Glottometrics 38
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. 38 (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 38 (2017) **ISSN 2625-8226**

Contents

Glottometrics 38, 2017, 1-20

What We Talk about When We Talk about Corpus: A Bibliometric Analysis of Corpus-related Research in Linguistics (2000-2015)

Shengyu Liao, Lei Lei[1]

Abstract: The study aims to conduct a bibliometric analysis of corpus-related studies in linguistics between 2000 and 2015. Results show that the output of corpus-related publications have significantly increased in the past 15 years. In addition, traditional scientific powers such as the United States play leading roles in the area, while developing countries such as China also exert their impact in the area. More importantly, findings reveal that corpora have permeated a wide range of research areas in linguistics and have changed, at least in terms of methodology, these areas.

Keywords: *Corpus, linguistics, bibliometric analysis*

1. Introduction

A corpus is a collection of natural language texts for language research (Sinclair 1991: 171). It has been widely used in linguistics since the 1960s (McEnery & Gabrielatos 2006), and has made great contributions to many areas, such as lexicography, second language learning and teaching, translation studies, and quantitative linguistics, though whether corpus linguistics is a theory or methodology still remains a controversy (e.g., Tognini-Bonelli 2001; McEnery & Gabrielatos 2006; Hardie & McEnery 2010; McEnery & Hardie 2012). For example, corpora have significant impacts on lexicography. Traditional dictionaries were "full of unnatural invented examples" (Hanks 2012: 404) and compiled primarily based on dictionary compilers' intuition and introspection (Hanks 2012). Then, the use of corpus in lexicography changed this situation. Corpus analysis provides opportunities for lexicographers to compile dictionaries based on empirical and authentic data (Hanks 2012), such as the first edition of the *Collins COBUILD English language dictionary* (1987), the third and subsequent editions of the *Longman dictionary of contemporary English* (1995), the first edition of the *New Oxford dictionary of English* (1998), and the tenth edition of the *Concise Oxford dictionary* (2000), etc.

[1] Correspondence address: Lei Lei, School of Foreign Languages, Huazhong University of Science and Technology, Luoyu Road 1037, Wuhan, Hubei, 430074, People's Republic of China.
E-mail: leileicn@126.com

Corpora have also exerted influence over second language learning and teaching, such as the compilation of teaching syllabi, reference works, and teaching materials, etc. (Römer 2011). As for the corpus applications in teaching syllabi, the design of *Collins COBUILD English course* (Willis & Willis 1989) was innovative in that it created a new, corpus-driven lexical syllabus (Römer 2011). The use of a corpus also helped facilitate the development of teaching materials such as word lists. Based on large scale corpora, an increasing number of teaching materials of English vocabulary and lexical bundles have been developed, such as the Academic Word List (Coxhead 2000), the Medical Academic Word List (Wang et al. 2008), the Academic Formulas List (Simpson-Vlach & Ellis 2010), the Academic Collocation List (Ackermann & Chen 2013), the Academic Vocabulary List (Gardner & Davies 2014), the New General Service List (Brezina & Gablasova 2015), and the New Medical Academic Word List (Lei & Liu 2016). These materials may have greatly facilitated the learning and teaching of English.

Another area that the corpus has played its role in is translation studies. Corpus-based translation studies have come to prominence since the 1990s and have made considerable progress in fields such as the compilation of corpora for translation studies, features of translational language, translator's style, translator training, and interpreting (Hu & Mao 2012). There are many corpora which have been created for translation studies, such as the Translational English Corpus (Baker 1995), the Corpus of Translated Finnish (Mauranen 2001), Hong Lou Meng Chinese-English Parallel Corpus (Z. Liu et al. 2008), the English- Chinese Parallel Corpus of Shakespeare's Plays (Hu & Zou 2009), and the Chinese-English Conference Interpreting Corpus (Hu & Tao 2013), etc., which are widely used in translation studies, such as Hu and Tao (2013), Baker (2000), and Bowker (1998).

In addition, corpora are also commonly used in quantitative linguistics. In a study that investigates the quantitative features of articles in the *Journal of Quantitative Linguistics*, a flagship in the area, Chen and Liu (2014) found that the word *corpus* appeared amongst the top 20 most frequently occurring key words from 1998. In fact, corpora are indispensable in quantitative linguistics and have contributed to important issues in the area such as statistic investigation of language, language modelling, and language law derivation (Chen & Liu, 2014).

Although the use of corpus has dramatically changed people's view towards language and the methodology in the study of language, few studies have been conducted to systematically review the development of corpus linguistics based on empirical data. For example, Feng (2006) reviewed the history and development of corpus linguistics in China, including earlier corpora, large-scale and authentic text corpora, national corpora, speech corpora, bilingual corpora, and corpora of minority languages in China as well as different processing techniques for corpora. However, Feng (2006) is limited in that he only reviewed corpus linguistics research in China and focused more on the corpora of Chinese language and more from the perspective of natural language processing. Scott (2012) reviewed the development and relationship of corpus linguistics and other disciplines. He argued that corpora can be used in a wide range of fields, such as social sciences, arts and humanities, natural sciences, and applied sciences. Römer (2011) provided an overview of corpus research applications in second language teaching based

on the publications in the area of corpus linguistics and language teaching. However, Römer (2011) only reviewed the applications of corpora in one area, i.e. second language learning and teaching. It is obvious that the aforementioned reviews largely used qualitative methods and relied heavily on the researchers' intuition and experience.

Another line of research pertinent to the review of corpus-related research relied on the data elicited from bibliometric techniques. For example, X. Liu et al. (2014) conducted a bibliometric analysis based on corpus linguistics articles from the Chinese Social Science Citation Index (CSSCI) database between 1998 and 2013. They found that corpus-related studies in China during the examined period are mainly in areas such as corpus-based interlanguage analysis, bilingual corpus-based translation studies, and contrastive linguistics. X. Liu et al. (2014) examined 708 articles from the CSSCI database only, and the data size is relatively small. Meanwhile, the scope of the study seems also limited, i.e. corpus-related studies only in the context of China was reviewed.

Another example is Ma and Chen (2016), which used co-word techniques to visualise the major research areas and their developmental trends in the field of corpus linguistics from 1971 to 2015 based on 23,078 literature records collected from the LLBA database (Linguistics and Language Behaviour Abstracts). Ten major research areas of corpus linguistics were identified in the study. However, two points pertinent to their research methods may be controversial. The first is related to their retrieval strategy. They searched the LLBA database with "corpus OR corpora", claiming that the 23,078 records retrieved should all be related to corpus research since the literature in LLBA database are all pertinent to language studies. However, after a quick search of the database, we found that some of the literature records retrieved via "corpus OR corpora" were about "corpus callosum", which is obviously a term in medicine and unrelated to corpus research in the area of linguistics or applied linguistics. Second, due to the limitation of the data offered by the database, they did not analyse the references cited by the records they retrieved, which are in fact invaluable to identify the knowledge base of corpus research.

Thus, this study aims to use bibliometric methods to analyse corpus-related studies in linguistics. To be specific, bibliometric information of corpus-related articles in linguistics published between 2000 and 2015 was downloaded and analysed from five perspectives, i.e., the annual research outputs, the research outputs by country, publication sources, highly-cited references, and co-citation networks. Hopefully, the study would provide both the state of arts and the developing trend of corpus-related studies in linguistics.

2. Methodology

All the data were collected from the Web of Science (WoS) database. The retrieval strategy used in the study was:

((TS = corpus) OR (TS = corpora)) NOT (TS = "corpus callosum"), Timespan = 2000-2015, Indexes = SSCI.

The results were then refined by: (Web of Science Categories = Linguistics) OR (Web of Science Categories = Language Linguistics) AND (Document Types = Article) AND (Languages = English).

Two points should be noted here. First, the starting year of investigation for the study was set as 2000 since the university library of the researchers only bought the WoS SSCI database from that year. Second, we refined the research categories to "Linguistics" and "Language Linguistics", and searched keywords such as "corpus callosum" in the downloaded data to ensure that no literature unrelated to language studies was in the examined data, which addressed possible limitations of Ma and Chen (2016). The bibliometric information of a total of 3,257 articles related to corpus research were retrieved and downloaded for further analyses. First, we investigated the annual research outputs, the countries, and the publication sources of corpus-related research articles between 2000 and 2015. Then, we conducted a co-citation analysis of cited references to identify the knowledge base and research areas of corpus studies.

3. Results and discussion

In this section, we report the results of the study. We first present findings such as the general bibliometric information about annual research outputs, research contributions by countries, and the publication journals of corpus-related research. Then, we report the findings of co-citation analysis of cited documents.

3.1. Annual research outputs

The annual research outputs of corpus-related articles are presented in Figure 1. The number of articles published from 2000 to 2006 was less than 100. From 2007, the number of publications increased rapidly, and it reached 469 articles in 2015.

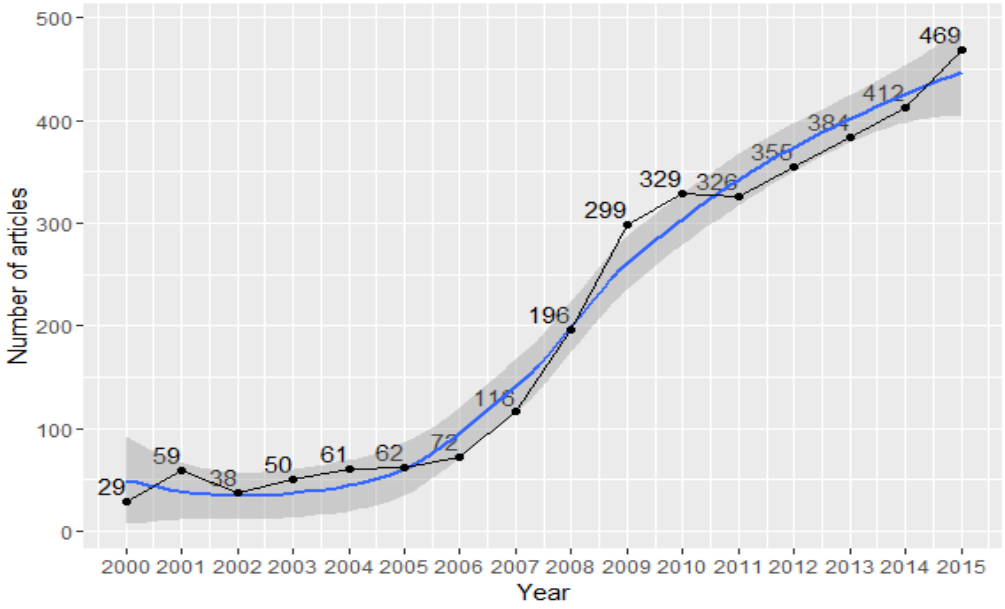


Figure 1. Annual outputs of corpus-related articles of all countries

The upward trend of corpus-related research showed the increasing concerns of corpus or corpora in linguistic studies. It is obvious that the examined studies in the present study are not all in the area of corpus linguistics. In fact, these studies spread in various areas in linguistics that are traditionally not directly pertinent to corpus linguistics, such as prag-

matics and language testing. However, research in these areas has begun to employ corpus as one of the important and most commonly used approaches to elicit research data. One of the most interesting examples is in the area of psycholinguistics, which traditionally depends heavily on behavioral experiments to elicit experiment data (e.g., Sonbul, 2015) and now tends to use corpus data to triangulate findings from behavioral experiments (e.g., Dominguez, Tracy-Ventura, Arche, Mitchell, & Myles, 2013; Vitevitch, 2012).

3.2. Research outputs by country

We list the countries whose authors published more than 30 corpus-related articles in Table 1. Researchers from the United States published 737 articles, which accounted for 22.63% of all publications. The United Kingdom ranked the second, whose researchers published 548 articles which made up 16.83% of the publications. Spain ranked the third, whose scholars published 296 articles which accounted for 9.09% of the total data collection, followed by Germany (291, 8.94%), China (269, 8.26%), and Belgium (219, 6.72%).

Table 1

Research outputs by country

Serial	Countries	No. of articles published	Percentage of all publications
1	USA	737	22.63%
2	UK	548	16.83%
3	Spain	296	9.09%
4	Germany	291	8.94%
5	China	269 (Hong Kong, 113; Mainland China, 87; Taiwan, 67; Macau, 2)	8.26%
6	Belgium	219	6.72%
7	Netherlands	120	3.68%
8	Australia	118	3.62%
9	Canada	113	3.47%
10	France	107	3.29%
11	Italy	100	3.07%
12	South Africa	88	2.70%
13	Sweden	60	1.84%
14	Japan	59	1.81%
15	New Zealand	54	1.66%
16	Finland	51	1.57%
17	Switzerland	48	1.47%
18	Norway	44	1.35%
19	Poland	41	1.26%
20	Austria	39	1.20%
21	Israel	33	1.01%

Two points are noteworthy from the foregoing findings. First, some countries seemed to be highly influential in corpus-related research in terms of their publications in the area. For example, researchers from the top six countries published approximately three quarters (72.47%) of the articles. That is, researchers in these countries had played a leading role in the area.

Second, it was also of interest to find that besides the developed countries such as the United States, the United Kingdom, and Germany, developing countries such as China, South Africa, and Poland, had begun to play their roles in this area. Take China for example. In the examined years, researchers from China published a total of 269 articles, which was responsible for 8.26% of all publications. The number of articles published from China had significantly increased during the period. See Figure 2 for the increasing trend of China's annual outputs.

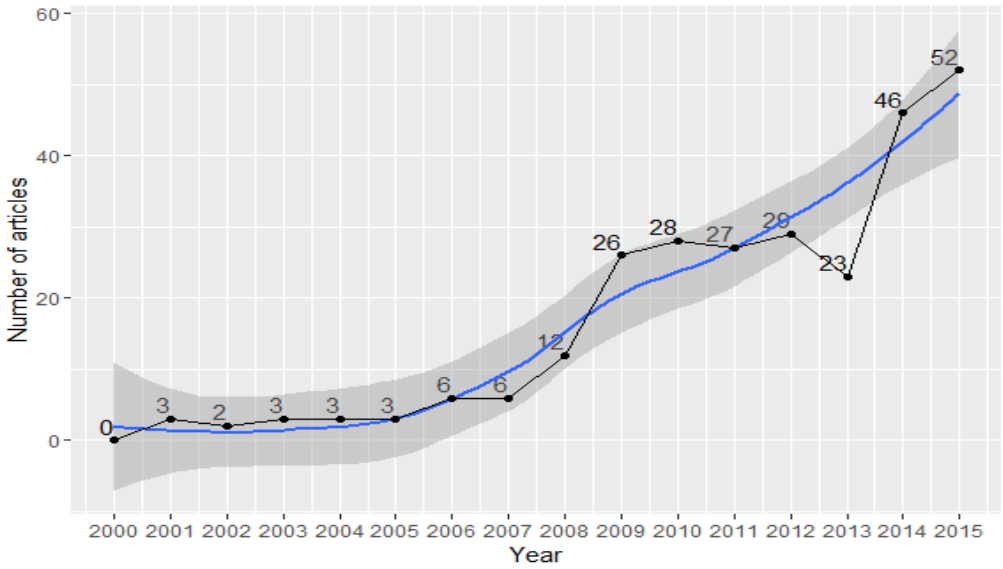


Figure 2. Annual outputs of corpus-related articles of China

Two reasons may explain China's development as one of leading countries in the area of corpus research. First, with its rapid economic development, China has attached more importance to its scientific development. Second, researchers from Hong Kong and Taiwan have made great contributions to China's research outputs in this area, which can be seen from Table 1. But except for the contributions from Hong Kong (113) and Taiwan (67), researchers from Mainland China published 87 corpus-related articles between 2000 and 2015, which could still be listed among the top prolific countries/regions in the area of corpus research over the examined period.

3.3. Publication sources

We list the journals that have published more than 30 corpus-related articles over the examined period in Table 2. *Journal of Pragmatics* published the most corpus-related articles (285) between 2000 and 2015, which account for 8.75% of all published articles.

It was followed by *International Journal of Corpus Linguistics* (133), *Computational Linguistics* (111), and *English for Specific Purposes* (111), making up 4.08%, 3.41%, and 3.41% of the published articles.

Table 2
Publications by journals*

Journal Title	**No. of articles published**	**Percentage of all publications**
Journal of Pragmatics	285	8.75 %
International Journal of Corpus Linguistics	133	4.08 %
Computational Linguistics	111	3.41 %
English for Specific Purposes	111	3.41 %
Literary and Linguistic Computing (*Digital Scholarship in the Humanities*)*	78	2.40 %
Corpus Linguistics and Linguistic Theory	74	2.27 %
Linguistics	74	2.27 %
Text & Talk	62	1.90 %
Journal of English for Academic Purposes	55	1.69 %
Applied Linguistics	52	1.60 %
Natural Language Engineering	51	1.57 %
Lingua	47	1.44 %
Journal of Quantitative Linguistics	46	1.41 %
Journal of Child Language	45	1.38 %
Language Sciences	44	1.35 %
Cognitive Linguistics	43	1.32 %
Lexikos	41	1.26 %
Pragmatics	41	1.26 %
English Language & Linguistics	40	1.23 %
Ibérica	39	1.20 %
Journal of English Linguistics	38	1.17 %
Journal of Phonetics	38	1.17 %
Language	38	1.17 %
International Journal of Lexicography	37	1.14 %
Terminology	37	1.14 %
Journal of Historical Pragmatics	35	1.08 %
English World Wide	34	1.04 %
Revista Española de Lingüística Aplicada/ Spanish Journal of Applied Linguistics	34	1.04 %
World Englishes	34	1.04 %
Poznań Studies in Contemporary Linguistics	31	0.95 %
TESOL Quarterly	31	0.95 %
System	30	0.92 %

* * *Literary and Linguistic Computing* has its title changed to *Digital Scholarship in the Humanities* since 3rd December, 2014.

It was interesting to find that *Journal of Pragmatics* published more corpus-related articles than *International Journal of Corpus Linguistics* and *Corpus Linguistics and Linguistic Theory*, two journals in corpus linguistics. Three points might explain the finding. First, the time the journals began to be listed in the WoS database was different. *Journal of Pragmatics* was indexed by WoS database before 2000, but *Corpus Linguistics and Linguistic Theory* was not indexed by WoS until 2008 and *International Journal of Corpus Linguistics* was first listed in 2009. Second, while *International Journal of Corpus Linguistics* published four issues a year since 2005 and *Corpus Linguistics and Linguistic Theory* was a semi-annual journal, *Journal of Pragmatics* published 12 or more issues each year since 1991. Third, many studies published in *Journal of Pragmatics* employed corpus-based methods to its data collection and analysis, which could be evidenced by the high percentage of corpus-related articles published in the journal (8.75%, 285 in a total of 3,257 articles in the examined period). In fact, 11.92% of the articles published in the journal were related to corpus (285 of the 2,390 published articles in total from 2000 to 2015).

In addition, this finding also served as an evidence to our previously discussed argument in Section 3.1, that is, an increasing number of articles in various areas of linguistics began to appeal to corpus as their research methods.

3.4. Highly-cited references

We present the highly-cited references with more than 50 citations in Appendix 1. Of the 42 highly-cited references, 32 are monographs, six are articles/book chapters, two are grammar books, and the remaining two are popular tools for corpus research (i.e., *WordSmith tools*, a concordancing programme, and *R*).

Interestingly, four of the highly-cited items (Scott 1996; Baayen 2008; MacWhinney 1991; R Development Core Team 2004) are pertinent to research tools, such as *WordSmith tools*, *R*, and the Child Language Data Exchange System (CHILDES) project. In particular, the inclusion of Baayen (2008), R Development Core Team (2004), and MacWhinney (1991) indicates the cross-fertilization of corpus-related research and the continuous developments of corpus tools. To be specific, Baayen (2008) focuses on the marriage of psycholinguistics and corpus linguistics, while MacWhinney (1991) reflects the combination of corpus linguistics and studies of child language acquisition.

The increasingly significant role of R should also be noted. *R* is “a language and environment for statistical computing and graphics” (R Development Core Team 2004), which provides tools and packages that facilitate statistical analyses and corpus data processing, such as measuring texts from the perspectives of readability and lexical diversity (Mizumoto & Plonsky 2016: 288). The wide use of *R* for corpus research may reflect important trends of methodology in the area of linguistics. On the one hand, researchers in the area traditionally heavily depended on statistic tools such as IBM SPSS, and they began to appeal to more robust tools such as *R* for statistical analyses. On the other hand, researchers may not be satisfied to be bound to the functionality of corpus tools such as *WordSmith tools* (Scott 1996) and *AntConc* (Anthony 2014) and they tended to use programming tools such as *R* for more personalised analyses of corpus data.

3.5. Co-citation networks of cited references

In this section, we report on the results of co-citation analysis of the cited references with VOSviewer (Van Eck & Waltman 2010). The procedure of the co-citation analysis is to be described as follows. First, we set the minimum number of citations of the cited references to 10. Of the 93,244 cited documents, 1,081 met the threshold. Then, we tried out different values for the parameter "resolution", which "determines the level of detail provided by the VOS clustering technique" (van Eck & Waltman 2016: 18), until we found the value of 0.80 most satisfactory. As a result, five clusters for the co-citation model were found, and the co-citation networks of the cited references was plotted and reported in Figure 3.

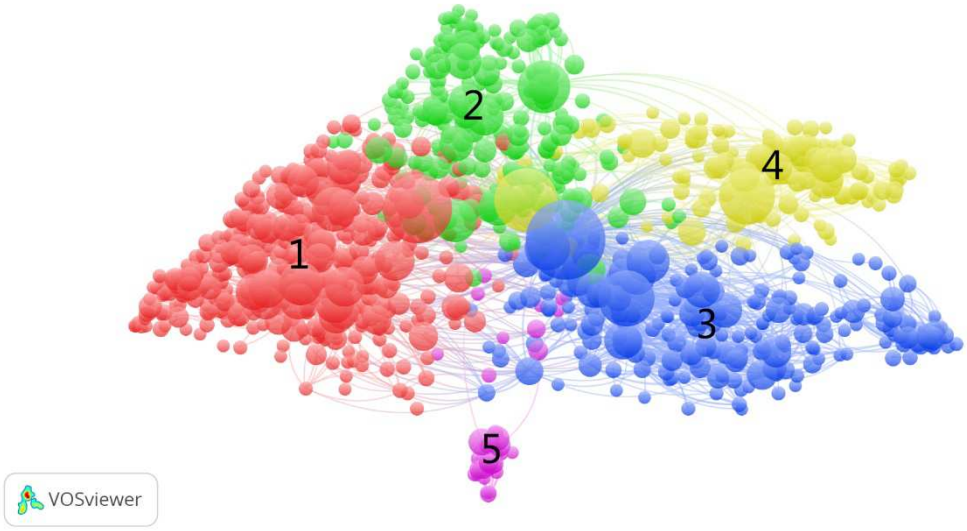


Figure 3. Co-citation networks of cited references

The detailed information for each cluster is reported in Appendix 2. The first column lists the cluster number and the number of documents each cluster contains. The second column provides the details of the 10 most highly-cited references for each cluster. The third column lists the number of citations of the corresponding items.

Cluster #1, the largest cluster, was denoted by the red colour and contained literature relevant to English grammar, cognitive linguistics (e.g., Lakoff 1987; Langacker 1987; Goldberg 1995, 2006), especially construction grammar (Goldberg 1995, 2006), and research tools (Baayen 2008; R Development Core Team 2004; MacWhinney 1991).

