerath-Altmann Law	Zhejiang Gongshang University	China
4	A Quantitative Investigation of the Genre Development of Modern Chinese Novels	the National Social Science Foundation of China	China
5	Golden section in Chinese Contemporary Poetry	the National Social Science Foundation of China	China
6	Comparison of vocabulary richness in two translated Hongloumeng	the National Social Science Foundation of China	China
7	Probability distribution of interlingual lexical divergences in Chinese and English: (dao) and said in Hongloumeng	the National Social Science Foundation of China	China
8	A diachronic study of Chinese word length distribution	the National Social Science Foundation of China	China
9	How do Local Syntactic Structures Influence Global Properties in Language Networks?	the National Social Science Foundation of China, the Communication University of China	China
10	Adnominal Constructions in Modern Chinese and their Distribution Properties	the National Social Science Foundation of China	China
11	Quantitative Studies in Chinese Language	the National Social Science Foundation of China	China
12	Mastering the measurement of text's frequency structure: an investigation on Lambda's reliability	the Fundamental Research Funds for the Central Universities and the MOE Project of the Center for GDUFS	China
13	Quantitative Aspects of RST Rhetorical Relations across Individual Levels	Department of Education of Zhejiang Province, China and the National Social Science Foundation of China	China
14	Vocabulary richness in Slovak poetry	the Czech Science Foundation	Czech R.

15	Fractal analysis of Poe's Raven	the Council of Czech Government	Czech R.
16	Word frequency and position in sentence	Project 1 ET 1011 20413 (Academy of Sciences of the Czech Republic)	Czech R.
17	Four reasons for a revision of the transitivity hypothesis	GAČR (Czech Science Foundation)	Czech R.
18	Word form and lemma syntactic dependency networks in Czech: a comparative study	GAČR (Czech Science Foundation)	Czech R.
19	Hidden communication aspects in the exponent of Zipf's law	the Future and Emerging Technologies program	Europe
	A psycholinguistic application of synergetic linguistics	the European Union in the framework of a Marie Curie Intra-European Fellowship	Germany
20	Predicting Attachment of the Light Verb –suru to Japanese Two-kanji Compound Words Using Four Aspects	the Japan Society for the Promotion of Science	Japan
21	A Database of Two-Kanji Compound Words Featuring Morphological Family, Morphological Structure, and Semantic Category Data	the 21st Century COE Program	Japan
22	Constructing a Large-Scale Database of Japanese Word Associations	the 21st Century COE Program	Japan
23	New Kango of the early Meiji era: Their survival and disappearance from Meiji to the present	"Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists" and "Grant in Aid for JSPS Fellows"	Japan
24	Zum Problem der Entstehung des syllabotonischen Versmaßsystems im europäischen Vers	dem Deutschen Akademischen Austauschdienst (DAAD) und in den Jahren 2003–2004 von dem Russischen Bildungsministerium	Russia
25	Some statistical investigations concerning word classes	VEGA	Slovakia
26	Discrete distributions connected by partial summations	VEGA	Slovakia
27	Distribution of complexities in the Vai script	VEGA	Slovakia
28	Some problems of musical texts	VEGA	Slovakia
29	Confidence intervals and tests for the h-point and related text characteristics	VEGA	Slovakia
30	Runes: complexity and distinctivity	VEGA	Slovakia
31	Some properties of the Ukrainian writing system	VEGA	Slovakia
32	Towards a model for rank-frequency distributions of melodic intervals	VEGA	Slovakia

33	The Meaning-Frequency Law in Zipfian Optimization Models of Communication	APCOM from MINECO (Ministerio de Economía y Competitividad), the grant 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya)	Spain

Table 2 shows the funding sources are mainly from government, foundations and professional organizations. In countries and regions like China and Czech Republic, the investigation in education and research is commonly seen as part of governmental strategy. For instance, studies of Liu's team in recent years have been largely supported by the National Social Science Foundation of China. All confirms the significance of QL beyond a researcher's personal interest and concern.

At the same time, the proportion of articles funded in *Glottometrics* is much lower than those of the top linguistics journals shown in the Appendix (e.g. *Applied Linguistics*: 96.38%; *Journal of Memory and Language*: 66.67%; *Bilingualism-Language and Cognition*: 78.50%). Admittedly, research funding concerns factors of social, economic and political aspects. Viewed from the sub-discipline itself, the low funding rate may result from relatively little attention in the linguistic circle. QL research in theory and application still needs more support in different forms on the way to embrace a more promising scenario.

3.3. Research Content

3.3.1. Keywords

The information of this field of 24 articles (10%) is missing. Keywords of the rest 218 articles (90%) are extracted from the self-built corpus. Results covering the time span of 2001~2017 are given in Table 3.

Table 3
A wordlist of keywords in general articles (2001~2017, freq. >= 5)

Rank	Frequency	Word
1	25	German
2	21	word length
3	21	Zipf's law
4	16	English
5	14	diversification
6	13	Piotrowski law
7	12	entropy
8	11	Russian
9	11	sentence length
10	10	word frequency
11	9	Chinese
12	8	borrowings
13	8	rank-frequency distribution
14	7	arc length

15	7	corpus
16	7	h-point
17	7	ranking
18	6	lambda
19	6	Slovak
20	6	stratification
21	6	text
22	5	rank frequency
23	5	repeat rate
24	5	vocabulary richness
25	5	Zipf

Aided by AntConc, we get four wordlists of keywords in different periods from the four sub-corpora in Table 4.

Table 4
Wordlists of keywords in general articles in four periods (freq. >= 3)

Period I	**Period II**	**Period III**	**Period IV**
Zipf's law	German	entropy	Russian
entropy	word length	stratification	compounds
German	English	word length	distance
ranking	diversification	Chinese	English
economy	Russian	diversification	German
information	h-point	German	Pushkin
language change	sentence length	lambda	
Piotrowski law	Zipf's law	rank-frequency distribution	
word frequency	arc length	English	
word length	borrowings	Piotrowski law	
	Chinese	arc length	
	Piotrowski law	binomial distribution	
	word classes	borrowings	
	word frequency	corpus	
		distribution	
		polysemy	
		rank frequency	
		repeat rate	
		sentence length	
		translation	
		verse length	
		vocabulary richness	
		word frequency	

Table 3 and Table 4 show the focuses and shifts of QL research over the years. The keywords including *Zipf's law, Piotrowski law, word length, word frequency, rank, rank-frequency, rank-frequency distribution* are shared by all the periods. It indicates that studies on

laws in languages have been canonical. Another group of key words like *German* are related to the languages being studied or as source of material. The popularity of *German* and *English* never fades, and *Chinese* and *Russian* also catch the eyes of researchers in the past decade. Chronically, the first period focuses on systems and laws. In Period II, studies on words (such as *word length*, *word class*) are emphasized, together with *borrowing*, *arc length*, *sentence length* and *diversification*, which are still popular in Period III. Meanwhile, keywords concerning translation and literature see an increase in the third and fourth periods.

3.3.2. Abstracts

A wordlist of the abstracts in (1 abstract missing) is provided below in Table 5.

Table 5
A wordlist of abstracts in general articles (2001~2017, freq. >= 15)

distribution	kanji	sentence	Piotrowski
word	English	complexity	size
law	linguistic	classes	theory
length	Altmann	lexical	entropy
frequency	semantic	corpus	laws
text	frequencies	structure	speech
texts	functions	Japanese	tests
words	vocabulary	modern	diversification
language	rank	features	information
Zipf	dependency	statistical	lengths
data	properties	theoretical	Russian
model	quantitative	logistic	syntactic
distributions	hypothesis	power	type
German	linguistics	system	units
languages	Chinese	indicators	

A list of N-Grams (N: 2~5) of abstracts are also extracted from this corpus. After manual selection, results are shown in Table 6.

Table 6
An N-Gram list of abstracts in general articles (2001~2017, freq. >= 10)

word length	rank frequency distribution
the distribution	natural languages
rank frequency	Poisson distribution
frequency distribution	power law
Piotrowski law	word classes
sentence length	compound words
logistic law	the logistic law
parts of speech	word frequency
frequency distributions	

Table 7 (lemmatized) illustrates differences and changes in four periods in a more specific way.

Table 7
Wordlists and N-Grams lists of abstracts in general articles in four periods

Period	**Wordlist (freq. > 10)**	**N-Grams (freq. > 5)**
Period I	law, Zipf, word, frequency, distribution, Kanji, text, language, data, number, length, German, model, Japanese, linguistics, semantic, structure, compound, term, lexical, network, property, quantitative	Zipf's law, compound word, Kanji compound word, word length, kanji stroke, natural language, word class, word frequency
Period II	distribution, law, word, frequency, language, Zipf, text, length, kanji, data, German, model, property, Japanese, rank, semantic, linguistics, analysis, statistical, lexical, model, natural, order, sentence, structure, English, modern, power, quantitative, theoretical, class, hypothesis, logistic, network, compound, letter, speech, system, unit, Altmann, empirical, feature, Piotrowski, size, test	distribution, word length, in German, rank frequency, natural language, power law, compound word, Poisson distribution, sentence length, frequency distribution, parts of speech, kanji compound word, the Piotrowski law, kanji stroke, language change, the h point, word class, word frequency
Period III	length, word, distribution, frequency, text, English, language, law, function, vocabulary, Chinese, model, data, Altmann, German, complexity, hypothesis	word length, frequency distribution, content word, length distribution, word length distribution, rank frequency distribution
Period IV	dependency, text, number, distribution, word, Altmann, frequency, length, speech, compound, corpus, function, lambda, language, complexity, information, type, vocabulary, crossing, distance, model, Popescu	code switching, inaugural address, number of crossings

Table 5 ~ Table 7 provide us more information about the developments of QL. As the findings from the study of keywords suggest, word length and frequency studies have gone along with the development of QL. Words like *language*, *text*, *word*, *vocabulary*, *lexical*,

semantic on the list imply the objects and material of investigation in QL as a branch of linguistics. Others like *empirical*, *hypothesis*, *law*, *model*, *data* and *test* indicate that QL research observes the paradigm of scientific research. And *frequency* and *lambda* are related to the indices in QL. In terms of shifts in different time, Japanese Kanji forms an issue for a number of studies especially in Period I and II. The third period still concerns quantitative studies on word level combined with textual research. Recently, researchers start to turn their eyes to syntactic and textual levels.

3.3.3. Objects Studied

Combined with the quantitative analysis of two corpora, we summarize and mark the object being studied in each general article. These objects can be classified into nine themes in reference to the taxonomy of linguistics:

(1) System: laws in language systems, properties of a system like economy or symmetry, and relations of levels or elements within a system;

(2) Phonology and phonetics: phonemes, prosody in literary works, sound symbolism;

(3) Morphology, lexicology and lexicography: word class, word frequency, word length, type-token relation, entropy, polysemy and synonym; affix, borrowing and compounding;

(4) Sentence and syntax: sentence length, syntactic complexity, syntactic network;

(5) Semantics and pragmatics: lexical semantics, information content in communication;

(6) Text: text genre and style, translation, text processing;

(7) Dialectology, typology, diachronics, psycholinguistics, language learning, computational linguistics;

(8) Script: script complexity, grapheme-phoneme relationship, letters;

(9) Others: overviews of QL, introductions to the scholars, etc..

We calculate the number of articles falling into the themes above, whose proportions are given in Figure 15.

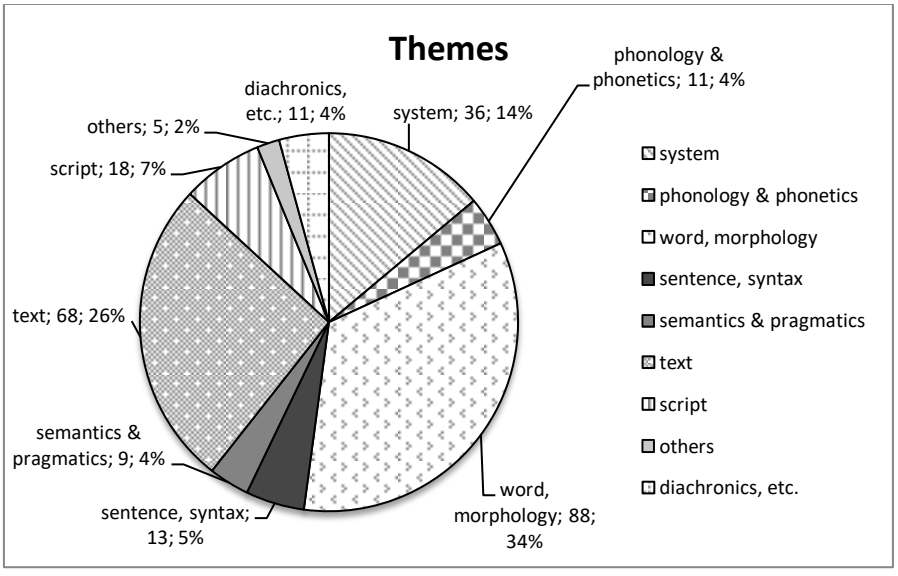

Figure 15. Proportions of research themes in general articles

Figure 15 shows 34% of general articles focus on the exploration of words and morphology. Textual research also constitutes approximately one third (26%) of the total followed by studies on system (14%). Other themes such as scripts, sentence and syntax take

up only a minor part.

Figure 16 further illustrates the percentage changes of the research themes over the years.

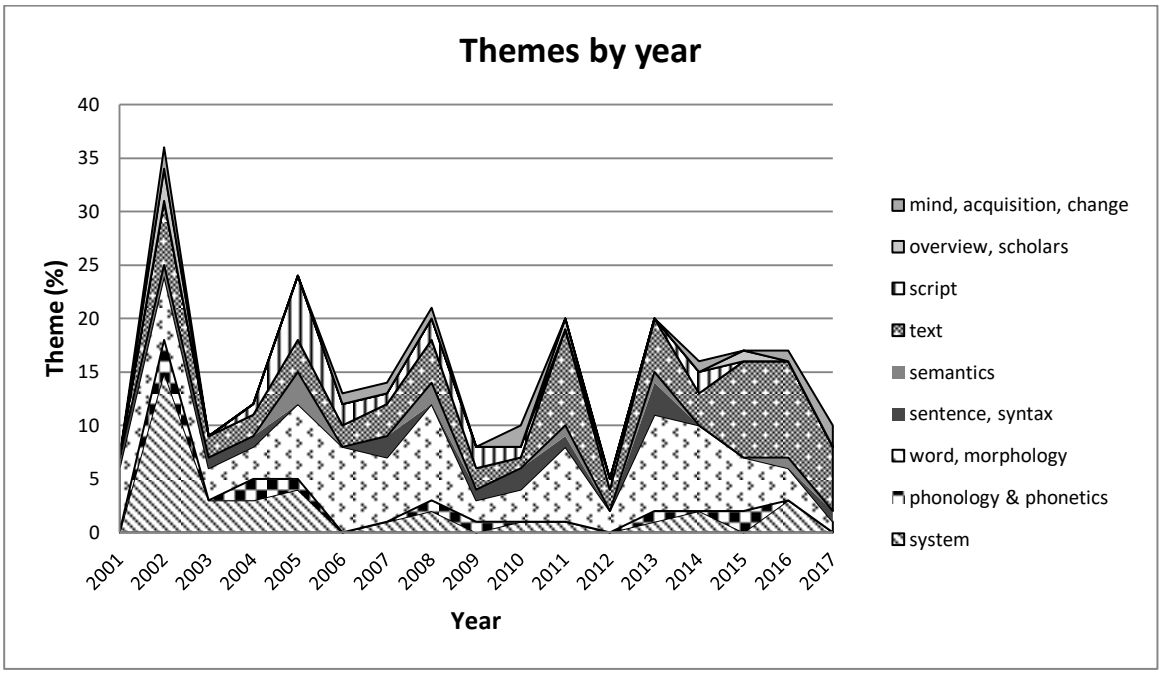

Figure 16. Proportions of research themes by year in general articles

As is illustrated, the theme "word and morphology" has constantly attracted researchers' attention throughout the years. Another canonical theme is "text", which gradually outnumbers "word and morphology" recently. Other themes have been paid attention to by a small part of articles.

Among enormous academic literature, a hot topic emerges when it has been focused on by a number of studies during a certain time span. Hot topics can be identified with citation analysis tools by detecting burst terms. In our study, we do manual analysis instead, setting the minimum frequency of appearance at 3 in two consecutive years for a hot topic.

Results show that there are 8 hot topics: law, word frequency, word class, word length, borrowing, indicator, text genre and style. Half of them deal with words and morphology. Specifically, some representative studies of each hot topic are given:

(1) Law: the application and modification of Zipf's law (Adamic & Huberman, 2002; Köhler, 2002; Popescu, 2003; Wheeler, 2002; Kromer, 2002; Li, 2002; Popescu, 2003; Wheeler, 2002), power law (Hřebíček, 2003; Köhler, 2002), etc.;

(2) Word frequency: aspects (Popescu & Altmann, 2006), relations to word order and position (Fenk-Oczlon & Fenk, 2002; Uhlířová, 2007), etc.;

(3) Word class: mathematical and statistical investigation (Vulanović & Canton, 2008; Wimmer & Altmann, 2001), diversification (Best, 2013; Tuzzi, Popescu, & Altmann, 2011), dynamics (Popescu, Best, & Altmann, 2007), investigations into parts of speech (including adnominal, adverbial, verb, noun, adjective), etc.;

(4) Word length: lengths of linguistic units (Best, 2011a); its distribution (Best, 2011b; Chen & Liu, 2014; Wang, 2013; Wilson, 2003), relations to sentence length (Fan, Grzybek, & Altmann, 2010), etc.;

(5) Borrowing: borrowing and Piotrowski law (Best, 2005, 2015) (too many to list here);

(6) Indicator: arc length (Popescu, Mačutek, & Altmann, 2008; Popescu, Zörnig, & Altmann, 2013; Zörnig, 2017), Lambda (Poiret & Liu, 2017; Popescu & Altmann, 2015);

(7) Text genre: quantitative analysis of a certain genre such as speech (Kubát & Čech,

2016), poem (Pan, Qiu, & Liu, 2015), musical texts (Mačutek, Švehlíková, & Cenkerová, 2011; Martináková, Popescu, Mačutek, & Altmann, 2008), etc.;

(8) Text style: stylistic analysis of literary work (Andreev, 2016; Bortolato, 2016; Levickij & Hikow, 2004).

Changes of the hot topics above in frequency are shown in Figure 17.

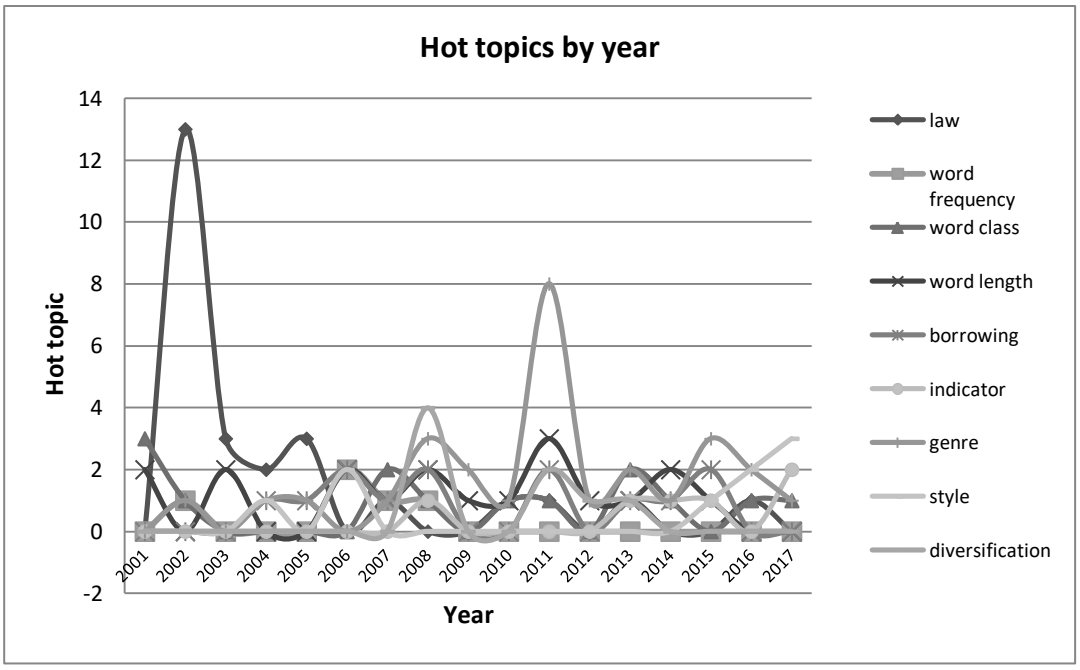

Figure 17. Hot topics by year in general articles

As Figure 17 shows, a conspicuous rise of "law" reached to a peak in 2002, becoming the hottest topic of that year whose popularity lasts in the following years. Genre studies also witnessed an obvious rise in 2011.

3.4. Citations

From the bibliometric view, references in a citation web are connected by two kinds of citation relations: citing and cited. Next, the citing articles and the cited references of the 330 source articles in *Glottometrics* are analyzed respectively.

3.4.1. Source Articles

In the databases such as Web of Science and Google Scholar, citation activity is easily tracked.

Unlike SCIE and SSCI, Journal Impact Factor[5] metrics for journals covered in ESCI are not calculated. Therefore, times cited is used here as one of the bibliometric indices to measure the academic influence of an article in the scientific community.

According to Web of Science, there are altogether 168 of 330 source articles (22.6%) in *Glottometrics* cited in the dataset. In terms of documents cited, it would have been at 54th

[5] In Web of Science, *Journal Impact Factor* is defined as "all citations to the journal in the current JCR year to items published in the previous two years, divided by the total number of scholarly items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years." (Thomson Reuters, 2017)

percentile in the linguistics journals in InCites[6].

The rank-frequency relation is given in Figure 16.

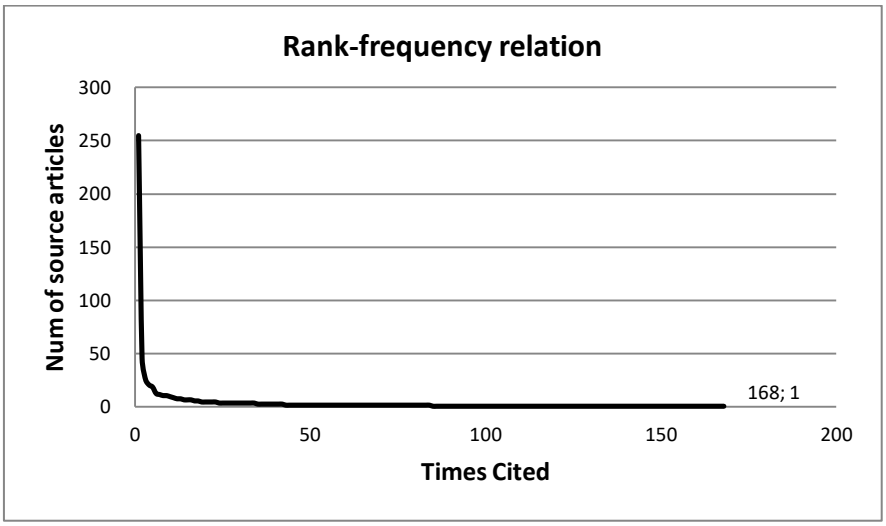

Figure 16. The rank-frequency curve for source articles according to Web of Science

As Figure 16 shows, all the 330 documents of the journal have 743 total cites in Web of Science, with an average of 2.25 cites per document and an *h*-index[7] of 10. In terms of times cited per document only, the journal may have ranked at the 70th percentile in linguistics journals in inCites (similar to those of *Anaphors in Text*, *Language-Meaning-Social Construction Interdisciplinary Studies*, *Primate Communication and Human Language: Vocalisation, Gestures, Imitation*, and *Determiners: Universals and Variation*).

A list of most cited source articles in the journal (freq. >= 5) is shown in Table 8.

Table 8
The most cited source articles in *Glottometrics* (according to Web of Science)

Rank	Author	Title	Year	Vol.	Times Cited
1	Adamic, L.A.; Huberman, B. A.	Zipf's law and the internet	2002	3	255
2	Li, W.	Zipf's law everywhere	2002	5	46
3	Popescu, I. I.	On a Zipf's Law extension to impact factors	2003	6	26
4	Kornai, A.	How many words are there?	2002	4	21
5	Liu, H.	Probability distribution of dependency distance	2007	15	19
6	Rousseau, R.	George Kingsley Zipf. Life, Ideas, his Law and Informetrics	2002	3	13

[6] From: https://incites.thomsonreuters.com/#/explore/0/funder//. The InCites dataset used here was updated on 2017-07-01, which includes Web of Science content indexed through 2017-03-31.

[7] In bibliometrics, *h*-index is an author-level metric that quantifies both the productivity and the citation impact of a scientist or scholar (from: http://www.pnas.org/content/102/46/16569). Journal *h*-index refers to journal's number of articles (*h*) that have received at least *h* citations over the whole period.

7	Popescu, I. - I.; Altmann, G.	Some aspects of word frequencies	2006	13	12
8	Balasubrahmanyan, V.; Naranan, S.	Algorithmic Information, Complexity and Zipf's Law	2002	4	11
8	Montemurro, M. A.; Zanette, D. H.	New perspectives on Zipfs law in linguistics: from single texts to large corpora	2002	4	11
10	Pauli, F.; Tuzzi, A.	The end of year addresses of the presidents of the Italian republic (1948-2006): Discourse similarities and differences	2009	18	10
11	Ferrer-i-Cancho, R.	Hubiness, length and crossings and their relationships in dependency trees	2013	25	9
12	Ferrer-i-Cancho, R.; Servedio, V. D.	Can simple models explain Zipf's law in all cases?	2005	11	8
12	Smith, R	Distinct word length frequencies: distributions and symbol entropies	2012	23	8
14	Best, K.-H.	Zur Haufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten (On the frequency of letters, spaces and other characters in German texts)	2005	11	7
14	Grzybek, P.	On the systematic and system-based study of grapheme frequencies: A re-analysis of German letter frequencies	2007	15	7
14	Popescu, I.-I.; Best, K.-H.; Altmann, G.	On the dynamics of word classes in text	2007	14	7
17	Altmann, G.	Towards a theory of language	1978	1	6
17	Best, K.-H.; Altmann, G.	Some properties of graphemic systems	2005	9	6
19	Altmann, G.	Script complexity	2004	8	5
19	Best, Karl-Heinz	Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes	2003	6	5
19	Grzybek, P.; Kelih, E.; Stadlober, E.	The relation between word length and sentence length. An intra-systemic perspective in the core data structure	2008	16	5
19	Kelih, E.	The type-token relationship in Slavic parallel texts	2010	20	5
19	Köhler, R.	Quantitative Untersuchungen zur Valenz deutscher Verben	2005	9	5

According to Google Scholar (up to July 8th, 2017), the *h*-index of *Glottometrics* is 14. A list of top 15 most cited references is shown in Table 9 (freq. >= 5).

Table 9
The most cited source references in *Glottometrics* (according to Google Scholar)

Rank	Times Cited	Article
1	715	Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics, 3*, 143-150.
2	94	Li, W. (2002). Zipf's Law Everywhere. *Glottometrics, 5*, 14-21.
3	61	Kornai, A. (2002). How many words are there? *Glottometrics, 4*, 61-86.
4	50	Popescu, I.-I. (2003). On a Zipf's Law Extension to Impact Factors. *Glottometrics, 6*, 61-64.
5	39	Popescu, I.-I., & Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics, 13*, 23-46.
6	34	Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics, 15*, 13-23.
7	27	Joyce, T. (2005). Constructing a Large-Scale Database of Japanese Word Associations. *Glottometrics, 10*, 82-98.
8	23	Montemurro, M. A., & Zanette, D. H. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics, 4*, 87-99.
9	22	Pauli, F., & Tuzzi, A. (2009). The End of Year Addresses of the Presidents of the Italian Republic (1948-2006): discoursal similarities and differences. *Glottometrics, 18*, 40-51.
10	21	Rousseau, R. (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics, 3*, 11-18.
11	16	Wheeler, E. S. (2002). Zipf's Law and why it works everywhere. *Glottometrics, 4*, 45-48.
11	16	Čech, R., & Mačutek, J. (2011). Word form and lemma syntactic dependency networks in Czech: a comparative study. *Glottometrics, 19*, 85-98.
13	15	Altmann, G. (2004). Script complexity. *Glottometrics, 8*, 68-74.
13	15	Best, K. H. (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics, 6*, 9-34.
14	14	Grzybek, P., Kelih, E., & Stadlober, E. (2008). The relation between word length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics, 16*, 111-121.
15	13	Körner, H. (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics, 7*, 25-49.
15	13	Altmann, G. (2002). Zipfian linguistics. *Glottometrics, 3*, 19-26.
15	13	Grzybek, P. (2007). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. *Glottometrics, 15*, 82-91.
15	13	Körner, H. (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics, 7*, 25-49.
19	12	Balasubrahmanyan, V. K., & Naranan, S. (2002). Algorithmic information, complexity and Zipf´s law. *Glottometrics, 4*, 1-26.
19	12	Martináková, Z., Popescu, I.-I., Mačutek, J., & Altmann, G. (2008).

		Some problems of musical texts. *Glottometrics, 16,* 63-79.
19	12	Liu, H., Zhao, Y., & Huang, W. (2010). How do Local Syntactic Structures Influence Global Properties in Language Networks? *Glottometrics, 20,* 38-58.
22	11	Gumenyuk, A., Kostyshin, A., & Simonova, S. (2002). An approach to the research of the structure of linguistic and musical texts. *Glottometrics, 3,* 61-89.
22	11	Hřebíček, L. (2002). Zipf's Law and Text. *Glottometrics, 3,* 27-38.
22	11	Kelih, E. (2009). Graphemhäufigkeiten in slawischen Sprachen: stetige Modelle. *Glottometrics, 18,* 52-68.
22	11	Popescu, I.-I., & Altmann, G. (2007). Writer´s view of text generation. *Glottometrics, 15,* 71-81.
22	11	Köhler, R. (2005). Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics, 9,* 13-20.
27	10	Mačutek, J., Popescu, I.-I., & Altmann, G. (2007). Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics, 15,* 45-52.
27	10	Ferrer-i-Cancho, R., & Servedio, V. (2005). Can simple models explain Zipf's law for all exponents? *Glottometrics, 11,* 1-8.
27	10	Popescu, I.-I., Best, K.-H., & Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics, 14,* 58-71.
27	10	Pawlowski, A. (2005). VI. Wincenty Lutoslawski-a forgotten father of stylometry. *Glottometrics, 8,* 83-89.
27	10	Best, K. H. (2005). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics, 11,* 9-31.
32	9	Grzybek, P., & Altmann, G. (2002). Oscillation in the frequency-length relationship. *Glottometrics, 5,* 97-107.
32	9	Vulanović, R. (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics 17,* 51, 65.
34	8	Best, K.-H. (2002). The distribution of rhythmic units in German short prose. *Glottometrics, 3,* 136-142.
34	8	Fan, F. (2006). Models for dynamic inter-textual type-token relationship. *Glottometrics, 12,* 1-10.
34	8	Popescu, I.-I., & Altmann, G. (2008). Zipf´s mean and language typology. *Glottometrics, 16,* 31-37.
34	8	Roelcke, T. (2002). Efficiency of communication: A new concept of language economy. *Glottometrics, 4,* 27-38.
34	8	Kazartsev, E. (2006). Zum Problem der Entstehung des syllabotonischen Versmaßsystems im europäischen Vers. *Glottometrics, 13,* 1-22.
34	8	Best, K. H. (2001). Zur Gesetzmäßigkeit der Wortverteilung in deutschen Texten. *Glottometrics, 1,* 1-26.
40	7	Ishida, M., & Ishida, K. (2007). On distributions of sentence lengths in Japanese writing. *Glottometrics, 15,* 28-44.
40	7	Kromer, V. (2001). Word length model based on the one-displaced Poisson-uniform distribution. *Glottometrics, 1,* 87-96.
40	7	Grzybek, P., & Kelih, E. (2004). Anton Semënovič Budilovič. *Glottometrics, 7,* 94-96.
40	7	Naumann, S., Popescu, I.-I., & Altmann, G. (2012). Aspects of

		nominal style. *Glottometrics, 23*, 23-55.
40	7	Fenk-Oczlon, G., & Fenk, A. (2002). Zipf's tool analogy and word order. *Glottometrics, 5*, 22-28.
40	7	Kelih, E. (2010). The type-token relationship in Slavic parallel texts. *Glottometrics, 20*, 1-11.
40	7	Köhler, R. (2002). Power law models in linguistics: Hungarian. *Glottometrics, 5*, 51-61.
40	7	Lehfeldt, W., & Altmann, G. (2002). Der altrussische Jerwandel. *Glottometrics, 2*, 34-44.
48	6	Hřebíček, L. (2002). Zipf's Law and Text. *Glottometrics, 3*, 27-38.
48	6	Jayaram, B. D., & Vidya, M. N. (2006). Word length distribution in Indian languages. *Glottometrics, 12*, 16-38.
48	6	Meyer, P. (2002). Laws and theories in quantitative linguistics. *Glottometrics, 5*, 62-80.
48	6	Antić, G., & Altmann, G. (2005). On letter distinctivity. *Glottometrics, 9*, 46-53.
48	6	Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics, 16*, 1-16.
48	6	Best, K. H. (2005). Turzismen im Deutschen. *Glottometrics, 11*, 56-63.
54	5	Best, K. H., & Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics, 9*, 29-39.
54	5	Tuzzi, A., Popescu, I.-I., & Altmann, G. (2011). Parts-of-speech diversification in Italian texts. *Glottometrics, 19*, 42-48.
54	5	Hisashi, M., & Joyce, T. (2005). Database of Two-Kanji Compound Words Featuring Morphological Family, Morphological Structure, and Semantic Category Data. *Glottometrics, 10*, 30-44.
54	5	Hilberg, W. (2002). The Unexpected Fundamental Influence of Mathematics upon Language. *Glottometrics, 5*, 29-50.
54	5	Peust, C. (2006). Script complexity revisited. *Glottometrics, 12*, 11-15.
54	5	Prün, C. (2002). Biographical notes on GK Zipf. *Glottometrics, 3*, 1-10.
54	5	Popescu, I. I., Čech, R., & Altmann, G. (2011). On stratification in poetry. *Glottometrics, 21*, 54-59.
54	5	Tamaoka, K., & Altmann, G. (2004). Symmetry of Japanese Kanji lexical productivity on the left-and right-hand side. *Glottometrics, 7*, 65-84.
54	5	Popescu, I. I., & Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics, 17*, 94-108.
54	5	Best, K. H. (2002). Der Zuwachs der Wörter auf -ical im Deutschen. *Glottometrics, 2*, 11-16.

Table 8 and Table 9 show that about half of the top 15 most cited articles are from a collection on the theme "Zipf's law" published in the earlier years. Top 1 on the lists is *Zipf's law and the Internet* (Adamic & Huberman, 2002). So far it is cited as high as 255 times by Web of Science and 715 times by Google Scholar. Other source articles have much fewer times cited, covering the canonical topics in QL including word frequency, word and sentence length, probability distribution, dependency syntax, syntactic network, script complexity and

text characteristics.

Despite of the high times cited of a few studies, the majority of the source articles have little contribution to the impact, especially in the recent decade. Whether the academic impact of QL research only displays after a longer period needs further exploration.

3.4.2. Citing Articles

Glottometrics is cited by a variety of references or citing articles, whose total number increases by year (data in 2017 not complete yet).

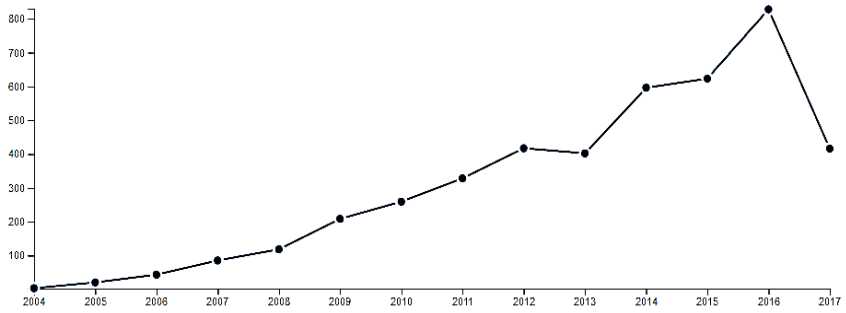

Figure 19. Citing frequencies by year

There is 1 among the citing articles marked as "highly cited article" in Web of Science, namely:

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in cognitive sciences*, *17*(7), 348-360.

It cites the following source article in *Glottometrics*:

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

A similar bibliometric analysis of these citing articles is conducted in Web of Science. Results are given in shortlist of Table 10~Table 17.

Table 10
References citing *Glottometrics*: type of article

Type of Article	Records	% of 638
Article	479	75.08%
Meeting	156	24.45%
Book	55	8.62%
Other	44	6.90%
Review	21	3.29%
Editorial	5	0.78%
Letter	2	0.31%

Table 11
References citing *Glottometrics*: categories

Category	Records	% of 638
Science Technology	477	74.77%
Technology	369	57.84%
Social Sciences	309	48.43%
Physical Sciences	201	31.51%
Life Sciences Biomedicine	103	16.14%
Arts Humanities	37	5.80%

Table 12
References citing *Glottometrics*: research areas

Research Area	**Records**	**% of 638**
Computer science	287	44.98%
Linguistics	172	26.96%
Mathematics	153	23.98%
Telecommunications	101	15.83%
Engineering	95	14.89%
Physics	77	12.07%
Information science library science	65	10.19%
Communication	59	9.25%
Science technology other topics	54	8.46%
Mathematical computational biology	48	7.52%

Table 13
References citing *Glottometrics*: journals

Journal	**Records**	**% of 638**
Journal of Quantitative Linguistics	44	6.90%
Glottometrics	27	4.23%
Lecture Notes in Computer Science	18	2.82%
Physica A Statistical Mechanics and Its Applications	17	2.67%
Handbücher zur Sprach- und Kommunikationswissenschaft	16	2.51%
Plos ONE	16	2.51%
Quantitative Linguistics Quantitative Linguistik	13	2.04%
Quantitative Linguistics	11	1.72%
Analyses of Script Properties of Characters and Writing Systems	9	1.41%
Scientometrics	9	1.41%
Physica A	8	1.25%
Physical Review E	8	1.25%
Physical Review E Statistical Nonlinear and Soft Matter Physics	7	1.10%
Journal of Informetrics	6	0.94%
Complexity	5	0.78%
European Physical Journal B	5	0.78%
IEEE Transactions on Parallel and Distributed Systems	5	0.78%

Table 14
References citing *Glottometrics*: conferences

No.	**Conference**	**Records**	**% of 638**
1	IEEE International Conference on Communications (ICC)	3	0.47%
2	15TH IEEE INTERNATIONAL SYMPOSIUM ON A	2	0.31%

	WORLD OF WIRELESS MOBILE AND MULTIMEDIA NETWORKS WOWMOM		
3	2016 IEEE TRUSTCOM BIGDATASE ISPA	2	0.31%
4	2ND INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS AND TECHNOLOGIES	2	0.31%
5	34TH IEEE CONFERENCE ON COMPUTER COMMUNICATIONS INFOCOM	2	0.31%
6	8TH INTERNATIONAL CONFERENCE ON HYBRID ARTIFICIAL INTELLIGENT SYSTEMS HAIS	2	0.31%
7	8TH POLISH SYMPOSIUM OF PHYSICS IN ECONOMY AND SOCIAL SCIENCES FENS	2	0.31%
8	IEEE GLOBAL COMMUNICATIONS CONFERENCE GLOBECOM	2	0.31%
9	IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE GLOBECOM 05	2	0.31%
10	IEEE GLOBECOM WORKSHOPS GC WKSHPS	2	0.31%

Table 15
References citing *Glottometrics*: authors

Author	**Records**	**% of 638**
Altmann G.	35	5.486
Liu H.	30	4.702
Kohler R.	20	3.135
Piotrowski R.	13	2.038
Ferrer-i-Cancho R.	12	1.881
Tassiulas L.	11	1.724
Popescu II.	9	1.411
Sourlas V.	8	1.254
Mačutek J.	8	1.254
Ausloos M.	8	1.254

Table 16
References citing *Glottometrics*: countries and regions

Country	**Records**	**% of 638**
China	169	26.49%
USA	96	18.81%
Germany	50	7.84%
Spain	39	6.11%
England	33	5.17%
Italy	32	5.02%
Belgium	26	4.08%
Japan	24	3.76%
Greece	20	3.14%
Canada	17	2.67%
Israel	16	2.51%
UK	16	2.51%

Table 17
References citing *Glottometrics*: affiliations

Affiliation	**Records**	**% of 638**
Zhejiang University	37	5.80%
Rutgers State University	13	2.04%
Polytechnic University of Catalonia	11	1.72%
Universidad Nacional Autonoma de Mexico	9	1.41%

University of Thessaly	9	1.41%
University System of Georgia	9	1.41%
Princeton University	8	1.25%
Sapienza University Rome	8	1.25%
University of London	8	1.25%
Beijing University of Posts Telecommunications	7	1.10%
Northwell Health	7	1.10%
Princeton University	7	1.10%

To our surprise, results in Table 11 and Table 12 clearly display that Glottometrics is more cited by references in "Science and Technology" than in "Social Science" (the categories it belongs to). In other words, its academic influence goes far beyond linguistics itself, more in natural sciences than in social science and art and humanities.

Table 12 shows the achievements and methods are often referred and applied in a wide ranges of research areas: Computer science, Linguistics, Mathematics, Telecommunications, Engineering, Physics, Information science library science, Communication, etc. Interdisciplinary studies attract much attention in the scientific community. As is mentioned in 3.4.1, for example, Zipf's law, a discovery originated in linguistics, has wide application "everywhere" in disciplines ranging from bibliometrics to physics (Li, 2002; Popescu, 2003); Syntactic network also provides another instance of complex network in statistical physics; achievements in text generation, analysis and classification are applied in natural language processing. In addition, the vitality of QL research is also facilitated by the research paradigm of QL, i.e., hypothesizing, data collection, statistical diagnostics, accepting or rejecting the hypothesis, and explanation (Köhler, Altmann, & Piotrowski, 2005). It is a well-established and widely accepted paradigm from the perspective of philosophy of science. As for the geographical distribution, the countries and regions with the most citing articles are from Europe and Asia, none from Australian or African countries. The top 3 countries and regions are China, US and Germany. And the top 3 institutions with most citing articles are Zhejiang University, Rutgers State University and Polytechnic University of Catalonia.

3.4.3. Cited References

Co-cited references form the research basis of studies. Given below are the top 30 cited references which frequently appear in the bibliographies of *Glottometrics*.

Table 18
Top 30 cited references in *Glottometrics*

Rank	**Freq.**	**Cited Reference**
1	32	Zipf G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley.
2	29	Wimmer, G. & Altmann, G. (1999). Thesaurus of Univariate Discrete Probability Distributions. Essen: Stamm.
3	23	Altmann, G. (1988). *Wiederholungen in Texten*. (*Quantitative Linguistics 36*). Bochum: Studienverlag Brockmeyer.
4	22	Zipf, G.K. (1935) The Psycho-Biology of Language. An Introduction to Dynamic Philology. Boston: Houghton-Mifflin.
5	20	Köhler, R. (1986) . Zur Linguistischen Synergetik: Struktur und

		Dynamik der Lexik. Bochum: Brockmeyer.
6	19	Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerung. In: Best, K.-H., Kohlhase, Jörg (Hrsg.): *Exakte Sprachwandelforschung. Theoretische Beiträge, Statistische Analysen und Arbeitsberichte* (S. 59-90). Göttingen: edition herodot.
6	19	Wimmer, G., Altmann, G. (2002). *Unified derivation of some linguistic laws*. Paper at the Graz Conference on Word Length, August 2002.
8	18	Popescu, I.-I., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., & Vidya, M.N. (2009). *Word Frequency Studies*. Berlin, New York: de Gruyter.
9	16	Best, K.-H. (2001). *Wo kommen die deutschen Fremdwörter her?* Göttinger Beiträge zur Sprachwissenschaft 5, 7-20.
9	16	Popescu, I.-I., Mačutek, Altmann, G. (2008a). *Aspects of word frequencies*. Lüdenscheid: RAM.
11	15	Altmann, G. (2005). Der Diversifikationsprozess. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics*, Art. 65: 646-658. Berlin: de Gruyter .
12	14	Hřebíček, L. (1997). *Lectures on text theory*. Prague, Oriental Institute.
13	13	Altmann, G. (1980). Prolegomena to Menzerath.s law. *Glottometrika 2*, 1-10.
13	13	Körner, Helle (2004). Zur Entwicklung des deutschen (Lehn-) Wortschatzes. *Glottometrics 7*, 25-49.
13	13	Rothe, Ursula (1991). Diversification Processes in Grammar. An Introduction. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar*: 3-32. Hagen: Margit Rottmann Medienverlag.
16	12	Altmann, G. (1991). Modeling diversification phenomena in language. In: Rothe, U. (Ed.), *Diversification Processes in Language: Grammar*: 33-46. Hagen: Rottmann.
17	11	Amano, N. & Kondo, K. (2000). Nihongo-no goi tokusei [Lexical properties of Japanese]. Tokyo: Sanseido.
17	11	Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt
17	11	Popescu, I.-I., Altmann, G., & Köhler, R. (2010). Zipf´s law – another view. *Quality and Quantity 44(4)*, 713-731.
20	10	Altmann, G. (1993). Phoneme counts. *Glottometrika 14*, 54-58.
20	10	Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
20	10	Hřebíček, L. (2000). *Variation in sequences*. (Contributions to general text theory). Prague: Oriental Institute.
21	10	Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
22	10	Hřebíček, L. (2000). *Variation in sequences*. (Contributions to general text theory). Prague: Oriental Institute.
23	10	Popescu, I.-I., Čech, R. & Altmann, G. (2011). *The Lambda-*

		structure of Texts. RAM-Verlag.
24	9	Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13*, 105- 120.
25	9	Köhler, Reinhard (2005), Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, Rajmund G. [ed.]: *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*: 760-775. (= HSK27) Berlin, New York: de Gruyter..
26	9	Ord, J. K. (1972). *Families of frequency distributions*. London: Griffin.
27	9	Pfeifer, Wolfgang (2000). *Etymologisches Wörterbuch des Deutschen*. 5. Auflage. München: Deutscher Taschenbuchverlag.
28	9	Popescu, I.-I. (2006). Text ranking by the weight of highly frequent words. In: *Exact methods in the study of language and text*, edited by Peter Grzybek and Reinhard Köhler: 555-566. Berlin/New York: Mouton de Gruyter.
29	9	Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (ed..), *Glottometrika 15*, 112-133. Trier: Wissenschaftlicher Verlag.
30	9	Wimmer, G., Köhler, R., Grotjahn, R. & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1*, 98-106.

Among the listed items, *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Zipf, 1935), *Human Behavior and the Principle of Least Effort* (Zipf, 1949) and *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik* (Köhler, 1986) are the classic references in which the basic conceptions, principles and theories of QL are proposed. Others focus on laws, word frequencies and length and probability distribution and so on. Several references are written in German, manifesting again the tradition of QL research in Germany.

We also calculate the proportions of the journals cited in the bibliographies of *Glottometrics*. Among the 1234 journals cited by *Glottometrics*, those cited more than 5 times are shown in Table 19.

Table 19
Journals cited by *Glottometrics* (freq. >= 5)

No.	**Journal Title**	**Freq.**	**Proportion**
1	Glottometrics	225	18.23%
2	Journal of Quantitative Linguistics	185	14.99%
3	Göttinger Beiträge zur Sprachwissenschaft	37	3.00%
4	Glottotheory	21	1.70%
5	Biometrika	16	1.30%
6	Physica A	15	1.22%
6	Quality and Quantity	15	1.22%
6	Science	15	1.22%
9	Information and Control	14	1.13%
10	Language	13	1.05%

11	Physical Review Letters	12	0.97%
12	Computers and the Humanities	10	0.81%
12	American Journal of Psychology	10	0.81%
14	Bell System Technical Journal	9	0.73%
15	Nature	8	0.65%
15	Linguistic Inquiry	8	0.65%
15	Folia Linguistica Historica	8	0.65%
15	Physical Review E	8	0.65%
15	Journal of Experimental Psychology	8	0.65%
20	Computational Linguistics	7	0.57%
20	Linguistics	7	0.57%
20	Behavior Research Methods, Instruments & Computers	7	0.57%
23	Literary and Linguistic Computing	6	0.49%
23	Information and Control	6	0.49%
25	Lingua	5	0.41%
25	Cognition	5	0.41%
25	Anzeiger für Slavische Philologie	5	0.41%
25	Europhysics Letters	5	0.41%
25	Theoretical linguistics	5	0.41%
25	Animal Behavior	5	0.41%
25	Cognitive Science	5	0.41%
25	Language and Cognitive Processes	5	0.41%
25	Scientometrics	5	0.41%

As Table 19 shows, bibliographies in *Glottometrics* cover various disciplines from natural to social sciences as a result of the broad spectrum of QL investigation. References from systems science, statistics and computation sciences are often quoted, which differs QL from other branches of linguistics in methodology. Table 19 also shows that the proportions of linguistics journals are comparatively lower in the bibliographies.

It is noticed that *Glottometrics* and *Journal of Quantitative Linguistics* take up about 30% of the journals cited. According to 2014 JCR Science Edition, 85% of the ESCI journals have self-citation rates of 15% or less[8]. The self-cited rate of 18.23% is slightly higher, thus reducing the diversity of source publications.

As is shown in Table 19, the top linguistics journals in Table 20 and Table 22 (in the Appendix), especially in the "mainstream" sense, are rarely quoted in *Glottometrics* (except for *Lingua* and *Linguistic Inquiry*).

4. Concluding Remarks

In this paper, we conduct a bibliometric study of *Glottometrics* by analyzing the metadata of 37 volumes during 2001~2017, based on data from the library and self-built corpora. Our analysis covers four main aspects: a. publication profile of the journal including publication frequency, type of work, length of article and language; b. authors, countries and regions,

[8] From: http://wokinfo.com/essays/journal-selection-process/

affiliations contributing to the journal as well as funding; c. research content of the articles including keywords, abstracts and objects studied; d. citations.

Results suggest that QL research is characterized by addressing linguistics problems by scientific approaches. It encompasses nearly all the sub-disciplines of theoretical and applied linguistics, as a confirmation and supplementation of Chen and Liu (2014)'s findings. In this sense, "the objects and the epistemological interest of QL research do not differ principally from those of other linguistic and textological disciplines, nor is there a principal difference in epistemological interest." (Köhler, 2012)

Since its first publication, *Glottometrics* has been serving as an unparalleled platform of QL research. With its academic impact, it undergoes revolutions in alliance with another authoritative journal, *Journal of Quantitative Linguistics*. Certainly, as a comparatively "younger" publication, *Glottometrics* still has its inadequacies. Next, some remarks and suggestions based on the results are presented for further improvements.

First, in terms of publication, the completeness of elements is expected to be improved. In our study, the reliability of bibliometric analysis is affected by a lack of data. Nowadays for a journal in modern sense, informative elements are required by almost all the citation databases including: journal title, year of publication, volume and/or issue number, page number, article title, abstract, keywords, author name(s), full address for every author, institution (name, city, country or region), fund or project, subject, research area, citations. To our delight, citation analysis reports will be generated after several clicks after its acceptation by Scopus in 2017, thus making bibliometric analysis more efficient in the near future. Besides, since timeliness of publication implies ongoing viability in the research area, *Glottometrics* needs to keep a regular publication and a steady flow of articles online or in print are of fundamental importance (53% of 2016 SSCI journals in the WOS category of Linguistics are quarterly). Also, for the sake of global academic communication, access to full texts in English is necessary. The recent rise of English in proportion just indicates the efforts made by the editors to be more international.