The second cluster, denoted by the green colour, was characterised by topics in the area of pragmatics, such as politeness (Brown & Levinson 1987), discourse markers (Schiffrin 1987), and conversation (Sacks et al. 1974; Grice 1975; Sacks & Jefferson 1992).

Cluster #3, denoted by the blue colour, consisted of publications mainly pertinent to phraseology (Sinclair 1991), which ranges from words (e.g., Hoey 2005) and lexical bundles (e.g., Biber et al. 2004; Wray 2002) to registers (e.g., Biber 1988), and texts (e.g., Sinclair 2004).

In fact, the most highly-cited document in this cluster (i.e. Biber et al. 1999) was also

the most highly-cited work in the whole dataset. In contrast to traditional grammar books such as Quirk et al.'s (1985), the most highly-cited in Cluster 1, Biber et al.'s (1999) grammar book has stood out and gained its popularity since it may be the first grammar book compiled largely based on empirical analyses of corpus data and accordingly provide detailed descriptions of the language (Hirst 2001). Biber et al. (1999) was assigned to this cluster, probably due to the fact that it spends much of its space in analysing phraseological items such as lexical bundles and has become one of the classic literatures in the area.

The fourth cluster, denoted by the yellow colour, focused on studies of academic English. Important topics such as systemic functional grammar (Halliday 1985, Martin & White 2005), genre analysis (Swales 1990, 2004; Bhatia 1993; Berkenkotter & Huckin 1995), and metadiscourse analysis (Hyland 1998, 2000, 2005) were included in the cluster.

Cluster #5, the purple-coloured cluster, focused on translation studies. The literature in this cluster indicated that Mona Baker (Baker 1993, 1995, 1996; Olohan & Baker 2000) played an important role in corpus-based translation studies. This cluster was also the smallest cluster, which seemingly showed that corpus-based translation studies was relatively young compared to other corpus-based research areas.

4. Conclusion

This study investigated the corpus-related studies in linguistics from 2000 to 2015. Four points of interest were found. First, the output of corpus-related publications has significantly increased in the past 15 years. The upward trend seems to indicate the increasingly important role of corpus as a research tool in the discipline of linguistics. This point was also supported by the finding that *Journal of Pragmatics*, a journal specialised in pragmatics, published more corpus-related articles than journals in corpus linguistics such as *International Journal of Corpus Linguistics* and *Corpus Linguistics and Linguistic Theory*.

Second, traditional scientific powers such as the United States, the United Kingdom, and Germany have played leading roles in the area from the perspective of the number of outputs. However, developing countries such as China, South Africa, and Poland have also begun to exert their impact in the area.

Third, corpus has been indispensable in many areas of linguistics research, particularly as a tool to facilitate data elicitation and analyses. Also, corpus has seemingly changed the ecology of methodology in linguistics research, which was evidenced by the fact that programming tools such as *R* has become increasingly popular in linguistics.

Last, the study also shows that the influence of corpus has spread over research areas such as English grammar, cognitive linguistics, pragmatics, phraseology, academic English, and translation studies. It is of particular interest to find that areas that traditionally largely relies on researchers' inspections such as English grammar, cognitive linguistics, pragmatics, and translation studies has also begun to embrace and use corpus as its methods.

To summarise, corpus has become increasingly popular for linguistics research in the

past fifteen years amongst researchers from both traditional scientific powers such as the United States and emerging economies such as China. In fact, corpus has permeated a wide range of areas in linguistics, and has already exerted much impact on these areas, and seemingly has changed, at least in terms of methodology, these areas.

The data used in the study is limited to the bibliometric information retrieved from the WoS SSCI database. Future research may employ data from other databases to verify the findings of this study. In addition, corpus has also been widely used in other areas or disciplines than linguistics, such as communication, natural language processing, and computer science. Thus, investigation in such areas may explore implications of corpus in such areas.

Acknowledgements

This work was supported by the National Social Science Fund of China (Grant No. 15BYY179).

References

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes 12(4), 235-247*.

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/. Accessed 6 April 2016.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair: 233-250*. Amsterdam: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target 7(2), 223-243*.

Baker, M. (1996). Corpus-based Translation Studies: The challenges that lie ahead. In H. Somers (eds.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager: 175-187*. Amsterdam: John Benjamins.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target 12(2), 241-266*.

Berkenkotter, C., & Huckin, T. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Lawrence Erlbaum.

Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics 25(3), 371-405*.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bowker, L. (1998). Using specialized monolingual native-language corpora as a translation resource: A pilot study. *Meta: Translators' Journal 43(4), 631-651*.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the *New General Service List*. *Applied Linguistics 36(1), 1-22*.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. New York: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34(2), 213-238*.

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

Dominguez, L., Tracy-Ventura, N., Arche, M. J., Mitchell, R., & Myles, F. (2013). The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism-Language and Cognition 16(3), 558-577*.

Feng, Z. (2006). Evolution and present situation of corpus research in China. *International Journal of Corpus Linguistics 11(2), 173-207*.

Gardner, D., & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics 35(3), 305-327*.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and semantics, vol. 3: Speech acts: 41-58*. New York: Academic Press.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography 25(4), 398-436*.

Hardie, A., & McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics 15(3), 384-394*.

Hirst, G. (2001). Review of *Longman grammar of spoken and written English*. *Computational Linguistics 27(1), 132-139*.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hu, K., & Mao, P. (2012). Corpus translation studies abroad: A critical review. *Contemporary Linguistics 14(4), 380-395+437-438 [in Chinese]*.

Hu, K., & Tao, Q. (2013). The Chinese-English Conference Interpreting Corpus: Uses and limitations. *Meta: Translators' Journal 58(3), 626-642*.

Hu, K., & Zou, S. (2009). The compilation and application of the English-Chinese parallel corpus of Shakespeare's plays. *Foreign Languages Research (5), 64-71+112 [in Chinese]*.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Langacker, R. (1987). *Foundations of cognitive grammar vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes 22, 42-53*.

Liu, X., Xu, J., & Liu, L. (2014). An overview of corpus linguistics research in China (1998-2013): A CiteSpace based analysis. *Corpus Linguistics 1(1), 69-77+112 [in Chinese]*.

Liu, Z., Tian, L., & Liu, C. (2008). The compilation of Hong Lou Meng Chinese-English parallel corpus. *Contemporary Linguistics 10(4), 329-339+379-380 [in Chinese]*.

Ma, X., & Chen, Y. (2016). Mapping the intellectual structure of corpus linguistics: A co-word analysis (1971-2015). *Corpus Linguistics 3(1), 41-54+116 [in Chinese]*.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

Martin, J., & White, P. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan.

Mauranen, A. (2001). Strange strings in translated language: A study on corpora. In M. Olohan (eds.), *Intercultural faultlines. Research models in translation studies, vol. 1: Textual and cognitive aspects: 119-142*. London: Routledge.

McEnery, T., & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts & A. McMahon (eds.), *The handbook of English linguistics: 33-71*. Oxford: Blackwell Publishing.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mizumoto, A., & Plonsky, L. (2016). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics 37(2), 284-291*.

Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures 1(2), 141-158*.

Pearsall, J. (eds.). (2000). *Concise Oxford dictionary* (Tenth edition). Oxford: Oxford University Press.

Pearsall, J., Hanks, P., et al. (eds.). (1998). *The new Oxford dictionary of English* (First edition; Second edition published in 2003, as *Oxford dictionary of English*). Oxford: Oxford University Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

R Development Core Team. (2004). *R: A language and environment for statistical computing*. http://www.R-project.org.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics 31, 205-225*.

Sacks, H., & Jefferson, G. (1992). *Lectures on conversation, vol. 2*. Cambridge: Blackwell.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language 50(4), 696-735*.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Scott, M. (1996). *WordSmith tools*. Oxford, UK: Oxford University Press.

Scott, M. (2012). Looking back or looking forward in corpus linguistics: What can the last 20 years suggest about the next? *Ibérica 24, 75-86*.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics 31(4), 487-512*.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, J., Hanks, P., et al. (eds.). (1987). *Collins COBUILD English language dictionary* (First edition). London: Collins.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism-Language and Cognition 18(3), 419-437*.

Summers, D., et al. (eds.). (1987). *Longman dictionary of contemporary English* (Second edition; Third edition 1995). London: Longman.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics 84(2), 523-538*.

Van Eck, N. J., & Waltman, L. (2016). *Manual for VOSviewer version 1.6.4*. http://www.vosviewer.com/download. Accessed 6 April 2016.

Vitevitch, M. S. (2012). What do foreign neighbors say about the mental lexicon? *Bilingualism-Language and Cognition 15(1), 167-172*.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes 27(4), 442-458*.

Willis, D., & Willis, J. (1989). *Collins COBUILD English course*. London: HarperCollins.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Appendix 1

Table 3
Highly-cited references with more than 50 citations*

Cited References	Citations
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.	377
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.	278
Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.	239
Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.	191
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.	190
Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. New York: Cambridge University Press.	165
Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.	124
Huddleston, R., & Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.	119
Scott, M. (1996). *WordSmith tools*. Oxford, UK: Oxford University Press.	114
Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.	113
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.	110
Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.	97
Langacker, R. (1987). *Foundations of cognitive grammar vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.	97
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.	88
Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.	87

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.	86
Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.	85
Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.	85
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.	84
Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.	81
R Development Core Team. (2004). *R: A language and environment for statistical computing*. http://www.R-project.org.	80
Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language 50(4), 696-735*.	78
MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Erlbaum.	77
Levinson, S. (1983). *Pragmatics* (First edition). Cambridge: Cambridge University Press.	75
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Boston: Harvard University Press.	73
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics 25(3), 371-405*.	72
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.	72
Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.	72
Swales, J. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.	70
Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and semantics, vol. 3: Speech acts: 41-58*. New York: Academic Press.	67
Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.	67
Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.	67
Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.	66

Stefanowitsch, A., & Gries, S. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics 8(2), 209-243*.	64
Hopper, P., & Traugott, E. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.	61
Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. London: Longman.	57
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34(2), 213-238*.	57
Stubbs, M. (1996). *Text and corpus analysis: Computer assisted studies of language and culture*. Oxford: Blackwell.	57
Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.	56
Leech, G. (1983). *Principles of pragmatics*. London: Longman.	56
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.	56
Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (eds.), *Language and communication: 191-226*. London: Longman.	56

* Different versions of *An introduction to functional grammar*, *WordSmith tools*, *R*, and *The CHILDES project: Tools for analyzing talk* have been merged into one reference and represented by the first version respectively.

Appendix 2

Table 4
Detailed information for clusters

Cluster	Representative cited literature	Citations
Cluster #1 (461 items)	Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London and New York: Longman.	278
	Huddleston, R., & Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.	119
	Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.	113

	Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.	97
	Langacker, R. (1987). *Foundations of cognitive grammar vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.	97
	Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.	87
	Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.	85
	R Development Core Team. (2004). *R: A language and environment for statistical computing*. http://www.R-project.org.	80
	MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.	77
	Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Boston: Harvard University Press.	73
Cluster #2 (236 items)	Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. New York: Cambridge University Press.	165
	Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.	110
	Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.	88
	Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.	85
	Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language 50(4), 696-735*.	78
	Levinson, S. (1983). *Pragmatics* (First edition). Cambridge: Cambridge University Press.	75
	Grice, H. 1975. Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and semantics, vol. 3: Speech acts: 41-58*. New York: Academic Press.	67
	Stubbs, M. 1996. *Text and corpus analysis: Computer assisted studies of language and culture*. Oxford: Blackwell.	57
	Leech, G. (1983). *Principles of pragmatics*. London: Longman.	56
	Sacks, H., & Jefferson, G. (1992). *Lectures on conversation, vol. 2*. Cambridge: Blackwell.	51
Cluster #3 (220	Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.	377

items)	Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.	190
	Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.	124
	Scott, M. (1996). *WordSmith tools*. Oxford, UK: Oxford University Press.	114
	Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.	86
	Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.	84
	Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.	81
	Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics 25(3), 371-405*.	72
	Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.	72
	Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.	66
Cluster #4 (131 items)	Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.	239
	Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.	191
	Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.	72
	Swales, J. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.	70
	Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. London: Longman.	57
	Martin, J., & White, P. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan.	53
	Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.	52
	Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.	47
	Berkenkotter, C., & Huckin, T. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Lawrence Erlbaum.	46
	Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.	46

Cluster #5 (33 items)	Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair: 233-250*. Amsterdam: John Benjamins.	45
	Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.	38
	Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target 7(2), 223-243*.	31
	Bowker, L., & Pearson, J. (2002). *Working with specialized languages: A practical guide to using corpora*. London: Routledge.	25
	Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.	25
	Olohan, M. (2004). *Introducing corpora in translation studies*. London: Routledge.	24
	Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing 8(4), 243-257*.	22
	Kenny, D. (2001). *Lexis and creativity in translation: A corpus-based study*. Manchester: St. Jerome.	22
	Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (eds.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager: 175-187*. Amsterdam: John Benjamins.	21
	Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures 1(2), 141-158*.	21

Glottometrics 38, 2017, 21-29

Lexicographic Problems of Collocations in Quantitative Linguistics

Fabienne Petrack (University of Trier)[1]
Hanna Gnatchuk (Alpen-Adria University/University of Trier)[2]

Abstract: The present article deals with an overview of the problem of collocations in quantitative linguistics. The aim of the article is to highlight problematic aspects of collocations which can be solved with the help of statistical methods. The main focus is on the overview of the research on collations carried out by Levitskij and under his supervision as well as his remarks on the analyzed problem.

Key words: *collocate, compatibility, word collocation, intensity/width of word compatibility, chi-squared test*.

1. Introduction: Some notes on lexicography

In the third to second century BCE, people of India would help their descendants to understand ancient texts by writing glossaries. These writing would eventually "evolve into what we would recognize as dictionaries today" (Halliday, 2004: 11). But not only Indians used dictionaries as a means of understanding texts, the Greeks would eventually write dictionaries using the works of Homer. In Chinese was the Amera Kosha, in Sanskrit the Abhidhana Kintamani and the Desinamamala; and during the Ming dynasty, giant books of up to 50,000 characters would appear; those books "took its place as part of the systematic description of language" (Halliday, 2004: 12).

But with the growth of the earth´s population and the time of exploration, monolingual dictionaries were not enough anymore. In the 11th century the Persians seemed to already be aware of these problems so they wrote one of the first bilingual dictionaries beneficial of learning Arabic and later Turkish. With the rise of education in Europe, bilingual dictionaries grew in popularity, especially in schools, where Latin was taught. In England, those glossaries would contain difficult Latin words. These lists of words with their Latin explanation would later be called "vocabulary" (Halliday, 2004: 14).

In 1440, Johannes Gutenberg invented the printing press and thus the "publish[ing] of dictionaries grew rapidly" (Halliday, 2004: 14). By now, these books would be printed in alphabetical order to insure an easier finding. Also, citations of popular works would begin to be used to give examples of the use of specific words.

[1] Fabienne Petrack. University Trier, Computational Linguistics and Digital Humanities, Universitätsring 15, s2fapetr@uni-trier.de
[2] Hanna Gnatchuk, MA. Alpen-Adria University, Institute of Slavic Studies, Universitätsstraße 65-67, Klagenfurt, Austria, hgnatchu@edu.aau.at

Only in 1604, the first monolingual dictionary was written in England. In Germany, the Grimm Brothers wrote the first dictionary in 1838. Besides the use of alphabetical order and citations, etymology, the study of the origins of the words would now also be a part (cf. Halliday, 2004: 14 ff./ Publikationsverfahren in den Geisteswissenschaften 2017).

To make a dictionary, one first must be aware that a word is not as easily defined as one might think. Every person capable of reading and writing knows what a word is, but where does a word start and end? Many would say: if there is a space between an alignment of letters, it's a word! But what about hyphens in between two aligned set of letters, would that be one word or two? Would one consider conjugated verbs all as the same word or should they be separate words? In linguistics it has been a controversial topic. In most dictionaries hyphened words are separate words or in a sub-group underneath a specific main word and conjugated words are considered as Lemmas, "a glossed word or phrase" (Merriam Webster 2017). The situation is still more complex in languages which do not use spaces between words, e.g. Chinese.

As we can already see in the definition, the word is only used in context with classifying words. The "lemma […] will serve as headwords in the dictionary" preventing the reader to have an overflow of information (cf. Garret, E et al. 2015: 55; Halliday et al. 2004: 1 ff.). The words collected in a dictionary usually come from various sources: books, newspapers, magazines, and transcripts to oral texts. In the end, the more sources a linguist can take into consideration, the better the dictionary, especially since it gives the linguist a greater range. Today, it has become much easier to write a dictionary inasmuch as the existence of online corpora. A corpus is any kind of text; Online corpora are databases with hundreds to thousands of texts. Often these texts are annotated so the linguist does not have to make the list of words himself (cf. Biber, 1998: 21 ff.). The annotation consists of: the lemma, the part of speech and the source of the word, therefore the citation. This way he is already half way through his entry, which, according to Halliday (2004: 5) normally consists of:

1. The headword or lemma, often in bold or some other special font;
2. Its pronunciation, in some form of alphabetic notation;
3. Its word class ("part of speech");
4. Its etymology (historical origin and derivation)
5. Its definition;
6. Citations (examples of its use).

Now, all that is missing is the pronunciation, which a phonologist can provide; etymology, which a historical linguist contributes and the definition. The definition is influenced by the word class, for which the citation in which the word was used is also important. Looking at the citation, one needs to examine the other words, called collocation. In many cases, one word is often written along with similar words in different sentences. Analyzing these can implement the word class and therefore the definition. It can also be important when choosing the most representative citation, so that foreign language speakers find it easier to use the word.

2. The problems of collocations in lexicography

2.1. General remarks on the nature of collocation in a dictionary/corpus

Looking at collocations, Biber et al. (1998) chooses the word "deal" to take a closer look at the frequency this word is used in which part of speech. When we look

"deal" up in the Longman Dictionary of Contemporary English, we can find 10 different entries. There is *deal* as a noun meaning "quantity or degree" (Procter, 1978: 282), or "the act of giving out cards" (Procter 1978: 282); it can be written as a verb "to give as one's share" (Procter, 1978: 282), or "to give out playing cards" etc. Online, one can find even more occurrences and definitions of the word. But how can one figure out which are the most representative uses?

Biber et al. (1998) compared frequencies of words used in the TACT corpus to enhance the understanding. While words such as "a" have a frequency of 478, "deal" seems less common, though there are more entries since there is not just one lemma. But does this classify as common or rare? In other corpora, one might find more occurrences while there are less in others. (cf. Biber et al. 1998: 28 ff.) It is important to be aware of what type of corpus one is dealing with: The size and the sources that are used. To make it clearer and more comparable, one needs to norm the counts, see Table 1.

Table 1
Normed count (source: Biber et al. 1998: 29)

Deal……182

Dealing…….....52

Deals…….25

Dealt……..31

Table 2

Frequencies of "Deal" (source: Biber et al. 1998: 29)

	Approx. number of words in sample	**Raw counts**		**Normed per million**	
		Noun	Verb	Noun	Verb
Total Sample	4.000.000	366	482	90	119
Fiction	2.000.000	217	127	107	63
Academic prose	2.000.000	149	355	74	176

With Table 1, Biber et al. (1998) concludes, that *deal* as a verb is more common than it is as a noun. Yet, comparing fiction and academic prose, the noun is more common in fiction and the verb is more common in the academic style. Therefore, Biber et al. makes it clear that "overall generalizations about English are often misleading" (Biber et al. 1998: 35). But we still could not answer the question whether the noun or the verb is more common and what is more representative for the dictionary.

To answer this, Biber et al. (1998) turns to the collocations of the word. He argues that looking at all the uses of *deal* might be challenging and to simplify the task, he first looks at *deal* as a noun. In the paper, he only looks at the left collocate, the one word before *deal* and the right collocate, the word after *deal*. The most common left collocates that he finds are adjectives such as *great, good* and *big*. To the right, he finds prepositions: *of, to, in, about*, etc. In Prose, the most common uses are with the left collocate *great* and *good.* Thus, the most common use of the noun in prose is "referring

to an amount" (Biber et al. 1998: 37). This coincides with the right collocate, which is, with a huge gap to the second most used, the preposition *of* (cf. Biber et al. 1998: 37 ff).

Using *deal* in academic prose, *great* and *good* are again the most used collocates and the right ones are *of, more, in* and *to*. Taking a closer look at some of the other right collocates, though, Biber et al. (1998) realizes, that the noun is used in the same sense that it is in prose: referring to an amount. Interestingly enough, both genres use the same noun most commonly. Yet we can spot differences when moving on to the less commonly used. Here, we find that the third most used collocate in prose, referring to a business transaction. This becomes clearer when Biber et al. (1998) further analyzes the less frequent collocates: "Many collocates that are not frequent individual are associated with […] a business transaction. […] together they show another important sense" (Biber et al. 1998: 38).

Another use of *deal*, referring to a type of word, can be found not so often but they "come from five different texts, showing that its use is relatively widespread" (Biber et al. 1998: 39). One can argue that this use is of importance or not. Different dictionaries have contrary views on this subject. Some have several definitions "var[ying] from two or three to twenty or thirty [entries]" (Biber et al. 1998: 39). For the user of the dictionary, especially a foreign language user, it is difficult to tell if speaking of an amount or speaking of a type of word is the most normal use of the word. When looking back at the entries of the Longman Dictionary, one of the definitions was also the distribution of cards. This would only be of importance when playing cards, yet this is a common entry in dictionaries. None of the texts Biber et al. (1998) has dealt with any indication of the use of this entry. Biber et al. (1998) highlights with this example, that sometimes "intuitions should complement corpus-based analysis" (Biber et al. 1998: 41).

But sometimes intuition is not enough. When looking through dictionaries one can often also find IDIOM or PHRASEME behind a word or sentence in the dictionary. Idioms or phrasemes are "the language peculiar to a people or to a district, community, or class" (Merriam Webster 2017). These usually have some kind of historical or political backgrounds.

2.2. The problems of collocations in quantitative linguistics

In the present article we deal with the problem of the ability of a word to be combined with other words in a certain corpus of texts (Agricola, 1975: 50). This phenomenon is known in linguistics as *lexical compatibility*. In general, Schippan (1987: 207-210) distinguishes three types of compatibility:

a) Denotative compatibility: i.e. the German word *bellen* is compatible with *Hund*, but it is not compatible with *Katze*;
b) Pragmatic compatibility presupposes the compatibility of communicative situations: i.e. *edel + Antlitz* is compatible, but *Fratze* and *Visage* are not compatible with *edel*);
c) Lexical compatibility: i.e. the compatibility of German *stark* and *kräftig* with certain words *Mensch, Tee, Kälte, Tier*).

According to Levic`kij (2012) different words in a text or a corpus of texts can be combined with an unequal number of "partners" (= words). The number of these "partners" can be called *the width of a word's compatibility*. One can express this width by means of absolute values, ranks or relative values. The number of a word's partners in a corpus can vary from 1 (a word is combined with only one word) to a great number

of values. Levic`kij (2012) states that a great number of "partners" are characteristic of the word with a wide semantics (i.e. good (Vendler, 1967)) or with a considerable number of meanings. This hypothesis has been confirmed in the works by Bistrova (1978), Moskovic` (1969), Ginka (1983). Nevertheless, there is a lack of the studies on the factors influencing the width of a word's compatibility.