Second, *Glottometrics* is on the way to embrace a wider research community. Over the past 17 years, the majority of contributors to *Glottometrics* are from the European universities where QL tradition is deeply rooted. With high productive scholars like Best and Altmann, the impact of Germany has long been unparalleled. This can be seen in the statistical results of language used in the manuscripts, language studied as objects of research as well as the uneven geographic distribution of contributors. However, recent years see a pleasing emergence of China and Brazil. Yet it is also noted that the author's nationality is largely related to the language studied or as source material, *German* and *English* as good examples. In this view, cooperation from more countries and regions is welcome to enrich language data. A journal with international focus always needs a diverse group of authors, editors and editorial advisory board members, especially for those with highly cited articles. Certainly, more funds in a variety of sources (e.g. private industry) are necessary for the development of the discipline.

Third, from the perspective of research content, besides the wide coverage of exploration, researchers turn their eyes from the canonical word studies to textual levels in the recent decade. Syntax and semantics need further investigation towards a higher stage of synergetic linguistics. Areas of applied linguistics in broad sense such as language acquisition and psycholinguistics almost remain untouched while these are supposed to be quite promising in this century. Another perspective may be called the activities of QL research: metrification, quantitative analysis and description, numerical classification, diagnostic comparison and trend detection, modelling, theory construction, explanation, extension, methodological elaboration, and practical application (Köhler et al., 2005). A possible way of doing QL research is to combine the two perspectives: to perform the activities mentioned on the

linguistic and textual objects above. Some of the approaches have already been used in a number of studies, some need to form a more specific procedure, and some are rarely tried for many reasons. In addition, it is highly recommended that the research problems should be proposed in such a way that they can arouse the interest of the "mainstream" linguists.

Finally, as for citation, an interesting phenomenon is that in contrast to a limited influence in linguistics, *Glottometrics* has its academic impact in other disciplines such as information sciences. Thanks to the endeavors made to promote interdisciplinary research, *Glottometrics* has kept its vitality by citation despite of high professionality in mathematics and statistics. Meanwhile, due to methodological consideration, the academic impact of QL remains restricted within a comparatively smaller circle. However, a better acceptation by a wider community both within and beyond linguistics itself is expected. Therefore, it is advised that the top journals in the linguistics community be cited more, and the journal's self-cited rate be controlled below 15%. After all, the essence of QL should be overshadowed by any theoretical gap or methodological divergence.

Acknowledgements

This work is supported by the National Social Science Foundation of China (Grant No. 17AYY021) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics, 3*, 143-150.

Andreev, S. (2016). Verbal vs. Adjectival Styles in Long Poems by A.S. Pushkin. *Glottometrics, 33*, 25-31.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in cognitive sciences, 17*(7), 348-360.

Best, K.-H. (2005). Turzismen im Deutschen. *Glottometrics, 11*, 56-63.

Best, K.-H. (2011a). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics, 21*, 1-13.

Best, K.-H. (2011b). Word length distribution in French. *Glottometrics, 22*, 44-56.

Best, K.-H. (2013). Diversifikation der starken Verben im Deutschen. *Glottometrics, 24*, 1-4.

Best, K.-H. (2015). Malay borrowings in English. *Glottometrics, 31*, 50-53.

Best, K. H. (2006). *Quantitative Linguistik: eine Annaeherung*. Göttingen: Peust & Gutschmidt.

Bortolato, C. (2016). Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature. *Glottometrics, 34*, 28-43.

Chen, H., & Liu, H. (2014). A diachronic study of Chinese word length distribution. *Glottometrics, 29*, 81-94.

Chen, R., & Liu, H. (2014). Quantitative Aspects of Journal of Quantitative Linguistics. *Journal of Quantitative Linguistics, 21*(4), 299-340.

Fan, F., Grzybek, P., & Altmann, G. (2010). Dynamics of word length in sentence.

Glottometrics, 20, 70-109.
Fenk-Oczlon, G., & Fenk, A. (2002). Zipf's Tool Analogy and Word Order. *Glottometrics, 5*, 22-28.
Ferrer-i-Cancho, R. (2013). Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics. 25*, 1-21.
Hřebíček, L. (2003). Some Aspects of the Power Law. *Glottometrics, 6*, 1-8.
Köhler, R. (2002). Power Law Models in Linguistics: Hungarian. *Glottometrics, 5*, 51-61.
Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin, New York: de Gruyter.
Köhler, R., Altmann, G., & Piotrowski, R. G. (2005). Quantitative Linguistik/Quantitative Linguistics: ein Internationales Handbuch/An International Handbook. Berlin & New York: de Gruyter.
Kubát, M., & Čech, R. (2016). Quantitative Analysis of US Presidential Inaugural Addresses. *Glottometrics, 34*, 14-27.
Levickij, V., & Hikow, L. (2004). Zum Gebrauch der Wortarten im Autorenstil. *Glottometrics, 8*, 12-22.
Li, W. (2002). Zipf's Law Everywhere. *Glottometrics, 5*, 14-21.
Mačutek, J., Švehlíková, Z., & Cenkerová, Z. (2011). Towards a model for rank-frequency distributions of melodic intervals. *Glottometrics, 21*, 60-64.
Martináková, Z., Popescu, I.-I., Mačutek, J., & Altmann, G. (2008). Some problems of musical texts. *Glottometrics, 16*, 63-79.
Pan, X., Qiu, H., & Liu, H. (2015). Golden section in Chinese Contemporary Poetry. *Glottometrics, 32*, 55-62.
Poiret, R., & Liu, H. (2017). Mastering the measurement of text's frequency structure: an investigation on Lambda's reliability. *Glottometrics, 37*, 82-100.
Popescu, I.-I. (2003). On a Zipf's Law Extension to Impact Factors. *Glottometrics, 6*, 61-64.
Popescu, I.-I., & Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics, 13*, 23-46.
Popescu, I.-I., & Altmann, G. (2015). A simplified lambda indicator in text analysis. *Glottometrics, 30*, 19-44.
Popescu, I.-I., Best, K.-H., & Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics, 14*, 58-71.
Popescu, I.-I., Mačutek, J., & Altmann, G. (2008). Word frequency and arc length. *Glottometrics, 17*, 18-42.
Popescu, I.-I., Zörnig, P., & Altmann, G. (2013). Arc length, vocabulary richness and text size. *Glottometrics, 25*, 43-53.
Thomson Reuters. 2017. InCites™ Journal Citation Report Help. Retrieved July 09th 2017 from
http://ipscience-help.thomsonreuters.com/incitesLiveJCR/overviewGroup/overviewJCR.html
Tuzzi, A., Popescu, I.-I., & Altmann, G. (2011). Parts-of-speech diversification in Italian texts. *Glottometrics, 19*, 42-48.
Uhlířová, L. (2007). Word frequency and position in sentence. *Glottometrics, 14*, 1-20.
Vulanović, R., & Canton, N. (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics, 17*, 51-65.
Wang, H. (2013). Length and complexity of NPs in Written English. *Glottometrics, 24*, 79-87.
Wheeler, E. S. (2002). Zipf's Law and why it works everywhere. *Glottometrics, 4*, 45-48.
Wilson, A. (2003). Word-Length Distribution in Modern Welsh Prose Texts. *Glottometrics, 6*, 35-39.
Wimmer, G., & Altmann, G. (2001). Some statistical investigations concerning word classes. *Glottometrics, 1*, 109-123.
Zörnig, P. (2017). On the arc length in quantitative linguistics: a continuous model.

Glottometrics, 36, 22-31.

Appendix: Citations of Top Linguistic Journals

In reference to in Journal Citation Report, the top linguistics journals in 2016 are as follows in Table 20.

Table 20
2016 top linguistics journals in JCR ranked by Journal Impact Factor

No.	Journal Title	Total Cites[9]	Journal Impact Factor	Eigenfactor Score[10]
1	Applied Linguistics	2797	3.593	0.00251
2	Journal of Memory and Language	8541	3.065	0.00923
3	Bilingualism-Language and Cognition	2210	3.010	0.00437
4	Journal of Fluency Disorders	968	2.714	0.00101
5	Computational Linguistics	2235	2.528	0.00101
6	Brain and Language	6186	2.439	0.00971
7	ReCALL	595	2.333	0.00081
8	Language Learning & Technology	1189	2.293	0.00115
9	International Journal of Language & Communication Disorders	1745	2.195	0.00321
10	Cognitive Linguistics	1010	2.135	0.00141
11	Computer Assisted Language Learning	976	2.121	0.00115
12	Annual Review of Applied Linguistics	723	2.083	0.00111
13	Language Learning	3198	2.079	0.00415
14	TESOL Quarterly	3174	2.056	0.00219
15	Studies in Second Language Acquisition	2274	2.044	0.00198
16	Applied Psycholinguistics	2095	1.970	0.00267
17	Language Teaching	849	1.913	0.00166
18	Research on Language and Social Interaction	1016	1.896	0.00301
19	Language Cognition and Neuroscience	413	1.852	0.00194
20	Journal of Speech Language and Hearing Research	6675	1.771	0.00125

Results of citation analysis of the top journals in Table 20 are provided below.

[9] In Web of Science, *Total Cites* or the total number of times that a journal has been cited by all journals included in the database in the JCR year (Thomson Reuters, 2017).

[10] In Web of Science, *Eigenfactor Score* is "based on the number of times articles from the journal published in the past five years have been cited in the JCR year, but it also considers which journals have contributed these citations so that highly cited journals will influence the network more than lesser cited journals." (Thomson Reuters, 2017)

Table 21
Results of citation analysis of the journals in Table 20

No.	Journal	Results found	*h*-index	Average citations per item
1	Applied Linguistics	849	61	15.73
2	Journal of Memory and Language	1452	119	48.02
3	Bilingualism-Language and Cognition	642	39	11.63
4	Journal of Fluency Disorders	797	40	9.26
5	Computational Linguistics	814	59	19.62
6	Brain and Language	5097	130	23.63
7	ReCALL	201	20	6.55
8	Language Learning & Technology	475	39	11.27
9	International Journal of Language & Communication Disorders	1357	47	9.86
10	Cognitive Linguistics	503	35	9.12
11	Computer Assisted Language Learning	306	23	6.84
12	Annual Review of Applied Linguistics	118	18	9.35
13	Language Learning	698	60	20.73
14	TESOL Quarterly	1224	55	11.1
15	Studies in Second Language Acquisition	760	44	8.5
16	Applied Psycholinguistics	816	59	16.89
17	Language Teaching	276	21	6.13
18	Research on Language and Social Interaction	370	37	17.11
19	Language Cognition and Neuroscience	380	11	2.39
20	Journal of Speech Language and Hearing Research	2661	109	26.38

Google also releases another list of top publications in the subcategory of "Language and Linguistics" in the 2017 version of Scholar Metrics.

Table 22
Top publications in Language and Linguistics[11] according to Scholar Metrics 2017

No.	Publication	*h*5-index[12]	*h*5-median
1	Language Learning	42	64
2	Journal of Memory and Language	39	60
3	Applied Linguistics	34	46
4	Natural Language & Linguistic Theory	30	51

[11] From: https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=hum_languagelinguistics. This release covers articles published in 2012-2016 and includes citations from all articles that are indexes in Google Scholar as of June 2017.

[12] In Google Scholar Metrics, *h*5-index means the *h*-index in the five years, and *h*5-median means *h*-median in the five years.

5	Language	28	48
6	Applied Psycholinguistics	28	41
7	Linguistic Inquiry	27	46
8	Lingua	27	39
9	Studies in Second Language Acquisition	26	49
10	Journal of Phonetics	26	36
11	International Journal of Bilingualism	24	32
12	Journal of Child Language	23	30
13	Language and Linguistics Compass	22	36
14	Journal of Neurolinguistics	21	34
15	Language Learning and Development	20	32
16	Language Sciences	20	31
17	Second Language Research	20	31
18	Language, Cognition and Neuroscience	20	30
19	First Language	18	29
20	Language and Speech	18	27

Glottometrics 39, 2017, 38-71

The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach

Ramon Ferrer-i-Cancho[1]

Abstract: The minimization of the length of syntactic dependencies is a well-established principle of word order and the basis of a mathematical theory of word order. Here we complete that theory from the perspective of information theory, adding a competing word order principle: the maximization of predictability of a target element. These two principles are in conflict: to maximize the predictability of the head, the head should appear last, which maximizes the costs with respect to dependency length minimization. The implications of such a broad theoretical framework to understand the optimality, diversity and evolution of the six possible orderings of subject, object and verb, are reviewed.

Keywords: *word order, gesture, information theory, compression, Hilberg's law*

1. Introduction

When producing an utterance speakers have to arrange elements linearly, forming a sequence. The same problem applies to users of a sign language or unconventional gesture systems (Goldin-Meadow 1999). Suppose that we have to order linearly a head and its dependents (complements or modifiers). In a verbal sequence made of subject, verb and object, we assume that the verb is the head. In a gestural sequence made of actor, action and patient, we assume that the action is the head. In general, what is the best placement of the head?

For the particular case of the ordering of the verb (i.e. the head) and the subject and the object (i.e. the complements), various sources of evidence suggest a preference for placing the verb last. First, the non-verbal experiments in (Goldin-Meadow et al 2008, Langus & Nespor 2010) where a robust strong preference for an order consistent with subject-object-verb (head last) was found even in speakers whose language did not have subject-object-verb as the dominant word order. Second, *in silico* experiments with neural networks have shown that subject-object-verb (head last) is the word order that emerges when languages are selected to be more easily learned by networks predicting the next element in a sequence (Reali & Christiansen 2009). Thirdly, the most frequent dominant word order among world languages is subject-object-verb (head last) (Drier 2013, Hammarström 2016). Table 1 shows that the total frequency of dominant orders increases as the head (V) moves from the beginning of the sequence (VOS/VSO) to the center (SVO/OVS) and finally to the end

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

(SOV/OSV). That fact suggests that postponing the verb (the head) is favored for some reason.

Table 1
The frequency of the placement of the ordering of the subject (S), verb (V) and object (O) in world languages showing a dominant word order. Frequency is measured in languages and in families.

Order	Languages		Families	
	Frequency	Percentage	Frequency	Percentage
SOV	2275	43.3	239	65.3
SVO	2117	40.3	55	15.0
VSO	503	9.6	27	7.4
VOS	174	3.3	15	4.1
OVS	40	0.8	3	0.8
OSV	19	0.4	1	0.3
No dominant order	124	2.4	26	7.1
**V	2294	43.7	240	65.6
V	2157	41.1	58	15.8
V**	677	13.9	42	11.5
All	5252		366	

V is used for verb final orderings (SOV and OSV), *V* is used for central verb placements (SVO and OVS) and V for verb initial orderings (VSO and VOS). Frequency is measured in languages and also in families. Absolute frequencies are borrowed from (Hammarström 2016). Percentages were rounded to the nearest decimal.

Here we will provide general information theoretic arguments that predict that the verb (or the head in general) should be postponed and eventually placed last to maximize its predictability. The outline of the argument is as follows. Consider two practically equivalent pressures: the minimization of the uncertainty about a target element, and the maximization of the predictability of a target element (they are equivalent for sequences of length three or longer as explained in detail in Section 2). A target element is a specific element of the sequence that has not been produced yet. For simplicity, suppose that the sequence consists of word forms and the target is a word form. This setup can be easily adapted to other contexts, e.g., in animal behavior research, the target could be a type (of behavior) and the sequence would be made of types (Section 2 presents a generalization of the setup). We may choose a target between a head and its dependents or between a verb and its arguments. These pressures predict that the target element should be placed last. This result is intuitive: adding more elements before the target element cannot hurt (a reduction in predictability would hurt), and in general will help to predict it or to reduce its uncertainty (Cover & Thomas 2006). Similarly, Fenk-Oczlon (1989) stated that, "*as a linguistic sequence progresses, the number of possible continuations becomes more and more restricted; that is, there is a reduction of uncertainty of the information*".

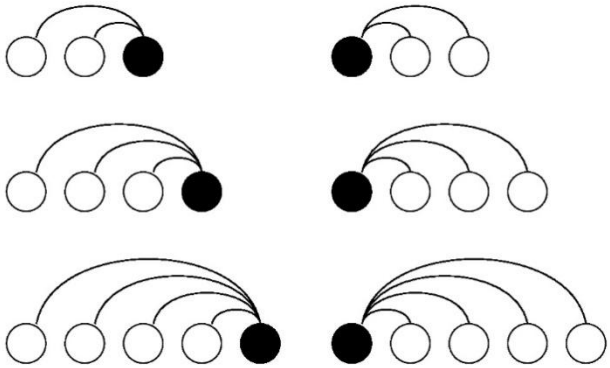

Figure 1. Optimal sequential placement of a head and its dependents (modifier/ complements) according to predictability maximization (or uncertainty minimizeation) for sequences of increasing length *m*.

The black circle indicates the head while the white circles indicate the dependents. Edges indicate syntactic dependencies between a head and its dependents. Top: $m = 3$. Center: $m = 4$. Bottom: $m = 5$. The left column indicates the optimal placements when the head is the target of predictability maximization. The right column corresponds to the optimal placement when the target are the dependents.

In case that the target element is the head, the result above implies that the head should be placed last (Fig. 1, left column). Assuming that the verb is the head the latter implies left branching. For the particular case of the subject-verb-object triple, the verb (or the action) then should be placed after the subject and the object (or the agent and the action). Interestingly, placing the verb at the center is not optimal but it is better than putting it first: postponing the verb is increasingly beneficial. In case that the target elements are the dependents, the head should be put first (Fig 1, right column). This implies right branching; the verb or the action should be the first element. The key is to understand why there should be a preference for the verb (or heads in general) to be the target.

These considerations notwithstanding, language is a multiconstraint engineering problem (Evans & Levinson 2009, Zipf 1949). Uncertainty minimization / predictability maximization are not the only relevant pressure in word order. An alternative well-established principle of word order is dependency length minimization (Liu 2017, Ferrer-i-Cancho 2015a). Suppose that we define the length of a dependency as the linear distance in words between the head and the dependent. If the head and the dependent are adjacent, the length is 1; if they are separated by one element, the length is 2; and so on…The principle of dependency length minimization consists of minimizing the sum of those dependencies. According to that principle, the optimal placement of a single head and its n dependents is at the center (Ferrer-i-Cancho 2015a). The length of the sequence is $m = n + 1$. If m is odd then there is only one possible central placement (Top and bottom of Fig. 2) while if it is even then there are two central placements (Center of Fig. 2). For this reason, the placement of the head is irrelevant when the se_quence only has two elements. The predictions of a central placement by the principle of dependency length minimization is exact if the dependents are atomic, i.e. made of just one word (Ferrer-i-Cancho 2015a), and is approximately valid when they are not (Ferrer-i-Cancho 2008, Ferrer-i-Cancho 2014). The argument can be refined (and generalized) supposing that the cognitive cost of a dependency increases as its length

increases, and that the target of the minimization is the sum of the costs of all dependencies. Again, the optimal placement of the head is at the center (Ferrer-i-Cancho 2015a, Ferrer-i-Cancho 2014). The argument can also be refined measuring length in letters or phonemes instead of words (Ferrer-i-Cancho 2015b).

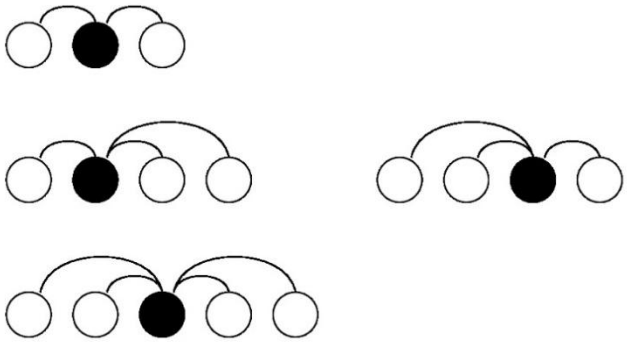

Figure 2. Optimal sequential placement of a head and its dependents according to dependency length minimization for sequences of increasing length *m*.

The black circle indicates the head while the white circles indicate the dependents. Edges indicate syntactic dependencies between a head and its modifier/complements. Top: $m = 3$ with only one optimal placement Center: $m = 4$ with two optimal placements. Bottom: $m = 5$ with only one optimal placement.

Interestingly, the principle of dependency length minimization is in conflict with the principle of predictability maximization / uncertainty minimization: while the former predicts that the head should be placed at the center of the sequence, the latter predicts that it should be placed at one of the ends. This article explores the implications of these conflicts and how they can be integrated into a general theory of word order.

The remainder of the article is organized as follows. Section 2 presents the mathematical arguments in detail. In our information theoretic approach, uncertainty is formalized as an entropy and predictability is formalized as a mutual information. We will show that uncertainty minimization has higher predictive power than mutual information maximization and we will also show that the former is equivalent to the latter for a sequence of at least three elements. This section is recommended to readers who lack the intuitions behind the results summarized above. Section 3 reviews the constant entropy rate and other information theoretic hypotheses since they are often regarded as reference theories. Section 4 presents a broad perspective on word order theory, incorporating the information theoretic approach elaborated in Section 2 and discussing implications for the ordering of subject, verb and object or its semantic correlates, i.e. actor, action and patient. Sections 2 and 3 can be skipped.

2. Information theory of word order

We aim to provide an information theoretic approach to word order that is consistent with other information theoretic approaches to language. Our guiding principle is that *"Scientific knowledge is systematic: a science is not an aggregation of disconnected information, but a system of ideas that are logically connected among themselves. Any system of ideas that is characterized by a certain set of fundamental (but refutable) peculiar hypotheses that try to fit a class of facts is a theory"* (Bunge, 2013, pp. 32-33).

For this reason, we will extend information theoretic principles that have been successful in explaining various linguistic phenomena: entropy minimization and mutual information maximization. A family of optimization models of natural communication is based on a combination of minimization of $H(S)$, the entropy of words of a vocabulary S, and the maximization of $I(S,R)$, the mutual information between the words (S) and the meanings (from a repertoire R). Here we extend and generalize this principles to be able to model word order phenomena. We refer the reader to Ferrer-i-Cancho (2017a) for a review of the cognitive and information theoretic justification of these principles. We also refer the reader to Chapter 2 of Cover & Thomas (2006) for further mathematical details about entropy, mutual information and conditional entropy.

We model a linguistic sequence (e.g. a sentence) as a sequence of elements X_1, X_2, X_3,...(e.g., the words of the sentence). First, let us consider $H(S)$. We proceed by replacing S (a whole vocabulary) by a target of a sequence Y and conditioning on elements of the sequence that have already appeared. This yields $H(Y|X_1,X_2, X_3,...)$. We postulate that this conditional entropy has to be minimized as $H(S)$. The next subsection presents the details of this minimization. We note that the minimization of entropy could be an axiom or a side-effect of compression. In the case of a vocabulary, the goal of compression is to minimize $L(S)$, the mean length of words. Interestingly, $L(S)$ is bounded below by $H(S)$ under the constraint of uniquely decipherability (Ferrer-i-Cancho, 2017a). Thus, minimizing $H(S)$ could be a consequence of pressure of the minimization of $L(S)$. The possibility that the minimization of $H(Y|X_1,X_2, X_3,...)$ is a side-effect of compression should be the subject of future research. The reason is that compression has the potential to offer a parsimonious explanation to various linguistic laws, including the popular Zipf's law for word frequencies (Ferrer-i-Cancho 2016b) and also Zipf's law of abbreviation (Ferrer-i-Cancho et al 2013b, Ferrer-i-Cancho et al 2015) and Menzerath's law (Gustison et al 2016).

Second, let us consider $I(S,R)$. As before, we proceed by replacing S (a whole vocabulary) by a target of a sequence Y and replacing R by elements of the sequence that have already appeared. This yields $I(Y; X_1, X_2, X_3,...)$. We postulate that this mutual information has to be maximized as $I(S,R)$. The next subsection presents the details of this maximization.

As we have been recently reminded, a model of Zipf's law for word frequencies should be able to make predictions beyond Zipf's law (Piantadosi 2014), and this is what applies to the family of optimization models above, which make successful predictions about the mapping of words into meanings (the principle of contrast), and vocabulary learning in children (Ferrer-i-Cancho 2017b). However, here we are going further: we are providing a set of general information theoretic principles, i.e. *a set of fundamental* (*but refutable*) *peculiar hypotheses* (as M. Bunge would put it), that can be used to build models in new domains, e.g., word order for the present article. Piantadosi's (2014) reminder falls short: the ultimate goal of a language researcher is not to design a model that predicts various properties of language simultaneously but to build a general theory for the class of linguistic phenomena.

2.1 The order that minimizes the uncertainty about the target or that maximizes its predictability

Suppose that a linguistic sequence (a sequence of words or a sequence of gestures) has m elements. The sequence can be represented by m random variables X_1, ...,X_i, ..., X_m, where X_i represents some information about the i-the element of the sequence. The setup is abstract and thus flexible: X_i could be the word type, the part-of-speech or the meaning of the i-th element of the sequence.

Suppose that the whole sequence consists of one target element and other $n = m - 1$ elements. For instance, the target element could be the head and the other elements could be the dependents (modifiers or complements). We use the random variable Y for the target and $X_1,...,X_i,...,X_n$ for the other elements. Again, Y could be the word type, the part-of-speech or the meaning of the target element of the sequence.

When i elements have been produced,

- The uncertainty about the target Y is defined as $H(Y|X_1, X_2,...,X_i)$, the conditional entropy of Y given $X_1, X_2,...,X_i$.
- The predictability of the target is defined as $I(Y|X_1, X_2,...,X_i)$, the mutual information between Y and $X_1, X_2,...,X_i$.

For instance,

- $H(Y|X_1, X_2,...,X_i)$ could be the uncertainty about the meaning of the target Y (e.g., the predicate representing the meaning of the target according to logical semantics) when the speaker has produced the words forms $X_1, X_2,...,X_i$.
- $I(Y|X_1, X_2,...,X_i)$ could be the predictability of the meaning of the target Y when the speaker has produced the word forms $X_1, X_2,...,X_i$.