Another parameter of word compatibility is its *intensity* – the degree of connection of the collocates. Porzig (1934) was the first to pay attention to this problem. Lyons ((1977), Mel´c`uk (1960) were engaged with this problem. According to Levickij (2012) the degree of compatibility (intensity) can be studied by means of psycholinguistic and contextual methods. In this case Dridze (1971) investigated 50 collocates of the Russian language of the type *вносить вклад* (Germ. *Beitrag leisten*, Engl. *to make a contribution*) using an associative experiment. The task of the informants was to name a missing component for a word *вносить* (Germ. *leisten*, Engl. *to make*). The statistical analysis of the frequencies of associative words (which were named by informants) has shown that there is a certain degree of fixedness of words in the analyzed word combinations. Dridze measured this degree by means of a relation of the number of informants, who name a certain word, to a number of all informants. For example, 28 informants out of 50 name a Russian word *вклад* (Germ. *Beitrag*, Engl. *contribution*) as a component of a verb *вносить* (Germ. *leisten*, Engl. *to make*). Similar experiments have been conducted in Ukrainian by Levickij (1975). The results of his research were in favor of the existence of lexical compatibility. For example, the Ukrainian adjective *сильний* (Engl. *strong*) is not compatible lexically with a word *чай* (Engl. *tea*): *сильний* ≠ *чай*, but *міцний* + *чай* (*strong tea*).

The study of the intensity of word compatibility (= collocates) can be undertaken by means of the frequencies of these collocates in a text. The measure unit of intensity can be represented by an absolute or relative value showing the repetition of the "partners" of words in the text. For instance, a word *read* is compatible with *book* in English texts (6,1 % of all combinations of *read*), Bible (3%), novel (1,5 %) (Miller, 1971). Nevertheless, Levic`kij (2012) considers that the highest frequency of *read* + *book* can be caused by either the highest frequency of *read* or *book*. In this case, it is relevant to compare theoretically the expected frequencies of a word collocation with empirical frequencies. This method is based on the occurrences of two words in a certain fragment of text.

The deviation of empirical frequencies from theoretically expected can be measured with the help of chi-squared test or coefficient of contingency. Tuldava (1988: 56) analyzed the distribution using contingency coefficient. In this case, Levic`kij (2012) points out that this research needs some correction, namely normalization of the coefficient. He states that the first experiments with chi-squared test and contingency coefficient turned out to be effective in order to determine the degree of word compatibility.

Using these methods, the compatibility of English adjectives *strong* and *weak*, English and German adjectives of qualities (i.e. *good*, *bad*), English and German adjectives denoting the appearance, the nouns with the meaning "Arbeit" in German, the verbs of speaking (*speak, tell, talk, utter, etc*) has been studied under Levitskij's supervision.

The frequencies of collocations are given in these investigations in alternative tables which have been treated by Bernoulli coefficient (1.1):

$$\Phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (1.1)$$

The significance of a coefficient is determined by means of chi-square. The significance is accepted at $X^2 = 3.84$ (with 1 DF); $p = 0.05$.

The data about the compatibility of words A and B are placed in the field (a) of the alternative table (Table 3, 4). Only positive coefficient Φ can be considered. This type of processing the data does not show that the highest frequencies of compatibility of two words are the evidence of the intensity of their semantic connection. For instance, a word *strong* is often combined with *man* in English prose, but the coefficient Φ for this pair does not have a statistical significance.

Table 3

The distribution of frequencies of collocate *strong + man* (Levitskij, 2012 : 255)

	Man	**Other nouns**	**Total**
Strong	27 (a)	637 (b)	664
Other adjectives	81 (c)	1674 (d)	1755
Total	108	2311	2419

$X^2 = 0.34$; Φ = -0.01 no significance.

Table 4

The distribution of frequencies of collocate *stout + man* (Levitskij, 2012 : 255)

	Man	**Other nouns**	**Total**
Stout	18 (a)	50 (b)	68
Other adjectives	90 (c)	2261 (d)	2351
Total	108	2311	2419

$X^2 = 79.44$; Φ = + 0.18; P = 0.001

In contrast, the lower frequencies can be more intensive: *stout + man* (Φ = 0.29), *feeble + glow* (0.2), *firm + lip* (0.15), *firm + hand* (0.11) (Bistrova, 1978).

The above enumerated collocates are designated by Agricola (1975) as *lose Wortverbindungen*. Levic`kij (1989: 48-49) suggests naming these collocations *stable collocates*. These stable word combinations coincide with the so-called typical word combinations in the Explanatory Dictionary.

Levic`kij (1989: 50-53) and Zadoroz`na (2003) studied the combination of words belonging to the human body and animals (i.e. *Kopf schütteln*, *Mund halten*). They have noticed that there are some word combinations in which the components/words are statistically significant $X^2 > 3.84$, but they are not fixed in the phraseological dictionary of German as a phraseological unit (i.e. *Ohren zuhalten*). Moreover, the collocations with the sums lower than 3.84 (no statistical significance) occur in the text and they are fixed in the phraseological dictionary as a fixed collocation (i.e. *Mund halten*). Levic`kij (2012) explains it by the incompleteness of the compilation of the dictionary or a sample for the dictionary is not enough.

In such a way, Levic`kij (2012) makes a conclusion that the compatible intensity of a word can be revealed by means of statistical methods. Although no proportional coincidence between a statistical value of Φ in a text and an intuitive determination of the degree of collocation in the dictionary has been revealed. The highest degree of compatibility of two elements can never be accidental: stable word combination is interpreted in lexicography as typical one. The possibility of delimitating the units

having different degrees of phraseological connection by means of statistical methods requires further study.

2.3. Empirical study of English word combinations

In this research we have decided to check the degree of combinations of English words "look", "glance", "stare" and "gaze" with adjectives as a modifier. We want to reveal whether there is a statistically significant connection between the above-mentioned English words in the Corpus of American English (COCA) and the fixedness of these collocations in the BBI Dictionary of English Word Combinations (1997).

The material and the procedure of the research. The COCA is a corpus which includes 520 million words taken from newspapers, academic articles, magazines, colloquial speeches and fictions. By means of COCA we were able to find the frequencies of word combinations. Let us consider a word "look", In order to find the possible combinations of "look" and the frequencies of these collocations, we gave the following request *[j*] look*. This means that [j*] is any adjective before *look*. In such a way, we have received the following frequencies: *closer look = 2370, good look = 1100, quick look = 987*, etc. In the BBI Dictionary of English Word Combinations (1997) we have found the following combinations of *look* which are considered as stable ones: *inquiring + look, searching + look, adoring + look, loving + look.* The corpus also contains the frequencies of these combinations, although they do not have a high frequency. In order to reveal whether these combinations "searching, adoring, loving + look) are statistically significant, we perform a chi-squared test (at http://www.socscistatistics.com/tests/chisquare2/Default2.aspx). Let us test the combination of "inquiring + look". Before testing it, we must make a contingency table with the following data:

Table 5

The distribution of frequencies of collocate *inquiring + look*

	look	**Other nouns**	**Total**
inquiring	10 (a)	79 (b)	89
Other adjectives	4457 (c)	48546 (d)	53003
Total	4467	48625	53092

(a) Contains the frequencies of *inquiring + look* in the corpus
(b) Contains the frequencies of adjectives *inquiring* with other nouns: i.e. inquiring minds (42), mind (28), glance (9);
(c) Contains the combination of *look* with other adjectives (closer 2370, good = 1100, quick = 987);
(d) Contains the combinations of other adjectives (which were combined with *look* in (c)) with other nouns (except *look*)

The data in Table 4 are processed with the help of chi-square calculator with p = 99%. $X^2_{0.001} = 0.591$ at df = 1 which is not significant for the corpus data, although it is fixed in the BBI Dictionary of English Word Combinations (1997). Nevertheless, the combinations of *searching + look* ($X^2 = 178.103$), *adoring * look* ($X^2 = 7.181$) and *loving + look* ($X^2 = 12.279$) have turned out to be statistically significant. They are also fixed in the BBI Dictionary of English Word Combinations.

The distribution of a word *glance* proved to be statistically significant together with an adjective *casual* ($X^2 = 134.7$). Although the combinations *amused* ($X^2 =$

0.7321*), wistful* (X^2 = 3.574), *imploring* (X^2 = 1.447) + *glance* do not have statistical significance having been fixed in the analyzed dictionary.

The distribution of *icy* (X^2 = 193.61) and *vacant* (X^2 = 96.53) + *stare* is extremely significant. Only *fixed* + *stare* is devoid of statistical significance in the corpus, although it is found in the dictionary.

The last distribution concerns the word "gaze" which has turned out to be significant with adjectives *steady* (X^2 = 24.6691) and *piercing* (X^2 = 130.0724). Only *intense* + *gaze* (X^2 = 2.0511) is not significant.

In such a way, our small-scaled study supports the conclusion that the combinations of words can be statistically significant in the corpus of 520 million of words, but they are not fixed in the dictionary of word combinations. Therefore, the further study on the given topic is required with the application of a large number of words which are to be combined.

References

Agricola, E. (1975). *Semantische Relationen im Text und im System*. Halle (Saale): Max Niemeyer Verlag

Benson, M., Benson, E., Ilson, R. (1997). *The BBI Dictionary of English word combinations.* John Benjamins Publishing Company.

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge.

Bistrova, L.V.(1978). Vyvchenn'a syntagmatychnyh zvjazkiv za dopomohoju statystychnyh metodiv. *Movoznavstvo, 44-48.*

Corpus of Contemporary American English. http://corpus.byu.edu/coca/

Dridze, T.A. (1971). Assotsiativnij eksperiment v konkretnom sotsiologicheskom issledovanii. In: *Semanticheskaja struktura slova.* Moskva: Nauka.

Garret, E., Hill, N.W., Kilgarriff, A., Vadlapudi, R., Zadoks, A. (2015). *The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries.* London.

Halliday, M.A.K., Teubert, W., Yallop, C., Cermáková, A. (2004). *Lexicology and Corpus Linguistics. An Introduction*. London.

Ginka, B.I. (1983). Leksiko-semanticheskaja gruppa sushchestvitel'nyh so znachenijem „trud, rabota" v sovremennom nemetskom jaz'ke. Odessa: AKD.

Levickij, V.V. (1975). Problem' eksperimentalnoj semasiologii**.** Kiev: ADD.

Levickij, V.V. (1989). Do metodyky inventyryzatsiji leksychnyh mikrosystem za dopomohoju slovnykovyh definitsij. *Movoznavstvo,* 1: 21-26.

Levickij, V.V. (2012). *Semasiologija*. Nova Knyha.

Lyons, J. (1977). *Semantics*. Vol. 1-2. London; New York; Cambridge University Press.

Mel´c`uk I .A. (1960). O terminah ustojchivost' i idiomatichnost'. *Voprosy Jazykoznanija 4,73 – 80.*

Miller, G.A. (1971). Empirical methods in the study of semantics. (ed. by D. Steinberg, A. Jakobovits), *Semantics. An interdisciplinary reader in philosophy, linguistics and psychology.* Cambridge: University Press, 569 – 585.

Moskovych, V. A. (1969). *Statistika i semantika*. Moskva: Nauka.

Merriam Webster (2017): Dictionary. – URL: https://www.merriam-webster.com/. [03-11-2017]

Porzig, W. (1934). „Wesenhafte Bedeutungsbeziehungen“. *Beiträge zur Geschichte der deutschen Sprache und Literatur. Bd. 58, 70-97.*

Zadorozhna, I. P. (2003). *Semantychni ta spoluchuvalni vlastyvosti komponentiv frazeolohizmiv u nimetskij movi*. АКД: L‘viv.

Procter, P. (1978): *Longman Dictionary of Contemporary English*. London.

Schippan, Th., Sommerfeld, K.-E. (1966). “Wort und Kontext”. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19(6).*

Tuldava, J. (1988). „Ob izmerenii sv’azi kachestvennyh priznakov v lingvistike (1): sopr’a zhennost‘ al’ternativnyh priznakov“. *Uch. Zap. Tartuskogo universiteta, Tartu, Vyp. 827, 146-162.*

Vendler, Z. (1967). „The grammar of goodness“. In: *Vendler Z. Linguistics in Philosophy: 172 – 195.* New York.

Glottometrics 38, 2017, 30-54

On Rank-Frequency Distributions in Poetry

Ioan-Iovitz Popescu[1]
Tayebeh Mosavi Miangah
Hanna Gnatchuk
Radĕk Čech
Alice Bodoc
Gabriel Altmann

Abstract: Some properties of poetic texts are investigated in English, German, Russian, Czech, Slovak, Hungarian, Ukrainian, Romanian, Persian and Maori. Some well known properties of texts (e.g. h-point, repeat rate) are computed and various hypotheses are tested. The results are presented in tables and figures hence there is a good possibility to compare the results both with other languages and other text types. Both the rank-frequency relations and the spectrum of frequencies are investigated and fitted by well known functions or distribution. The translations of a Hungarian poem into six languages are investigated and compared.

1. Introduction

Ranking whatever type of linguistic data we obtain some regularity. If we rank the frequency of words or word forms, the obtained regularity expresses a kind of structuring. In case of lemmas, some indicators derived from the given distribution give hint at the vocabulary richness of the text. If we rank the frequencies of any other entities, we obtain an image of structuring in the given domain. It must be remarked that ranking is bound to some kind of measurement (at least ordinal), even if one takes only frequencies into account.

The number of indicators capturing vocabulary richness is enormous and their number increases (cf. Wimmer, Altmann 1999, Popescu et al. 2011, 2013) yearly. We may start from the assumption that the richer is a text, the more different lemmas occur in it. Of course, an infinitely long text could contain all words of the language hence taking text length into account yields a very colorful characterization. Seen from another perspective, the more uniform is the rank-frequency distribution, i.e. the more similar are the frequencies of words in the text, the richer is the text. This view automatically leads to the idea that the greater the entropy of frequencies, the richer is the text. Since entropy can be transformed into other indicators, e.g. the repeat rate, the

[1] Address correspondence to: I.-I. Popescu: iovitzu@gmail.com; T. Mosavi Miangah, Linguistics Department Payame Noor University, I.R. of Iran, mosavit@pnu.ac.ir;
Anna Gnatchuk: Alpen-Adria University, Slawistik: Universitätsstraße 65-67, 9020 Klagenfurt am Wörthersee, Austria: agnatchuk@gmail.com;
Radek Čech, Faculty of Arts, University of Ostrava, Reln 3, Ostrava 701 03, Czech Republic: cechradek@gmail.com;
Alice Bodoc, Faculty of Letters, Transilvania University of Brasov, Romania Eroilor Street, No. 29, alice_bodoc@yahoo.com

smaller the repeat rate, the more uniform is the distribution, consequently the richer is the text. The more concentrated is a text, the steeper the rank-frequency distribution, hence the smaller is the entropy and the larger is the repeat rate. Or, the smaller is Gini's coefficient, the richer is the text. The smallest Gini's coefficient is attained if the Lorenz function is a straight line lying on the diagonal of the Euclidean coordinate system, i.e. the plain between the Lorenz curve and the diagonal is minimal.

One can interpret various simple statistical indicators in terms of text richness. Thus text richness is a property having a link to various other properties and is part of the Köhlerian (2005) control cycle. Hence it may be used as an integral part of the Köhlerian or other cycles which will be created in the future.

In the present study we shall be interested in (1) the form of the rank-frequency distribution of lemmas restricted to poetic texts, (2) the parameters of the distribution, perhaps characterizing types of poetry or types of languages, or individual writers, (3) the possible development of the function interpreted as the change performed by the writer or the change in the language.

We shall use lemmas – not word forms – because word forms yield automatically different results for different languages, even if a text is translated. Strongly synthetic languages have more word forms than strongly analytic languages.

To this purpose we shall use poetry in various languages. Poetic texts are a special text type. Surely, there will be differences between rhymed and non-rhymed poems, between strictly metric and free poems, between lyric and epic poems, etc. We shall merely try to show some aspects of this possible research domain. We shall not touch other properties, but only lemmas, because the domain of text analysis is infinite.

An analysis applying many indicators to various aspects of poetic texts has been performed concerning the complete work of the Romanian writer M. Eminescu (cf. Popescu et al. 2015). The literature concerning analysis of individual texts is enormous.

Though G. Herdan often (e.g. 1963) criticized the rank-frequency distribution because rank is no observed variable, it can be shown that the spectrum of frequencies can easily be transformed in it and vice versa. Besides, there are no "given" variables, all linguistic units and all of their properties are defined by us in such a way that they can be used either in qualitative descriptions of rules or in quantitative modeling of other regularities. If ranking yields a possibility to set up hypotheses, to derive them from a theoretical background and to test them, then it is a legal variable hinting at a kind of order in data. In all sciences, we create some kind of order in our "data" if we introduce some conventional definitions, criteria, etc., i.e., if we organize the "data" in our sense. The attained order can be modeled and characterized by a number of functions and indicators.

A point concerning modeling should be emphasized here: Ordered data can be modeled both by functions, series, or distributions. All of these mathematical means have been constructed by Man in order to give us an exact instrument for capturing *our* view of the reality, to draw exact conclusions but not to express "the truth" (cf. Bunge 1963: 269). As repeated many times in quantitative linguistics: things merely exist, but their definition and treatment is our human invention.

Lemmatization is a special problem that can be solved in many different ways. The simplest way is to look for the basic form of a word form in a dictionary but even here our decision is important. For example, in many languages there are articles differing according to gender, case, number; there are pronouns, e.g. "I", "me", "my", "we", "our", "us" which may be considered as belonging to one lemma (in German there are "ich", "mich", "mir", "mein", "wir", "uns", "unser"), quite different forms of the verb "be", etc. in which suppletivism is active. Thus the lemma is not given, it is

constructed by us – just as any other data. One may adhere to a simple lemmatization ignoring merely grammatical categories; one may ignore the POS differentiation, e.g. in. German "sie nehmen" vs. "das Nehmen" or "Das Licht war hell" vs. "Die Lampe leuchtete hell", i.e. verb and noun are in some cases considered identical, or adjective and adverb are in some cases identical, etc. Of course, one can lemmatize only to a certain degree or "drastically", omitting also affixes and clitics but no program performs it mechanically.

As far as lemmatizations of English, German, Ukrainian and Russian languages are concerned, we followed the following principle: we left the derivatives to be separate lemmas by eliminating only inflectional affixes. We have done everything by pen and paper without using any programs. The same has been done also in the other languages scrutinized. In Maori, no lemmatization was necessary. In Czech we used the program QUITA (Kubát, Matlach, Čech 2014).

To lemmatize Persian words, we reduced inflectional forms and derivationally related forms of a word to a common base form. All this was carried out manually without any special software or program. There are many problems with Persian lemmatization. For one thing, there are many words in this language which may contain two subject and object pronouns as suffix. Consider, for instance the Persian word '*khordamash*' in which '*khord*' is a past tense stem of the verb '*eat*' with two suffixes '*am*' and '*ash*' as 1st person subject pronoun and 3rd person object pronoun, respectively, with the meaning 'I ate it' as a whole. This is the same case with some nouns in Persian named "broken plurals" conveying the concept of plural but not showing plural signs in their surface form according to Persian plural-making structures (Mosavi Miangah, 2006). In such cases automatic lemmatization fails due to inconsistencies in language morphology.

In strongly analytic languages one can consider the words without any problem because the forms are equal to lemmas, but the more synthetic a language, the more problems arise. The question is how to manage the problem, especially if one applies a ready-made program. As a matter of fact, the solution is simple: use that lemmatization that fully corresponds to the derived hypothesis! But to this end one must already have a hypothesis. If one does not have it, then both lemmatization/data and the hypothesis may/must be "corrected" and one continues in this way until the data in all languages behave according to the hypothesis. It is both inductive and deductive work. In many cases, the given text type may differ from other ones (in the same language) to such an extent that either new hypotheses must be derived or the boundary conditions must be inserted in the topical hypothesis. Thus the analysis of texts taking into account the rank-frequency distribution of lemmas is an eternal problem whose solution will go through mountains and valleys. A good example is the problem of length of units in texts which began in the 19th century, went through dozens of models until a unique model has been found (cf. Popescu et al. 2014).

The rank-frequency study of whatever entity has a sore point, namely the statistical testing. Since it is a distribution, one tests the adequacy of a model using the chi-square test. Unfortunately, the majority of frequencies in the tail of the distribution are ones and the theoretical frequencies get smaller than one, hence many classes must be pooled. One may add 1 as a constant to the theoretical distribution but the result of testing depends on the pooling criterion. Some statisticians use 5, others admit 1 (i.e. the size of the theoretical class must be at least 1). A kind of remedy is the fitting of a model up to the last frequency 2, and consider all other as a straight line $y = 1$. If one does not strive for modeling, one may compute various indicators (cf. e.g. Popescu et al. 2014; Altmann, Köhler 2015) characterizing the distribution. Another remedy is to

apply a continuous function and test its adequacy only on the (integral) ranking points. This technique has been used ever more in the last time, on grounds mentioned above: a mathematical model is no truth, it is a scheme expressing the reality *for us*. It has the advantage that no classes need to be pooled, the determination coefficient is a sufficient indicator of adequacy. As far as the chi-square test is concerned, it is well known that the chi-square increases linearly with the sample size, hence the larger the sample the "worse" is the fitting. One tried to avoid this problem by introducing coefficients relativized by the sample size, but they decrease with sample size.

As to modeling the rank-frequencies of lemmas, one can go three ways: (a) to choose the models yielding the best fit for each text individually or (b) to search for a model which is common to all, even if the fitting is in some cases not quite satisfactory. In the latter case, one can try to find the boundary conditions and express them by adding a further parameter. (c) A further possibility is to apply different models to the individual languages and text types.

Here we shall try to (1) present the data (numbers), (2) derive the fitting function and (3) test each case. We shall prefer here a continuous function. The step between continuous and discrete modeling has been described in Mačutek, Altmann (2007).

For this investigation we used Czech, English, German, Hungarian, Maori Persian, Romanian, Russian, Slovak and Ukrainian data. Ten languages is a very modest sample but we hope that the work can be continued by other researchers. The identification of the texts is given in each table.

2. Characterization

The term "characterization" can be specified in various ways and in various domains of linguistics. Since text is the basic linguistic phenomenon, there can be numbers of characterizations beginning with phonetics up to levels which have not yet been constructed. There is no linguistic book that would contain at least the naming of the possible characterizations. Here we shall stay at the level of word frequencies which is, of course, only one of the aspects of texts.

We begin the text characterization with the h-point (Popescu 2007) which depends also on text length. The h-point has been defined as

$$h = \begin{cases} r & \text{if there is an } r = f(r) \\ \dfrac{f(i)*r_j - f(j)*r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \tag{2.1}$$

where $f(i)$ is the frequency at rank i; r_i is the rank of the i-th lemma. For example, for the text in Table 3.1 the rank $r = 4$ is identical with its frequency, hence $h = 4$.

It can be shown that the h-point depends almost linearly both on N and on V (increasing). Its dependence on V/N is on the whole decreasing but the oscillation is too strong. The texts are too short to obtain a better expressed relationship.

In short texts, the h-point is usually very small. Ordering the texts in Table 3.2 according to V/N, we obtain an oscillating decreasing line of h presented in Figure 3.1.

Since *h* increases with *N* or *V*, Popescu et al. (2009) proposed a relativization in form of two indicators

$$a = \frac{N}{h^2} \quad (2.2)$$

and

$$b = \frac{V}{h^2}. \quad (2.3)$$

For rank-frequencies, *N* is the sum of all frequencies; for spectra, *V* may be considered the number of different lemmas or forms. However, considering the above poems, no systematic results follow. This may be caused by the lemmatization or by the fact that the indicators cannot distinguish languages and age. The data are presented in Table 2.1 and displayed in Figure 2.1a and 2.1b.

Table 2.1
The h-point of individual poems

Poet	Poem	V	N	h	a	b
Minulescu, Ion (Romanian)	*In asteptare* (1936)	65	103	4.5	5.09	3.21
	Drum crucial (1939)	38	69	3	7.67	4.22
	Peisaj umed (1943)	48	83	4	5.19	3
Blaga, Lucian (Romanian)	*Izvorul nopții* (1919)	38	53	3	5.89	4.22
	Vară (1920)	52	65	3	7.22	5.78
	Pluguri (1922)	58	70	2.5	11.20	9.28
Stanescu, Nichita (Romanian)	*Emoție de toamnă* (1964)	53	76	3.5	6.20	4.33
	Inima (1965)	39	97	6	2.69	1.08
	Inchinare (1967)	33	37	2	9.25	8.25
E.Bachletová (Slovak)	*Aby spriesvitnela (2004)*	52	63	3	7	5.78
	Iba neha (2008)	71	130	5	5.2	2.84
A. Sládkovič B. (Slovak)	*Ctibor (1842)*	391	665	8	10.39	6.11
	Kykymora (1861)	314	559	9.25	6.53	3.70
Sadi (Persian)	*Couplet (Z) (13th century)*	300	534	7.5	9.49	5.33
Hafez (Persian)	*Sonnets (14th century)*	298	598	7	12.20	6.08
Shahryar (Persian)	*Ode (19th century)*	265	550	10	5.5	2.65
Mehdi Soheili (Persian)	*Farewell (1966)*	103	158	4.67	7.24	4.72
	Love (1953)	122	233	7	4.76	2.49

E. Muir (English)	*Suburban Dream (1946)*	106	150	4.5	7.41	5.23
C. Day Lewis	*In the Heart of Contemplation (1938)*	82	119	3.5	9.71	6.69
Ivan Franko (Ukrainian)	*Чого являєшся мені у сні? (1896)*	96	154	4.5	7.60	4.74
V. Simonenko (Ukrainian)	*Є в коханні і будні, і свята... (1961)*	59	92	4	5.75	3.69
A. Malisko (Ukrainian)	*Пісня про рушник (1959)*	55	120	5	4.8	2.2
S. Esenin (Russian)	*Мне грустно на Тебя смотреть (1923)*	74	98	4.57	4.69	3.54
	Sobake Kačalova (1925)	86	138	4.67	6.33	3.94
A. Pushkin (Russian)	*K***(1825)*	52	100	4	6.25	3.25
E.W. Lotz (German)	*Spät über den Häusern (1914)*	83	132	5	5.28	3.32
Th. Kramer (German)	*An einen unbekannten Buchhändler*	103	179	4.67	8.21	4.73
H. Hesse (German)	*Stufen (1941)*	93	153	5	6.12	3.72
J.W.v.Goethe, (German)	*Der Erlkönig (1778)*	225	183	5.5	7.43	6.05
J. Seifert *Město v slzách* 1921 (Czech)	*Modlitba na chodníku*	266	467	8	7.30	4.16
	Monolog bezrukého vojáka	168	290	6	8.06	4.67
	V předměstské uličce	212	370	6	10.28	5.89
	Báseň plná odvahy a víry	175	318	7.33	5.92	3.26
	Prosinec 1920	178	309	6.33	7.71	4.44
http://folksong.org.nz/ waiata.html (Maori)	*E pā tō hau*	77	147	5	5.88	3.08
http://folksong.org.nz/ waiata.html (Maori)	*E to matou Matua*	52	92	5	3.80	2.08
J. Orten, *Čítanka jaro.* 1939 (Czech)	*Na tomto místě*	64	108	4.5	5.33	3.16
	Pod každým slovem	63	99	5	3.96	2.52
	Ohlédnutí	54	65	3	7.22	6
	O malých dětech	60	82	3.67	6.09	4.45
	Lítost	67	113	4	7.06	4.19
	Hostinka	30	40	2	10	7.5
	Hrob	57	81	3	9	6.33
	Udělej černou myš	54	72	3	8	6
	Zavřené oči	80	128	4.5	6.32	3.95
	Poloha	30	32	2	8	7.5
	Zasnění	112	177	5.67	5.51	3.48
	V nazlátlém vidění	63	86	3	9.56	7
	Nepojmenovatelná báseň	58	83	3.5	6.78	4.73

	O oči civící	71	119	4	7.44	4.44
	Až nebe	55	76	3.33	6.85	4.96
	Déšť	47	63	3	7	5.22
	Pohádka	32	38	2	9.5	8
	Odhoď šat svůj	62	76	3	8.44	6.89
	Utonulá	58	76	3	8.44	6.44
	Dívka v ohni	81	102	3	11.33	9
J. Orten, *Cesta k mrazu* 1940 (Czech)	*Báseň pozdního léta*	39	48	2.5	7.68	6.24
	Jablka	57	70	2.75	9.26	7.53
	Malá elegie	102	160	4.5	7.90	5.04
	Končiny štěstí	48	61	3	6.78	5.33
	Dívka a strom	101	167	5	6.68	4.04
	Nahá dívka	116	194	5	7.76	4.64
	Hrdlo písně	56	85	3.5	6.94	4.57
	Modlitba	78	111	4	6.94	4.88
	Listopad	65	81	3	9	7.22
	Takový smutný den	70	96	4	6	4.38
	Návrat	68	90	3	10	7.56
	Rty písně	66	84	3	9.33	7.33
	Holub na střeše	55	71	3	7.89	6.11
	Bílý obraz	80	152	5	6.08	3.2
	Ústí písně	70	119	5	4.76	2.80
	Sklaě se k trávě	54	81	3	9	6
	Báseě z kamene	84	136	4	8.50	5.25
	V mamince	34	40	2	10	8.6
	I. Tolik to zní	86	143	5	5.72	3.44
	II. Veliký pohyb	55	99	4	6.19	3.44
J. Orten, *Ohnice* 1941 Czech	*První báseň*	120	170	4	11.19	7.50
	Čí jsem?	58	85	4	5.31	3.62
	Kouří se…	112	165	5	6.6	4.48
	Život	61	83	3	9.22	6.78
	Na starých…	70	113	4.5	5.58	3.46
	Na zimu	70	102	3.5	8.33	5.71
	Zem…	94	120	4	7.5	5.88
	Vždycky se…	38	49	2	12.25	9.5
	Když tma…	81	109	4	6.81	5.06
	Po smrti	74	104	4	6.5	1.62

	První noc	86	105	2.5	16.8	13.76
	Báseň léta	85	118	4	7.38	5.31
	Podzim	47	63	3	7	5.22
	Ta krajina	167	318	7.67	5.40	2.84
	V noci	37	43	2	10.75	9.25
	Potkal jsem…	46	72	3.75	5.12	3.27
	Báseň šera	77	119	3.5	9.71	6.29
	Jsem na polibek…	35	73	4	4.56	2.19
	Co mi řekl kanárek	41	65	3.5	5.14	3.35
	Co jsem odpověděl kanárkovi	45	61	3	6.78	5

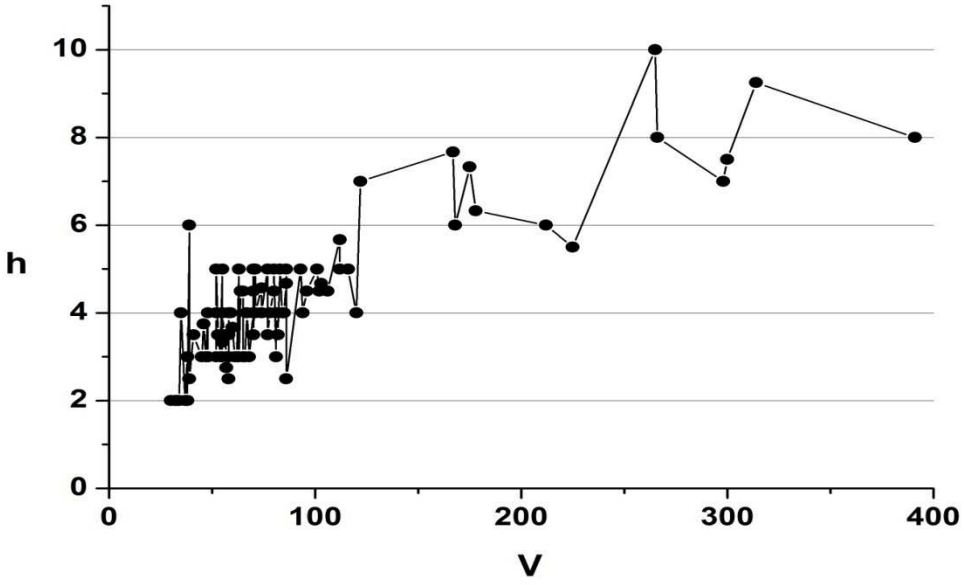


Figure 2.1a. The relation between V and h

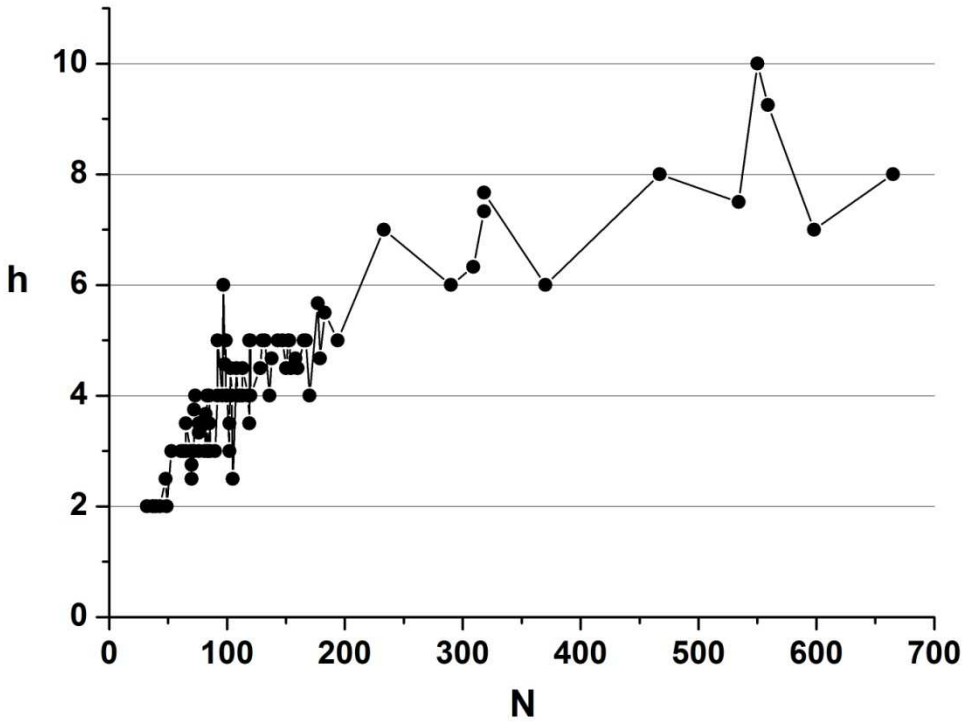


Figure 2.1b. The relation between N and h

In both cases the increasing tendency is evident but the oscillation is too strong. This is caused by some boundary conditions which could be further investigated. The relation (N,a) yields a still more oscillating function, as shown in Figure 2.2a.

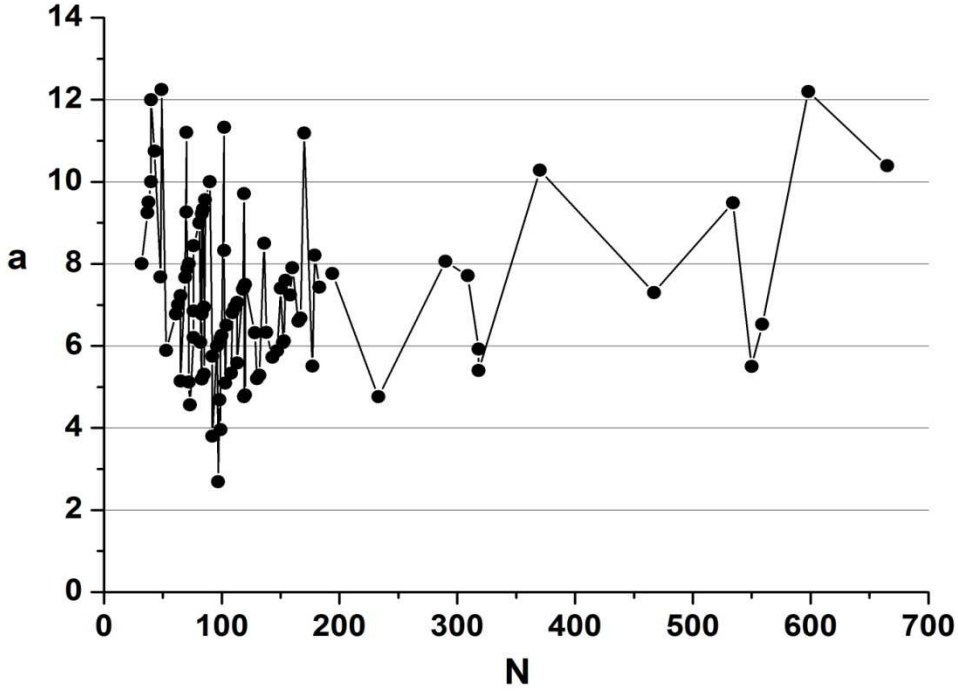


Figure 2.2a. The relation (*N*,*a*)

Still more irregular is the relation (2.3) as shown in Figure 2.2b. Evidently, there are texts deviating strongly from a "regular" course. The deviations may be captured by literary interpretation – if one knows the creator`s procedure: (s)he may have corrected, changed parts of the text; the editors changed something; the writer made a long pause between the writing of the given and the preceding text; or on the contrary, (s)he wanted the next text to be quite different, etc. Placing the results in separate figures according to author, date, text type, language, etc. would require much more individual data and in all cases the interrogation of the authors which is not possible in most cases. Here merely a first picture of a possible regularity should be shown.

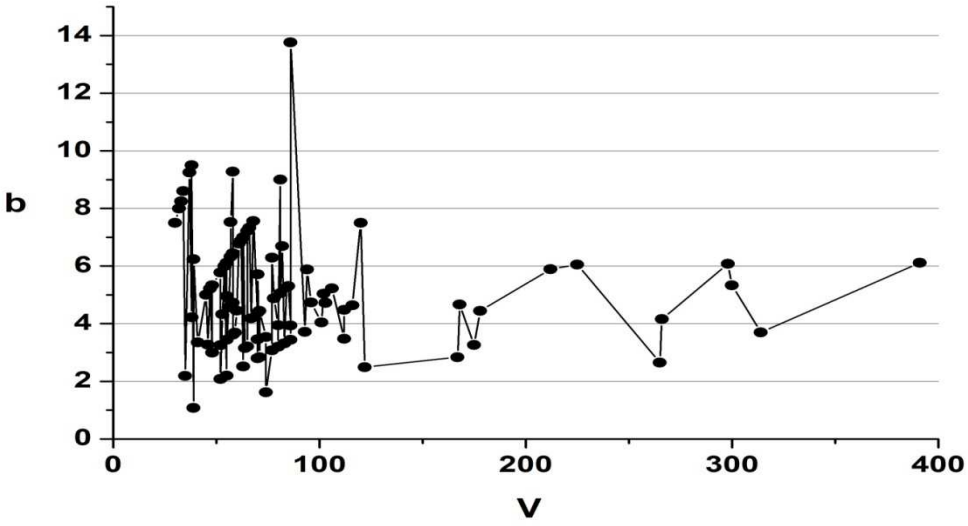


Figure 2.2b. The relation (V,b)

The indicators (2.2) and (2.3) could be replaced by other ones having in the denominator h^3 or other power of *h*. But even in that case, shorter texts display a strong oscillation, longer texts oscillate around a horizontal line. Evidently, a lot of work is necessary to find the boundary conditions which may be different in various texts and languages. Nevertheless, one could perform a kind of smoothing (lags, filtration) usual in time series analysis in order to obtain a simple function. Here we shall renounce searching for an appropriate smoothing and wait until one processes much more poetic texts in every language. Then two or three steps must be made: (1) The study of individual authors, (2) the study of individual languages, (3) the study of historical change in these categories.

Needless to say, the oscillation could be captured by Fourier analysis but an introduction of a third and further variables representing other characteristics of texts would be theoretically more appropriate.

3. Modeling

For modeling we apply the most common approach. Since the data are ordered and it does not play any role whether we use a series, a function or a normalized formula, we apply the unified theory (Wimmer, Altmann 2005). We consider the relative rate of change of the frequency, i.e. dy/y and put it equal to $[g(x)/h(x)]dx$. Since the frequencies cannot be smaller than 1, we may modify the ratio and set $dy/(y\text{-}1)$. On the right hand side, we consider the function $g(x) = -c$, a constant, and $h(x) = b+x$, where *b* is the controlling force. One can interpret *c* as a language constant or as the constant effort of the speaker/writer; *b* is in any case the constant effort of the hearer. We obtain

$$(3.1) \quad \frac{dy}{y-1} = -\frac{c}{b+x}dx$$

which yields the famous Zipf-Mandelbrot formula

$$(3.2) \quad y = 1 + \frac{a}{(b+x)^c}$$

with added constant 1, where *y* are the frequencies, *x* is the rank-order, and *a* is the integration constant. However, the Zipf-Mandelbrot function yields sometimes enormous parameters *a*. This can be the sign of overestimating some boundary conditions. In order to unify the numbers we shall use also the simple Zipf (zeta/power)-function

$$(3.3) \quad y = 1 + ax^c$$

taking into account the constant 1 or not. Function (3.3) can be obtained in the same way as (3.1), with $b = 0$.

For the sake of illustration we apply (3.2) without the constant 1 to the poem *In asteptare* of the Romanian poet Ion Minulescu and obtain the results presented in Table 4.1.

Table 3.1

Fitting (4.3) without 1 to the Romanian poem *In asteptare* by I. Minulescu (1936)

Rank	**Fr**	**Z**		**Rank**	**Fr**	**Z**		**Rank**	**Fr**	**Z**
1	5	6.41		23	1	1.51		45	1	1.11
2	5	4.66		24	1	1.48		46	1	1.09
3	5	3.86		25	1	1.45		47	1	1.08
4	4	3.38		26	1	1.42		48	1	1.07
5	4	3.05		27	1	1.40		49	1	1.06
6	3	2.80		28	1	1.38		50	1	1.05
7	3	2.61		29	1	1.35		51	1	1.04
8	3	2.45		30	1	1.33		52	1	1.03
9	3	2.32		31	1	1.31		53	1	1.03
10	2	2.21		32	1	1.29		54	1	1.02
11	2	2.12		33	1	1.28		55	1	1.01
12	2	2.04		34	1	1.26		56	1	1.00
13	2	1.96		35	1	1.24		57	1	0.99
14	2	1.90		36	1	1.23		58	1	0.98
15	2	1.84		37	1	1.21		59	1	0.98
16	2	1.78		38	1	1.20		60	1	0.97
17	2	1.73		39	1	1.18		61	1	0.96
18	2	1.69		40	1	1.17		62	1	0.95
19	2	1.65		41	1	1.15		63	1	0.95
20	2	1.61		42	1	1.14		64	1	0.94
21	2	1.57		43	1	1.13		65	1	0.93
22	1	1.54		44	1	1.12				
a = 6.1117, c = -0.4617, R^2 = 0.8770										

For Zipf + 1 we obtain $R^2 = 0.79$, for Zipf-Mandelbrot $R^2 = 0.94$, and for Zipf-Mandelbrot + 1, $R^2 = 0.94$. As expected, the fitting is very good. Our aim is to reduce the number of parameters to minimum. In general, we consider the fitting as sufficient, if the determination coefficient $R^2 > 0.8$.

The results for poems in various languages are presented in Table 4.2.

Table 3.2
Fitting the Zipf function to rank-frequencies
of lemmatized poems

Poet	Poem	a	c	R^2
I. Minulescu (Romanian)	*In asteptare* (1936)	6.4117	-0.4617	0.88
	Drum crucial (1939)	6.7173	-0.5498	0.84
	Peisaj umed (1943)	8.8409	-0.6180	0.93
L. Blaga (Romanian)	*Izvorul nopții* (1919)	4.2665	-0.4431	0.89
	Vară (1920)	3.9472	-0.4106	0.88
	Pluguri (1922)	3.6593	-0.3811	0.85
N. Stanescu (Romanian)	*Emoție de toamnă* (1964)	6.0666	-0.5397	0.94
	Inima (1965)	11.3928	-0.6020	0.87
	Inchinare (1967)	10.9818	-0.5980	0.91
A. Sládkovič (Slovak)	*Ctibor (1842)*	27.5786	-0.6192	0.98
	Kykymora (1861)	29.8805	-0.6476	0.94
Sadi (Persian)	*Couplet (13th century)*	28.3947	-0.6474	0.97
Shahryar (Persian)	*Ode (19th century)*	30.1913	-0.6248	0.92
Hafez (Persian)	*Sonnets (14th century)*	27.7399	-0.6395	0.94
Mehdi Soheili (Persian)	*Farewell (1966)*	12.3321	-0.6439	0.96
	Love (1953)	16.2787	-0.6046	0.92
E. Muir (English)	*The Angel and the Girl* (Z+1)	14.3174	-1.2642	0.95
E. Muir (English)	*Suburban Dream*	16.2951	-0.8260	0.91
C.D. Lewis (English)	*In the Heart of Contem-plation* (1938)	12.1928	-0.7424	0.91
I. Franko (Ukrainian)	*Чого являєшся мені у сні? (1896)*	11.8537	-0.6175	0.94
V. Simonenko (Ukrainian)	*Є в коханні і будні, і свята... (1961)*	9.6239	-0.6577	0.95
A. Mališko (Ukrainian)	*Пісня про рушник (1959)*	15.8925	-0.7626	0.97
S. Esenin (Russian)	*Мне грустно на тебя смотреть (1923)*	8.4815	-0.5947	0.90
S. Esenin (Russian)	*Sobake Kačalova* (1925)	9.8781	-0.5697	0.95
A. Pushkin (Russian)	*K**** (1825)	9.3998	-0.6251	0.95
E.W. Lotz (German)	*Spät über den Häusern*	14.2892	-0.7391	0.96
Th. Kramer (German)	*An einen unbekannten Buchhändler*	25.3709	-0.9511	0.94
H. Hesse (German)	*Stufen (1961)*	12.5911	-0.6380	0.95
J. Seifert: *Město v slzách*. 1921 (Czech)	*Modlitba na chodníku*	24.4857	-0.6223	0.92
	Monolog bezrukého vojáka	15.5156	-0.5722	0.94
	V předměstské uličce	18.6411	-0.5878	0.95
	Báseň plná odvahy a víry	17.9487	-0.5917	0.94
	Prosinec 1920	18.65483	-0.5234	0.97
http://folksong.org.nz/waiata.html (Maori)	*E pā tō hau*	17.1944	-0.7386	0.93
http://folksong.org.nz/waiata.html (Maori)	*E to matou Matua*	9.4319	-0.6122	0.91
J.Orten, *Čítanka jaro.* 1939 (Czech)	*Na tomto místě*	9.1718	-0.5797	0.91
	Pod každým slovem	7.3967	-0.5221	0.82
	Ohlédnutí	3.8137	-0.4083	0.84
	O malých dětech	5.9181	-0.5076	0.85
	Lítost	8.9551	-0.5596	0.94