We are interested in the placement of the target where its uncertainty is minimized or its predictability is maximized. Further mathematical details can be found in Appendix A. Here we explain the bulk of the arguments.

The problem of the optimal placement of the target can be formalized as follows. The solutions of

$$\text{argmin}_{1\leq i\leq n} H(Y|X_1, X_2,..., X_i) \quad (1)$$

yield the optimal placements according to uncertainty. For instance, if the solution was $i=n$ then the minimum would be reached when the target is placed last. If the solution was $i=0$ then minimum would be reached when the target is placed first. Similarly, the solutions of

$$\text{argmax}_{1\leq i\leq n} I(Y|X_1, X_2,..., X_i) \quad (2)$$

yield the optimal placements according to predictability. It can be shown that Eqs. 1 and 2 have at least one solution, i.e. $i=n$. Put differently, the optimal placement of the target is at least in the last position in a real linguistic sequence: real linguistic sequences exhibit long-range correlations both at the level of letters and at the level of words (Montemurro & Pury 2002, Ebeling & Pöschel 1994, Alvarez-Lacalle et al 2006, Moscoso del Prado Martín 2011, Altmann et al 2012). The argument relies on two crucial properties (Appendix A):

$$H(Y|X_1, X_2,..., X_{i-1}) \geq H(Y|X_1, X_2,..., X_i) \quad (3)$$

for $i \geq 1$, and

$$I(Y|X_1, X_2,..., X_{i-1}) \leq I(Y|X_1, X_2,..., X_i) \quad (4)$$

for $i \geq 2$. Equality in Eqs. 3 and 4 appears only in some particular cases (Appendix A).

The result in Eq. 3 and Eq. 4 allow one to understand why postponing the target (producing more elements of the sequence) is optimal. In general, the uncertainty about the target reduces as the target is postponed, and implies that the minimum uncertainty is reached

when it appears last (Eq. 3). Similarly, the predictability of the target improves, in general, as the target is postponed and the maximum predictability will be reached at least when it is placed at the end of the sequence. Therefore, in the absence of further knowledge about a sequence, the optimal strategy is to put the target last.

To sum up, the minimization of the uncertainty of the target or the maximization of its predictability leads to a final placement of the target. Interestingly, the element that has to be put last depends on the target. For instance, if the target is the head then its dependents should appear first. In contrast, if the target is one of the dependents (e.g., the object of a verb) then the head should not appear last.

The argument can be refined considering the problem of the minimization of the energetic cost associated to the uncertainty or to predictability. In this case, we define two functions, i.e. g_H and g_I, that translate, respectively, entropy and mutual information into an energetic cost from the perspective of uncertainty or predictability (thus these cost functions do not take into account dependency length minimization costs). In particular, g_H is a strictly monotonically increasing function while g_I is a strictly monotonically decreasing function.

Then the optimal placement according to uncertainty is given by

$$\text{argmin}_{1\leq i\leq n} g_H[H(Y|X_1, X_2,..., X_i)]. \quad (5)$$

while the optimal solution according to predictability is given by

$$\text{argmax}_{1\leq i\leq n} g_I[I(Y|X_1, X_2,..., X_i)]. \quad (6)$$

Again the optimal strategy in general is to put the target last in the absence of any further information.

g_H and g_I play the same role as the function g that has been used to investigate the optimal placement of the head according to dependency length minimization (Ferrer-i-Cancho 2015a, Ferrer-i-Cancho 2014). In the latter case, g is a strictly monotonically increasing function that translates an edge length into its energetic cost.

We have presented uncertainty minimization and predictability maximization as equivalent (Section 1). However, Eqs. 5 and 6 show that uncertainty minimization has a broader scope because $m \geq 2$ suffices to decide that the target should be placed last (when $m = 1$ there is no decision to make). In contrast, predictability maximization needs $m \geq 3$. Therefore, uncertainty minimization can operate on smaller sequences than predictability maximization. Hereafter we will use uncertainty minimization by default bearing in mind that it is equivalent to predictability maximization when $m \geq 3$.

2.2 A conflict between uncertainty minimization and dependency length minimization

Suppose that a sequence consists of a head and $n = m - 1$ dependents. According to the principle of minimization of uncertainty, the optimal placement of the head is extreme: at the end if the target is the head or at the beginning if the target are the dependents seen as a block of consecutive elements (in the latter case, the dependents have to be placed last which implies that the head is placed first). In contrast, the optimal placement of the head is at the center according to the principle of dependency length minimization (Ferrer-i-Cancho 2015a), as illustrated in Fig. 2. If m is even there only one central placement that is optimal. If m is odd there are two central placements (Fig. 2).

This implies that these two order principles are in conflict provided that $m \geq 3$. To see that no conflict exists when $m < 3$ notice that no word order problem exists when $m < 2$. When m is even, there are two central positions and if $m = 2$ any position is therefore optimal for dependency length minimization (Ferrer-i-Cancho 2015a). Therefore, one expects that word order is determined by uncertainty minimization when $m = 2$.

To understand the severity of the trade-off, notice that an extreme head placement (head first or head last), thus an optimal placement of the target according to uncertainty minimization, maximizes the cost of dependency lengths (Ferrer-i-Cancho 2015a). While an extreme placement of the head yields a maximum sum of dependency lengths that is (Ferrer-i-Cancho 2015a)

$$D = \binom{m}{2} = \frac{m(m-1)}{2}, \tag{7}$$

a central placement of the heads gives a minimum sum of dependency lengths that is (Ferrer-i-Cancho 2015a)

$$D = \frac{1}{4}(m^2 - m \bmod 2). \tag{8}$$

In sum, the best case for uncertainty minimization is the worst case for dependency length minimization.

Interestingly, the converse does not hold: the best case for dependency length minimization is not the worst case for uncertainty minimization. When the head is the target and it is placed at the center, it is preceded by some elements that may have helped to reduce its uncertainty. When the target is the dependents and the head is placed at the center, the head helps to reduce the uncertainty of the dependents that have not appeared yet.

3. Constant entropy rate and related hypotheses

3.1. An introduction

Here we compare our arguments about word order against the constant entropy rate (CER) and related hypotheses (Genzel & Charniak 2002, Levy & Jaeger 2007, Jaeger 2010), These hypotheses are argued to explain various linguistic phenomena, e.g., syntactic reduction (Levy & Jaeger 2007, Jaeger 2010) and the frequency of word orders (Maurits et al 2010). We review them here because they are considered as a reference theory to any alternative information theoretic approach to language by some language researchers. The importance of these hypotheses is evident from the number of citations, the impact factors of the journals, and the institutions from which they are broadcast.

The core of these hypotheses is the existence of a "*preference to distribute information uniformly across the linguistic signal*" (Jaeger 2010, p. 23). In greater detail, the hypothesis could be formulated as (Jaeger 2010, p. 24)

> "*Human language production could be organized to be efficient at all levels of linguistic processing in that speakers prefer to trade off redundancy and reduction. Put differently, speakers may be managing the amount of information per amount of linguistic signal (henceforth information density), so as to avoid peaks and troughs in information density.*"

3.2. The origins of the hypotheses

This idea was introduced by August and Gertraud Fenk (1980, 2nd paragraph from the bottom of page 402):

> "*A communication system, which is supposed to deliver messages without loss, should not only be required to have a certain average level of redundancy (not exceeding the short term memory capacity), but also, that the information is distributed as uniformly as possible across small time spans.*"[2]

and developed in a series of articles (see Fenck-Oczlon (2001) for a review). Figure 1 of Jaeger (2010) and the figure in p. 403 of Fenk & Fenk (1980) are similar in terms of the axes' names and the shape of the curves. The work by G. Fenk predates by about 30 years what are considered to be the core articles (Jaeger 2010, Jaeger & Levy 2007) and by about 20 years the foundational articles of this family of hypotheses (Genzel & Charniak 2002, Aylett & Turk 2004).

There is a general reference to Fenk-Oczlon (2001) in Jaeger (2010), detached from the context of uniform information density. The relevant passages of section "2.2 Frequency and the constant flow of linguistic information" of Fenk-Oczlon (2001) are not mentioned. In the following, we will use the label "constant flow hypothesis" to refer to the original formulation. The following sections are focused on the developments of the later hypotheses of Section 3.1.

3.3 Their formal definition and their real support

Constant entropy rate and related hypotheses are popular among cognitive scientists working on language. However, they are generally unknown to quantitative linguists and the physicists who started investigating the statistical properties of symbolic sequences in the 1990s (e.g., Ebeling & Pöschel 1994). This is not very surprising given the lack of contact between these different disciplines, but also given the large gulf that separates the formal statements of these hypotheses and the statistical properties of real language.

Suppose that $H(X_i|X_1, X_2,...,X_{i-1})$ is the entropy of X_i, the i-th type of the sequence, knowing the types that precede it. In mathematical detail, the constant entropy rate (CER) hypothesis states that $H(X_i|X_1, X_2,...,X_{i-1})$ should remain constant as i increases, i.e. (Genzel, D. & Charniak 2002)

$$H(X_1) = H(X_2|X_1) = ... = H(X_i|X_1,...,X_{i-1}) = ... = H(X_m|X_1,...,X_{m-1}). \quad (9)$$

To a quantitative linguist familiar with Hilberg's law (Hilberg 1990), it is obvious that Eq. 9 does not hold since that law states that

$$H(X_i|X_1,...,X_{i-1}) \approx \mathrm{a}i^{-\gamma}, \quad (10)$$

where $\gamma \approx 0.5$ and a is a positi.ve constant. A more plausible version of the law has been proposed, by Dębowski (2015), namely

$$H(X_i|X_1,...,X_{i-1}) \approx \mathrm{a}i^{-\gamma} + b, \quad (11)$$

where a and b are positive constants.

[2] We owe this translation from the original German version to Chris Bentz.

Therefore, real texts do not satisfy Eq. 9. However, Eq. 9 is satisfied when $X_1,...,X_i,...,X_m$ are independent identically distributed (i.i.d.) variables. Thus a text consistent with the constant entropy rate hypothesis is easy to generate: take a real text and scramble it at the desired level (e.g., letters or words). The random text that you will produce will fit CER beautifully at the level chosen.

Consider a concrete sequence $x_1,x_2,...,x_i,...,x_m$. A related hypothesis is the uniform information density (UID) hypothesis, that is defined on $p(x_i|x_1,...,x_{i-1})$, the probability of the i-th element of a sequence conditioned on the previous elements. The hypothesis states that (Levy & Jaeger 2007)

$$p(x_1) = p(x_2|x_1) = ... = p(x_i|x_1,..., x_{i-1}) = ... = p(x_m|x_1,..., x_{m-1}). \quad (12)$$

While testing the validity of CER is easy, as we have seen above, refuting UID is more difficult *a priori* because it is poorly specified. For this reason, more specific hypotheses have been defined from Eq. 12 (Ferrer-i-Cancho et al 2013a). The strong UID hypothesis states that Eq. 12 should hold in every sequence of length m that can be produced. The full UID hypothesis is a particular case of strong UID where the set of sequences that can be produced are all possible sequences (i.e. the Cartesian product of the sets of symbols available at every position). Strong UID is a particular case of CER and therefore both versions of UID suffer from all the limitations of CER. The full UID is a particular version of the strong UID that implies a sequence of independent elements.

A challenge for CER and UID is that they hold in situations that are incomepatible with language. A scrambled text satisfies CER, i.e. Eq. 9, with $H(X_i|X_1, X_2,...,X_{i-1}) = H(X)$ for $1 \leq i \leq m$, where $H(X)$ is the entropy of the words of the text. Other sequences also satisfy CER (Eq. 9) with $H(X_i|X_1, X_2,...,X_{i-1}) = 0$ for $1 \leq i \leq m$:

- A homogenous sequence, e.g., "*aaaaaa*…" (another example of a sequence of i.i.d. variables, notice $p(x_i|x_1,\ldots,x_{i-1}) = 1$ for $i \geq 1$ and for every $x_1,\ldots,x_{i-1},x_i$ in the support set).
- A perfect periodic sequence, i.e. a sequence of that consists of the repetition of a block of T different types, e.g.,"*abcabcabc*…". When $T = 1$ we have a homogenous sequence and thus the interesting case is $T > 1$. If we assume that $H(X_i|X_1, X_2,...,X_{i-1})$ is the entropy of the i-th element of the sequence given all the preceding elements then we have $H(X_i|X_1, X_2,...,X_{i-1}) = 0$ for $1 \leq i \leq m$ because the first element is always the same and the next element can always be predicted perfectly knowing the last element. If we relax the definition of $H(X_i|X_1, X_2,...,X_{i-1})$ as the entropy of the i-th element of an arbitrary subsequence of the original sequence given the preceding elements in that subsequence then we have quasi CER, namely $H(X_i|X_1, X_2,...,X_{i-1}) = \log T$ for i = 1 and $H(X_i|X_1, X_2,...,X_{i-1}) = 0$ for $2 \leq i \leq m$. The reasons is that the first element is one of the block chosen uniformly at random and the next element can still be predicted perfectly knowing the last element. A perfect periodic sequence with T > 1 shows that CER does not imply independence between elements.

Notice that a scrambled text and a homogeneous sequence are examples of sequences of independent and identically distributed (i.i.d.) elements. CER holds for any i.i.d. process but is not limited to them as the example of a perfect periodic sequence with $T > 1$ indicates.

Therefore, CER is satisfied by sequences that include the best case (a perfect periodic sequence) and the worst case (a sequence of identically distributed elements) for predicting the next element of the sequence. As a principle of word order, CER includes sequences that lack any order.

3.4 The justification of the hypotheses

The main argument used to justify the uniform information density and related hypotheses is the phenomenon of reduction, namely "*more predictable instances of the same word are on average produced with shorter duration and with less phonological and phonetic detail*" (see Jaeger 2010, p.23 for a review of the literature on this phenomenon). This context-dependent reduction is reminiscent of the tendency of more frequent words to be reduced regardless of their context (Fenk-Oczlon 2001). We will refer to the latter as 1st order reduction and to the former as higher order reduction.

Standard information theory is concerned about 1st order reduction. Suppose that p_i and l_i are, respectively, the probability and the length of the code of the i-th type and then

$$\sum_i p_i = 1. \tag{13}$$

Within coding theory, the goal of solving the problem of compression is to minimize the mean length of the codes assigned to each type (Cover & Thomas 2006, p. 110), i.e.

$$L = \sum_i p_i l_i, \tag{14}$$

under a certain coding scheme (typically uniquely decipherable codes). Put differently, coding theory is concerned about reducing the length of "words" as much as possible. Under the scheme of uniquely decipherable codes or non-singular codes, optimal coding successfully predicts Zipf's law of abbreviation, namely the tendency of more likely elements to be shorter (Ferrer-i-Cancho et al 2015). Therefore, standard information theory is concerned with reduction of more likely elements without context. Interestingly, standard information theory can be easily extended to reduction with context, namely higher order reduction. Suppose that we focus on the reduction of a concrete word y.

We may define the mean length of a type in combination with a previous context of n consecutive words as

$$L_n = \sum_{x_1,x_2,\ldots,x_n,y} p(x_1, x_2,\ldots, x_n, y) l(x_1, x_2,\ldots, x_n, y), \tag{15}$$

where $p(x_1,x_2,\ldots,x_n,y)$ and $l(x_1,x_2,\ldots,x_n,y)$ are, respectively, the probability and the length of the type y when it is preceded by the sequence of types $x_1,x_2,\ldots,x_n$. We assume

$$\sum_{x_1,x_2,\ldots,x_n,y} p(x_1, x_2,\ldots, x_n, y) = 1, \tag{16}$$

when $n = 0$, L_n becomes L as defined in Eq. 14. Again, optimal coding predicts a generalized Zipf's law of abbreviation: the tendency of more frequent type-context combinations to be shorter (Ferrer-i-Cancho et al 2015). A prediction under non-singular coding or uniquely decipherable encoding is that the minima of L_n satisfy

$$\tau(p(x_1, x_2,\ldots, x_n, y), l(x_1, x_2,\ldots, x_n, y)) \leq 0, \tag{17}$$

where $\tau(\ldots,\ldots)$ is the Kendall tau correlation (Ferrer-i-Cancho et al 2015). This general result has strong implications for research on the reduction of a target type, e.g. "that" as in Levy and Jaeger (2007). In particular, a target type is expected to be shorter in contexts that are more likely. Put differently, compression predicts that types that appear in more predictable contexts have to be reduced.

To see it from a complementary perspective, we may define L_n equivalently as

$$L_n = \sum_y L_n(y), \tag{18}$$

with

$$L_n(y) = \sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2,\ldots, x_n, y) l(x_1, x_2,\ldots, x_n, y). \tag{19}$$

Renormalizing locally, i.e. dividing $L_n(y)$ by $p(y)$, we obtain

$$M_n(y) = \sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2,\ldots, x_n|y) l(x_1, x_2,\ldots, x_n, y), \tag{20}$$

where $p(x_1,x_2,\ldots,x_n|y)$ is the probability of the block $x_1,x_2,\ldots,x_n$ knowing that it is followed by y.

Notice that

$$\sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2,\ldots, x_n|y) = 1. \tag{21}$$

$M_n(y)$ can be seen as a particular case of L where the set of types is defined by all the contexts of length n that can precede a concrete type y. The minimization of $M_n(y)$ predicts that y should be shorter in more likely contexts (Ferrer-i-Cancho et al 2015). Again, a prediction is that the minima of M_n (y) satisfy

$$\tau(p(x_1, x_2,\ldots, x_n|y), l(x_1, x_2,\ldots, x_n, y)) \leq 0. \tag{22}$$

Therefore, one does not need uniform information density and related hypotheses to explain reduction. The principle of compression can suffice.

A potential difference between first order compression and higher order compression could be that the latter may allow for types of length 0, namely full reduction, thanks to the preceding context or the function that the words that undergo total reduction perform. For instance, function words such as the conjunction "that" are easier to remove than content words. Such a tolerance to function word removal is the basis of telegraphic speech (Akmajian et al 2001, p. 23). In 1st order compression, non-singular coding implies codes of length greater than zero.

Interestingly, the case of full reduction and telegraphic speech could be regarded as cases of lossy compression. The critical question is: if lossless or lossy compression may account for reduction, why should CER or UID be necessary?

3.5. The link with standard information theory

A very important feature of a scientific field is that it must be

> "*a component of a wider cognitive filed, i.e. there is at least one other (contiguous) research field such that (a) the general outlooks, formal backgrounds, specific backgrounds, funds of knowledge, aims and methodics of the two fields have non-empty overlaps and (b) either the domain of one field is included in that of the other, or each member of the domain of one of them is a component of a system belonging to the other domain*" (Bunge 1984).

Research in the field of CER/UID and information theory overlap. The domain of CER/UID – human language – is a subset of the domain of information theory, that is also concerned with artificial systems as well as other means of information storage and transmission of information such as genomic sequences (e.g., Naranan & Balasubrahmanyan 2000) or animal behavior (e.g., McCowan et al 1999, Suzuki et al 2006). A very important component of a

scientific theory is a formal background, namely "*a collection of up-to-date logical or mathematical theories (rather than being empty or formed by obsolete formal theories)"* (Bunge 1984). Followers of CER/UID employ jargon from standard information theory such as "*noisy channel*", "*channel capacity*" (e.g., Jaeger 2010, Piantadosi et al 2011), and posit strong links between information theory and uniform information density

> "*The hypothesis of Uniform Information Density links speakers' preferences at choice points during incremental language production to information theoretic theorems about efficient communication through a noisy channel with a limited bandwidth* (Shannon, 1948)" (Jaeger 2010, p. 25)

Does it mean that information theory is actually the formal background of CER/UID in a Bungean sense?

Mentions of standard information theory such as the ones given above could be neglected if CER/UID were not considered reference theories to alternative approaches based on information theory, such as ours. However, since they are widely considered as such it is worth scrutinizing in more detail their actual links with information theory. As we have shown above, followers of CER/UID fail to identify the phenomenon of reduction as a manifestation of compression, thus missing a link with standard coding theory. The loose connection with standard information theory can be understood further when revising the predictions of CER/UID on the efficiency of language.

One of the major problems of CER/UID and related hypotheses is that they are presented as arising from efficiency considerations, but the exact link with optimization is unclear. One example is an article that makes strong claims about the efficiency of language but does not specify the cost function that is being minimized (Piantadosi et al 2011). A complete argument about optimization requires at least three fundamental components:

1. A cost function
2. A theoretical insight linking the minimization of that function and statistical properties of the system.
3. A baseline

In standard coding theory, L (Eq. 14) is the cost function. If L is minimum then it is well-known that

$$l_i = \lceil -\log p_i \rceil \tag{23}$$

for uniquely decipherable encoding (Cover & Thomas, 2006). The three components are found in an extension of coding theory for research on Zipf's law of abbreviation in natural communication systems (Ferrer-i-Cancho et al 2013b, Ferrer-i-Cancho et al 2015): there the cost function is the generalization of L and various mathematical arguments are used to show the relationship between Zipf's law of abbreviation and the minimization of that cost function. The baseline is defined by a randomization of the mapping of probabilities into lengths.

The theoretical insight is crucial. Without it, it is easy to make wrong inferences. Finding a strong correlation between a measure of "information content" and length does not imply that speakers are making optimal choices involving the contexts where words appear (Piantadosi et al 2011): a linear dependency between these two variables may simply arise internally, from the units making a word (e.g., letters) as random typing shows simply (Ferrer-i-Cancho & Moscoso del Prado Martín 2011). Paradoxically, destroying a text by scrambling the text sequence (at the level of words or at the level of characters) will produce a sequence of i.i.d. words that exhibits perfect agreement with CER. Furthermore, finding a correlation

between "information content" and word length that is stronger than the correlation between frequency and length of Zipf's law of abbreviation (Piantadosi et al 2011) does not imply that the former correlation is the outcome of a higher degree optimization: in case of optimal coding, a perfect correlation between frequency and length is not expected due to ties of length in optimal codes (Ferrer-i-Cancho et al 2015). For instance, Eq. 23 implies that all types with the same probability should have the same length and frequency ties are many in real texts (frequency ties are beautifully described by Zipf's number-frequency law, Zipf 1949). The lack of a cost function and a theoretical understanding of its predictions can lead to wrong inferences.

As far as we know, the only attempt to derive mathematically uniform information density from cost minimization can be found in Levy & Jaeger 2007. The attempt is partial for two reasons: it depends on a parameter k and cost is only minimized for $k < 1$. The value or the range of values of k that are suitable for human language are unclear. Second, it does not concern CER, which is a more general condition than strong UID (Section 3.3).

3.6. CER and UID versus our word theory order

It is important to notice that both CER and our entropy minimization principle for word order are hypotheses on conditional entropies. However, there are some differences between our word order theory and CER/UID that are worth reviewing:

1. While in CER the target of conditioning is moving (Eq. 9), in our case the target is constant (Eq. 1).
2. While CER applies even to sequences that lack any order (namely to sequences of independent and identically distributed elements), our approach relies heavily on statistical dependencies among elements of the sequence (in sequence of independent elements, postponing the target will not help to predict it).
3. While the major statement of CER is a hypothesis which real language does not satisfy, our hypothesis is based on a basic truth, that "*conditioning reduces entropy*" in general, and this predicts the optimal placement of costly elements. The latter is not an opinion, conjecture or a hypothesis, but a mathematical fact. The same applies to the optimal placement according to dependency length minimization and the conflict between uncertainty minimization and dependency length minimization. Notice that the original "constant flow hypothesis" was also based on the fact that "conditioning reduces entropy" (recall the quote of Fenk-Oczlon (1989) in the introduction).
4. While CER and UID are presented as primary overarching principles, a conflict between principles is at the core of our theoretical approach. CER and UID are concerned about the trade-off "between redundancy and reduction" (Jaeger 2010), but only in the periphery of the argument. In contrast, the core of our theory defines word order as multiconstraint satisfaction problem where a principle of entropy minimization is in conflict with the principle of dependency length minimization (Section 2). Because of the secondary importance of distorting factor and conflicts between principles in CER and UID (Jaeger 2010, Levy & Jaeger 2007), these hypotheses are seen as incomplete (Ferrer-i-Cancho et al 2013a). Classic examples of linguistic theory where conflicts are at the core are G. K. Zipf's, whose view is based on conflicts between hearer and speaker needs (Zipf 1949), as well as R. Köhler's synergetics (Köhler, 1987; Köhler, 2005). A spin-off of Zipf's view are model of

Zipf's law for word frequencies that are based on the conflict between the minimization of the entropy of a vocabulary and the maximization of the mutual information between words and meanings (Ferrer-i-Cancho 2005, Ferrer-i-Cancho & Solé 2003).

5. UID and related hypotheses are concerned with a trade-off "between redundancy and reduction" (Jaeger 2010) that are symmetric terms: one is simply the opposite of the other. In contrast, our theory is concerned with a trade-off between two non-symmetric principles: uncertainty minimization and dependency length minimization. In standard information theory, there are no trade-offs between redundancy and reduction *per se* but trade-offs between different goals. In the terminology of information theory, goals define problems (Cover & Thomas 2006). Roughly speaking, the solution to the problem of transmission leads to increased redundancy while the solution to the problem of compression leads to reduction (Cover & Thomas 2006).
6. While CER and related hypotheses appear disconnected from optimization models of communication (Ferrer-i-Cancho 2017a), our approach to word order extends the domain of application of two fundamental principles of these models, i.e. entropy minimization and mutual information maximization. These optimization principles are relevant for their capacity to shed light on the origins of Zipf's law for word frequencies, the principle of contrast and a vocabulary learning (Ferrer-i-Cancho 2014).
7. While the connection between standard information theory and the UID/CER hypotheses is weak, various connections are already available between optimization models of communication and information theory through the problem of compression or model selection (Ferrer-i-Cancho 2017a).
8. CER and related hypotheses suffer from a psychological bias: they stem from a view where linguistic phenomena (word order in particular) are caused by absolute constraints of the human brain. In contrast, our framework is open to other causes: certain word order features may simple increase the survival over time of dominant word orders (Ferrer-i-Cancho 2015a). These other causes my exploit constraints of the human brain and then these constraints would not be the ultimate reason for the observed phenomena.