	Hrob	12.9859	-0.9740	0.83
	Udělej černou myš	4.5906	-0.4355	0.87
	Zavřené oči	8.6165	-0.4848	0.89
	Zasnění	14.6865	-0.6725	0.94
	V nazlátlém vidění	7.4174	-0.5970	0.85
	Nepojmenovatelná báseň	8.5626	-0.6586	0.87
	O oči civící	16.8371	-0.8489	0.97
	Až nebe	8.9215	-0.7222	0.88
	Déšť	5.6257	-0.5501	0.92
	Pohádka	3.3322	-0.4439	0.78
	Odhoď šat svůj	4.6303	-0.4619	0.83
	Utonulá	3.7948	-0.3595	0.80
	Dívka v ohni	4.1939	-0.3707	0.87
J. Orten, *Cesta k mrazu.* 1940 (Czech)	*Báseň pozdního léta*	3.1948	-0.3714	0.83
	Jablka (Z+1)	4.6237	-1.1190	0.83
	Malá elegie	12.4423	-0.6420	0.94
	Končiny štěstí	3.4772	-0.3632	0.83
	Dívka a strom	13.6227	-0.6525	0.97
	Nahá dívka	12.8088	.0.5890	0.94
	Hrdlo písně	8.3605	-0.6295	0.95
	Modlitba	6.7903	-0.4998	0.93
	Listopad	4.3357	-0.4153	0.86
	Takový smutný den	6.4247	-0.5123	0.92
	Návrat	4.7941	-0.4221	0.90
	Rty písně	4.6311	-0.4301	0.89
	Holub na střeše	5.1768	-0.4985	0.86
	Bílý obraz	8.6776	-0.4703	0.90
	Ústí písně	11.1749	-0.6369	0.95
	Skloň se k trávě	9.1272	-0.6951	0.91
	Báseň z kamene	13.3504	-0.6964	0.96
	V mamince	2.7633	-0.3489	0.80
	I. Tolik to zní	9.0598	-0.5213	0.91
	II. Veliký pohyb	11.1884	-0.6766	0.97
J. Orten, *Ohnice.* 1941 (Czech)	*První báseň*	11.8343	-0.6086	0.96
	Čí jsem?	11.3554	-0.8013	0.93
	Kouří se…	11.2793	-0.6052	0.94
	Život	5.4758	-0.4807	0.91

	Na starých…	10.0093	-0.6174	0.96
	Na zimu	5.9219	-0.4550	0.90
	Zem…	6.7856	-0.5156	0.87
	Vždycky se…	3.0214	-0.3304	0.79
	Když tma…	7.6223	-0.5634	0.92
	Po smrti	7.0827	-0.5303	0.94
	První noc	5.2069	-0.4547	0.78
	Báseň léta	8.9805	-0.6075	0.93
	Podzim	4.2147	-0.4211	0.89
	Ta krajina	21.5411	-0.6363	0.89
	V noci	2.7227	-0.3368	0.79
	Potkal jsem…	9.1618	-0.6979	0.87
	Báseň šera	7.2129	-0.4937	0.92
	Jsem na polibek…	6.7169	-0.4757	0.90
	Co mi řekl kanárek	6.4662	-0.5551	0.95
	Co jsem odpoveděl kanárkovi	5.7771	-0.5652	0.92

In our data, there were only two cases in which we used the Zipf-Mandelbrot function. Peculiar enough, both are written in modern Slovak (namely by E.Bachletová yielding for *Aby spriesvitnela* (2004) a = 2.7387, b = -0.9169, c = 0.3163, R^2 = 0.91, and *Iba neha* (2008) a = 13.6288, b = 0.3558, c = 0.6705, R^2 = 0.98) though in older Slovak, with the poems by A. Sládkovič which were created in the 19th century, the usual Zipf-function is sufficient. The question whether it depends on the development of language or on the author will not be scrutinized here.

It can be shown that many other functions could be fitted but here we do not want to avoid a well known procedure and strive for a simple model. We choose, if possible, the simplest model yielding sufficient corroboration. For example, for the data in the Slovak poem *Kykymora* by A. Sládkovič we could obtain the following determination coefficients: Zipf = 0.94, Zipf +1 = 0.89, Zipf-Mandelbrot = 0.97, Zipf-Mandelbrot + 1 = 0.99. For our purposes, the simple Zipf is mostly sufficient.

Since we have results from seven languages, we may conjecture that ordering the lemmas in poetry obeys the well-known Zipf- and Zipf-Mandelbrot law. We further conjecture that Zipf-Mandelbrot results from some boundary conditions which are not yet known. Nevertheless, the regularity can be considered as corroborated. Though the fitting is in all cases sufficient, using one of the functions we obtained an enormous parameter a. In that case we used rather another variant.

Needless to say, there are a number of other functions that could be satisfactorily fitted to all this data.

The rank-frequencies of the lemmas can be used for various text characterizations. We obtain other values than with word-forms because here the synthetism of a language is eliminated. Computations based on lemmas yield, e.g. more genuine estimation of vocabulary richness.

4. Translation

A comparison of languages can be performed scrutinizing an original poem and its translations. Here we shall consider the Hungarian poem *Szeptember végén* by S. Petöfi (1847) and its translations that can be found on the Internet (s. Table 4.1). Some translations are not rhymed but very exact. None of the languages is typologically similar to Hungarian.

Table 4.1
Translations from Hungarian: Characterization
(a and *b* are defined in (2.2) and (2.3))

Poem/Language	**V**	**N**	**h**	**V/N**	**a**	**b**
Szeptember végén	102	140	3.5	0.729	8.33	11.43
English translation	123	212	6	0.58	3.42	5.89
French translation	120	223	5.67	0.538	3.73	6.94
Romanian translation	98	158	5	0.62	3.92	6.32
German translation 1	114	194	5	0.588	4.56	7.76
German translation 2	108	195	4.67	0.554	4.95	8.94
Slovak translation	115	163	4	0.706	7.19	10.19
Italian translation	98	162	3	0.587	10.89	18.00

Table 4.2
Fitting the functions to translations of the Hungarian Poem: *Szepember végén* by S. Petöfi
(two parameters = Zipf; three parameters = Zipf-Mandelbrto)

Language/Translator	**a**	**b**	**c**	**R^2**
Hungarian: *Szeptember végén*	6.7268	-0.5795	1.0682	0.99
German translation: Neugebauer	151.7254	1.7770	1.8341	0.98
German translation: M. Remané	64.5010	1.1875	1.4201	0.91
English translation: G. Szirtes (Z)	20.9550	-	-0.7471	0.96
French translation: M. de Polignac (Z)	20.7720	-	-0.7211	0.97
Italian translation: A. Preszler	79.2276	1.7033	1.6323	0.94
Slovak translation: J. Smrek	86.2002	2.8180	1.6540	0.96
Romanian translation: Bandi	40.9338	1.5358	1.3600	0.96

Considering the individual parameters, in all cases the Hungarian has the smallest ones. For English and French we used the simple Zipf in which *c* is very small. The influence of the person of the translator can be seen in German for which we have two translations. The parameter *a* of Zipf-Mandelbrot yields the sequence: Hungarian – Romanian – German – Italian – Slovak – German; the parameter b yields: Hungarian – German – Romanian - Italian – German – Slovak. The parameter c for Zipf-Mandelbrot yields the sequence: Hungarian – Romanian – German – Italian – Slovak – German.

There are some similarities but one cannot generalize. In all cases, Hungarian has the smallest value.

Translating rhymed poetry, every translator brings in his own style, world view, national spirit, etc. Thus German can be found always in two places in the above ordering. But perhaps, the study of more translations would stabilize its position.

5. The spectrum

Even if we do not work with the rank-frequency distribution, the function used (i.e. zeta) can be transformed in a spectrum, in which one counts the number of lemmas occurring once, twice, etc. For probability distributions, the transformation is given e.g. in Haight (1966, 1969), Zörnig, Boroda (1992), Baayen (1989), Wimmer et al. (2003:119 ff.) etc. Here we shall use again a function with as few parameters as possible. Consider e.g. the data in Table 1. Counting the numbers of lemmas occurring 1,2,3,… times we obtain the result presented in Table 5.1

Table 5.1
The spectrum in the Romanian poem *In asteptare* by I. Minulescu (1936) (cf. Table 1)

Occurrence	Number
1	44
2	12
3	4
4	2
5	3

The number of functions or distributions that can be applied here is enormous. The use of the known transformation does not help to strengthen the theory until ranking is not definitively accepted.

The characteristic feature of the spectrum is the possibility of making statements about the lemmatic richness, the possibility of comparing and classifying texts, the characterization of text types, etc. Needless to say, the longer the text, the flatter will be the sequence because ever more words will occur once. But the same phenomenon can appear also in short texts in which each word can be different. And if each word is different, the spectrum has only one class to which no function or distribution can be fitted. In cases where there were only three classes, we added a fourth class with x = 0 in order to obtain a zeta-function. The respective cases are marked with two asterisks (**).

Table 5.2
Spectra of texts (power function)

Author	**Text**	**a**	**b**	**R^2**
I. Franko (Ukrainian)	*Чого являешся мені у сні?* (1896)	74.8713	-2.9920	0.990
E.Bachletová (Slovak)	*Aby spriesvitnela*(2004)	46.9907	-4.2530	0.998
	Iba neha (2008)	44.4227	-1.7819	0.988
Minulescu (Romanian)	*In asteptare* (1936)	44.0636	-1.9706	0.977
	Drum crucial (1939)	19.3639	-0.9286	0.760
	Peisaj umed (1943)	46.9961	-2.2539	0.998

L. Blaga (Romanian)	*Izvorul nopții* (1919)	26.9652	-2.2989	0.998
	Vară (1920)	42.9954	-2.8235	1.000
	Pluguri (1922)	49.0221	-2.9061	0.999
N. Stanescu (Romanian)	*Emoție de toamnă* (1964)	37.9709	-2.5916	0.999
	Inima (1965)	18.9168	-1.4510	0.977
	Inchinare (1967) **	29.9960	-3.7956	1.000
A.Sládkovič (Slovak)	*Ctibor (1842)*	281.1978	-2.2773	1.000
	Kykymora (1861)	235.2492	-2.4612	0.999
Sadi (Persian)	*Couplet (13th century)*	222.9085	-2.5191	1.000
Shahryar (Persian)	*Ode (19th century)*	178.5335	-2.0936	0.998
Hafez (Persian)	*Sonnets (14th century)*	218.0492	-2.3984	1.000
Mehdi Soheili (Persian)	*Farewell (1966)*	83.0499	-2.3470	0.998
	Love (1953)	82.2729	-1.8845	0.995
E. Muir (English)	*The Angel and the Girl*	84.2562	-2.3045	0.996
	Suburban Dream (1946)	92.9908	-3.5070	0.999
C.D. Lewis (English)	*In the Heart of Con-templation* (1938)	71.9970	-2.8477	1.000
V. Simonenko (Ukrainian)	*Є в коханні і будні, і свята... (1961)*	46.9554	-3.0098	0.997
B. Mališko (Ukrainian)	*Пісня про рушник (1959)*	27.6930	-1.2985	0.845
S. Esenin (Russian)	*Мне грустно на тебя смотреть (1923)*	58.9635	-2.3816	1.000
	Sobake Kačalova (1925)	62.9648	-2.3539	0.999
B. Pushkin (Russian)	*K**** (1825)	43.2211	-2.0442	0.986
E.W. Lotz (German)	*Spät über den Häusern (1916)*	67.9658	-3.1504	0.998
Th. Kramer (German)	*An einen unbekann-ten Buchhändler*	79.9285	-2.7304	0.999
H. Hesse (German)	*Stufen (1961)*	68.0375	-2.2349	0.999
J. Seifert: *Město v slzách*. 1921 (Czech)	*Modlitba na chodníku*	201.0335	-2.5574	0.999
	Monolog bezrukého vojáka	121.8956	-2.40812	0.998
	V předměstské uličce	121.8956	-2.4081	0.999
	Báseň plná odvahy a víry	123.9313	-2.2915	0.999
	Prosinec 1920	127.1701	-2.2499	0.998
http://folksong.org.nz/ waiata.html (Maori)	*E pā tō hau*	55.0012	-2.3171	0.998
http://folksong.org.nz/ waiata.html (Maori)	*E to matou Matua*	35.9366	-2.1036	0.995
J. Orten, *Čítanka jaro.* 1939 (Czech)	*Na tomto místě*	42.4926	-1.6961	0.968
	Pod každým slovem	48.9170	-2.9548	0.985
	Ohlédnutí	46.9825	-3.3972	0.999
	O malých dětech	49.9566	-3.3188	0.993
	Lítost	46.1311	-2.0395	0.985
	Hostinka	20.4684	-1.7009	0.899
	Hrob	48.9738	-3.3723	0.997
	Udělej černou myš	42.9539	-2.6816	0.998
	Zavřené oči	50.6490	-1.7606	0.989
	Poloha	28.0036	-3.9122	1.000
	Zasnění	89.9685	-2.8618	0.998
	V nazlátlém vidění	54.9839	-4.6958	0.987
	Nepojmenovatelná báseň	49.0029	-3.0684	0.998

	O oči civící	57.9585	-3.2853	0.997
	Až nebe	47.9925	-3.5287	0.999
	Déšť	38.0133	-2.7418	0.999
	Pohádka	27.9933	-3.1639	0.999
	Odhoď šat svůj	55.8826	-3.7384	1.000
	Utonulá	44.9817	-2.4078	0.995
	Dívka v ohni	46.0386	-2.2538	0.999
J. Orten, *Cesta k mrazu.* 1940 (Czech)	*Báseň pozdního léta*	32.0079	-2.6943	0.999
	Jablka	49.9987	-3.2672	0.998
	Malá elegie	77.0033	-2.4980	0.998
	Končiny štěstí	38.0209	-2.4743	0.998
	Dívka a strom	73.9993	-2.3780	0.999
	Nahá dívka	83.1829	-2.2057	0.997
	Hrdlo písně	41.1071	-2.2383	0.996
	Modlitba	59.9507	-2.5043	0.999
	Listopad	54.9794	-3.1077	0.999
	Takový smutný den	56.9757	-2.9595	0.999
	Návrat	52.9957	-2.3947	1.000
	Rty písně	53.9866	-2.7179	1.000
	Holub na střeše	45.9992	-2.9672	0.998
	Bílý obraz	43.9500	-1.4772	0.969
	Ústí písně	50.1128	-2.2587	0.995
	Skloň se k trávě	40.1461	-2.1283	0.989
	Báseň z kamene	64.9677	-2.6585	0.999
	V mamince	29.0121	-2.9343	1.000
	I. Tolik to zní	61.8570	-2.4250	0.996
	II. Veliký pohyb	36.9289	-2.0484	0.993
J. Orten, *Ohnice.* 1941 (Czech)	*První báseň*	93.0035	-2.5541	1.000
	Čí jsem?	49.9900	-3.5607	0.999
	Kouří se…	91.0069	-2.9662	0.999
	Život	46.1864	-2.2461	0.990
	Na starých…	52.9107	-2.6814	0.997
	Na zimu	51.9919	-2.3581	0.996
	Zem…	80.9947	-3.1336	0.999
	Vždycky se…	27.2321	-2.0092	0.975
	Když tma…	68.9653	-3.5360	0.998
	Po smrti	58.9657	-2.7867	0.999
	První noc	73.0483	-2.8684	0.998
	Báseň léta	70.9486	-3.2962	0.998
	Podzim	36.0013	-2.3558	1.000
	Ta krajina	119.0593	-1.9355	9.987
	V noci	32.0089	-3.0597	1.000
	Potkal jsem…	37.9784	-3.4201	0.994
	Báseň šera	51.2904	-1.7955	0.993
	Jsem na polibek…	15.1781	-1.0253	0.822
	Co mi řekl kanárek	28.1055	-2.0179	0.995
	Co jsem odpověděl kanárkovi	36.9733	-3.0371	0.998

First we shall characterize the texts applying some indicators and order the texts according to their values. We shall use here the *Repeat rate* defined as

$$(5.1) \quad RR = \sum_i p_i^2 = \sum_i \left(\frac{f_i}{N} \right)^2 = \frac{1}{N^2} \sum_i f_i^2 ,$$

i.e. the sum of squared frequencies divided by the square of the sum (N = number of lemmas). In Table 5.1 we have N = 65, hence

$$RR = (44^2 + 12^2 + 4^2 + 2^2 + 3^2)/65^2 = 2109/4225 = 0.4992.$$

The variance of RR is approximately

$$(5.2) \quad Var(RR) = \frac{4}{N} \left[\sum_i p_i^3 - RR^2 \right],$$

and the relative *RR* is

$$(5.3) \quad RR_{rel} = \frac{1 - RR}{1 - 1/k}$$

where k is the highest occurrence. There are many cases in the spectrum with zero frequency. But k is here the value of the greatest frequency in the spectrum – here usually that of x = 1, i.e. the number of ones in the rank-frequency distribution. For the above case we would obtain RR_{rel} = (1-0.4992)/(1 – 1/5) = 0.6260. For the variance of *RR* we obtain $Var(RR) = 4[(44^3+12^3+4^3+2^3+3^3)/65^3 - 0.4992^2]/65 = 0.0042$.

Here we shall use McIntosh's relative repeat rate defined as

$$(5.4) \quad RR_{rel} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{k}}$$

It is to be remarked that for the rank-frequency distribution the notation would be slightly different, namely V would be N, and k would be V respectively.

The richer is the text, the smaller is the repeat rate, i.e. the lemmas are repeated less frequently. The results are presented in Table 5.3.

Table 5.3
The repeat rate of spectra

Author	**Text (year)**	**RR**	**Var(RR)**	**RR_{rel}**
E.Bachletová (Slovak)	*Aby spriesvitnela* (2004)	0.8203	0.0051	0.1886
	Iba neha (2008)	0.4414	0.0031	0.5192
I. Minulescu (Romanian)	*In asteptare* (1936)	0.4992	0.0541	0.4876
	Drum crucial (1939)	0.4086	0.0015	0.7216
	Peisaj umed (1943)	0.5684	0.0048	0.4452
L. Blaga (Romanian)	*Izvorul nopții* (1919)	0.5895	0.0086	0.4644
	Vară (1920)	0.6990	0.0060	0.3279
	Pluguri (1922)	0.7289	0.0051	0.2925

N. Stanescu (Romanian)	*Emoţie de toamnă* (1964)	0.6202	0.0068	0.3844
	Inima (1965)	0.3380	0.0047	0.7075
	Inchinare (1967)	0.8310	0.0074	0.2091
A.Sládkovič (Slovak)	*Ctibor (1842)*	0.5444	0.0008	0.3628
	Kykymora (1861)	0.5836	0.0010	0.3182
Sadi (Persian)	*Couplet (13th century)*	0.5720	0.0011	0.3372
Shahryar (Persian)	*Ode (19th century)*	0.4867	0.0011	0.4076
Hafez (Persian)	*Sonnets (14th century)*	0.5586	0.0011	0.3496
Soheili (Persian)	*Farewell (1966)*	0.5658	0.0028	0.3833
	Love (1953)	0.4100	0.0019	0.5149
E. Muir (English)	*The Angel and the Girl* (year)	0.6067	0.0026	0.3736
E. Muir (English)	*Suburban Dream* (1946)	0.7761	0.0028	0.2012
C.D. Lewis (English)	*In the Heart of Contemplation* (1938)	0.6996	0.0036	0.2960
I. Franko II. (Ukrainian)	*Чого являєшся мені у сні? (1896)*	0.6186	0.0038	0.3432
V. Simonenko (Ukrainian)	*Є в коханні і будні, і свята... (1961)*	0.6455	0.0061	0.3160
C. Mališko (Ukrainian)	*Пісня про рушник (1959)*	0.3613	0.0017	0.6171
S. Esenin (Russian)	*Мне грустно на тебя смотреть (1923)*	0.6551	0.0042	0.3448
	Sobake Kačalova (1925)	0.5527	0.0038	0.4124
A.Pushkin (Russian)	*K**** (1825)	0.5276	0.0042	0.4624
E.W. Lotz (German)	*Spät über den Häusern (1916)*	0.6972	0.0041	0.2789
Th. Kramer (German)	*An einen unbekannten Buchhändler (year)*	0.6182	0.0034	0.3306
H. Hesse (German)	*Stufen (1961)*	0.5643	0.0033	0.3999
J. Seifert: *Město v slzách*. 1921 (Czech)	*Modlitba na chodníku*	0.5907	0.0013	0.3313
	Monolog bezrukého vojáka	0.5495	0.0020	0.3784
	V předměstské uličce	0.5185	0.0016	0.4408
	Báseň plná odvahy a víry	0.5277	0.0018	0.4103
	Prosinec 1920	0.5388	0.0013	0.3890
http://folksong.org.nz/ waiata.html (Maori)	*E pā tō hau*	0.5358	0.0042	0.4020
http://folksong.org.nz/ waiata.html (Maori)	*E to matou Matua*	0.2934	0.0027	0.8292
J. Orten, *Čítanka jaro*. 1939 (Czech)	*Na tomto místě*	0.5107	0.0033	0.4822
	Pod každým slovem	0.6201	0.0055	0.3845
	Ohlédnutí	0.7647	0.0056	0.2510
	O malých dětech	0.7039	0.0056	0.3220
	Lítost	0.5253	0.0042	0.4652
	Hostinka	0.5556	0.0033	0.8694
	Hrob	0.7470	0.0055	0.2714
	Udělej černou myš	0.6509	0.0061	0.3864
	Zavřené oči	0.4419	0.0034	0.5666
	Poloha	0.8756	0.0062	0.2195
	Zasnění	0.6593	0.0031	0.3023
	V nazlátlém vidění	0.7697	0.0047	0.2453
	Nepojmenovatelná báseň	0.6748	0.0057	0.3230
	O oči civící	0.6755	0.0050	0.2864
	Až nebe	0.7679	0.0055	0.2237

	Déšť	0.6713	0.0068	0.3268
	Pohádka	0.7754	0.0087	0.2826
	Odhoď šat svůj	0.7601	0.0058	0.2319
	Utonulá	0.6284	0.0052	0.4904
	Dívka v ohni	0.6818	0.0039	0.3486
J. Orten, *Cesta k mrazu.* 1940 (Czech)	*Báseň pozdního léta*	0.6923	0.0077	0.3974
	Jablka	0.7784	0.0049	0.2786
	Malá elegie	0.5917	0.0032	0.3710
	Končiny štěstí	0.6519	0.0062	0.4557
	Dívka a strom	0.5614	0.0032	0.3761
	Nahá dívka	0.5361	0.0026	0.4017
	Hrdlo písně	0.5702	0.0052	0.4139
	Modlitba	0.6124	0.0042	0.3933
	Listopad	0.7264	0.0049	0.2954
	Takový smutný den	0.6751	0.0049	0.3014
	Návrat	0.6328	0.0045	0.4090
	Rty písně	0.6864	0.0048	0.3430
	Holub na střeše	0.7131	0.0057	0.3111
	Bílý obraz	0.3697	0.0019	0.6302
	Ústí písně	0.5420	0.0043	0.4080
	Skloň se k trávě	0.5912	0.0048	0.4180
	Báseň z kamene	0.6159	0.0041	0.3459
	V mamince	0.7422	0.0084	0.3277
	I. Tolik to zní	0.5403	0.0039	0.4259
	II. Veliký pohyb	0.4894	0.0052	0.4858
J. Orten, *Ohnice.* 1941 (Czech)	*První báseň*	0.6210	0.0027	0.3408
	Čí jsem?	0.7491	0.0055	0.2273
	Kouří se…	0.6728	0.0030	0.2780
	Život	0.6082	0.0044	0.3982
	Na starých…	0.5886	0.0051	0.3743
	Na zimu	0.5767	0.0046	0.4352
	Zem…	0.7524	0.0032	0.2399
	Vždycky se…	0.5924	0.0056	0.5449
	Když tma…	0.7317	0.0041	0.2443
	Po smrti	0.6501	0.0046	0.3273
	První noc	0.7372	0.0033	0.2828
	Báseň léta	0.7057	0.0040	0.2702
	Podzim	0.6134	0.0065	0.4336
	Ta krajina	0.4877	0.0013	0.4319
	V noci	0.7604	0.0076	0.3028
	Potkal jsem…	0.6909	0.0075	0.3052
	Báseň šera	0.4943	0.0030	0.5371
	Jsem na polibek…	0.2996	0.0019	0.7649
	Co mi řekl kanárek	0.5086	0.0066	0.4847
	Co jsem odpověděl kanárkovi	0.6869	0.0075	0.3097

6. Order according to RR

If there is no year or approximate date, we omit the text (e.g. Maori). It can be observed that some writers have a monotone sequence, other ones a concave or convex succession. Again, one could smooth the historical sequence but we have both a small number of languages and a small number of poems in each. In any case, we present the relative Repeat rate according to the date of their creation in Figure 6.1.