Points 5-7 are very important because a scientific theory should be more than a collection of disconnected ideas, as Bunge (2013) reminds us.

3.7. Ways to improve CER

Proponents of these hypotheses may argue that the disagreement with Hilberg's law does not reject their hypothesis because their true definition is that languages should tend towards to CER or UID (whether they actually reach Eq. 9 or Eq. 12 is irrelevant or secondary). However, such a disagreement implies two fundamental questions:

1. Why should languages tend towards CER or UID (Eq. 9 or Eq. 12)?
2. Why are languages not reaching CER or UID?

As we have seen in Section 3.5, the answer to Question 1 is unclear because a sufficiently developed theory is not available: the cost function that should accompany any claim on

optimization and other fundamental components of a real theory are missing in general. Such a theoretical understanding is also lacking for Question 2. With such incompleteness and under -specification (to the extent of being fully satisfied by i.i.d. processes, Section 3.3), it is rather straightforward to fit these hypotheses to a wide range of phenomena. Would the explanatory power of these theories remain constant if they were specified in greater detail? More importantly, are these hypotheses really necessary? We have argued that compression is an alternative hypothesis with higher predictive power (Section 3.4). A further challenge for their need will be provided below.

The disagreement between CER and Hilberg's law forces one to see CER as a tendency and this has the drawback of reducing the precision of the hypothesis. The constant entropy rate hypothesis can be relaxed with precision in a way that does not contradict that law. The goal is to avoid peaks of information by reducing the conditional entropy from the very beginning (Fenk-Oczlon 1989). The problem can be formalized as the minimization of the following cost function:

$$\max_i H(X_i|X_1, X_2, \ldots, X_{i-1}). \quad (24)$$

This is equivalent to minimizing $H(X_1)$ in a real linguistic sequence thanks to Hilberg's law (Appendix B). Put differently, peaks could be reduced simply with a bias to minimize the entropy of the initial elements assuming Hilberg's law.

Then, we do not need to invent a new principle: the minimization $H(X_1)$ can be seen as an example of a general principle of entropy minimization that has been applied to shed light on the origins of Zipf's law and that could be an indirect consequence of the minimization of L, namely compression (Ferrer-i-Cancho 2017a). Therefore, our attempt to improve CER adds another reason to not need CER (recall Section 4.3).

In sum, there is no objective reason to regard CER and related hypotheses as reference theories.

4. Discussion

4.1 The word order predicted by minimizing uncertainty

Section 2 provides a general argument for the placement of a target element of the sequence: it should be placed last to minimize its uncertainty. The argument is general but under-specified till we choose a target. We may choose the target between the head and its dependents or between the predicate and its arguments.

The word order problem has two symmetric solutions depending on the target:

- If the target is the head, uncertainty minimization predicts that the target should be placed last.
- If the targets are the non-head elements (the arguments), the prediction is that the non-head elements should be placed last, which implies that the head should be placed first.

Therefore our findings have implications for branching direction theory (Dyer 2011): left-branching minimizes the uncertainty of the head and right-branching minimizes the uncertainty of the dependents (complements/modifiers). Notice dependency length minimize-

ation can also produce consistent branching once the main verb has an extreme placement (Ferrer-i-Cancho 2008, Ferrer-i-Cancho 2015a).

Our general argument is that the most costly element should be the target, allowing one to break the initial symmetry between targets. Some costs that may determine the choice of the target will be presented below.

Let us consider the particular case of the ordering of the triple defined by S (subject), V (verb) and O (object). The following discussions assume that the verb is the head and is also valid for their semantic correlates: actor, action and patient (e.g., Goldin-Meadow et al 200; Langus and Nespor 2010). First, we consider that the target is either the head or the dependents. This yields that

- SOV and OSV are optimal when the target is the verbal constituent.
- VSO, VOS are optimal when the target is the non-verbal constituent.

The statistics of word orders suggest that verbs and their arguments are not symmetric targets. 89% of world languages that show a dominant word order do not put the verb first (Table 1). The *a priori* symmetry between verb initial and verb final languages can be broken in favour of verb final languages taking into account that verbs are harder to learn than nouns (Saxton 2010), which are the heads of the verbal complements (subject and object). For children, nouns are easier to learn than verbs (e.g., Imai et al 2008, Casas et al 2016), and actions (typically represented by verbs) are harder to pick up, encode and recall than objects (typically represented by nouns) (e.g., Gentner 1982, Gentner 2006, Imai et al 2005). Verb meanings are more difficult to extend than those of nouns (e.g., Imai et al 2005). Also, see McDonough et al 2011 for an overview of arguments on the difficulty of verbs as compared to nouns. Furthermore, arguments for the greater difficulty of verbs for infants can easily be extended to adults beyond the domain of learning. For these reasons, a communication system that aims at facilitating the processing and the learning of the most difficult items, i.e. verbs, may favour the strategy of minimizing the uncertainty about the verb (leading to verb last) over the strategy of minimizing the uncertainty about the nouns (leading to head first). The suitability of a verb last placement is supported by computer and eye-tracking experiments which indicate that the arguments that precede the verb help to predict it (Konieczny & Döring 2003).

The argument can be refined splitting dependents (arguments) into subjects and objects. By considering each of the elements as the target we get the optimal orderings (orderings that either minimize the uncertainty about the target or maximize the predictability about the target):

- The orders SOV and OSV are the optimal when the verb is the target.
- The orders SVO and VSO are the optimal when the object is the target.
- The orders VOS and OVS are the optimal when the subject is the target.

Again, the symmetry can be broken taking into account that verbs are harder to learn. In this case, SOV or OSV are expected. Interestingly, SOV is a verb final order that

- Covers 43.3 % of dominant orders in languages (65.3% in families) according to Table 1.
- Is hypothesized to prevail in early stages of evolution of spoken (Gell-Mann & Ruhlen 2011, Newmeyer 2000, Givon 1979) and signed languages (Sandler 2005, Fisher 1975).

- Is recovered in experiments of gestural communication (Goldin-Meadow et al 2008, Langus & Nespor 2010).
- Appears in *in silico* experiments under pressure to maximize predictability (Reali & Christiansen 2009).

The problem is that our optimality argument also predicts OSV, that covers only 0.4% of dominant orders in languages (0.3% in families) according to Table 1.

Our argument predicting verb final placement can be refined to yield only SOV in four ways:

- Assuming a hierarchy of multiple targets, namely the verb is the main target and the object is a secondary target. That would give the subject is placed first for not being a target and that the verb is put last for being the main target. Then SOV would follow. The idea is reminiscent of the standard approach to word order in typology that consists of assuming pairwise word order preferences (Cysouw 2008).
- Postulating an agent or subject bias that determines that the subject is placed first, the so-called agent first pragmatic rule (Schouwstra & de Swart 2014).
- As more frequent elements are put first (due to some psychological preference), the subject would be put first (Fenk-Oczlon, 1989). The argument is interesting for connecting the frequency effects that are used to justify the minimization of entropy in optimization models of communication (Ferrer-i-Cancho 2017a) with word order and thus this option has the potential to yield a compact theory of language with respect to the two preceding alternatives.
- An indirect effect of a hidden attraction towards SVO (see Section 4.6).

These possibilities should be the subject of further research.

Let us move to the problem of the optimal placement of dependents within the nominal constituents. For simplicity, we consider that the target are the head or the dependents. This yields that

- Placing the dependents before the nominal head is optimal when the target is the nominal head.
- Placing the dependents after the nominal head is optimal when the target are the dependents.

Thus the principle of uncertainty minimization could contribute to explain why no language consistently splits its noun phrases around a central nominal pivot, with half of the modifiers to the left and half to the right, as expected from the principle of dependency length minimization (Ferrer-i-Cancho 2015a). Support for this possibility comes from the complex interaction between dependency length minimization and other factor at short ranges (Gulordava et al 2015). However, uncertainty minimization does not need to be the only reason for this phenomenon: we have argued that the actual placement of modifiers could be the result of competition between dominant orders struggleing for survival (Ferrer-i-Cancho 2015a). However, believing that predictability maximization is the only reason why dependents of nominal heads tend to be put at one side of their head is theoretically naïve, because the principle of dependency length minimizeation at global scale predicts that those dependents are placed before the nominal head in verb final languages and after the nominal head in verb initial languages (Ferrer-i-Cancho 2008; Ferrer-i-Cancho 2015a). Therefore, dependency length minimization and uncertainty minimization can collaborate to yield an

asymmetric placement of dependents at short ranges and may explain the origins of consistent branching in languages.

In Section 2, we have provided some mathematical results to understand the placement of heads in single head structures. More realistic scenarios should be investigated. Our theoretical framework should be extended to the case of the multiple head structures that are found in of complex sentences.

4.2 The optimality of word orders

Integrating the arguments of Section 4.1 with the predictions of dependency length minimization one obtains the following optimality map:

- SOV and OSV are optimal according to the minimization of the uncertainty about the verb, 43.7% of dominant orders in languages (65.6% in families) according to Table 1.
- SVO and OVS are optimal according to the minimization of dependency lengths, 41.1% of dominant orders in languages (15.8% in families) according to Table 1.
- VSO, VOS are optimal according to the minimization of the uncertainty about the non-verbal constituents, 13.9% of dominant orders in languages (15.9% in families) according to Table 1.

More precise optimality arguments can be built splitting the non-verbal constituents into subjects and objects or assuming a hierarchy or targets as explained in Section 4.1.

Our findings on the optimality of word orders and on the properties of the adaptive landscapes that optimality principles define (Section 2.3) are particularly relevant for researchers who have "no evidence that SOV, SVO, or any other word order confers any selective advantage in evolution" (Gell-Mann & Ruhlen 2011). Interpreting the diversity of word orders (Table 1) or the rather large proportion of languages lacking a dominant (about 2.4% of languages according to Table 1 but 13.7% according to Dryer 2013) as an absence of principles or adaptive value is theoretically naïve: it may simply reflect the difficulty for complying with incompatible constraints (Ferrer-i-Cancho 2014). The diversity of word orders would not be a manifestation of arbitrariness but an inevitable consequence of a multiconstraint optimization problem where the availability of word orders is constrained by a word order permutation ring.

4.3. Word order conflicts

The simple optimality map presented above clearly shows that there are at least two conflicts between principles: one internal to uncertainty minimization, i.e. the optimal order depends on the target of uncertainty minimization and another external, between dependency length minimization and uncertainty minimization.

The external conflict is due to the fact that the principle of dependency length minimization predicts that the head should be placed at the center while the principle of uncertainty minimization predicts that it should be placed at one of the ends (Section 2).

It is worth considering the interplay between dependency length and uncertainty minimization with the head as the target as the head moves from the beginning of the utterance to the end. Postponing the head minimizes its uncertainty but the cost of dependency lengths will depend on its placement (Ferrrer-i-Cancho 2015a). The cost of dependency lengths decreases as the head is postponed before the center of the sequence. From then on, dependency length costs will increase as the head is postponed. Put in technical

terms, the landscape of dependency lengths as a function of the position of the head is quasi-convex for the case of a single head (Ferrer-i-Cancho 2015a). Put differently, dependency length minimization and postponing the head to minimize its uncertainty are 'allies' during the first half of the sequence and 'enemies' in the second half. In contrast, dependency length minimization and bringing the head forward to minimize the uncertainty about dependents are 'enemies' in the first half of the sequences and 'allies' in the second half.

Notice that the external conflict above arises in the context of the optimal placement of one head and its dependents. The problem is more complex if one considers further levels of organization: e.g., from the head verb to the heads of its complements and then from the heads of these complements to their dependents. In Section 4.1, we have shown that the principles can be in conflict at the level of the optimal placement of the verb but collaborate at the level of the placement of dependents of nominal heads.

G. Heyer and A. Mehler (2009) made us notice that the conflict between predictability (uncertainty) and dependency length minimization could be seen as a conflict between long term memory, that stores the probabilistic information underlying the definition of uncertainty or predictability, and online memory, where pressure to minimize dependency lengths originates. To Heyer & Mehler, conflicts between principles are reminiscent of conflicts between time cost and memory cost in algorithmic theory (Cormen et al 2009).

Since we have argued that any word order can be optimal *a priori* (e.g., any placement of V with respect to S and O is optimal for some reason), and that different orders are in conflict (e.g., putting the V at one of the two ends is against dependency length minimization), it is tempting to conclude that "*anything goes*" (any word order is valid). However, this is not our view. We believe that word order is determined at least by the experimental or ecological conditions (and previous history as we will see below). Examples of ecological conditions are the proportion of L2 speakers and the proportion of deaf individuals of the community. Examples of experimental conditions are the length the sequences to be uttered or gestured or the amount of pressure to maximize predictability as we will see immediately.

If additional pressure for predictability maximization is added *in silico*, experiments show that a verb final language (SOV) emerges, as expected from the theoretical arguments above (see Ferrer-i-Cancho 2014 for further details about this experiment). In general, verb final or verb initial languages are more likely in simpler sequences while verb medial languages are expected in more complex sequences (Ferrer-i-Cancho 2014). The case of verb initial languages could be special because they originate from verb medial languages (Gell-Mann & Ruhlen 2011) and therefore their sequential complexity does not need to be as low as that of the verb final languages that are typically found at early stages of word order evolution. Finally, recall that on top of theoretical arguments indicating that any word order can be optimal for some reason, we have added a further factor that could break the symmetry between word orders, such as the higher intrinsic difficulty of certain words, that would increase the chance that they are chosen as targets, or the recency or frequency effects, by which subjects would be put first (Section 4.1).

The explanatory power and potential of word order conflicts is illustrated by their capacity to shed light on the origins of word order diversity, on the phenomenon of languages lacking a dominant word order, on word order reversions in historical developments, and on alternative orders with a verb at the center (Ferrer-i-Cancho 2014). Furthermore, they also allow one to understand why real sentences do not achieve the minimum sum of dependency lengths that is expected if dependency length minimization was the only principle (Ferrer-i-Cancho 2004). Subsection 4.4 provides an updated account on word order diversity.

4.4. Word order diversity

Word order diversity can be interpreted in two ways: externally, comparing the variation of dominant word orders across languages, and internally, looking at the word orders that are adopted within a language.

Concerning external diversity, the optimality map presented in Section 4.2 shows that all verbal placement are optimal for some reason *a priori*. Adding that word orders are in conflict one expects that there is no single winner. Indeed, the six possible orders of subject, verb and object are found in languages (Table 1). We believe that the principles of word order and their conflicts have the potential to explain word order diversity (in combination with other components such as the word order permutation ring that will be reviewed later on). Although all verbal placements are optimal according to some word order principle, word order could be biased towards verb medial due to the increasing pressure for dependency length minimization as linguistic complexity increases (Ferrer-i-Cancho 2014), and also towards verb final due the higher complexity of verbs (McDonough 2011, Gentner 2006). We do not mean that the counts of word orders are unequivocally indicative of the degree of optimality of a word order because of word order evolution. A full explanation of the diversity of dominant word orders requires acknowledgement of the fact that word order evolution is a *path dependent process* where the initial word order is critical. Section 4.4 sheds some light on how the bulk of word order diversity can be generated, step by step.

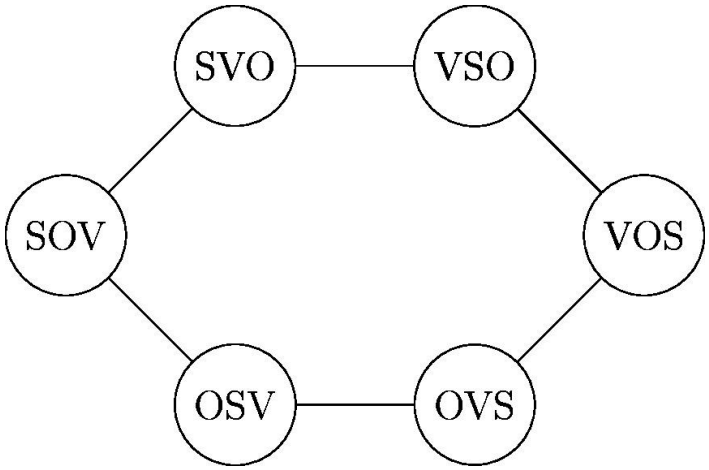

Figure 3. The permutation ring defined by all the 6 possible orderings of subject (S), verb (V), and object (O). Two orderings are connected if one leads the other after swapping two adjacent elements.

Internal word order diversity has been hypothesized to be constrained by a word order permutation ring that determines how a new word order can be generated from another (Ferrer-i-Cancho 2016). The *a priori* probability of a variant is hypothesized to be a monotonically decreasing function of the distance between the variant of the dominant order in a permutation ring (Fig. 3). The word order permutation ring beats the standard model of typology in explaining the composition of the couples of primary alternating word orders (Ferrer-i-Cancho 2016).

The power of the permutation ring to explain the evolution of the dominant word order will be reviewed in Section 4.5.

4.5 Word order evolution

Here we revisit the framework for word order evolution that has been presented in a series of articles for the evolution of the dominant ordering of subject, verb and object in languages (Ferrer-i-Cancho 2008, Ferrer-i-Cancho 2014, Ferrer-i-Cancho 2015a, Ferrer-i-Cancho 2016a). This framework has two major components: an early or initial order and transitions between orders.

Converging evidence supports SOV (or its semantic correlate actor, patient, action) as an initial or early stage in evolution (Gell-Mann & Ruhlen 2011, Langus and Nespor 2010, Pagel 2009, Goldin-Meadow et al 2008, Sandler et al 2005, Newmeyer 2000, Givon 1979, Fisher 1975). The early or initial word order is determined by conditions that facilitate the dominance of maximization about the predictability of the verb over either (1) dependency length minimization or (2) the minimization of the uncertainty about the other components. We have argued that the victory of the maximization of the predictability of the verb is likely to be determined by a series of factors:

- *The length of the sequences*. At early stages, linguistic sequences (of words or gestures) are expected to be shorter (Ferrer-i-Cancho 2014). This is easy to see in the extreme case of sequences of length two: the placement of the head is irrelevant for dependency length minimization but to minimize the uncertainty about the head, the verb should appear last. The size of the sequence where dependency length minimization can be neglected may be determined by the capacity of short term memory, i.e. about four elements (Cowan 2000).
- *Morphology*. Case marking facilitates the processing of SOV structures (Lupyan & Christiansen 2002).

In section 4.1 we have provide arguments for a preference for SOV over OSV.

Transitions are hypothesized to be constrained by the structure of the space of possible transitions and conditions that help one principle to dominate in the struggle between dependency length minimization and uncertainty minimization (or predictability maximize-ation). The space of possible transitions has been hypothesized to be determined by the minimum number of swaps of adjacent constituents that are needed to reach a word order from the current word order (Ferrer-i-Cancho 2008, 2015, 2016a). The *a priori* probability of a transition is hypothesized to be a monotonically decreasing function of the distance between the source order and the destination order in a permutation ring (Fig. 3). The transition from SOV to SVO is more likely *a priori* than the transition from SOV to OVS (the former requires only one swap; the latter requires two swaps). This is known as the *word order permutation ring hypothesis*. Further conditions may operate on this permutation ring, possibly distorting the predictions that can be made if the ring was the only constraint.

We will use the main path for word order evolution namely the transition from SOV to SVO and the transition from SVO to VSO/VOS to illustrate how these conditions apply (Gell-Mann & Ruhlen 2011). A striking feature of these transitions is that they involve source and destination orders that are adjacent or almost adjacent in the word order permutation ring (Fig. 3).

Table 3
Predictions on the most likely transition from SOV. Yes and No indicate presence or absence of the feature indicated in header of the corresponding column.

Word order permutation ring	Dependency length minimization	Most likely destination
Yes	No	SVO and OSV
No	Yes	SVO and OVS
Yes	Yes	SVO

According to the permutation ring, the most likely transitions from SOV are SVO and OSV. However, the typical destination from SOV is SVO (Gell-Mann & Ruhlen 2011). We hypothesize that the tie is broken in favour of SVO by the principle of dependency length minimization that predicts that the head is placed at the center and factors that may favour SOV over OSV exposed in Section 4.1. However, this opens a new problem since there are two orders with the verb at the center, i.e. SVO and OVS. Interestingly, OVS is farther away from SOV in the word order permutation ring. Thus, we conclude that SVO is the most likely transition. A summary of the argument is provided in Table 3.
Since we have argued above that a main raison for SOV to be the initial or early stage is the victory of the minimization of the uncertainty about the head over other principles, it is reasonable to think that SOV will be abandoned when the sequence complexity (sequence length) increases. That increase facilitates the victory of dependency length minimization (Ferrer-i-Cancho 2014). The chances of success of the transition increase under further conditions that prevent regression to SOV:

- SVO languages that put adjectives after the noun are more likely to stabilize because this relative placement of adjectives is neutral for SVO but inconvenient for SOV from the perspective of dependency length minimization (Ferrer-i-Cancho 2015). Interestingly, the number of SVO language with that peculiar placement of adjectives is above chance.
- Case marking facilitates the learning of SOV structures (Lupyan & Christiansen, 2002). The need of case marking for a more efficient processing of SOV is supported by Greenberg's universal 41, stating that SOV languages almost always have case marking (Greenberg 1963). Thus, regression to SOV could be harder from SVO languages lacking case marking. In turn, as languages with a high proportion of L2 speakers tend to lose case marking (Bentz & Winter 2013), the proportion of L2 speakers is likely to be one of the factors that determines the stabilization of a dominant SVO order, expanding the predictions of the Linguistic Niche Hypothesis (Dale & Lupyan 2012) to the domain of word order.

Once a language is SVO why should it become VSO/VOS? Again the permutation ring and certain conditions can explain the transition. Once a system has reached SVO the permutation ring offers two main possibilities: to come back to SOV or to move forward towards VSO (or VOS with less probability). Adding pressure to minimize the uncertainty about the nominal heads then VSO appears as the most like solution. VOS is among the second best solution for being only one step farther in the permutation ring with respect to SVO and putting the verb first optimally as VSO. A summary of the argument is provided in Table 4.

Table 4
Predictions on the most likely transition from SVO. Yes and No indicate presence or absence of the feature indicated in header of the corresponding column

Word order permutation ring	Minimization of the uncertainty about the nominal constituents	Most likely destination
Yes	No	SOV and VSO
No	Yes	VSO and VOS
Yes	Yes	VSO

The likelihood of the transition to VSO/VOS is increased by adaptations in SVO that prevent regression to SOV that preadapt SVO for VSO/VOS: placing adjectives after the nominal head is convenient for VSO/VOS (Ferrer-i-Cancho 2015a) but not for SOV.

The scenario of word evolution presented above strongly suggests that word order evolution is a *path dependent process* (Ferrer-i-Cancho 2016a, Ferrer-i-Cancho 2015a, Dunn et al 2011).

It is worth noting that the number of languages (or the number of families) with a certain dominant word order decreases as one moves in the permutation ring in a clock-wise sense (Table 1, Fig. 3). It has been argued that word order evolution may not have reached a steady state or equilibrium (Gell-Mann & Ruhlen 2011). With our arguments above, we do not mean that SOV or SVO may not exist anymore, as dominant word orders, in the future (if no new languages were created). From our arguments above it follows that the dominance of dependency length minimization (SVO) is easier to achieve but harder to abandon, because of the time elapsed since the birth of these languages and the ecological conditions of many linguistic communities. As for the former, the length and the complexity of sentences has probably increased over time (it is unlikely that in the birth of a language from scratch long sentences are used). The adoption of a writing system or access to higher education are relevant ecological variables for this growth as they facilitate the creation of longer and more complex sentences where dependency length minimization is critical. If these conditions remain, there is no reason to believe that in the future languages will tend to go back to extremely short sentences where SOV is easier to handle. Due to this fundamental pressure for dependency length minimization and the importance of the verb as a target, transitions beyond SVO (and SOV) could be secondary.

4.6. Word order diversity in the light of evolution

The fact that SOV and SVO cover the overwhelming majority of dominant orders in languages (Table 1) could result from a tree-fold combination

1. The initial preference for SOV a word permutation ring constraining possible moves.
2. A bias to reduce the uncertainty of the head.
3. Dependency length minimization.

In this view, the initial preference for SOV and the word permutation ring are crucial to understand the evolutionary history of word order as a path dependent process (Ferrer-i-Cancho 2016a, Ferrer-i-Cancho 2015a, Dunn et al 2011). Following the three-fold hypothesis,

we revise the frequency of six possible orders classifying them according to the position of the verb:

- SVO, OVS (central verb) is a compromise between dependency length minimizeation and the minimization of the uncertainty of the head: the placement of the verb at the center is optimal according to dependency length minimization and in-between its best placement (last) and its worst placement (first) according to minimization of the uncertainty of the head (Section 2.2). The low frequency of OVS could be explained by the evolutionary history.
- SOV, OSV (verb last) satisfies the optimality of the principle of postponing the head. The low frequency of OSV could be explained by three facts 1): the initial preference for SOV 2) an attraction towards SVO due to pressure for dependency length minimization 3) OSV is farther from SVO than SOV according to the permutation ring.
- VSO/VOS (verb first) should have lower frequency because placing the verb first is the worst case for both principles. Additionally, they might be under represented due to the evolutionary history.

4.7. A general theory of word order and beyond

In this article, we have made one step forward to building a coherent theory of word order. The major components of the theory are

- A subtheory of word order from the dimension of dependency length minimization (Ferrer-i-Cancho 2008, 2015a,b).
- A subtheory of word order from the dimension of uncertainty minimization or predictability maximization (this article).
- An integrated subtheory of word order that explains how these principles interact: their conflict and the factors that determine the dominance of one over the other (this article and Ferrer-i-Cancho 2014).
- A subtheory of word order variation, both internal, i.e. within a language (Ferrer-i-Cancho 2016a) and also externally, i.e. across languages (this article and Ferrer-i-Cancho 2014).
- A subtheory of word order evolution (Ferrer-i-Cancho 2014, Ferrer-i-Cancho 2016a).