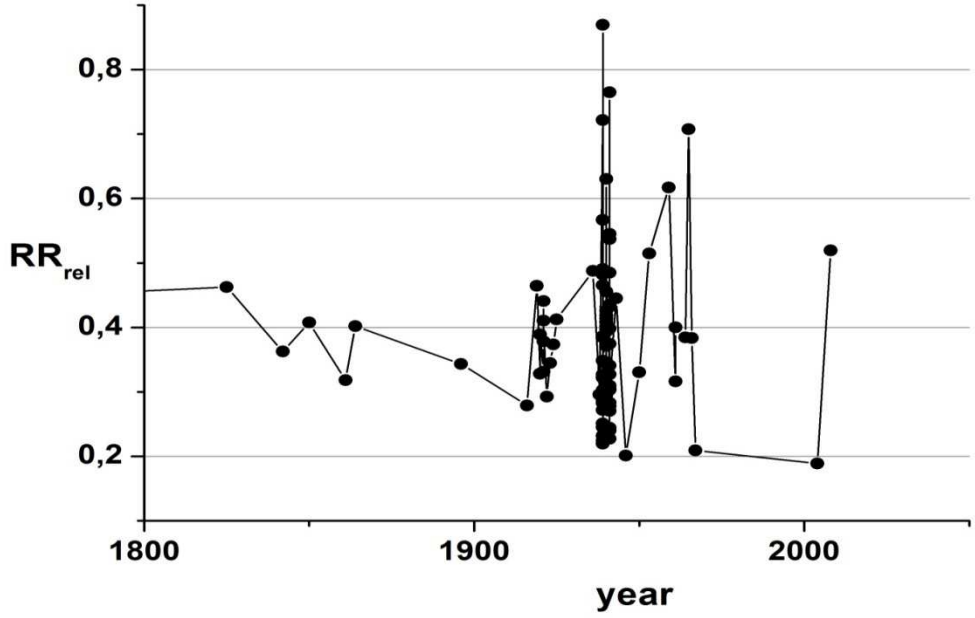


Figure 6.1. Relative Repeat rate according to the date of creation

It can be expected that after having many texts written in the same year and taking their mean of relative repeat rate we obtain a simple linear dependence. Hence, linguistically, repeat rate may be a constant but textologically it is full of boundary conditions whose exploration should belong to any text analysis.

The number of texts (and languages) is not sufficient to make even conjectures. Most probably, this work is an endless enterprise. The results may be different even for a single author.

7. Conclusions

In spite of irregularities which, e.g. in physics, can be explained by measurable boundary or initial conditions, we may conjecture that we did not analyze enough properties of poetry hence some of the irregularities will be captured by other linked properties later on.

In some cases, e.g. in Figures 3.1a and 3.1b one can observe an oscillating increase, but in other ones we are persuaded that we do not know some intrinsic conditions leading to the given oscillation. In order to find them, we recommend applying all indicators shown in the references to the given texts. A final result cannot be expected but we hope that step by step we can approach the reality.

The article shows merely a modest choice of possibilities. The literature about text analysis is very extensive. We can deduce from the results that mixing of data may lead to oscillations, rather chaotic images in which some fixed points must be searched for. In spite of the extensive literature, this discipline develops very slowly and yields sometimes very deviating results.

References

Altmann, G., Köhler, R. (2015). *Forms and Degrees of Repetition in Texts. Detection and Analysis.* Berlin/Munich/Boston: de Gruyter.

Baayen, R.H. (1989). *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation.* Amsterdam: Centrum voor Wiskunde en Informatica.

Bunge, M. (1983). *Treatise on Basic Philosophy, Vol. 6. Epistemology & Methodology II: Understanding the World.* Dordrecht/Boston/Lancaster: Reidel.

Haight, F. (1966). Some statistical problems in connection with word association data. *Journal of Mathematical Psychology 3, 217-233.*

Haight, F. (1969). Two probability distributions connected with Zipf's rank-size conjecture. *Zastosowania matematyki 10, 225-228.*

Herdan, G. (1963). Mathematical models of linguistic distribution functions. *Etudes de linguistique appliquée 2, 47-64*.

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

Kubát, M., Matlach, V., Čech, R. (2014). *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modeling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 83-94.*

McIntosh, R.P. (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology 48, 392-404.*

Mosavi Miangah, T. (2006). Automatic lemmatization of Persian words. *Journal of Quantitative Linguistics 13(1), 1-15.*

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

Popescu, I.-I. et al. (2009). *Word Frequency Studies.* Berlin: de Gruyter.

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

Popescu, I.-I., Čech, R., Altmann, G. (2011). Vocabulary richness in Slovak poetry. *Glottometrics 22, 62-72.*

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative Analysis of Poetic Texts.* Berlin/Boston: de Gruyter Mouton.

Popescu, I.-I., Zörnig, P., Altmann (2013). Arc length, vocabulary richness and text size. *Glottometrics 25, 45-53.*

Wimmer, G., Altmann, G. (1999). On vocabulary richness. *Journal of Quantitative Linguistics 6, 1-9.*

Wimmer, G., Altmann, G., Hřebiček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* Bratislava: VEDA.

Zörnig, P., Boroda, M. (1992). The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent text. *Glottometrika 13, 205-218.*

Glottometrics 38, 2017, 55-76

Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries

Jacques Savoy[1]

Abstract. To study the evolution of the rhetoric and language style of the American presidents from 1789 to 2017, we have analyzed all the annual *State of the Union* (SOTU) and inaugural addresses. Those speeches present the intentions and indicate the legislative priorities of the Chief of the Executive. Based on this relatively fixed form, this analysis corroborates several studies and emphasizes new trends and methods. With the years, the mean sentence length becomes shorter to facilitate comprehension by a larger audience. To further improve this goal, the vocabulary becomes less complex, and the percentage of big words decreases, indicating a preference for common words or expressions. The rhetoric evolves towards a more assertive and intimate tone with an increase in the occurrence of pronouns, particularly with the lemma *we*. Based on automatic classification techniques, we also observe that every president after 1961 generally has a clean and distinct style. When inspecting all presidents, some figures show a distinct break with their predecessors such as Lincoln, T. Roosevelt, Wilson, F.D. Roosevelt, or Kennedy. A closer analysis of presidencies since 1961 reveals that, with time, governmental speeches include more words related to humans and emotional terms, as well as references to God and symbolic expressions (*America*, *country*, *freedom*).

Keywords: *Governmental speeches, automatic classification, language models, authorship attribution, stylometry*.

1. Introduction

The presidential function has considerably changed from the time of the young republic under the presidency of Washington to Trump. To identify the broad trends of this evolution, several studies have analyzed presidential speeches. The number of such addresses remains low until the early twentieth century (Tulis, 1987), and their volume rose sharply after 1945 reaching an average of one speech per day under Carter's presidency (Hart, 1984). This evolution can be explained by the growing importance of journalists, media, and in particular, television. Beside their number, the content and the style of the presidential addresses has also evolved during the last two centuries. For political scientists J. Caesar et al. (1981), "speaking is governing" indicating the first role played nowadays by the governmental speeches. The president must convince the Congress as well as the citizens of his choice and persuade them (Neustadt, 1990) that the proposed policies are the most appropriate ones.

All addresses do not however have the same importance, and their type and audience vary. Faced with the impossibility of analyzing them all, past studies have limited their investigation to speeches considered as essential. For the United States, such analyses focus on the annual *State of the Union* (SOTU) allocutions and/or the inaugural addresses uttered during the swearing-in procedure. To supplement this selection, certain studies add some remarks delivered in a crisis context (e.g., address to the nation after the attacks of Sept. 11th, 2001).

[1] Computer Science Dept., University of Neuchatel, Switzerland, Jacques.Savoy@unine.ch

Based on both past studies and using our proposed methods, this article describes the main changes in the US presidential rhetoric and style from Washington to Trump using computational tools. In this view, rhetoric is defined as the art of effective and persuasive speaking, the way to motivate an audience, while language style is present as pervasive and frequent forms used by an author for mainly aesthetical value (Biber & Conrad, 2009).

The rest of this paper is composed as follows. The second section outlines the value and characteristics of the SOTU and inaugural speeches. To analyze this corpus, the third section describes previous studies and the main methods used. In the fourth, a set of overall stylistic measurements describes the main trends of the presidential rhetoric evolution. In the fifth section, the part-of-speech (POS) distribution is used to differentiate the presidents' styles. In the sixth section, automatic classification is applied to generate a map showing the differences between presidents based on their stylistic preferences. The last section provides a more detailed analysis of presidential rhetoric and style since Kennedy (1961).

2. Selection of the Governmental Speeches

To analyze the presidential rhetoric and its evolution, the majority of studies are based on a subset of the *State of the Union* (SOTU) addresses. Using computer technology, the entire set can be analyzed which includes 228 speeches given by 43 presidents from Washington (Jan. 8th, 1790) to Trump (Feb. 28th, 2017). This SOTU address is required by the US Constitution (Article II, Section 3) where it is mentioned that the president must provide information to the Congress about the state of the Union and "*measures as he shall judge necessary and expedient*". Such an address provides both an analysis of the current situation, indicates the president's priorities and presents the legislative agenda for the coming year. All of them are available on the internet (e.g., www.presidency.ucsb.edu or MillerCenter.org) and an annotated version of the 20th century addresses was recently published (Kalb & Peters, 2007).

The following reasons explain the importance of the *State of the Union* (SOTU) addresses. First, the United States occupies a position of global importance, consequently the president's vision transcends the interests of a single country. Second, this set of governmental allocutions covers a period spanning more than two centuries allowing us to analyze the evolution of the rhetoric and style. Moreover, they are delivered in a relatively stable institutional context, reducing some factors of variation. Fourth, several of these speeches have outlined significant political positions such as the Monroe Doctrine (1823), the four freedoms (F. D. Roosevelt in 1941), or the war against poverty (L. B. Johnson in 1964). More recently, this allocution was an opportunity to introduce new phrases such as "*axis of evil*" (G. W. Bush in 2002). Several studies describe in detail the institutional and political context of these speeches (Kolakowski & Neale, 2006), (Hoffman & Howard, 2007), (Shogan & Neale, 2012).

As a second major source of US government speeches, previous studies have considered the 58 inaugural addresses uttered at the beginning of each term by each of the 40 elected presidents. This set begins with the first allocution (Apr. 30th, 1789) uttered by Washington and ends with the allocution of Trump (Jan. 20th, 2017). For all of them, the oral form was chosen and the general form and topics are more diverse than with the SOTU speeches. They often present the main objectives for their term in the White House, expose their intentions regarding foreign policy, and give broad guidelines fixed for the new administration. However, their lengths show some variability. For example, the second inaugural speech of Washington (Mar. 4th, 1793) includes only

four sentences (145 words) while that of W. Harrison (Mar. 4th, 1841) was the longest with 8,356 tokens. A complete list of the selected speeches is provided in Table A.1 in the Appendix.

One can consider that (oral) speeches delivered by the presidents correspond to an oral communication form while (written) messages (e.g., sent to the Congress) must be categorized as a distinct text genre. However, as mentioned by Biber & Conrad (2009, p. 262):

> "Language that has its source in writing but performed in speech does not necessarily follow the generalization (written vs. oral). That is, a person reading a written text aloud will produce speech that has the linguistic characteristics of the written text. Similarly, written texts can be memorized and then spoken".

If one considers the president as the author of a speech, one does not take this literally. It is known that behind each important politician one can usually find a team of ghostwriters (Humes, 1997). For example, under Kennedy's presidency, the main ghostwriter was Sorensen (Carpenter & Seltzer, 1970), Madison & Hamilton behind Washington, and Favreau[2] & Keenan behind Obama. However, though some presidents were actually the author of their speeches (e.g., Lincoln), the tenant of the White House is involved more or less intensively in drafting their important speeches (Hume, 1997). The website *YouTube.com*[3] provides some video showing the preparation of some of Obama's SOTU addresses.

Finally, this analysis ignores two tenants of the White House (W. Harrison (1841), J. Garfield (1881)) because they just uttered one inaugural allocution, without any SOTU addresses, as their terms were limited to a few months. Moreover, in this study one can find only two speeches uttered by Trump (with a length of 6,252 tokens, the smallest over all remaining presidents); therefore, the findings related to the last US president must be taken with prudence.

3. Related Work and Methods

To analyze the rhetoric and style of presidential writings, the first quantitative linguistics studies focused on the word usages and their frequencies (Baayen, 2008, Jockers, 2014, Popescu et al., 2009). As the English language has a relatively simple morphology, working on inflected forms (e.g., *we*, *us*, *ours*, or *wars*, *war*) or lemmas (dictionary entries such as *we* or *war* from the previous example) often leads to similar conclusions. If the definition of a lemma is clear, the term "word" is usually ambiguous. The expression word token (or simply token) refers to an occurrence or instance of a word type (or type). For example, the sentence "I saw a man with a saw" counts seven tokens for five word types (I, saw, a, man, with), and six lemmas (I, (to) see, a, man, with, saw). In this study, the statistics are usually computed based on lemmas.

For Biber & Conrad (2009), a stylistic study should be based on ubiquitous and frequent forms. As an operational definition, our analysis is based on the k most frequent word types or lemmas, with k = 200 or 300, values that have been justified in author attribution studies (Burrows, 2002), (Savoy, 2015a).

Simple frequency analysis may report interesting aspects of the evolution of presidential style. For example, Figure 1 illustrates the evolution of the relative fre-

[2] J. Favreau comments his ghostwriter job at www.youtube.com/watch?v=zFbaesLEa4g

[3] See at www.youtube.com/watch?v=FxwcJx0-21E the behind the scene of *State of the Union* address of 2012, or at www.youtube.com/watch?v=BaR2jXboVQ0 for the SOTU 2014.

quency of the definite determiner *the* and the lemma *we*. Until Taft's presidency, some stability can be observed with a maximum for the determiner reached by Q. Adams (9.9%). For the pronoun *we*, a maximum is reached under Carter's presidency (4.7%), and this frequency remains relatively high and stable after this extreme value.

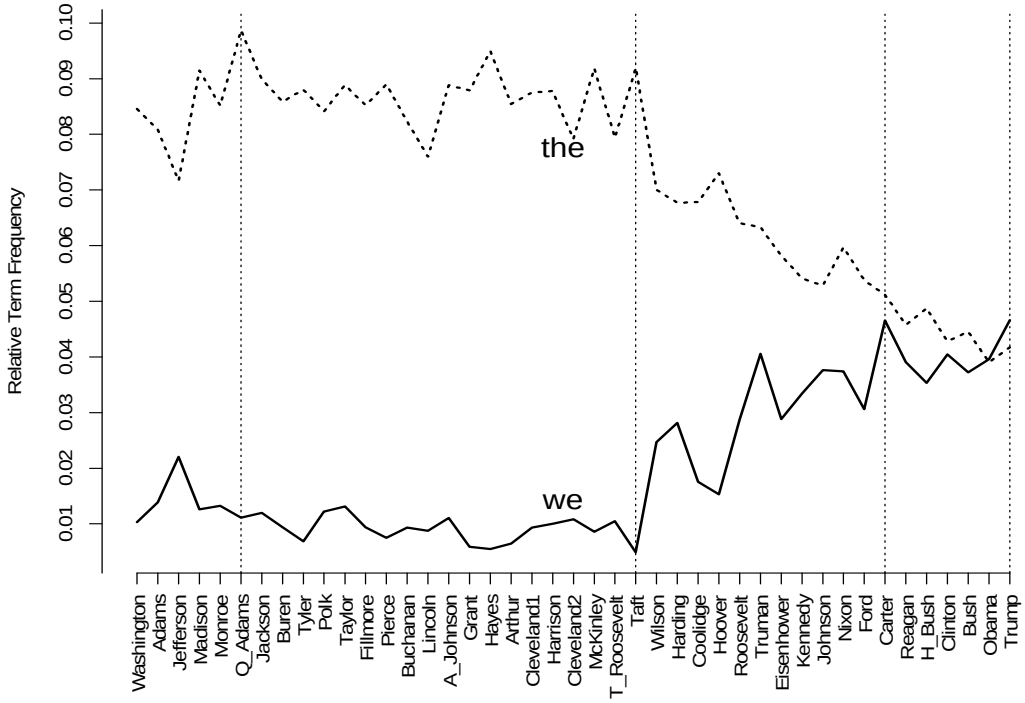


Figure 1. Evolution of the relative occurrence frequency of the lemmas *the* and *we*

In addition, a simple study may focus on the mean sentence length (number of tokens). The presence of long sentences indicates a substantiated reasoning or specifies the presence of detailed explanations. Even if a long sentence is required, its length is usually not conducive to an easy understanding.

Word length is another indicator of a message's complexity (Hart, 1984), (Tausczik & Pennebaker, 2010), the longer the words, the higher the complexity. Of course, a simple count of the number of letters to indicate the word complexity should be taken with caution. The letters are not the direct constituent of the word (a role played by the syllables or the morphemes). Moreover, the graphophonic relationship is not direct and simple in the English language. Nevertheless, Lakoff and Wehling's study (2012) indicates a relationship between the word length and word complexity analyzed be y the receiver.

> "One finding of cognitive science is that words have the most powerful effect on our minds when they are simple. The technical term is basic level. Basic-level words tend to be short. … Basic-level words are easily remembered; those messages will be best recalled that use basic-level language." (Lakoff & Wehling, 2012, p. 41).

Based on this finding, the frequency of word types composed of six letters or more indicates the use of a rich and sophisticated vocabulary (Tausczik & Pennebaker, 2010). Of course, this limit of six letters is arbitrary and another value can be used with

an alphabet-based language. For the Chinese language based on sonograms, or for the Korean based on syllabic system (called Hangul), this limit can be fixed at one or two characters[4]. In fact, Lee et al. (1999) found more than 80% of Korean nouns were composed of one or two Hangul characters, and for Chinese, Sproat (1992) reported a similar finding.

In conclusion, a text or a dialogue with a high percentage of big words tends to be more complex to understand. L. B. Johnson recognized this rhetoric problem by specifying to his ghostwriters: "I want four-letter words, and I want four sentences to the paragraph." (Sherrill, 1967).

Our two last measurements (i.e. mean sentence length and word length) are not fully independent and a relationship does exist between them, known as the Arens' law (Grzybek et al., 2008). However, the words are not the direct constituents of a sentence but only through phrases or clauses. Therefore, the relationship is less direct than expected.

To complement these frequency measures, Muller (1992) demonstrates the significance of studying the specific vocabulary for a given author (or a period), by defining a measure of lexical specificity. This approach was used for analyzing the political vocabulary in France, Quebec, and Canada (Labbé & Monière, 2003; 2008). Using this measure, the lexical specificity of the different French presidents or prime ministers can be detected. Moreover, this measure is not limited to words, but can include punctuation marks, or larger components such as word bigrams, trigrams or noun phrases (Savoy, 2010; 2016).

Other studies focus more on the Part-Of-Speech (POS) categories and their distribution. In this case, the automatic morphological tagging software (Manning & Schütze 1999) can be used to assign to each word token a syntactic and morphological label. For example, in the tagged sentence "It/PRP begins/VBZ with/IN our/ PRP$ energy/NN" we find labels attached to nouns (NN, common noun, singular, NNS common noun, plural), proper names (NNP), verbs (VB dictionary entry, VBZ for present tense, 3rd person, etc.), adjectives (JJ), pronouns (PRP), prepositions (IN), or adverbs (RB). This information can be employed to impose useful constraints for generating noun-phrases (bigrams or trigrams). For example, the pattern JJ-NN or NN-NN can extract the bigrams *clean energy*, *new jobs* or *nuclear weapons*. Considering only the most frequent bigrams results in uninteresting patterns such as *of the*, *in the*, …

Finally, several studies have suggested measuring the distance between two texts based on a predefined list of words, or the *k* most frequent word types or lemmas (with $k = 50$ to 1,000, (Burrows, 2002)) or simply according to the whole vocabulary (Labbé, 2007). Such measures have been used to determine the real author of a given document or text excerpt (Labbé, 2007), (Savoy, 2014). Knowing the distance between documents, an automatic classification algorithm (Baayen, 2008), (Jockers, 2014), (Arnold & Tilton, 2015) can visualize similarities between texts (or sets of texts) written by the same author. This allows one to draw maps showing the stylistic similarities between presidents (as shown in Section 6) or draw graphic affinities according to the main topics of their speeches (Savoy, 2015b). Other studies offer similar approaches to detect and monitor topics over time (Rule et al., 2015) or across scientific journals (Mimno, 2012).

[4] Instead of considering directly the sinogram or the Hangul character, one can count the number of strokes required to draw the corresponding character.

As another form of rhetoric analysis, several studies attempt to regroup several word types under a semantic tag. For example, under the category *Symbolism*[5], the DICTION system (Hart, 1984), (Hart et al., 2013) includes a list of words related to country (e.g., *America, nation*), ideology (e.g., *freedom, peace, rights*), or generally, political institutions or concepts (e.g., *government, law*). In a similar way, the system LIWC (Linguistic Inquiry & Word Count) (Tausczik & Pennebaker, 2010) regroups terms under syntactical, emotional or psychological categories. Such word classes may correspond to specific grammatical categories (e.g., first person singular denoted *Self* (*I, me, mine, my*)), broader ones (e.g., personal pronouns) as well as more complex ones (verbs in the future tense, auxiliary verbs). With some semantics, the LIWC system defines the inclusive class (under the label *Incl*) (e.g., *add, and, both*) or exclusive category (*Excl*) (e.g., *except, or, but, without*). As more pertinent categories, we may encounter positive emotions (*Posemo*) (e.g., *happy, hope, peace*) or negative (*Negemo*) (e.g., *humiliat*, war*), *Cognition* (e.g., *admitted, perceive*) or terms related to *Human* (e.g., *family, friend, child**). More details are given in (Tausczik & Pennebaker, 2010), (Hart, 1984). Those semantic categories will be used in Section 7 to analyze the rhetoric of the presidencies since 1961.

As examples of the usefulness of such categories, we reproduce below a passage having a high density in the category *Posemo* (words in italics, G. W. Bush), and a second text excerpt showing a high degree in the category *Affect* (Reagan).

> "For the *brave* Americans who bear the risk, no victory is *free* from sorrow. This Nation fights reluctantly, because we know the cost and we dread the days of mourning that always come. We seek *peace*. We strive for *peace*. And sometimes *peace* must be defended." G. W. Bush, *State of the Union*, Jan. 28th, 2003.

> "People of the Soviet, President Dwight Eisenhower, who *fought* by your side in World *War* II, said the essential *struggle* "is not merely man against man or nation against nation. It is man against *war*." Americans are people of *peace*. If your government wants *peace*, there will be *peace*. We can come together in *faith* and *friendship* to build a *safer* and far *better* world for our children and our children's children." R. Reagan, *State of the Union*, Jan. 25th, 1984.

4. Overall Stylistic Measurements

To define an overall measurement of the style, various studies have proposed different measures. As a first indicator, one can consider the mean sentence length (MSL) reflecting a syntactical choice. The sentence boundaries are defined by the POS tagger (Toutanova & Manning, 2000) and correspond to "strong" punctuation symbols (namely periods, question and exclamation marks). Usually, a longer sentence is more complex to understand, especially in the oral communication form. Using the *State of the Union* addresses given by the Founding Fathers, this average value is 43.9 tokens/sentence while the mean over all presidents is 33.3, as depicted in the last column of Table A.2 (shown in the Appendix). With Trump, the mean sentence length decreases to 20.5 tokens/sentence. These examples, together with Figure 3 (see below), clearly indicate that the style is changing over time. Currently, the preference goes to a shorter formulation, simpler to understand for the audience.

[5] In this study, measures, rhetorical, or stylistic indicators are shown in italic and are capitalized.

As a second global stylistic measurement, the frequency of big words (composed of six letters or more, and denoted BW) can be analyzed (Tausczik & Pennebaker, 2010). A text or a dialogue with a high percentage of big words tends to be more complex to understand. With this measurement, Eisenhower has the highest value (36.1%) over all presidents while G. H. Bush presents the lowest (25.6%) (values depicted in Table A.2 in the Appendix). As a general trend, one can see that recent presidencies (from Reagan) tend to employ less big words, in an attempt to simplify their formulations and produce less complex explanations.

Figure 2 depicts these first two stylistic measurements with the mean sentence length (y-axis, computed according to the number of tokens) and the percentage of big words in the addresses delivered by all presidents (x-axis). On the top, one can find the Founding Fathers (with Madison depicting the highest mean sentence length (48.3 tokens/sentence)), and below them the presidents of the 19th century. On the bottom left, we see the contemporary presidents (Obama, G. H. Bush, and Clinton) opting for short sentences and few big words. On the extreme right with a large percentage of big words, we discover Hoover (1929-1933) and Eisenhower (1953-1961) depicting the largest percentage of big words (36.1%). Trump presents the second smallest mean sentence length (20.5 tokens/sentence), slightly more than G. H Bush (18.9).

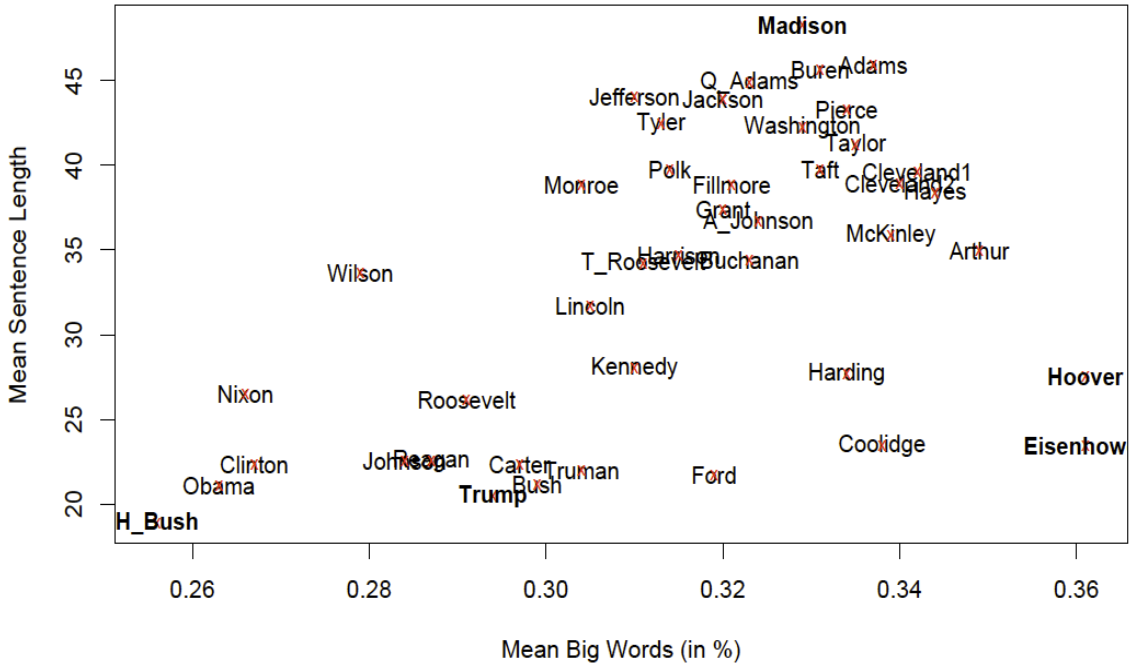


Figure 2. Relationship between mean sentence length (MSL, y-axis) vs. percentage of big words (BW) (> 5 letters, x-axis) based on the *State of the Union* and inaugural addresses

As a third stylistic indicator, the lexical density (denoted LD) can be used to reveal the informativeness of a text (Biber et al., 2002), (Hewings et al., 2005). The formulation is shown in Equation 1 where the variable *n(t)* indicates the total number of tokens (or the text length) of a text *t*, *function words(t)* the number of function words in *t*, *lexical word(t)* the number of lexical words in *t*. This latter set is composed of nouns, names, adjectives, verbs, and adverbs. On the other hand, function words regroup all other grammatical categories, namely determiners (e.g., *the*, *this*), pronouns (e.g., *you*, *us*), prepositions (e.g., *to*, *in*), conjunctions (e.g., *and*, *or*), modal verbs and auxiliary verb forms (e.g., *has*, *would*, *can*). The list of functional words for the English language

contains 273 entries. As depicted in Equation 1, this LD value is given in percentage over the number of tokens.

$$LD(t) = \frac{lexical\ word(t)}{n(t)} = 1 - \frac{functional\ word(t)}{n(t)} \quad (1)$$

A relatively high LD percentage indicates a more complex text, containing more information. Over all presidencies, the LD values varies from 50.3% (Eisenhower) who focusses more on topical forms and expressions to a minimal value of 41.5% (Wilson). For the last presidents (since 1980), this value is relatively stable and around 47% as shown in Figure 3.

The TTR (Type-Token Ratio) or the relationship between the vocabulary size and the number of word types (Baayen, 2008), (Popescu et al., 2009), (Mitchell, 2015) corresponds to our last global stylistic measure. High values indicate the presence of a rich vocabulary showing that the underlying text exposes many different topics or that the author tends to present a theme from several angles with different formulations. To compute this value, one can divide the vocabulary size (number of types) by the text length (number of tokens). This estimator has the drawback of being unstable, tending to decrease with text length (Baayen, 2008). To avoid this problem, a better computation is provided in (Covington & McFall, 2010) or (Popescu, 2009) that suggest taking the moving average of TTR denoted MATTR. This computation technique has been adopted.

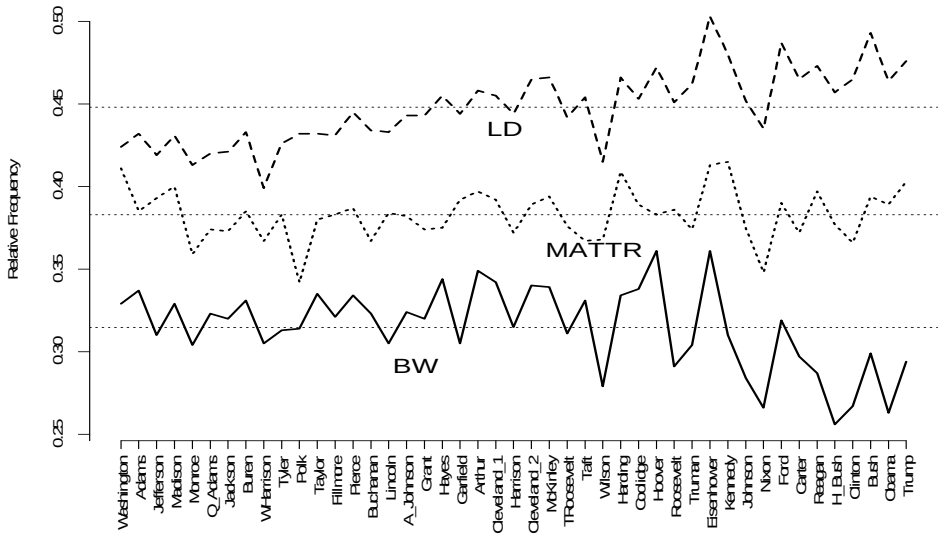


Figure 3. Evolution of lexical density (LD), big words (BW) and moving average type-token ratio (MATTR)

An overview of the values of these four stylistic indicators is reported in Table A.2 for some presidents. In this table, the largest values are depicted in bold, while the smallest are shown in italic. From this data, Washington exhibits a high MATTR value (41.2%) compared to Wilson (1913-1921) (36.8%). As other presidents presenting a rich vocabulary, one can mention Kennedy (41.2%) or Eisenhower (41.1%). On the opposite side, Wilson employs a simple vocabulary with a low MATTR value (36.1%), as well

as Nixon (34.8%) or Monroe (35.9%). For this stylistic measurement, the evolution is not related to the time but corresponds more to a personal choice.

5. Part-Of-Speech Distribution

After applying the Stanford POS tagger (Toutanova et al., 2003) over the inaugural and SOTU addresses, Table 1 shows the percentage of each grammatical category for some selected presidents. In this table, the largest values are depicted in bold, while the smallest are shown in italic. The last line under the label "Other" regroups other categories such as punctuation marks, numbers, symbols, dollar signs, or foreign words.

A first analysis indicates that nouns and adjectives represent a high percentage of Eisenhower's addresses. This emphasis on nouns can be justified by a real need for explanations. A lower frequency characterizes Lincoln's or Trump's speeches. Names (second row) are used frequently by Trump. Pronouns are very frequently used by Clinton and Obama, especially with the lemma *we* (see Figure 1). The verb and adverb categories also occur frequently in Obama's speeches, indicating that these remarks are more oriented towards action (these two findings give us a new perspective on the sentence: "*yes, we can*"). To a lesser extent, this finding can be applied to Clinton while Kennedy (JFK) tends to use this part-of-speech less. A high usage of verbs indicates dynamic thinking characterizing a person who analyzes a new problem from its historical forces or developing perspective (Pennebaker, 2009). A high occurrence frequency of determiners and prepositions characterizes the first presidency (Washington), a finding that can also be seen in Figure 1. These two categories characterize longer syntactic constructions. Since the '80s (Reagan's presidency), their use decreases and the adopted style favors shorter sentences with a higher occurrence of pronouns. This facet corresponds to a more direct tone, trying to establish a close personal relationship with the audience.

Table 1
Percentage of various POS for some selected presidencies

	Wash.	**Linc.**	**Wilson**	**Roos.**	**Eisen.**	**JFK**	**Reagan**	**Clinton**	**Obama**	**Trump**
noun	19.9	*18.2*	19.8	20.9	**22.7**	21.5	20.1	19.5	19.6	18.7
name	3.0	4.1	*1.8*	3.0	3.6	3.4	3.9	3.9	3.5	**5.8**
pron.	5.7	*4.8*	7.8	6.7	5.5	6.5	8.2	**9.5**	9.1	8.9
adj.	7.4	7.8	7.9	8.5	**9.4**	8.4	7.5	7.0	*6.5*	6.6
verb	14.9	14.6	15.0	13.8	13.3	*12.8*	14.9	15.4	**16.5**	15.1
adverb	3.8	5.0	4.7	4.1	*3.8*	4.2	4.8	4.6	**5.2**	4.7
det.	**12.9**	12.3	11.0	11.0	10.3	10.1	9.0	8.9	8.8	*8.1*
prep.	**19.3**	17.6	17.5	16.7	15.7	14.7	14.0	14.0	13.3	*12.5*
coor.	*3.5*	4.0	**4.9**	4.1	3.7	4.7	4.1	3.8	4.1	4.5
other	9.5	11.5	9.5	11.3	12.0	13.7	13.6	13.5	13.3	15.1

To verify how each presidency can deviate from an expected average style, we generate a centroid distribution over all POS tags by computing the mean distribution over all 43 presidents. Using the chi-square test (Conover, 1971), we found that each presidency deviates significantly ($p < 0.001$) from this centroid distribution. The closest presidency to this mean profile is J. Adams (1797-1801).

To better visualize the relationships between the POS categories and the presidents, a principal component analysis (PCA) (Baayen, 2008) has been applied on the data depicted in Table 1 (with some additional presidencies) using the R software (Jockers, 2014). In Figure 4, the horizontal axis emphasizes the contrast between the pronouns (and punctuation) appearing on the extreme right, and a group composed by a combination of prepositions and determiners (the two labels are also superposed) depicted on the left. At the ends of this axis, we can observe the opposition between Trump (and, to a lesser extent, Clinton and G. H. Bush) on the one hand and, on the other hand, T. Roosevelt (and, in part, Washington). As indicated, this first principal component axis represents 44.1% of all variability, while the vertical axis adds 19.3%. Thus, Figure 4 illustrates 63.4% of the total variance. Along this second axis, the verb appears on the top with Jackson and Obama as representatives. On the bottom, we encounter the group of nouns and adjectives, with Kennedy and Eisenhower as figureheads.

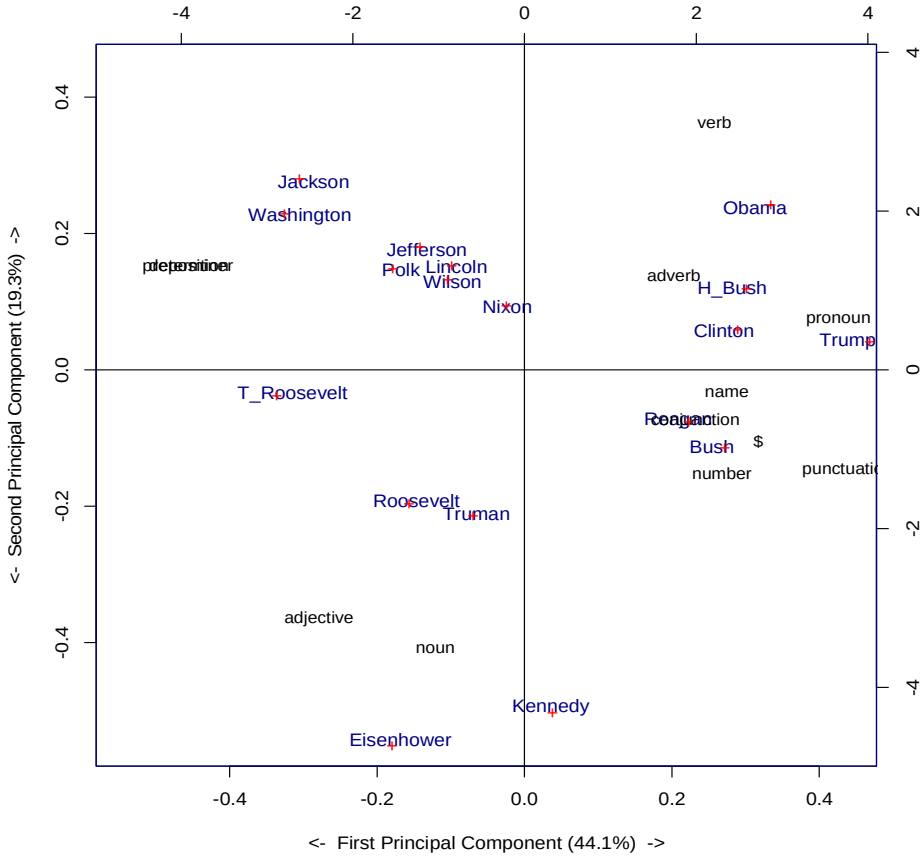


Figure 4. Principal component analysis (PCA) based on the POS distribution based on the *State of the Union* and inaugural addresses.

On the right part of Figure 4, we can observe the recent presidents with Clinton, G. H. Bush, Obama, and Trump, and just below Reagan, and G. W. Bush. These presidencies are using pronouns, verbs, adverbs, numbers, and punctuation symbols more frequently (their sentences are shorter, implying more full stops).

On the top left part, the early presidencies are regrouped with Washington, Jefferson, and Jackson, using more determiners and prepositions. The average speech, located in the center of the figure, does not have a clear representative. In this neighbor-

hood, one can see Nixon, and Wilson. Figure 4 also indicate that Eisenhower's and Kennedy's presidencies were clearly different from the others. Both are located on the bottom corresponding to speeches presenting more noun-phrases (nouns and adjectives).

Finally in this figure, the affinities between presidents do not follow a political party affiliation. For example, we find groups formed by a representative of each party (e.g., G. H. Bush – Clinton or Eisenhower – Kennedy). The subdivision into clusters seems to correspond better to a temporal proximity, a finding confirmed by Rule et al.'s study (2015).

6. Automatic Clustering Based on Stylistic Features

To globally compare the different presidential styles, the intertextual distance between two texts have been computed according to Labbé's method (2007). With this metric, the returned value depends on the overlapping between the two texts and varies between 0.0 and 1.0. Between these two extremes, the distance depends on the number of lemmas in common on both texts and their frequencies. In this study to represent each presidency, a profile is generated by concatenating all addresses delivered by a given president.

According to Labbé's definition, the intertextual distance between Profile A and Profile B is given by Equation 2 in which V_A (or V_B) indicates the vocabulary of Profile A, tf_{iA} (respectively tf_{iB}) denotes the term occurrence frequency of the ith word type in Profile A, and n_A (respectively n_B) the length of Profile A (number of tokens).

$$dist(A,B) = \frac{1}{2n_A} \sum_{i \varepsilon V_A \cup V_B} |tf_{iA} - tf_{iB}| \quad (2)$$

This formulation assumes that both texts have the same length ($n_A = n_B$). This is however rarely the case, and one needs to reduce the largest text (assuming it is Profile B) to the size of the smallest one (Profile A in our example). To achieve this, the term frequency of each word type belonging to the largest text is modified as follows:

$$tf'_{iB} = tf_{iB} \cdot {}^{n_A}/_{n_B} \quad (3)$$

To reflect only the stylistic aspects, each profile is represented by the top 300 most frequent lemmas occurring in the SOTU and inaugural addresses. Applying this distance measurement for each pair of profiles, we obtain a symmetric matrix (43 x 43 = 1,849 values). Just showing all these values does not provide a useful picture. On the other hand, this information can be used to apply an automatic classification (Baayen, 2008). The result, depicted in Figure 5, allows us to discover the clusters generated based on similar stylistic profiles.

In Figure 5, the distance between each president is visualized by a technique derived from genomic trees (Paradis, 2011), more precisely using the nj() function (Rzhrtsky and Nei, 1993), (Gascuel and Steel, 2006) available in R (Jockers, 2014). In this picture, the line length joining two presidents is proportional to the distance between them. For example, to go from Lincoln to Kennedy, we must travel a greater distance than between Lincoln and Roosevelt. Finally, to be on the left or the right, up or down, does not matter. This position is selected to allow a better overall visualization.

In this figure, the longest distance (0.337) can be found between Obama and Q. Adams, and the second largest (0.334) links Clinton to Q. Adams. The shortest distance (0.066) connects Jackson with his successor van Buren, while the second (0.075) joins the two Cleveland presidencies (1885-1889 and 1893-1897).

Following a movement from bottom to top, the presidents are placed almost in chronological order. On the bottom of the figure, the first group includes the Founding Fathers (Washington, Adams, Jefferson, Madison, and Monroe). This group covers the period from 1790 to 1825. A closer inspection reveals that Monroe is a little bit further apart from the kernel formed by the first four presidents. Although the stylistic distance is still small among these five presidents, the political vision of Jefferson or Madison (limited federal power) is different from that shared by Washington and J. Adams (strong federal government).

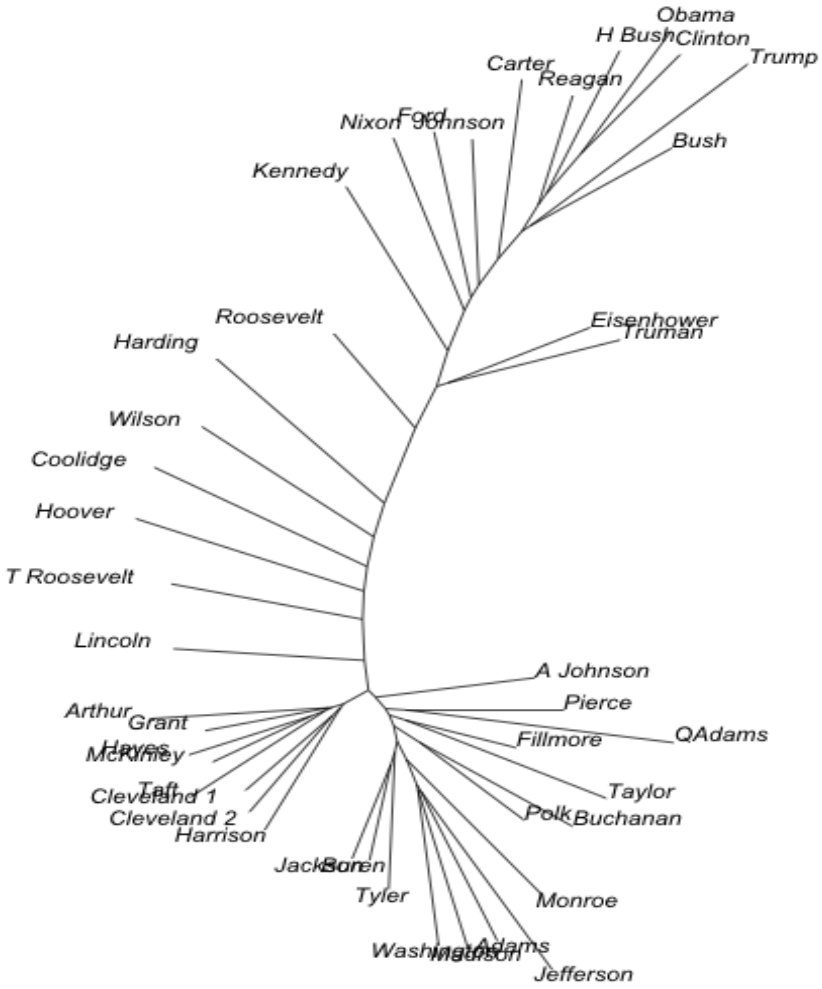


Figure 5. Tree representation of stylistic distances between the president profiles (with k = 300 most frequent lemmas) based on the *State of the Union* and inaugural addresses.

The second period comprising the years 1829-1845 is located just on the left, formed by the Democratic trio Jackson-Van Buren-Tyler. On the right, and closely related, we can find the cluster Polk (1845-49) - Buchanan (1857-1861) and the Whig duo Taylor-Fillmore (1849-1853). Just above, the pair Q. Adams (1825-1829) and Pierce (1853-1857) share a common style but with some temporal distance. Finally, one can see A. Johnson (1865-1869) as an isolated figure having a district style.