These subtheories are not collections of disconnected ideas. The subtheory of word order evolution relies on the assumption that word order evolution operates on constraints on word order variation (Ferrer-i-Cancho 2016a). These subtheories are unified through the word order permutation ring. In turn, this ring and the principle of dependency length minimization stem from a general principle of distance minimization (Ferrer-i-Cancho 2016a). Beyond word order the theory is connected with the theory of Zipf's law for word frequencies: both the minimization of uncertainty and predictability maximization follow from a general principle of entropy minimization and mutual information maximization that can be applied to shed light on the origins of Zipf's law for word frequencies.

The theory is articulated by key traversal concepts:

- *Intrinsic conflicts*: word order principles are intrinsically in conflict as we have seen in this article. These conflicts may underlie the diversity of dominant word orders found

across languages as well as the lack of a dominant order in certain languages (Ferrer-i-Cancho 2014).

- *Coexistence*: different word order principles can dominate simultaneously in a language. For instance, SOV languages suggest that the ordering of the triple is determined by the need of minimizing the uncertainty about the verb. Besides, the tendency of these languages to put adjectives before nouns (or auxiliaries after verbs) is a prediction of the principle of dependency length minimization (Ferrer-i-Cancho 2008, Ferrer-i-Cancho 2015a). The converse may happen to SVO languages where the placement of the object is optimal with respect to dependency length minimization but the placement of adjectives before nouns in the nominal constituents could be driven by the principle of minimization of the uncertainty about the head. A beautiful example of coexistence of principles is provided by languages that are not SVO but have SVO as alternative order (Greenberg 1963). There SVO could arise to compensate for a suboptimal choice from the perspective of dependency length minimization (Ferrer-i-Cancho 2015a).
- *Cooperation*: coexistence makes emphasis on the diversity of word orders that may result from conflicting constraints, e.g., a couple of primary alternating orders instead of one (Ferrer-i-Cancho 2016a). The idea of cooperation emphasizes the possibility that two word orders interact to produce the same word order pattern. Take the principle of minimization of the predictability of the head versus dependency length minimization. Suppose that the former beats the latter for the placement of the verb that is then put last. In this context, assuming that the relative placement of adjectives has to be consistent for both the subject and the argument, it follows that dependency length minimization will lead to adjectives before nouns. This is also expected by the principle of head uncertainty minimization for the particular case of nominal heads (Ferrer-i-Cancho 2015a).
- *Neutrality*: certain placements may have literally no clear advantage for the brain with respect to a certain word order principle, e.g., *a priori* adjectives can either follow or precede nouns in SVO languages according to the principle of dependency length minimization (Ferrer-i-Cancho 2015a). Functional pressures do not imply that some orders are better than others in all cases.
- *Word order survival or the recipient of benefits*: certain placements may not be advantageous for the brain with respect to at least one word order principle (they can be neutral as we have seen above). Instead, they could be explained as a result of competition for survival among dominant word orders. For instance, dominant SVO orders may increase their survival by choosing a relative placement of adjectives with respect to nouns that is inconvenient for SOV languages (Ferrer-i-Cancho 2015a). A very important point is that then a certain placement would not be explained by its benefit for the brain but for its benefits for the survival of a word order (Ferrer-i-Cancho 2015b). This represents a radical shift of perspective with respect to the exclusive focus of word order research in cognitive science on benefits for the brain. This hypothesis should be evaluated considering an alternative hypothesis, namely that such a relative placement might be due to the coexistence of a principle to reduce the uncertainty about the nominal heads, that predicts that the nominal head should be put first. However, this alternative is less likely given that heads are normally more costly and thus should be the target.

- *Conditional word order biases or Kauffman's adjacent possible*: word order variation and word order change can be highly determined by the current state of the system (its dominant word order), overriding prior unconditional biases (Ferrer-i-Cancho 2016a).
- Word order evolution as a *path dependent process*: the next steps of word order evolution are determined by history (Dun et al 2011). For instance, running away from the attraction of SOV preadapts SVO languages to become VSO/VOS languages (Ferrer-i-Cancho 2016a). Again, interpreting any word order configuration as arising exclusively from absolute brain costs, as it is customary in cognitive science, can be misguiding. History matters.
- *Symmetry breaking*: to understand word order it is important to understand how the tie between alternative orders could be broken. Some examples are the following:
    - The symmetry between the minimization of the uncertainty about the verb and the uncertainty about its arguments (the nominal constituents) is broken by the fact that verbs are harder to learn.
    - The relative placement of adjectives in SVO languages reviewed above.
    - When the current state is SOV, pressure for dependency length minimizeation predicts two as orders as the most likely: SVO and its symmetric OVS. The word permutation ring hypothesis breaks the symmetry towards SVO.
    - The conflict between dependency length minimization and uncertainty minimization could be broken by the length or the scale in favor of the former. The conflict between principles needs at least three elements and uncertainty minimization starts operating with just two elements (Section 2.2). Then, the former would tend to dominate in longer sequences or at higher scales while the latter would tend to dominate in shorter sequences or at short ranges (Ferrer-i-Cancho 2014, Ferrer-i-Cancho 2015a). Such a division of labour is linked to the concept of coexistence.
- *The shape of the adaptive landscape*: the adaptive landscape of dependency lengths for the case of a single head is quasi-convex under some general assumptions (Ferrer-i-Cancho 2015a). Further research should be carried out to determine if it is also convex and to shed light on the shape of the complex landscape that results when uncertainty minimization is integrated.

The case of Mandarin Chinese can help us to see how the concepts above can be applied. That language has SVO as dominant order and tends to put adjectives before nouns (Dryer & Haspelmath 2013). As we have seen above (*Word order survival or the recipient of benefits*), the dominance and the survival of SVO is enhanced by placing adjectives in a relative position with respect to nouns that is inconvenient for SOV, namely, after nouns. This is not the case of Mandarin Chinese and that may explain the *coexistence* of SOV and SVO in that language (Gao 2008).

The concept of *intrinsic conflicts* and the concept of *collaboration* can be seen as instances of Morin's (1990) dialogic principle, according two principles (dependency length minimization and uncertainty minimization in our case) could be at the same time antagonistic and complementary. This is an example of how the philosophical and epistemeological approach of "general complexity" can be unified with the mainly scientific and methodological approach of "restricted complexity" (Malaina 2015).

Our theoretical results on the conflict between word order principles provide an answer to the question of the "relative roles of worktye ing memory principles", i.e. dependency length minimization in our framework, "and principles of information theory

accounts of sentence processing such as surprisal", i.e. uncertainty minimization/predictability maximization in our setup (Lewis et al 2006). Notice that we do not view information theoretic principles as necessarily external to working memory in our approach.

Early in the article (Section 2), we justified the principles of word order based on extensions of principles that are defined over individual words in optimization principles of communication (Ferrer-i-Cancho 2017a). These lexical principles probably apply beyond lexical elements and then support the "fast content-addressed access to item information" involved in processing of sequences (Lewis et al 2006). It is time to close the circle in the opposite direction. The principle of dependency length minimizeation is a principle of word order that has words as units. Length could be measured with more precision in syllables or phonemes (Ferrer-i-Cancho 2015b). In that way, the length of a dependency would be a function of the length of the words defining the dependency and that of the words in-between. Therefore, word lengths should be minimized to minimize dependency lengths (Ferrer-i-Cancho 2017c). Put differently, dependency length minimization predicts the principle of compression, linking dependency length minimization with the origins of Zipf's law for word frequencies (Ferrer-i-Cancho 2016b, Ferrer-i-Cancho 2017a), Zipf's law of abbreviation (Ferrer-i-Cancho et al 2013b) and Menzerath's law (Gustison et al 2016). Therefore, dependency length minimization predicts reduction, a phenomenon that has been used to justify the uniform information density and related hypotheses (Section 3.4). The need for the need for uniform information density and similar hypotheses as independent standalone hypotheses is seriously challenged. However, we do not mean that dependency length is the only reason for compression. For instance, in small sequences where dependency length minimization is irrelevant or can be neglected (Section 4.5), compression *per se* still matters.

We hope that our sketch of a general theory of word order and beyond stimulates further research. Notice that the scope of the main theoretical results presented above goes beyond linguistics. Uncertainty minimization makes predictions about the optimal placement of target elements for any sequence *a priori*. Dependency length minimiztion requires that there is some structure, e.g., there must be a hub element, the equivalent of a head in a linguistic context. For these reasons our results could be applied to genomic sequences (Searls 1992) or animal behavior sequences (Kershenbaum et al 2016).

APPENDIX A

Property

Suppose that g_H and g_I are two functions whose domain and co-domain are real numbers: g_H is a strictly monotonically increasing function while g_I is a strictly monotonically decreasing function. One has

$$g_H[H(Y|X_1, X_2,..., X_{i-1})] \geq g_H[H(Y|X_1, X_2,..., X_i)] \quad \text{(A.1)}$$

for $i \geq 1$, and

$$g_I[I(Y|X_1, X_2,..., X_{i-1})] \leq g_I[I(Y|X_1, X_2,..., X_i)] \quad \text{(A.2)}$$

for $i \geq 2$.

Proof:

One has that

$$H(Y) \geq H(Y|X_1). \quad \text{(A.3)}$$

with equality if and only if Y and X_1 are independent (Theorem 2.6.5, p. 29, Cover & Thomas 2006). This is obtained by a straightforward application of the fact that "*conditioning reduces entropy*" (in general) or that "*information cannot hurt*" (Cover & Thomas 2006, p. 29).

We would like to prove the general case

$$H(Y|X_1, X_2,..., X_{i-1}) \geq H(Y|X_1, X_2,..., X_i) \quad \text{(A.4)}$$

for $i > 1$ (the case $i = 1$ corresponds to Eq. A.3). The conditional mutual information between Y and X_i knowing $X_1,\ldots X_{i-1}$ is

$$I(Y; X_i|X_1, X_2,..., X_{i-1}) = H(Y|X_1, X_2,..., X_{i-1}) - H(Y|X_1, X_2,..., X_i) \quad \text{(A.5)}$$

Then Eq. A.4 is equivalent to

$$I(Y; X_i|X_1, X_2,..., X_{i-1}) \geq 0 \quad \text{(A.6)}$$

Lemma 3.1 of Wyner (1978) warrants that the inequalities in A.4 and A.6 hold with equality if and only if X_i, $X_1,\ldots X_{i-1}$ and Y define a Markov chain.

The properties of g_H give Eq. A.1. A parallel conclusion can be reached for $I(Y; X_1, ..., X_i)$.

Multiplying by -1 in Eq. A.6 one gets

$$-H(Y|X_1, X_2,..., X_{i-1}) \leq -H(Y|X_1, X_2,..., X_i), \quad \text{(A.7)}$$

Adding $H(Y)$ one gets

$$H(Y) - H(Y|X_1, X_2,..., X_{i-1}) \leq H(Y) - H(Y|X_1, X_2,..., X_i) \quad \text{(A.8)}$$

and finally Eq. A.2 for $i \geq 2$ (notice that $I(Y; X_1, X_2,..., X_{i-1})$ is not defined when $i = 1$) as we wanted to prove.

The property above allows one to conclude easily that placing the target last is optimal, namely

$$n \in \text{argmin}_{1 \leq i \leq m} g_H[H(Y|X_1, X_2,..., X_i)] \quad \text{(A.9)}$$

and

$$n \in \text{argmax}_{2 \leq i \leq m} g_I[I(Y|X_1, X_2,..., X_i)], \quad \text{(A.10)}$$

although not necessarily the only optimum. Therefore, in the absence of any further information, placing the target last is the most conservative strategy and thus it is the optimal in general.

APPENDIX B

It is easy to show that

$$\max_{1 \leq i} H(X_i|X_1, X_2,..., X_{i-1}) = H(X_1), \quad \text{(B.1)}$$

assuming Zipf's law. Notice that Hilberg's law (Eq. 11) implies $a = H(X_1)$ and also that Eq. B.1 is approximately equivalent to

$$\max_{1 \leq i} a i^{-\gamma} = a. \quad \text{(B.2)}$$

as γ is strictly positive.

Acknowledgements

The present article is an evolved version of some of the arguments of the unpublished manuscript "Optimal placement of heads: a conflict between predictability and memory". The major result of the article was the conflict between the principle dependency length minimization and predictability maximization. The article was submitted for publication in September 2009 after being presented as "Memory versus predictability in syntactic dependencies" in the Kickoff Meeting "Linguistic Networks" (Bielefeld University, Germany) in June 5, 2009. We thank the participants of the Kickoff Meeting, specially G. Heyer and A. Mehler for valuable discussions. A more advanced version was presented in 2011 as "Word order as a constraint satisfaction problem: A mathematical approach." In the workshop "Complexity in Language: Developmental and Evolutionary Perspectives" (Collegium de Lyon, May 23 - 24).

Since 2009, at least R. Levy, F. Jaeger, E. Gibson, S. Piantadosi and R. Futrell have had access to different versions of the unpublished manuscript. Evolved versions of various components of the unpublished manuscript have already appeared (Ferrer-i-Cancho 2014, 2015a).

For the present article, we thank C. Bent and S. Semple for their careful revision. We are also grateful to Ł. Dębowski, G. Fenk-Oczlon, F. Moscoso del Prado Martín, M. Wang, and Eric Wheeler for helpful comments and discussions, to S. Wichmann for pointing us to Hammarström's work, and to Y. N. Kenett for pointing us to Cowan's work. This research was funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and also the APCOM project (TIN2014-57226-P) from MINECO (Ministerio de Economia y Competitividad).

REFERENCES

Akmajian, A., Demers, R.A., Farmer, A.K. & Harnish, R.M. (2001). *Linguistics: an introduction to language and communication.* 5th edition. Cambridge, MA: MIT Press.

Altmann, E. A., Cristadoro, G. & Esposti, M. D. (2012). On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences USA, 109:11582–11587, 2012.*

Alvarez-Lacalle, E., Dorow, B. & Eckmann, J.-P. & Moses, E. (2006). Hierarchical structures induce longrange dynamical correlations in written texts. *Proceedings of the National Academy of Sciences USA 103, 7956–7961.*

Aylett, M. & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech 47(1), 31-56.*

Bentz, C. & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change 3, 1-27.*

Bentz, C., Verkerk, A., Kiela, D., Hill, F., Buttery, P. (2015) Adaptive communication: languages with more non-native speakers tend to have fewer word forms. *PLoS ONE 10(6), e0128254.*

Bunge, M. (1984). What is pseudocience? *The Skeptical Inquirer 9, 36-46.*

Bunge, M. (2013). *La ciencia. Su método y su filosofía.* Pamplona: Laetoli.

Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-Fernández, A. & Baixeries, J. (2016). The polysemy of the words that children learn over time. http://arxiv.org/abs/1611.08807

Cormen, T. H., Leiserson, C.E., Rivest, R.L. & Stein, C. (2009*). Introduction to Algorithms* (3rd ed.). Cambrige, MA: MIT Press.

Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory*, 2nd edition. Hoboken, NJ: Wiley.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24, 87-185.*

Cysouw, M. (2008). Linear order as a predictor of word order regularities. A reply to Ferrer-i-Cancho (2008). *Advances in Complex Systems 11 (3), 415-420.*

Dale, R. & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems 15, 1150017.*

Dębowski, Ł. (2015). The relaxed Hilberg conjecture: a review and new experimental support. *Journal of Quantitative Linguistics 22 (4), 311–337.*

Dryer, M. (2009). The branching direction theory of word order correlations revisited. In: S. Scalise, E. Magni, and A. Bisetto (Eds.), *Universals of Language Today: 185–207*. Berlin: Springer.

Dryer, M.S. (2013). Order of subject, object and verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info/chapter/81, Accessed on 2017-04-12).

Dryer, M. S. & Haspelmath, M. (eds.) (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info, Accessed on 2017-06-15).

Dunn, M., Greenhill, S. J., Levinson, S. C. & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature 473 (79), 79-82.*

Ebeling, W. & Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters, 26(4), 241-246.*

Evans, N. & Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32:429492.*

Fenk, A. & Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Pshychologie XXVII (3), 400–414.*

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27, 517-556.*

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: J. Bybee and P. Hopper (eds.), *Frequency and the emergence of linguistic structure: 431-448.* Amsterdam: John Benjamins

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

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

Ferrer-i-Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems 11 (3), 394-414.*

Ferrer-i-Cancho, R. (2014). Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In: THE EVOLUTION OF LANGUAGE - Proceedings of the 10th International Conference (EVOLANG10), Cartmill, E. A., Roberts, S., Lyn, H. & Cornish, H. (eds.). Evolution of Language Conference (Evolang 2014). Vienna, Austria, April 14-17. pp. 66-73.

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

Ferrer-i-Cancho, R. (2016a). Kauffman's adjacent possible in word order evolution. In: S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher & T. Verhoef (eds.). *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).* New Orleans, USA, March 21-24.

Ferrer-i-Cancho, R. (2016b). Compression and the origins of Zipf's law for word frequencies. *Complexity, 21 (S2), 409-411.*

Ferrer-i-Cancho, R. (2017a). Optimization models of natural communication. *Journal of Quantitative Linguistics,* in press. http://arxiv.org/abs/1412.2486

Ferrer-i-Cancho, R. (2017b). The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics 36, 1-16.*

Ferrer-i-Cancho, R. (2017c). A commentary on "The now-or-never bottleneck: a fundamental constraint on language", by Christiansen and Chater (2016). *Glottometrics 38, 116-120.*

Ferrer-i-Cancho, R., Dębowski, Ł. & Moscoso del Prado Martín, F. (2013a). Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics, L07001.*

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

Ferrer-i-Cancho, R. & Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics, L12002.*

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

Ferrer-i-Cancho, R., Bentz, C. & Seguin, C. (2015). Compression and the origins of Zipf's law of abbreviation. http://arxiv.org/abs/1504.04884

Fisher, S. (1975). Influences on word order change in American sign language. In: Li, C.N. (ed.). *Word order and word order change: 1-25.* University of Texas, Austin.

Gao, Q. (2008). Word order in Mandarin: reading and speaking. *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20). Volume 2.* Edited by Marjorie K.M. Chan and Hana Kang. Columbus, Ohio: The Ohio State University, pp. 611-626.

Gell-Mann, M. & Ruhlen, M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences USA 108(42), 17290-17295.*

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In: Kuczaj, S. (Ed.), *Language development: Vol. 2. Language, thought, and culture 301-334.* Lawrence Erlbaum Associates: Hillsdale, NJ.

Gentner, D. (2006). Why verbs are hard to learn. In: Hirsh-Pasek, K.; Golinkoff, R., (eds.), *Action meets word: how children learn verbs: 544-564.* Oxford: Oxford University Press.

Genzel, D. & Charniak (2002). Entropy rate constancy in text. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), 199–206.*

Givon, T. (1979). *On understanding grammar*. New York: Academic Press.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. Trends in Cognitive Sciences 3 (11), 419-429.

Goldin-Meadow, S., Chee So, W., Ozyurek, A., & Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the USA, 105(27), 9163-9168.*

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

Gulordava, K., Merlo, P. & Crabbé, B. (2ß15). Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases,. *Annnual Meeting of the Assoc. for Computational Linguistics, ACL 2015,* Beijing.

Gustison, M.L., Semple, S., Ferrer-i-Cancho, R. & Bergman, T. J. (2016). Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA 113 (19), E2750-E2758.*

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution 1 (1), 19-29.*

Heyer, G. & Mehler, A. (2009). Personal communication.

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten: eine Fehlinterpretation der Shannonschen Experimente? *Frequenz 44, 243-248.*

Imai, M., Haryu, E. & Okada, H. (2005). Mapping novel nouns and verbs onto dynamic action events: are verb meanings easier to learn than noun meanings for Japanese children? *Child Development 76, 340–355.*

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R.M. & Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanese-speaking children. *Child Development 79, 979-1000.*

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology 61 (1), 23-62.*

Kershenbaum, A., Blumstein, D. T. & Roch, M. A. et al. (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews 91 (1), 13-52.*

Konieczny, L. & Döring, P. (2003). Anticipation of clause-final heads: evidence from eye-tracking and SRNs. In: P.P. Slezak (ed.), *Proceedings of the ICCS/ASCS-2003 Joint International Conference on Cognitive Science: 330-335*. Sydney: University of New South Wales.

Köhler, R. (1987). System theoretical linguistics. *Theoretical Linguistics, 14 (2-3), 241-247.*

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch, Quantitative Linguistics: An International Handbook: 760-775.* Berlin: Walter de Gruyter.

Langus, A. & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology 60(4), 291-318.*

Lewis, R. L., Vasishth, S. & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10(10), 447-454.*

Levy, R & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems.*

Liu, H., Xu, C. & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, accepted.

Lupyan, G. & Christiansen, M. H. (2002). Case, word order, and language learnability: insights from connectionist modeling. In: Wayne D. Gray and Christian D. Shunn

(eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum, pp. 596-601.

Malaina, A. (2015). Two complexities. The need to link complex thinking and complex adaptive systems science. *Emergence: complexity and organization 17(1), 1-9.*

Maurits, L., Perfors, A. A. & Navarro, D. (2010). Why are some word orders more common than others? A uniform information density account. *Advances in Neural Information Processing Systems 23, 1585-1593.*

McCowan, B., Hanser, S. F., & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour 57, 409-419.*

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R.M. & Lannon, R. (2011). An image is worth a thousand words: why nouns tend to dominate verbs in early word learning. *Developmental Science 14, 181-189.*

Montemurro, M. & Pury, P. A. (2002). Long-range fractal correlations in literary corpora. *Fractals 10, 451-461.*

Moscoso del Prado Martín, F. (2011). The universal "shape" of human languages: spectral analysis beyond speech. Available from Nature Procedings http://hdl.handle.net/10101/npre.2011.6097.1

Moscoso del Prado Martín, F. (2013). The missing baselines in arguments for the optimal efficiency of languages. In: *Proceedings of the 35th annual conference of the Cognitive Science Society, pp. 1032-1037.*

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

Naranan, S. & Balasubrahmanyan, V.K. (2000). Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. Part I: Efficiency of the genetic code of DNA. *Journal of Quantitative Linguistics 7 (2), 129-151.*

Newmeyer, F. J. (2000). On the reconstruction of 'proto-world' word order. In: Chris Knight, James R. Hurford, and Michael Studdert-Kennedy (eds.), *The Evolutionary Emergence of Language, 372-388.* Cambridge: Cambridge University Press.

Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics, 10(6), 405-415.*

Piantadosi, S. T., Tily, H. & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences USA, 108 (9), 3526-3529.*

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

Reali, F. & Christiansen, M.H. (2009). Sequential learning and the interaction between biological and linguistic adaptation in language evolution. *Interaction Studies 10, 5-30.*

Sandler, W., Meir, I., Padden, C. & Aronoff (2005). The emergence of grammar: systematic structure in a new language. *Proceedings of the National Academy of Sciences USA 102, 2661-2665.*

Saxton, M. (2010). Child language. Acquisition and development. Chapter 6: the developing lexicon: what's in a name? Los Angeles, CA: SAGE. pp. 133-158.

Searls, D. (1992). The Linguistics of DNA. *American Scientist 80, 579-591.*

Schouwstra, M. & de Swart, H. (2014). The semantic origins of word order. Cognition 131 (3), 431-436.

Suzuki, R., Buck, J. R., & Tyack, P. L. (2006). Information entropy of humpback whale songs. *Journal of the Acoustical Society of America 119(3), 1849-1866.*

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

Glottometrics 39, 2017, 72-87

Belza-Chains of Adnominals

Sergej Andreev[1], Mihaiela Lupea[2], Gabriel Altmann

Abstract. Russian texts are rewritten in form of adnominals, everything else is omitted. Then Belza-chains, i.e. uninterrupted sequences of sentences containing the given class of adnominals, are stated and their length is computed. One obtains a distribution which can be modeled. Russian text are characterized and compared.

Keywords: Russian, adnominals, Belza-chains

Introduction

The study of Belza-chain of adnominals is a high-level abstraction which can be performed in different ways. The basic, elementary requirement is a definition of relevant entities but definitions are no truth, merely conventions. The situation will be critical especially in texts not having a fixed mark for sentence – or, on the contrary, there are too many possibilities and one must decide. For example, in Russian texts, one may consider the end of sentence symbolized by a dot, a colon, a question mark, an exclamation mark, a quotation mark, etc. But one may find texts in languages having no punctuation and the study of chains must be performed in different ways. This can lead to divergences when one begins to construct theories – but one must begin somewhere.

We consider – for Russian – a unit of the Belza-chain a sentence marked by dot, dots, a question mark or exclamation mark. A Belza-chain (cf. Belza 1971, Skorochod´ko 1981, Chen, Altmann 2015) is an uninterrupted sequence of sentences containing – in the simplest case – the same word. But even the word must be defined: do we consider also prepositions, conjunctions, synonyms, hypernyms, etc. or only words defined in a special way? One can set up Belza-chains of any kind of entity starting from syllables, morphemes, grammatical categories, words, phrases, parts-of-speech, etc. What does such a segmentation mean? If the chains are long – a property that can be expressed quantitatively – the text is concentrated in the given domain. If they are short, the text is rather variable. Frequently, it is not easy to capture the given property. Sometimes, special chains tell something about language, not about the text. For the time being, the "highest" abstraction is a Belza-chain constructed in terms of hrebs. Hrebs are sets of all sentences containing the same concept – either directly or

[1] Sergej Andreev, Smolensk State University, 214000 Przhevalskijstr. 4, Smolensk, Russia. Email: smol.an@mail.ru
[2] Lupea Mihaiela, Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania. Email: lupea@cs.ubbcluj.ro

as a synonym, a pronoun, a reference, etc. (cf. Ziegler, Altmann 2002). They can be changed into Belza-chains if one subdivides the hrebs into direct sequences. Long and many chains mean here a strong denotative concentration.

In general, Belza-chains may express phonic, grammatical, semantic, thematic or stylistic concentration of the text. In poetry, one directly strives for some phonic repetitions, e.g. in rhyme or assonance; in scientific texts one describes an entity in long sequences of sentences, etc. But if we take into account adnominals we may speak only of stylistic concentration/inertia. Adnominals may be words, phrases, clauses and they may be classified in various classes. For Russian, we stated the following ones (cf. Andreev, Popescu, Altmann 2017):

A – adjective (*Бледное* лицо – Pale face; Человек *спокойный* – *Man calm).
ADV – adverb (Комната *наверху* – Room upstairs; *Назад* козырьком – *With the backwards peak).
AO – adjective in an elliptical construction (У меня есть один красный карандаш и один *синий*. – *I have one red pencil and one blue).
AP – apposition (Его *костюм, галстук, рубашка* – вся одежда была абсолютно новой – His suit, tie, shirt – all clothes were brand new; Незнакомец, *мужчина среднего возраста*, подошел ко мне – The stranger, a middle-aged man, came up to me).
APAJ - type of apposition based on adjoinment type of connection with the head word, i.e. its syntactic links with the head word are not based on either agreement, or government (Гостиница «*Байкал*»; слово «*привет*» – The hotel Baikal; the word ‘hello’).
APX – type of apposition expressed by a proper name which agrees in number, case and gender with the appositive (Хирург *Иванов*, капитан *Смоллетт* – Surgeon Ivanov, Captain Smollett).
AY – adjectival phrase (*Бледное от волнения* лицо – *Pale from anxiety* face; Лицо, *бледное от волнения* – Face pale from anxiety).
CN – compound word with attributive relations of two stems, one of which is a modifier (*Страдальцы*-мальки – Sufferers-fries; Спортсмен-*чемпион* – sportsman-champion).
DAT – dative case (Письмо *другу* – Letter to a friend).
DETF – demonstrative pronoun (*Этот* дом – This house; Книга *эта* – моя. – *Book this is mine).
DETH – indefinite pronoun (*Какие-то* книги – Some books; Книги *какие*-то – *Books some).
DETN – negative pronoun (*Никакой* ошибки – No mistake; Знакомств *никаких* не желаю – *Acquaintances any I do not want).
DETQ – *qualifying* pronoun (*Все* книги – All the books; Книги *все* – *Books all).
DETS – possessive pronoun (*Его* друг – His friend; Книги *мои* здесь – *Books mine are here).
DETV – relative pronouns (Я спросил, *какая* книга пропала – I asked which book was missing; Интересно, экономия *какая* будет – It is interesting economy what will happen).

DETW – interrogative pronoun (*Какая* книга пропала? – Which book is missing?; А машина *какая* там была? – *And car which was there?)
G – genitive case (*Отца* брат – *Of the father brother; Книга *брата* – Book of the brother).
I – infinitive (*Поехать* желание было, *собирать* вещи желания не было – *To go there was a wish, to pack things – there was no wish; Желание *узнать* – Wish to learn).
INSTR – instrumental case (Восхищение *книгой* – Fascination with the book).
PR – prepositional noun (*на плече* чехол – On the shoulder a cover; Книга *для детей* – Book for children).
PT – participle (*Разбитый* стакан – Broken glass; Чудеса *невиданные* – Miracles unseen).
PTY – participial construction (*Разбитый* на куски стакан – Broken to pieces glass; Книга, *потерянная несколько дней назад* – Book lost a few days ago).
RC – subordinate clause (Это тот человек, *который может нам помочь* – This is the man who can help us; Вот план, *что делать дальше* – Here is a plan what to do next; Это – место, *где мы встретились* – This is the place where we met).

Now, since the adnominals are classes, we may transcribe the text in form of symbols omitting everything else that does not belong to one of the given classes. If we subdivide the text into sentences, then we may study the inertia of individual classes of adnominals. First of all, we state the length of sequences (uninterrupted chains) in which a given adnominal occurs. One may omit those that occur only in one sentence, that means, f_x for x = 1 does not occur but one may consider them in the counting. Counting the length of sequences we obtain a distribution which is characteristic for the given text. Using the given distribution, we may characterize it with some indicators, e.g. mean, variance, repeat rate, h-indicator etc.

It must be remarked that within a given sequence another sequence may begin or end. The text is so to say, interwoven by chains. The simplest problem that can be solved is the distribution of chain lengths. We conjecture that there is a background law controlling the forming of chains but one needs many investigations in order to find it. The other problem is that of hierarchy: do "higher" units form different regularities? The most extensive problem is the relation of the given "chain-law" to other laws known already in linguistics. One can approach these levels only step by step.

For the sake of illustration we take text No. 8 (see Appendix) and rewrite it in terms of adnominals as shown in Table 1. In order to save place, we wrote them in a table. The mark "//" divides the sentences. The computation is simple: In the first sentence only "A" occurs and does not occur in the second sentence. Hence the length of the chain is 1. The same holds for DETH in the second sentence; DETS occurs in two subsequent sentence, hence the length of the chain is 2; we have further two "A" but they occur in the same sentence, hence they give lengths 1; PR occurs in sentence two and three, hence the length of the chain is 2, etc. There are of course very long

chains (e.g. for "A") in the text. If one took a symbol into account and computed the length of the chain, one can eliminate all concerned symbols in the chain. That means e.g. in the third sentence one counts the chain of A only once, not twice. Practically, the same symbol in the same sentence is taken only once into account. The computation may be made "by hand" but the possibility of making errors is greater.

Table 1
Adnominals in Text 8
(// divides the sentences)

A// DETH// DETS,A,A,PR,PT,G// DETS,PR,DETH// DETQ// A,AO// DETW,DETS// APAJ// PR,G,A,PR,A// RC,AP,G,AP//
A,A,G,A,A,A,A,DETS,A,G,A,PR,G// A,A,A,PR,A,A,A,PR,DETQ,APX,A// DETS// AO,AO,G,PR,AO,AO// RC// RC// A// APAJ,PTY,PT//
G,A,A,G,PT// A,A,G,APAJ,G,G,APX,A,PR,G,A,A,PR// G,G// DETQ,A// A,A,G// DETF,A,G,RC,G// RC,PTY,A,RC,DETF,A,A,G//
A,G,A,A,A,PR,A,RC,A,A// ADV,A,G,AP,DETQ,PR,PR,PR,PR,AP,A,G,A,G// A,PR,DETS,A,PTY,A,APAJ,RC,A// DETS,RC// DETS,A,A,A,G,A,PT,A,G//
A,A,A// G,A,G,A,A,A,G// CN// G,A,G// DETF,A,A,PTY,A,PR,A,A,A,DETF,DETQ,PT,G,RC,A,PTY,A,PTY,A,G//
APX// DETS,A// A// A,A,A,G,A,PR,PTY,PT// DETN// DETN// PTY,G,A,A,PR// A,G// PR// PT,AY,A,DETS,A,A,DETF,A,G,G//
A,A,A,DETF,G// A,A,AY,A,A,A,DETS,A,A,PR// A,PT,PT,A,G// G,PTY,G,RC,A,A// A,PR,PTY,A,G,G// A,G,A,A,A,G,G// A// A// A//
PT,A,PR// G// DETF,A// A,A,A,G// A,PR,G// DETS// A,A,A,A// DETF,G// A// A,G,ADV,ADV// A,A// A// DETH,A// A//
DETS,A,A,AY,A,A,A,A,RC// A,A// A// DETH,A,A,A// DETF,A,A,G// A,A// DETF,DETQ,A// PTY,A,G,DETF,A,A,A,A,PTY// AP//
AP,AP,AP,A,A// DETF,A// AO,AO,PT,A,DETS// A,G// A// A,A,A,A,PTY// PR,A// PT,G,AO,A,RC// A,A,RC// PT,A,A,A,A,A,PR,A,A,G// DETS//
A,A,G,AY,A,A,PTY,PTY,A// DETF,A,PR,G,RC,DETF,A,G,A,G,I,DETS,A,A// DETS// DETS,A// DETF,DETS,A,DETS,A,A,PR,A,A,PR// DETF// A,G//
A,G// DETQ// DETS,A,APX// A,A,APAJ// A,AP,APX,A,G,G// A// DETF,A,A// A,DETS,A,G,G,PR,PT,G// DETV// A,A,A,DETS,A,DETH,A//
DETS,DETN// A,A,A// A// A// A// DETF,A// A,A,A,G// DETS// I,A,A,APAJ// DETW// AP,RC// A,A,A// A,A// DETF,PR//
DETS,A,RC,G,A,G,PR,RC,A// PR,G// PT// PT,G// A,A,A,G// A,A,A,PR,A,PR,PR,G,DETS,A,A// G,A,A// A// A// A//
A,A,G,DETF,PT,A,RC,G// A,A,PR,PR// PTY,PT,G// DETW,A,A,DETQ,DETS// A,DETF,A,G,AP,A,G// A,A,PT,DETH,PR//
PR,PR// APX,RC,G,PTY,A,PTY,A// PT,G,PTY// PTY,DETS,PT,G// DETF,CN// A,RC,A,G// RC,A,A,G,A,A,PR// A,G,A,I//
DETF,A,A,PR,PR,DETF,DETQ,G// A,G,A,DETS// A// A// DETS// A,I,DETF,PR,DETS// DETF// DETS// DETS// A,APAJ,A// A// A,A//
DETS,DETF,PT,DETS// A,DETS,I,A,A,CN// G,PR,PR// APAJ,A,G,PR// A// DETS// DETQ,A,PR// DETS,APAJ,DETS,A// AO// DETS,DETH,A,G//

A,PT,PR,PT// PR// A,G// A,A,A,A,G,RC,AY// A,A,G,G// A// A// CN// G,DETQ,PR// DETS// A,PR,G// A,G// A,G,AY,DETS,A// PR,G,AO// PT// A,A,PR,DETS,A,PR// A,DETQ// A// A,PT,A,G// DETF,A,DETQ,PTY,RC,PT,G// PT,A// DETQ// A,PR,A,PR,RC,A// G,A,G,G// A,A,G// APX,DETS,A,DETS,PR// A,A,G,DETH// PR,G// G,G// A,A,PR,DETF// PR// A// G,A,PT,G,PTY// G,G// DETF,PR,RC,A// PR// DETF,CN,A,A,PR// G,A,G,DETF,DETQ,DETF,DETQ,A,RC// A,A,G// A,AY,PR,G// G,DETH,PT,PT// DETS,DETS// PT// PT,G,PTY,PT,A,PTY,PR,A// A,A,A,G,A,AP,A,AP// DETF,DETH,A,A// DETS// DETS,DETS// PR// A,PR,G// A,PR,A,RC,A,AP,G,AP,AP,APX,G,G// A,PR,G,G,APX,RC// DETH,A,A,G,A,DETF,A,G,PT,A,G// G,DETF// DETS,DETQ,I,DETS// AY,A,AP,DETQ,A,DETQ// G,DETS// DETN// DETV,G,A,DETQ,DETS,A,A,A// CN,G,AP// A,A// G,DETQ,A,G,A,DETQ,G//

Each sentence is separated by // from the next one. For the sake of illustration, let us note the length of all chains beginning with "A":

[1,1,1,1,2,1,2,7,3,2,3,2,11,3,1,14,10,2,2,2,6,1,6,1,2,1,8,3,1,7,1,3,1,2,2,2,5,3,6,5,1,2,1,4,3,4,1,1,2].

The individual chain lengths are presented in Table 2. Here chains of all adnominals are considered, e.g. x = 1 contains chains of length 1 of all adnominals. Needless to say, if the chain length is presented in the above way, one can compare also the representation of individual adnominals; one can form motifs of chains, etc.

Table 2
Lengths of adnominal Belza-chains in Text 8

x	$\mathbf{f_x}$
1	252
2	59
3	19
4	8
5	5
6	3
7	2
8	1
10	1
11	1
14	1

Since we are concerned with length, we apply the general model of length as used for any kind of length in texts (cf. Popescu, Best Altmann 2014), namely the Zipf-Alekseev function defined as

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

Since (1) may be derived from a differential equation which is part of the unified theory (cf. Wimmer, Altmann 2005), one may consider parameter *a* as a constant of language, *b* as the expression of the speaker/writer who in case of length works with logarithmic values (remembering the Weber-Fechner law), and *c* is the control parameter (of the hearer/reader) regulating the respective lengths; in this case, it is associated with the frequency of the smallest length. We added to (1) mostly 1 because the zero values have been omitted. In female texts we used the simple formula only in two cases.

The results of counting and fitting are presented in Table 3. As can be seen, only in three cases (T 9, T 17, T 22) the original Zipf-Alekseev function without added 1 has been used. One can consider cases of this kind as containing some boundary condition but they do not impair the results which are in all cases very satisfactory.

Table 3
Fitting the Zipf-Alekseev function to lengths of adnominal Belza-chains

	T 1		T 2		T 3		T 4	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	214	214.00	235	234.98	234	234.01	262	262.00
2	57	56.95	52	52.34	50	49.74	36	35.75
3	19	19.37	18	17.03	6	8.35	10	12.57
4	9	8.18	8	7.20	7	2.24	11	6.49
5	4	4.18	2	3.77	6	1.24	5	4.14
6	2	2.55	1	2.37	4	1.05	1	3.02
7	-	-	1	1.73	1	1.01	2	2.40
8	-	-	-	-	1	1.00	1	2.03
9	-	-	-	-	-	-	1	1.79
10	1	1.16	2	1.15	-	-	-	-
a	-1.4120		-1.7578		-0.7370		-3.0324	
b	-0.7453		-0.6212		-2.1930		0.1782	
c	213.0012		233.9830		233.0104		261.0001	
R^2	1.0000		0.9998		0.9987		0.9994	

	T 5		T 6		T 7		T 8	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	211	210.96	155	155.00	225	225.00	252	252.01
2	42	42.86	42	42.03	52	51.93	59	58.88
3	18	14.82	8	7.66	17	17.37	19	19.40
4	5	6.85	1	2.15	8	7.49	8	8.18
5	3	3.88	2	1.22	4	3.97	5	4.21
6	2	2.57	-	-	1	1.46	3	2.59
7	1	1.92	-	-	-	-	2	1.85
8	1	1.57	-	-	-	-	1	1.48

9	1	1.37	-	-	-	-	-	-
10	1	1.25	-	-	-	-	1	1.18
11	-	-	1	1.00	-	-	1	1.11
14					1	1.04	1	1.03
a	-2.0703		-0.2825		-1.7193		-1.6692	
b	-0.3698		-2.3452		-0.6024		0.6456	
c	209.9643		153.9985		224.0025		251.0061	
R^2	0.9996		0.9999		1.0000		1.0000	

	T 9		T 10		T 11		T 12	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	195	195.03	260	260.03	246	246.00	283	283.01
2	40	39.46	53	52.26	43	42.89	64	63.75
3	12	13.52	11	14.82	11	11.40	15	16.24
4	6	5.94	10	5.63	4	4.29	6	5.39
5	4	3.04	3	2.81	3	2.23	5	2.49
6	1	1.71	3	1.79	3	1.52	-	-
7	1	1.04	2	1.38	1	1.24	2	1.22
8	3	0.67	3	1.19	-	-	1	1.10
9	2	0.45	-	-	-	-		
10	-	-	-	-	-	-		
11	-	-	-	-	-	-		
12	-	-	2	1.02	-	-		
13	-	-	1	1.01	1	1.01		
14	-	-						
18	1	0.04						
a	-2.0923		-1.7720		-1.9872		-1.3337	
b	-0.3071		-0.8155		-0.8090		-1.2036	
c	195.0260		259.0288		245.0042		282.0113	
R^2	0.9996 (ZI-AL)		0.9993		0.9999		0.9999	

	T 13		T 14		T 15		T 16	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	277	277.00	195	194.99	281	280.99	265	264.99
2	60	60.06	52	52.25	44	44.24	51	51.25
3	14	13.38	18	17.10	15	14.02	19	18.17
4	2	4.04	7	6.98	6	6.29	8	8.65
5	4	1.86	2	3.51	3	3.56	6	4.98
6	1	1.28	1	2.16	1	2.39	2	3.30
7	1	1.10	3	1.58	2	1.82	-	-
8	1	1.04	1	1.31	-	-	2	1.93

9	-	-	-	-	2	1.34	1	1.64
10	-	-	1	1.10	2	1.23	1	1.45
11	-	-	1	1.06	-	-	1	1.33
12	1	1.00			-	-	2	1.25
13					-	-	-	-
14					-	-	1	1.14
15					1	1.05		
a	-1.1963		-1.3302		-2.5273		-2.2323	
b	-1.4830		-0.8514		-0.2418		-0.2321	
c	275.9979		193.9860		279.9925		263.9891	
R^2	0.9999		0.9997		0.9999		0.9999	

	T 17		T 18		T 19		T 20	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	229	228.99	257	256.97	215	215.00	205	204.99
2	58	58.25	45	45.77	36	36.09	33	33.14
3	19	17.49	17	12.98	10	9.64	12	12.32
4	4	6.22	1	5.11	4	3.74	9	6.49
5	2	2.52	1	2.66	1	2.03	2	4.17
6	1	1.12	1	1.76	1	1.44	-	-
7	2	0.54	-	-	1	1.20	-	-
8	1	0.29	-	-	-	-	-	-
9	1	0.15	-	-	-	-	1	1.77
10	1	0.09	-	-	-	-	2	1.60
11	1	0.05	-	-	1	1.02	-	-
12	2	0.03	-	-	-	-	-	-
13	-	-	-	-	1	1.01	-	-
14	-	-	-	-			1	1.28
16	-	-	1	1.01				
26	1	0.0002						
a	-1.3486		-2.0507		-2.0739		-2.7252	
b	-0.9036		-0.6702		-0.7711		0.0852	
c	228.9900		255.9738		213.9966		203.9922	
R^2	0.9996 (ZI-AL)		0.9993		1.0000		0.9997	

	T 21		T 22		T 23		T 24	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	219	218.98	226	226.00	249	249.01	110	110.00
2	44	44.53	36	35.89	40	39.63	24	24.04
3	17	14.89	3	4.28	12	13.11	8	7.42
4	5	6.64	4	0.59	6	6.15	1	3.20
5	4	3.67	3	0.10	5	3.61	4	1.87

6	1	2.40	1	0.02	1	2.47	-	-
7	1	1.79	4	0.00	3	1.90	-	-
8	1	1.48	2	0.0010	-	-	2	1.09
9	-	-	2	0.0003	-	-	1	1.05
10	2	1.20	-	-	4	1.28	-	-
11	-	-	-	-	1	1.21	-	-
12	-	-	-	-	2	1.15	-	-
13	-	-	-	-			-	-
14	-	-	-	-			-	-
15	1	1.03	-	-			-	-
16			1	0.00			-	-
17			-	-			-	-
18			-	-			-	-
19			-	-			-	-
20			1	0.00			-	-
21			1	0.00			-	-
22							-	-
23							1	1.00
a	-2.0127		-1.0199		-2.5708		-1.6684	
b	-0.4493		-2.3584		-0.1615		-0.8281	
c	217.9783		226.0033		248.0139		109.0002	
R^2	0.9997		0.9989 (ZI-AL)		0.9998		0.9989	

	T 25		T 26		T 27		T 28	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	225	224.97	320	319.99	150	150.00	342	342.00
2	30	31.06	55	55.21	34	33.99	83	83.08
3	15	11.48	19	18.36	6	6.08	27	25.73
4	4	6.23	8	8.39	2	1.87	6	9.96
5	5	4.15	6	4.71	-	-	6	4.71
6	3	3.12	1	3.08	2	1.04	6	2.70
7	1	2.54	-	-	1	1.01	3	1.84
8	-	-	-	-	-	-	1	1.44
9	2	1.94	-	-	-	-	1	1.25
10	-	-	1	1.38	-	-	2	1.14
11	-	-	-	-	-	-	1	1.09
12	1	1.56	1	1.20	1	1.00	-	-
13	-	-	-	-			1	1.03
14	1	1.43	-	-			1	1.02
15	-	-	-	-			1	1.01
16	1	1.35	1	1.07			1	1.01
a	-3.0850		-2.3981		-0.6368		-1.4842	
b	0.2707		-0.2292		-2.2195		-0.8229	
c	223.9680		318.9917		149.0005		341.0012	
R^2	0.9995		0.9999		0.9999		0.9997	

	T 29		T 30		T 31		T 32	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	217	217.07	199	198.96	158	158.00	146	146.01
2	52	50.54	23	24.41	32	31.95	24	23.55
3	10	16.47	12	9.76	12	12.39	7	8.45
4	13	6.91	10	5.91	7	6.48	4	4.36
5	5	3.60	2	4.36	4	4.06	5	2.81
6	3	2.26	1	3.57	4	2.89	-	-
7	3	1.66	4	3.12	1	2.25	2	1.70
8	2	1.37	2	2.83	-	-	-	-
9	-	-	2	2.64	-	-	1	1.34
10	1	1.13			1	1.47		
11	-	-						
12	-	-						
13	1	1.04						
14	-	-						
15	-	-						
16	-	-						
17	-	-						
18	1	1.01						
a	-1.6543		-3.4938		-2.2644		-2.6542	
b	-0.6789		0.5972		-0.1127		-0.0441	
c	216.0657		197.9564		156.9999		145.0148	
R^2	0.9979		0.9989		0.9998		0.9996	

	T 33		T 34		T 35		T 36	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	230	230.00	163	163.01	227	227.02	215	214.99
2	37	36.87	28	27.78	46	45.45	49	49.27
3	11	11.87	8	8.53	11	13.81	16	14.63
4	7	5.43	3	3.80	9	5.65	4	5.66
5	3	3.15	-	-	3	2.96	3	2.83
6	1	2.17	5	1.60	2	1.93	1	1.80
7	-	-	-	-	2	1.47	-	-
8	2	1.43	-	-	1	1.26	-	-
9			-	-	2	1.15	1	1.10
10			1	1.07	-	-		
11					1	1.06		
12					-	-		
13					1	1.02		
a	-2.5038		-2.2660		-1.8908		-1.5365	
b	-0.2463		-0.4772		-0.6572		-0.8827	
c	229.0030		162.0099		226.0207		213.9883	
R^2	0.9999		0.9994		0.9995		0.9999	

	T 37		T 38		T 39		T 40	
x	f_x	Comp	f_x	Comp	f_x	Comp	f_x	Comp
1	180	179.96	221	220.93	206	205.97	151	151.00
2	47	47.82	32	33.81	50	50.52	35	35.06
3	20	16.34	17	12.76	20	18.41	11	10.86
4	3	6.99	9	6.90	8	8.53	5	4.46
5	-	-	2	4.54	5	4.72	1	2.40
6	1	2.29	2	3.36	1	3.02	-	-
7	-	-	1	2.69	2	2.17	2	1.30
8	2	1.37	1	2.28	1	1.72	-	-
9	2	1.22	-	-			1	1.09
10	-	-	-	-				
11	1	1.08	1	1.68				
12			-	-				
13			-	-				
14			1	1.43				
a	-1.4186		-2.8801		-1.7156		-1.5598	
b	-0.7444		0.1951		-0.4815		-0.8356	
c	178.9600		219.9342		204.9722		149.9971	
R^2	0.9988		0.9991		0.9998		0.9998	

Characterization

If there are many short chains, the stylistic adnominal inertia is small, the author variegates the linguistic means. But this means automatically that the mean of the distribution is small, hence we can characterize the situation using the average of lengths.

Now, since inertia evidently depends also on the longest chains, one may use the length of the arc formed by the frequencies as a characteristic. Further, since all this involves also the variance and the form of the curve, one may use Ord's criterion containing the first three moments. If the x = 1 value is strongly represented, then the inertia is small, hence the usual Repeat rate is great and can be used for characterization. On the contrary, if the entropy of the distribution is great, then the inertia is great.

Needless to say, there are many other possibilities but the above ones have been frequently used in quantitative linguistics.

Ord's criterion (Ord 1972) is defined as

$$(3)\quad I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}$$

where m'_1 is the average and m_2, m_3 are the second and third central moments respectively, and the result is always a figure (I,S). The figure displaying the (I,S) relation is given in Figure 1.

Table 4
Relative arc length and Ord's (I,S) (female texts)

Text	I	S		Text	I	S
T 1	0.5655	2.1622		T 11	0.8716	6.0470
T 2	0.8150	4.6461		T 12	0.6041	3.2451
T 3	0.7897	3.4389		T 13	0.7169	5.5292
T 4	0.8608	3.9521		T 14	1.0826	5.0470
T 5	0.9610	4.5852		T 15	1.3479	7.4136
T 6	0.6428	5.6162		T 16	1.6503	7.2496
T 7	0.7048	4.1739		T 17	2.4585	14.5337
T 8	1.1950	5.5957		T 18	0.8233	9.6622
T 9	1.7073	8.8279		T 19	1.0458	7.3530
T 10	1.4647	6.7267		T 20	1.3238	7.3583

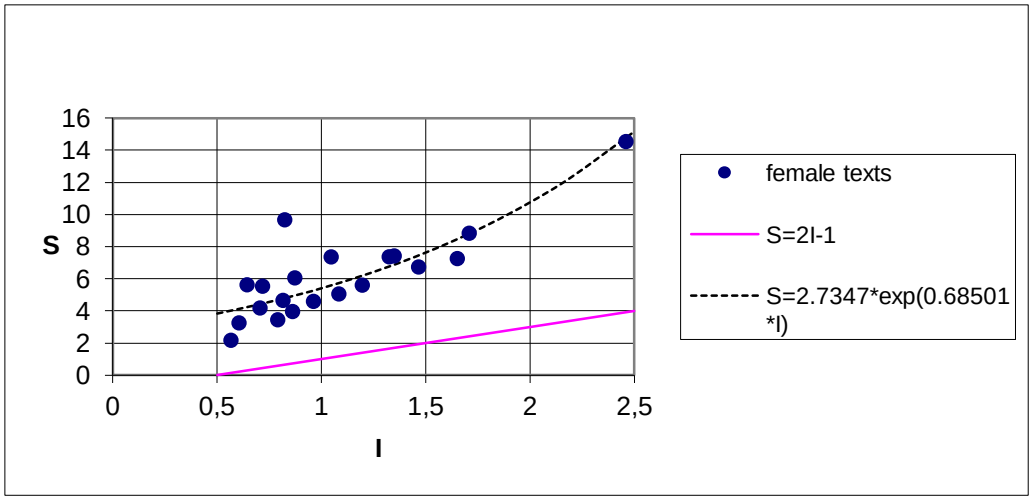

Figure 1. Ord's (I,S) for female texts

If one omits Text T 18, one obtains for female texts a simple exponential dependence y = a*exp(b*x), i.e. S = 2.7347exp(0.6850I) yielding R^2 = 0.8545. Text T 18 contains, evidently, a conscious style or a posteriori change of the text. The observed values of S of the female texts are always placed over the Ord's line S = 2I - 1.

Table 5
Relative arc length and Ord's (I,S) (male texts)

Text	I	S		Text	I	S
T 21	1.3289	7.3517		T 31	1.0063	3.9005
T 22	3.0241	13.2657		T 32	0.9360	4.0535
T 23	1.7552	6.9321		T 33	0.6783	3.6266

T 24	2.7746	15.2211		T 34	0.8699	4.6365
T 25	1.8904	8.7359		T 35	1.3836	6.2697
T 26	1.1323	8.1120		T 36	0.5843	3.3766
T 27	0.9133	6.4361		T 37	1.0764	4.3006
T 28	1.8489	8.3535		T 38	1.2482	6.7824
T 29	1.7796	8.3793		T 39	0.7601	3.0499
T 30	1.2699	4.5624		T 40	0.7658	3.9861

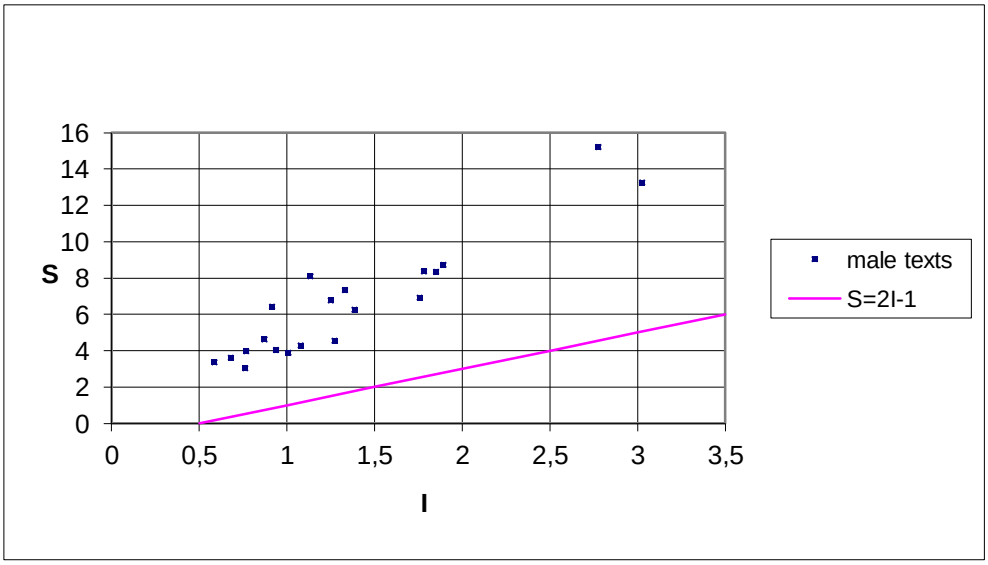

Figure 2. Ord's (I,S) for male texts

The comparison of individual texts and their ordering could be performed applying either individual tests or one of the 500 classification procedures. We are content with capturing the significant capturing of S = f(I).

When comparing male and female texts, one may use any indicator but we restrict ourselves to the presentation of both in one figure, marking female texts with a circle and male texts with a square. As can be seen in the Figure 3, the difference is not relevant.

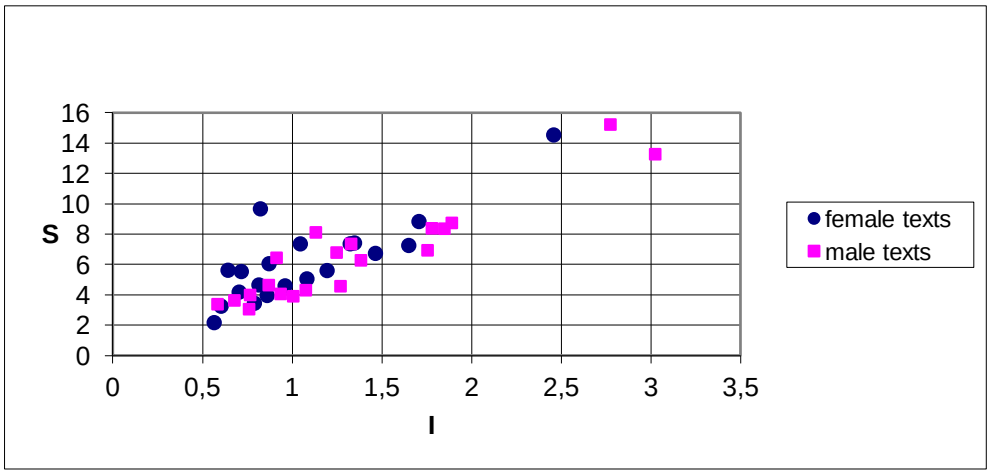

Figure 3. Ord's criterion for female and male texts

However, considering S as a direct indicator, one may compute its mean and perform the t-test for the difference of two means. In the above case, the means are almost equal and the difference is minimal, hence we may state that there is no difference between the S-values of female and male texts.

Other types of characterization will be postponed until one has data from several languages.

Conclusions

It would be very interesting to consider only one text and analyze the Belza-chain at all possible levels beginning from the phonetic one up to hrebs, in order to obtain a picture of inertia, its change through levels from phonetics up to stylistics. To solve this problem one will need many years and great teams. It is not sure that the results will be similar in all languages, all text types, all times, at all levels of language, etc., hence the problem can be developed as a special branch of stylistics.

References

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych textov. In: *Semantičeskie problemy avtomatizacii i informacionnogo potoka: 58-73*. Kiev.

Chen, R., Altmann, G. (2015). Conceptual inertia in texts**.** *Glottometrics 30, 73-88.*

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

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

Skorochod´ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten*. Bochum: Brockmeyer.

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

Appendix

Female authors

Author	Title	Year	Words in the abstract	Ad-nominals
T 1 S. Demidova	***Rubinovaja vernost'*** (*Ruby fidelity*). Novel.	2007	3723	614
T 2 D. Dontsova	***Kleopatra s parashjutom*** (*Cleopatra with a parachute*). Novel.	2013	4294	612

T 3 D. Dontsova	***In', Jan' i vsjakaja drjan'*** (*Yin-yang and various stuff*). Novel.	2008	4559	556
T 4 D. Dontsova	***Prodjuser koz'ej mordy*** (*Producer of dirty tricks*). Novel.	2008	4082	600
T 5 A. Marinina	***Kazn' bez zlogo umysla*** (*Execution without bad intentions*). Novel.	2015	4053	616
T6 A. Marinina	***Stechenie obstojatel'stv*** (*Coincidence of circumstances)* Novel.	1992	2591	370
T7 A. Marinina	***Ukradennyj son*** (*Stolen dream*) Novel.	1994	4605	637
T 8 D. Rubina	***Belaja golubka Kordovy*** (*White dove of Cordova*). Novel.	2009	4352	848
T 9 D. Rubina	***Poslednij kaban iz lesov Pontevedra*** (*The last boar from the woods of Pontevedra*). Novel.	1998	3055	653
T 10 D. Rubina	***Topolev pereulok*** (*Topolev alley*). Long story.	2015	3835	858
T 11 V. Tokareva	***Lavina*** (*Avalanche*). Long story.	1955	4532	508
T 12 V. Tokareva	***Moi muzhchiny*** (*My men*). Long story.	2015	4565	652
T 13 V. Tokareva	***Tihaja muzyka za stenoj*** (*Soft music behind the wall*). Long story.	2012	4537	644
T 14 L. Tret'jakova	***Damy i gospoda*** (*Ladies and gentlemen*). Novel.	2008	3180	574
T 15 L. Tret'jakova	***Krasavitsy ne umirajut*** (*Beatiful women don't die*). Novel.	1998	2982	734
T 16 L. Ulitskaja	***Zelenyj shater*** (*Green marquee*). Novel.	2011	4437	884
T 17 L. Ulitskaja	***Iskrenne vash Shurik*** (*Yours truly Shurik*). Novel.	2006	3796	957
T 18 T. Ustinova	***Oligarh s Bol'shoj Medveditsy*** (*Oligarch from the Big Dipper*). Novel.	2004	4749	587
T 19 T. Ustinova	***Vselenskij zagovor*** (*Cosmic conspiracy*). Novel.	2016	4228	467
T 20 T. Ustinova	***Moj general*** (*My general*). Novel.	2002	4076	567

Male authors

Author	Title	Year	Words in the abstract	Ad-nominals
T 21 B. Akunin	***Table-Talk.*** Story.	2006	3966	608
T 22	***Pikovyj valet*** (*Jack of spades*).	1999	4043	650

B. Akunin	Long story.			
T 23 B. Akunin	***Turetskij gambit*** (*Turkish gambit*). Novel.	1998	5225	777
T 24 A. Bushkov	***Piran'ja. Vojna oligarhov*** (*Piranha. War of oligarchs*). Novel.	2007	2573	354
T 25 A. Bushkov	***Piran'ja protiv vorov*** *(Piranha against thieves)*. Novel.	2001	3834	673
T 26 A. Bushkov	***Tanets Beshenoj*** (*The dance of the rabid*). Novel.	2001	4839	767
T 27 M. Veller	***Laokoon.*** Story.	1993.	2438	375
T 28 M. Veller	***Marina.*** Long story.	1993	7557	1292
T 29 M. Veller	***Pjatiknizhie*** (*The Torah*). Long story.	2009	3461	762
T 30 S. Dovlatov	***Inostranka*** (*A foreign woman*). Long story.	1986	2802	461
T 31 V. Erofeev	***Russkaja krasavitsa*** (*Russian beauty*).	1990	4206	577
T 32 D. Koretskij	***Antikiller.*** Novel.	1995	2948	422
T 33 D. Koretskij	***Antikiller-5.*** Novel.	2014	3937	528
T 34 D. Koretskij	***Antikiller-6.*** Novel.	2016	2815	414
T 35 V. Pelevin	***Operatsija "Burning Bush"*** (*Operation "Burning Bush"*).	2010	3392	676
T 36 V. Pelevin	***Assasin.***	2008	3565	487
T 37 V. Pelevin	***Grecheskij variant*** (*Greek variant*). Novel.	**1977**	2891	616
T 38 Z. Prilepin	***Obitel'*** (*Convent*). Novel.	2014	4523	618
T 39 Z. Prilepin	***Patologii*** (*Pathologies*). Novel.	2005	3968	**664**
T 40 Z. Prilepin	***Sher amin'*** (*Cher amen*). Story.	2016	3774	457

Glottometrics 39, 2017, 88-92

Ukrainian Compounds in the Texts of Computer Science

Denys Ishutin[1]
Hanna Gnatchuk[2]

Abstract: The present investigation deals with a quantitative study of Ukrainian compounds in Books "The Fundamentals of Programming" (Osnov' Programuvannja) by T. V. Kovaljuk (2005) and "Informatics" (Informatyka) by J. Ryvkind (2010). We concentrate our attention on a quantitative study of Ukrainian compounds by taking into account their types in computer texts. In such a way, the material of our study is represented by 2 books "Osnov' programuvannja" and "Informatyka" belonging to the sphere of the Exact Sciences. Each page of the book in question has been studied in order to reveal the behavior (models) of Ukrainian compounds in the text of Computer science.

Keywords: *Ukrainian, compounds, technical texts*

1. Introduction: linguistic features of Ukrainian compounds

An intensive development of Ukrainian compounds is indebted to a dynamic development of information technology and the spheres of communication. Different ways of building words have been available in the Ukrainian language for a very long time. This tendency was also characteristic of the Proto-Slavic, Old-East Slavic and Ruthenian languages. Before dealing with the study of Ukrainian compounds, it is necessary to clarify the term of "composition" (= compounding). In particular, Pljushch (2000) understands the composition as a way of forming complex words by combining two or more basic words or shortened (contacted) lexemes. In such a way, the author distinguishes three types (ways) of compositions in the Ukrainian language:

- *The composition of basic words (Osnovoskladannya)* presupposes combining several basic words by means of interfixes «о», «е» (*праце**е**здатний, добро**о**зичливий*) or without these interfixes (триповерховий). In this case, the basic words are combined according to the types of subordinate (*близькоспоріднений*) and coordinative (*природничо-географічний*) relations.
- *The composition of words* (*Slovoskladannja*) or *Juxtaposition*: the combination of several words or forms into one complex word (*салон-перукарня, місто-гігант*). In this case, this compound denotes one notion.
- *Abbreviation* foresees combining words with shortened basic words (*профком, ЗМІ = засоби масової комунікації*).

In the present investigation, we shall look at noun and adjective compounds in so far as the data of our empirical investigation is represented by these word classes. As far as noun-compounds are concerned, Pljushch (2000) distinguishes 7 types:

[1] Denys Ishutin, Ternopil National Pedagogical University by V. Hnatjuk, Department of Translation Studies, vul. M. Kryvonosa 2. Email: shutndenis@mail.ru
[2] Hanna Gnatchuk, Universität Trier, Computational Linguistics and Digital Humanities, Universitätsring 15. Email: agnatchuk@gmail.com or s2hagnat@uni-trier.de

1) The first type consists of two nouns of masculine gender: *увіз-вивіз, імпорт-експорт, купівля-продаж, генерал-майор, грам-калорія;*
2) The second type is represented by nouns (of all genders) made up of a verb and a dependent noun. In most cases, the combination of words results in attaching a suffix to the second basic word: *сталевар, газомір, турбобудівник, криголам, картоплесортування, мовознавство, діалектологія, землезрошення, хлібопостачання.*
3) The third type of nouns consists of the combination of noun and adjective (with an attribution relation): *чорнослив = чорна слива, жовтоцвіт, довгоносик, дрібнолісся*.
4) The fourth type deals with a noun made up of a verb and a dependent adverb: *скоропис, вільнодум, гуртожиток, всюдихід.*
5) The fifth type of nouns deals with the combination of a numeral and a noun. This type presupposes adding a suffix to the second basic word in the compound: *семикласник, одноліток, двовладдя, двокрапка, століття.*
6) The sixth type includes the combination of a noun (of a verbal origin) with a dependent pronoun: *самоаналіз, всесвіт, самоконтроль, собівартість, самоцвіт, самоскид.*
7) The seventh type is represented by a combination of verbs of imperative mood. According to this old model general words were formed which later became surnames and geographical names: *перекотиполе, вертихвістка, Борислав, Убийвовк*.

It is also worth mentioning that Ukrainian deals with three types of abbreviation: syllabic, mixed and initial:

1) **Syllabic group** presupposes combining contracted parts of words into one word: *лісгосп = лісове господарство*;
2) **Mixed group** is represented by the combination of the initial shortened words and the whole word: *медучилище, райземвідділ, міськпромрада*. The first component of the compound can be represented by the morphemes of a foreign origin: *авіа, авто, фото, аеро, гідро* being applied to the whole word: *фотограф, гідростанція, автотранспорт, аеросани.*
3) **Initial group** foresees forming shortened words from initial letters or sounds: *вуз – вищий учбовий заклад, ООН – організація об´єднаних націй, ноп – наукова організація праці.* Nevertheless, it is relevant to distinguish two subgroups: 1) The abbreviated words pronounced as a common word and 2) the abbreviated word pronounced like the letters in the alphabet: *ЧНУ – Чернівецький національний університет* (че-не-у).

As far as adjective-compounds are concerned, the composition of basic words (*Osnovoskladannya*) is considered to be a dominant way of forming qualitative and relative adjectives. On the whole adjective compounds are of 6 types:

1. The combination of two or more adjectives: *українсько-німецький, шахово-шашковий.*
2. The combination of adverb and noun: *легкоатлетичний, народногосподарський*.
3. The combination of adverb and adjective (participle): *загальноприйнятий, багатонаціональний.*
4. The combination of numeral and noun with an adjectival suffix: *стокілометровий, багатонаціональний.*
5. The combination of a pronoun and an adjective (participle): *самовдоволений, всенародний.*

6. The combination of a noun and a verb (participle) with a suffix: *волелюбний, працездатний.*

2. A quantitative study of Ukrainian compounds in the texts of computer science

The aim of our study is to reveal the frequencies of types for Ukrainian compounds in order to see the order of their distribution (cf. Mačutek, Altmann, 2007) in Ukrainian computer texts.

The material of our study. We have analyzed two books belonging to the computer science: "The Fundamentals of Programming" ("Osnov' Programuvannya") by T.V. Kovaljuk (2005) and "Informatics. The 10th Class" ("Informatyka. 10 Klas") by J. Ryvkind et al. (2010).

The procedure of the present research foresees analyzing each page of the above-mentioned books. In such a way, we conducted a systematic sampling. As a result, our sample includes 118 compounds. Therefore, we present the results of Ukrainian types for noun and adjective compounds:

Noun compounds

1) The combination of two nouns: *інтернет-ресурс, веб-сторінка, веб-дизайн, веб-сайт, введення-виведення, зчитування-запис, джойстик, кеш-пам´ять, чипсет, клієнт-сервіс, скріншот, чит-код, сорс-код, веб-документ, інтернет-адрес, шоу-проект, експрес-таблиця, експрес-стиль, рок-музикант, сертифікат-нагорода, користувач-початківець, веб-камера, програма-відеостудія, програма-програвач, веб-колекція, фільм-розповідь, веб-інтерфейс, комп´ютор-сервер, інтернет-провайдер, веб-пошта, лист-відповідь, дзвінок-виклик, документ-заготовка*.
2) Abbreviation. Mixed group: *кілобайт, гіперпосилання, гігабайт, міні-комп´ютор, мікросхема, відеоінформація, мегабайт, міні-додаток, радіокнопка, інфографіка, інтерфейс, міні-панель, автозаміна, фотоапарат, фотоальбом, мультимедіа, відеофрагмент, макрокоманда, діапроектор, відеофільм, фоторобот, аудіозапис, відеокомпозиція, аудіокомпозиція, відеодані, відеоредактор, відеостудія, відеокамера, відеомагнітофон, відеофайл, аудіокнига, відеоефект, аудіоефект, відеофрагмент, відеофільм, аудіофайл, аудіоформат, медіапрогравач, медіафайл, відеодиск, аудіодиск, відеооб´єкт, відеокліп, відеоефект, відеоперехід, відеодоріжка, відеомонтаж, телеконференція, відеодзвінок, відеозв´язок.*
3) Abbreviation. Initial group: *HTML-файл, DVD-програвач, IM-служба*.
4) The combination of a verb and a dependent noun: *дисковод, місцезнаходження, металообробка, звукозапис.*
5) The combination of a numeral and a noun: *двостороння*
6) The combination of a noun (of a verbal origin) with a dependent pronoun: *всесвіт.*

Adjective compounds

1) The combination of two adjectives: *арифметико-логічний, літерно-цифровий, науково-технічний, структурно-семантичний, дослідно-виробничий, соціально-економічний, організаційно-розпорядчий, художньо-технічний.*
2) The combination of adverb and adjective (participle): *багаторазовий, загальноприйнятий, електрообчислювальний, багатооборотний, багаторівневий, багатосторінковий, багатошаровий, малонасичений, широкоформатний, повноекранний, багатоцифровий.*
3) The combination of a numeral and a noun with an adjectival suffix: *триадресний, однойменний, двовимірний, одноразовий*.

4) The combination of a noun and a verb (participle): *файлообмінний*
5) Abbreviation. Mixed type: *відеооптичний, монохроматичний*.

In such a way, we have detected 11 types (models) for Ukrainian compounds in the computer texts. We conjecture that the relative rate of change of frequency with increasing rank is $y'/y = log(b)$. Integrating both sides, we obtain $In(y) = x\ ln(b) + k$, where k is an integrating constant. Taking antilogarithms and reparametrizing we obtain $y = ab^{\wedge}x$, a very simple function.

In such a way, we make a table of rank-frequency distribution of both Ukrainian noun and adjective compounds in the texts of computer science (cf. Table 1).

Table 1
Rank-frequency distribution of Ukrainian compounds in computer texts

Rank	***Pattern***	**Frequency**	**Computed values**
1.	Abbreviation. Mixed group. (Noun compound)	50	51.20
2.	Noun + Noun (noun compound)	33	28.30
3.	Adverb + Adjective (adjective compound)	11	15.63
4.	Adjective + Adjective (adjective compound)	8	8.64
5.	Verb + dependent noun (noun compound)	4	4.80
6.	Numeral + Noun (adjective compound)	4	2.64
7.	Abbreviation. Initial group. (Adjective compound)	3	1.50
8.	Abbreviation. Mixed group. (Adjective compound)	2	0.80
9.	Noun + Verb (participle). (Adjective compound)	1	0.44
10.	Numeral + Noun (noun compound)	1	0.24
11.	Noun with a dependent pronoun (noun compound)	1	0.13
a = 92.5318273, b = 0.552837164, R^2 = 0.9791 (97.91%)			

We have conducted a quantitative study of Ukrainian compounds in the texts of computer science. We have found 11 types (models) of compounds available in these texts where the models "Abbreviation – Mixed group of nouns" and "Noun + Noun" turned out to be the most productive in the analyzed texts. The results have been captured applying a simple power function with an excellent fitting R^2 = 0.9791 (97.91%). The study can be extended by studying the behavior of compounds in other languages as well as in different functional styles (sorts of texts) (cf. Gnatchuk, 2015). Needless to say, the simple exponential function would yield the same result hence a more stable result can be attained only after many languages have been examined.

Automatically, several questions arise: (1) Does the regularity found hold true in all languages or only in Ukrainian? (2) Does the regularity hold true specially for this text type or is it general? (3) Whatever the answer, one can ask the question "why it is so?", e.g. why is the relative rate of change a constant? Are there other rules for other languages? (4) What is the place of this regularity on Köhler's control cycle (1986, 2005)? That is, what are the properties having influence on the formation of this regularity?

The answers to these questions may bring us nearer to the possible theory of compound formation.

References

Gnatchuk, Hanna. (2015). A statistical analysis of English compounds in the newspaper style. *Mathematical Linguistics 1(1)*, 81-90.

Kovaljuk, T. V. (2005). *Osnov' programuvannya.* Vudavnutstvo: grupa BHV (in Ukrainian).

Köhler, R., Altmann, G. (1986). Synergetische Aspekte der Linguistik. *Sprach-wissenschaft* 5, 253-265.

Köhler, R. (2005). Synergetic linguistics. In: *Quantitative Linguistik. Ein internationales Handbuch* (Hrsg. Köhler, R., Altmann, G., Piotrowski, R.G.), 27, Berlin: Walter de Gruyter

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1)*, 81-94.

Pljushch, M. J., Bevzenko, S. P., Hrypas, N. J. (2009). *Suchasna ukrajinska literaturna mova.* 7th edition. Vyshcha shkola (in Ukrainian).

Ryvkind, J. J., Lysenko, T. I., Chernikova, L. A., Shakot´ko, V. V. (2010). *Informatyka. 10 Klas.* Kyjiv „Geneza" (in Ukrainian).

Book Review

Kubát, Miroslav. *Kvantitativní analýza žánrů [A Quantitative Analysis of Genres]*. Ostrava: Ostravská univerzita, 2016, 141 pp.
Reviewed by **Michal Místecký**

Kvantitativní analýza žánrů (*A Quantitative Analysis of Genres*), a published dissertation by Ostrava-based researcher Miroslav Kubát, is an accomplished combination of sound fieldwork and responsible analysis. Founded upon the work by Karel Čapek, a versatile Czech writer of both fiction and non-fiction, the study treats utility of various methods and indices in text genre classification; to this end, moving average type-token ratio (MATTR), moving window type-token ratio distribution (MWTTRD), three types of thematic concentration measurement (TC; the standard, the secondary, and the proportional ones), verb distances (VD), average token length (ATL), activity (Q), and author's multilevel n-gram profile (AMNP) altogether with most frequent words (MFW) analysis are exploited, with advantages and disadvantages being realistically discussed in all cases. As to the corpus, Kubát investigated various instances of Čapek's novels, studies, short stories, fairy tales, poems, travelogues, newspaper columns, and letters.

The first parts of the text deal with an introduction to the topic, giving balanced views on qualitative and quantitative approaches to language, linguistic units, available software, and text definitions. Besides, there is a brief summary of the author's previous research, which was focused on a quantitative analysis of the Czech and Czechoslovak presidents' addresses; this is intended to be an exemplar of the practical use of the methods presented in the book. The fact that the methodological explanations form a considerable part of the dissertation underlines the researcher's pensiveness and breadth of knowledge.

As for the results, two ways are proposed to assess the efficiencies of individual measurements: the total sum of the significant u-test values, and the number of u-test-based significant differences. Both the methods yielded the same outcome – whereas MATTR and TC proved to be of less help when genre differences are to be found out, Q, ATL, and VD seem to give very decent data about the searched-for distinctions. However, the most fitting output for the genre classification was obtained by AMNP (the 84-percent fit) – which, as Kubát admits, is compensated for by uneasiness of linguistic interpretations of the n-grams used in the analysis. If individual text types are to be evaluated, the genres of study, novel, and travelogue were the most discernible ones, whilst poems and short stories did not show enough distinctive features to be sorted out as separate units (moreover, AMPN was unable to distinguish anything like a fairy tale, which also puts it into an unfavourable position). This finding, assessed as counter-intuitive by the author, challenges the deep-rooted literary-criticism assumptions and calls for meticulous structural, metrics-oriented analyses.

Miroslav Kubát's publication brings about a lot of food for thought: first, it is a complex, coherent and intersubjective proof of the justifiability of the notion of genre; second, it elucidates the situation in the extensive literary production of Karel Čapek; third, it pronounces evidenced verdicts on the utility of the genre-analysis methods; fourth, it pushes scholars into deeper reflections on the validity of certain text types;

and last but not least, it does not succumb to the essentialist trap of the what-is-x questions, replacing them consistently with courageous attempts to put across pragmatic definitions of the studied notions. All these features point at the potential that is to be found in the contemporary Czech quantitative linguistics.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

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

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4*. 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.