On the left, beginning with Arthur (1881-1885), we can find a group of presidents (Grant, Hayes, McKinley, Taft, Cleveland, and B. Harrison) covering the period 1869-1913. Except for the two terms of Cleveland (1885-1889, 1893-1897), all are Republicans and form a fairly homogeneous group based on stylistic considerations.

Above this large cluster, we find a sequence of presidents having a style significantly different from each other. The first in this series is Lincoln (1861-1865) who owns a style usually considered as the most beautiful. In Figure 5, this president is clearly located further away from presidents covering the same time period. A. Johnson (1865-1869), who succeeded Lincoln, also presents a particular style that is closer to his predecessors. From a stylistic evolution point of view, A. Johnson's presidency marks a step backward.

Within the first half of the 20th century, three presidents clearly stand out from a stylistic point of view. First, T. Roosevelt (1901-1909) depicts a large distance with his predecessor (McKinley) and his direct successor (Taft), both appearing in the group of presidents covering the period 1869-1913. Second, Wilson (1913-1921) modernizes the presidency; the United States become a great power wishing to play a global role (Nye, 2013). Wilson adopts a clearly distinctive style to help him in achieving this goal. The third innovative president is F.D. Roosevelt (1933-1945) whose style stands out clearly from its predecessors (Coolidge and Hoover) both located between T. Roosevelt and Wilson. Harding (1921-1923) also appears as owning a distinct style compared to the others, but his presidency was judged as one of the worst in US history (Riding et al., 1997).

From 1946, each presidency depicts a pretty distinct style that the next section will describe with more detail. As exceptions, we can find the binomial Truman-Eisenhower or the strong similarity between Clinton and Obama. A quick inspection reveals that the presidents appear in almost perfect chronological order; the first exception is L. B. Johnson appearing after Nixon-Ford, the second is G. W. Bush located closer to Carter, and the third with Trump who is located close to G.W. Bush. Figure 5 also highlights the impact of the style of Kennedy, inaugurating a brighter presidential style located at a farther distance from his predecessors.

7. Analysis of the Presidential Styles Since 1961

This last section analyzes with more detail the rhetoric and presidential style since 1961 (Caesar et al., 1981; Hart, 1984; Tulis, 1987; Neustadt, 1990; Gelderman, 1997; Kalb & Peters, 2007; Greenstein, 2009). To limit this analysis, Ford and Carter's presidencies will be ignored. The main indicators used in our comparison are reported in Table 2, with the mean over these nine presidencies indicated in the last column.

In this analysis, a selected set of semantic categories defined by the DICTION (Hart, 1984; Hart et al., 2013) or LIWC system (Tausczik & Pennebaker, 2010) have been used. Such lists may regroup specific grammatical categories such as *Self* defined by the word tokens (*I, me, my, mine*), tokens related to a given topic (e.g., *Human* with (e.g., *child**, *family*, *friend*), or *Social* with (e.g., *societ**, *speak*, *tell*, *team*)), terms denoting an emotion (*Posemo* with (e.g., *hope*, *win*, *best*), *Negemo* with (e.g., *fear*, *tear*, *sadness*)) or other rhetoric aspects such as *Cognitive* mechanism (e.g., *cause*, *think*, *organize*, *realis**), *Concreteness* words (e.g., *bank*, *college*, *troop*, *police**) or *Tentative* language forms (e.g., *maybe*, *perhaps*, *appear*).

From these indicators, one can build a centroid reflecting the average president by computing the mean over all measurements. Applying the chi-square test (Conover, 1971), we found that each presidency described in Table 2 deviates significantly

($p < 0.001$) from the centroid distribution. The closest to this average presidency is Reagan.

Kennedy's presidency corresponds to an intellectually brilliant, but impersonal, style. His rhetoric will inspire and motivate a nation and, thanks to the absence of excessive patriotism (the lowest *Symbolim* value: 1.8%, see Table 2), this motivation wins throughout the Western world. For Pennebaker (2011), Kennedy owns a complex thinking, able to convey complex problems and ideas in a rhetoric that the people can understand (highest value in the category *Tentative*: 1.8%, see Table 2). As reported in Table 1, this presidency presents a higher percentage of noun phrases (nouns and adjectives), and a low percentage of verb phrases (verbs and adverbs). Kennedy's tone is also reflected by the frequent use of words belonging to the *Exclusive* category (e.g., *but, rather, without*) (2.4% indicated in Table 2, with a mean over the nine presidencies of 2.1%), *Causal* (e.g., *because, effect*), and negation (e.g., *not, never*, 1.4%, mean: 1.2%). On simple measures, the Kennedy presidency is also characterized by greater complexity (longer sentences and a relatively high percentage of big words as shown in Figure 3). His Type-Token Ratio (MATTR) is higher (41.5%) than the average (38.3% reported in Table A.2) indicating a richer vocabulary. As for Truman, the lemma *we* (*we, us, ours*) is significantly over-used (see Figure 1) compared to previous presidents.

Table 2
Percentages of different semantic categories for some presidents

	JFK	Johnson	Nixon	Reagan	H Bush	Clinton	Bush	Obama	Trump	Mean
Self	*0.7*	1.6	1.3	1.1	**2.0**	1.5	0.9	1.2	0.9	1.2
Incl.	9.0	9.5	*8.6*	9.5	9.7	9.5	10.3	9.3	**11.0**	9.6
Excl.	2.4	1.9	2.0	2.0	2.2	2.1	*1.5*	**2.6**	1.8	2.1
Negat.	1.4	1.0	1.1	1.3	1.2	*1.0*	*1.0*	**1.5**	1.1	1.2
Symbo	*1.8*	2.4	3.2	2.8	2.4	2.5	3.1	2.2	**3.4**	2.6
Cogn.	19.5	20.1	*18.3*	19.4	**21.0**	20.4	20.2	20.7	20.5	20.0
Humar	*5.7*	7.1	6.4	7.5	7.5	8.8	7.9	7.9	**9.2**	7.6
Concr.	4.3	5.0	*4.3*	4.7	4.5	5.3	5.1	5.4	**6.1**	5.0
Tenta.	**1.8**	1.7	1.4	1.4	1.5	1.5	1.4	**1.8**	*0.8*	1.5
Social	*8.4*	9.8	9.3	10.4	10.8	12.1	11.4	10.8	**12.4**	10.6
Posemo	4.2	*3.9*	4.4	4.9	4.4	4.4	**5.2**	4.0	4.7	4.5
Negem	2.3	1.9	*1.6*	1.8	1.7	*1.6*	**3.0**	1.7	2.0	1.9

With L. B. Johnson, the presidential style becomes more direct, simple, and popular. His lexicon complexity goes down to a LD value of 45.2% (compared to 48% for Kennedy or 50.3% for Eisenhower). At the level of personal pronouns, we can detect a clear increase in the first singular pronoun (*I, me, mine*, 1.6% as reported in Table 2), a characteristic that will also be overserved under G. H. Bush's presidency.

With Nixon, the presidency becomes imperial according to the bestselling title "*The Imperial Presidency*" (written by A. M. Schlesinger). The rhetoric becomes clearly assertive, optimistic (*Posemo* of 4.4% comparted to 3.9% for Johnson), and relies on familiar words. The MATTR value is 34.8%, the second lowest value over all US presidents (Polk presents the minimal MATTR value with 34.2%). Moreover, the LD value decreases to 43.5% compared to 45.2% for Johnson. Another facet of Nixon's style is the recurrent use of terms belonging to the *Self* category (shown in italics in the following example):

> "*I* know these have no ideology, no race. *I* know America. *I* know the heart of America is good. *I* speak from *my* own heart, and the heart of *my* country, the deep concern we have for those who suffer and those who sorrow. *I* have taken an oath today in the presence of God and *my* countrymen to uphold and defend the Constitution of the United States. To that oath *I* now add this sacred commitment: *I* shall consecrate *my* Office, *my* energies, and all the wisdom *I* can summon to the cause of peace among nations." R. Nixon, first Inaugural Address, Jan. 20th, 1969.

The pronoun *I* is usually more frequent in the SOTU addresses (mean: 1.3%) than in the inaugural allocutions (mean: 0.9%). Thus the over-use of the *Self* category in this passage extracted from his inaugural address is independent of the speech type. Moreover, in Nixon's addresses, the verbs are conjugated more frequently in the present tense (see previous example). Hart (1984) also points out that Nixon can be seen as a demagogue, being the president using familiar terms more frequently, using a limited vocabulary and promoting *Symbolism* (3.2% in Table 2 compared to 1.8% for Kennedy or 2.4% for Johnson).

With Reagan, America discovers the Great Communicator (but not a great orator or a great style master) (Hart, 1984). This president fits perfectly on television, accompanying his speeches with a voice and physical presence which provides an undeniable emotional embellishment. Seen as honest, sincere, and believing in simple values, the president knows how to adapt his speech to the context. Using fewer adjectives than his predecessors (see Table 1), he will emphasize simple phrases ("*we've come to a turning point*" or "*we've tried to fight inflation*") using a rather limited lexicon (a low MATTR value of 39.7% compared to 41.5% for Kennedy). This familiar vocabulary includes a large proportion of terms related to *Human* (e.g., *child, parent*). For this category, Table 2 indicates a percentage of 7.5% for Reagan compared to 6.4% for Nixon, of 5.7% for Kennedy. In addition, Reagan's addresses contain more symbolic (e.g., *freedom, America*), and religious expressions (e.g. *temple*) as shown by this example.

> "If *we* do that, if *we* care what *our children* and *our children's children will say* of *us*, if *we* want them one day to be thankful for what *we* did here in these *temples* of *freedom, we will* work *together* to make *America* better for *our* having been here, not just in this year or this *decade* but in the next century and beyond." R. Reagan, *State of the Union*, Jan. 25th, 1983.

One of Reagan's features is the higher proportion of verbs (in Table 1, 14.9% for Reagan compared to 12.8% for Kennedy), and words indicating *Action* (e.g., *achieve, deliver, recommend, teach*). With Reagan's presidency, references to God, or religion in general, become significantly more frequent, a rhetoric aspect followed by his successors as illustrated in the following passage:

> "I ask you to bow your heads. *Heavenly* Father, we bow our heads and thank *You* for *Your* love. … Make us strong to do *Your* work, *willing* to heed and hear *Your* will, and write on our hearts these words: "Use power to help people." For we are given power not to *advance* our own *purposes*, nor to make a great show in the world, nor a name. There is but one just use of power, and it is to serve people. Help us *remember, Lord. Amen*." G. H. Bush, Inaugural address, Jan. 20th, 1989.

In fact, Reagan is the president using the term *God* most often with a relative frequency of 1.15‰, followed by G. H. Bush (0.57‰), G. W. Bush (0.49‰), and Obama

(0.47‰). This reference introduced more frequently under Reagan's presidency is now part of the vocabulary of contemporary presidents.

According to Hart et al. (2013), the language of G. H. Bush's presidency is accompanied by a greater emphasis on patriotism, religious language (see example shown above), and references to citizens and people. The rhetoric is mainly assertive with an absence of doubt. The president employs the pronoun *I* frequently (*Self*: 2% compared to a mean of 1.2%) and presents the highest value in the category *Cognitive* (21% in Table 2). We can observe a slightly larger proportion of verbs (15.5%), a grammatical category that will increase with the next presidents (as indicated in Table 1). Overall, G. H. Bush's presidency is also characterized by a low percentage of big words (25.6% compared to a mean of 31.5%, see Table A.2), and a very short mean sentence length (18.8 tokens/sentence), the smallest value over all US presidencies (see Figure 2).

With Clinton, America knows the birth of the digital economy, but also a president who is faced with an impeachment proceeding (Lewinsky scandal). Clinton remains however one of the most popular presidents with a rhetoric combining a realistic tone (colloquial, concrete, with an interest in the *Human* (8.8% in Table 2 compared to a mean of 7.6%)), that avoids complex formulas. The MATTR indicates one of the lowest values (36.6%) since 1945, signaling a limited vocabulary and a clear tendency to repeat the same formulations and terms. For Americans, the president speaks a language they understand, he is one of them and they gain a sense of confidence in him (despite his lies in the Lewinsky affair) (Hart et al., 2013). The following passage indicates a high percentage of terms related to the Collectives thematic (e.g., *country*, *family*, *economy*):

> "The *world economy*, the *world* environment, the *world* AIDS crisis, the *world* arms *race*: they affect us all. Today, as an older order passes, the new *world* is more free but less stable. Communism's collapse has called forth old animosities and new dangers. Clearly, America must continue to lead the *world* we did so much to make." B. Clinton, first Inaugural Address, Jan. 20th, 1993.

The arrival of G. W. Bush marks the advent of a much more partisan presidency. The presidency wants to closely monitor the political agenda, conceives in secret the needed policies, and then sells them with enough authoritarianism and largely ignores the press (Jacobson, 2008). The presidential picture is created around the adjectives "*arrogant, critical, messianic*". Table 2 clearly indicates that this presidency can be characterized by using more emotional words, both positive (5.2%) and negative (3%). In the following passage, words related to the emotion categories are indicated in italics:

> "We will build our *defenses* beyond *challenge*, lest *weakness* invite *challenge*. We will *confront weapons* of mass *destruction*, so that a new century is spared new *horrors*. The *enemies* of *liberty* and our country should make no *mistake*: America remains *engaged* in the world, by history and by choice, shaping a balance of power that *favors freedom*. We will defend our allies and our *interests*. We will show purpose without arrogance. We will meet *aggression* and *bad faith* with *resolve* and *strength*." G. W. Bush, first Inaugural address, Jan. 20th, 2001.

During the 2004 re-election campaign, Kerry's and Bush's programs prove to be close (social security, global warming, embezzlement of large companies) (Slatcher et al., 2007). Overall, Bush's rhetoric is marked by an optimism and the frequent use of symbolic terms (3.1% in Table 2).

The election of Obama was the first to be marked by a wide use of the Internet and the social networks (Facebook, YouTube, blogs, Twitter). The first years passed in a difficult context (financial crisis and unemployment). The presidential tone, however, remains optimistic and the president insists on some key issues by using many repeats. The recourse to *Tentative* (1.8%), *Exclusive* (2.6%), and *Negation* (1.5%) are frequent, and these three categories present the highest values over the nine presidencies. He also chooses to return to a rhetoric emphasizing the *Concreteness* (5.4% as reported in Table 2) and *Human* (7.9%) terms. The tone is however less emotional than his predecessors (both categories *Posemo* and *Negemo* are below the mean, see Table 2). The percentage of big words (26.3% indicated in Table A.2) is low indicating a real concern to avoid complex formulations. Obama's rhetoric is also characterized by a higher frequency of story-telling to concretely illustrate a number (unemployment rate) or an action. Both are shown in the following example:

> "Today in America, a teacher spent extra time with a student who needed it, and did her part to lift America's graduation rate to its highest level in more than three decades. An entrepreneur flipped on the lights in her tech startup, and did her part to add to the more than 8 million new jobs our businesses have created over the past four years." B. Obama, *State of the Union*, Jan. 28th, 2014.

The 2016 US presidential election was characterized by two figures, H. Clinton & D. Trump, both unloved by the majority of Americans. Ignoring every norm of American politics and hoping to reflect the silent majority, Trump says what he thinks, and thus appears sincere and authentic. His rhetoric is centered around the high values for the categories *Inclusive* (11%, see Table 2, e.g., *together*, *with*) and *Symbolism* (3.4%) terms (e.g., *America*, *country*, *freedom*). In order to be understood by everybody, the mean sentence length is rather short (20.5 token/sentence, see Table A.2), the second lowest value over all US presidencies. Moreover, a high percentage of *Concreteness* terms (6.1%) boosts a direct formulation. Finally, this presidency shows the highest value in the categories *Social* (12.4%) and *Human* (9.2%). The following example illustrates these two aspects with words depicted in italics.

> "But to create this future, *we* must work with, not against, the *men* and *women* of law enforcement. *We* must build bridges of cooperation and trust, not drive the wedge of disunity and division. Police and sheriffs are members of *our community*. They are *friends* and *neighbors*, *they* are *mothers* and *fathers*, *sons* and *daughters*, and *they* leave behind *loved* ones every day *who* worry whether or not *they*'ll come home safe and sound. *We* must support the incredible *men* and *women* of law enforcement. And *we* must support the victims of crime." D. Trump, *State of the Union* Address, Feb. 28th, 2017.

8. Conclusion

The eloquence of the president, his power of persuasion (Neustadt, 1990) and, in general, his rhetoric and stylistic choices allow him to explain his positions and to justify his actions. For the United States before T. Roosevelt (1901-1908), the presidency does not fully match that vision. Simply by looking at the number of presidential speeches per year (Tulis, 1987), the president appears very infrequently in public to address his remarks, that are often limited to thanks. With the emergence of a strong presidential power (Nye, 2013), governmental speeches become more frequent and important.

By analyzing the general evolution of the presidential rhetoric over the past two hundred years, some of our measurements indicate a trend towards simplification. The sentences become shorter, the percentage of big words decreases (see Figure 3), and complex reasoning disappears. For other measures such as the MATTR or the lexical density (see Figure 3) do not corroborate such a simplification. The last presidents tend to employ more verbs and pronouns, while nouns and adjectives are reduced (see Table 1). The vocabulary is opened to a more poetic tone as well as presenting more abstract expressions and includes more religious vocabulary (Lim, 2002) (see examples in Section 7). The presidents offer an optimistic vision of the future, while being themselves more assertive; doubt tends to disappear from their addresses. The emotional terms tend to be more frequent and references to family and human beings occur more frequently (see Table 2). While being mostly enthusiastic, they tend to establish a dialogue, or, at least, a relationship with citizens and the people. The language style aims to be more intimate with a greater frequency of pronouns like *we*, or *I* (see Figure 1). The adoption of story-telling reinforces this tendency.

There is however a difference between what the president says and what he does or what he gets. Indeed, the essential purpose of the SOTU address is to explain the intention and legislative propositions of the White House and to justify budget requests to the Congress. On this last point, for the period from 1965 to 2002, the success rate of demands for credit by the president was 52% during his first term, and 39% during his second term (Hoffman & Howard, 2006). This finding indicates that if the presidential speech is the vehicle of the government's priorities, and a prerequisite for its actions, it is far from being imperial...

References

Arnold, T., & Tilton, L. (2015). *Humanities Data in R. Exploring Networks, Geospatial Data, Images, and Text*. Heidelberg: Springer-Verlag Press.

Baayen, H.R. (2008). *Analyzing Linguistic Data. A Practical Introduction Using R*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Leech, G. (2002). *The Longman Student Grammar of Spoken and Written English*. London: Longman.

Burrows, J.F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17, 267-287.

Caesar, J.W., Thurow, G.E., Tulis, J., & Bessette, J.M. (1981). The Rise of Rhetorical Presidency. *Presidential Studies Quarterly*, 11, 2, 158-171.

Carpenter, R.H., & Seltzer, R.V. (1970). On Nixon's Kennedy Style. *Speaker and Gavel*, 7, 41.

Gascuel, O., & Steel, M. (2006). Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23, 1997-2000.

Gelderman, C. (1997). *All the Presidents' Words. The Bully Pulpit and the Creation of the Virtual Presidency*. New York: Walker & Co.

Conover, W.J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Greenstein, F.I. (2009). *The Presidential Difference: The Leadership Style from FDR to Barack Obama*. Princeton: Princeton University Press.

Grzybek, P., Kelih, E., & Stadlober, E. (2008). The Relationship between Word Length and Sentence Length: An Intra-Systemic Perspective in the Core Data Structure. *Glottometrics*, 16, 111-121.

Hart, R.P. (1984). *Verbal Style and the Presidency*. Orlando: Academic Press.

Hart, R.P., Childers, J.P., & Lind, C.J. (2013). *Political Tone. How Leaders Talk & Why*. Chicago: Chicago University Press.

Hoffman, D.R., & Howard, A.D. (2006). *Addressing the* State of the Union. *The Evolution and Impact of the President's Big Speech*. Boulder: Lynne Rienner.

Humes, J.C. (1997). *Confessions of a White House Ghostwriter: Five Presidents and Other Political Adventures*. Washington: Regnecy Publ.

Jacobson, G.C. (2008). *A Divider, not a Uniter: George W. Bush and the American People*. New York: Pearson-Longman.

Jockers, M.L. (2014). *Text Analysis with R for Students of Literature*. Heidelberg: Spinger-Verlag.

Kalb, D., & Peters, G. (2007). State of the Union. *Presidential Rhetoric from Woodrow Wilson to George W. Bush*. Washington CQ Press.

Kolakowski, M., & Neale T.H. (2006). The President's *State of the Union* Message: Frequently Asked Questions. *Congressional Research Service*, n0 RS20021.

Labbé, D. (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14, 1, 33-80.

Labbé, D., & Monière, D. (2003). *Le discours gouvernemental, Canada, Québec, France (1945-2000)*. Paris : Champion.

Labbé, D., & Monière, D. (2008). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois : 1960-2005*. Montreal : Monière-Wollank.

Lakoff, G., & Wehling, E. (2012). *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. New York: Free Press.

Lee, J.J., Cho, H.Y., and Park, H.R. (1999). N-gram-Based Indexing for Korean Text Retrieval. *Information Processing & Management*, 35, 427-441

Lim, E.T. (2002). Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, 32, 2, 328-348.

Manning, C.D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *ACM Journal on Computing and Cultural Heritage*, 5, 1-19.

Mitchell, D. (2015). Type-Token Models: A Comparative Study. *Journal of Quantitative Linguistics*, 22, 1-21.

Muller, C. (1992), *Principes et méthodes de statistique lexicale*. Paris : Champion.

Neustadt, R.E. (1990). *The Presidential Power and the Modern Presidents. The Politics of Leadership from Roosevelt to Reagan*. New York Free Press.

Nye, J.S. (2013). *Presidential Leadership and the Creation of the American Era*. Princeton: Princeton University Press.

Paradis, E. (2011). *Analysis of Phylogenetics and Evolution with R*. New York: Springer.

Pennebaker, J.W. (2009). What is "I" saying? (guest post). The Language Log. http://languagelog.ldc.upenn.edu/nll/?p=1651 (access Feb. 19th., 2016).

Pennebaker, J.W. (2011). *The Secret Life of Pronouns. What our Words Say About us*. New York: Bloomsbury Press.

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: De Gruyter Mouton.

Ridings, W. J., & McIver, S. B. (1997). *Rating the Presidents: A Ranking of U.S. Leaders, from the Great and Honorable to the Dishonest and Incompetent*. Secaucus: Carol Publishing.

Rule, A., Cointet, J.-P., & Bearman, P.S. (2015). Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse. 1790-2014. In: *Proceedings of the National Academy of Sciences*, 112, 35, 1-8.

Rzhrtsky, A., & Nei, M. (1993). A Simple Method for Estimating and Testing Minimum-Evolution Trees. *Molecular Biology and Evolution*, 10, 1073-1095.

Savoy, J. (2010). Lexical Analysis of US Political Speeches. *Journal of Quantitative Linguistics*. 17, 2, 123-141.

Savoy, J. (2014). Authorship Attribution using Political Speeches. In: *Proceedings Qualico*, June, Reprint in Tuzzi, A., Benešová, M., & Mačutek, J. (eds.), 2015, *Recent Contributions to Quantitative Linguistics*. Berlin: De Gruyter Mouton.

Savoy, J. (2015a). Comparative Evaluation of Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30, 2, 246-261.

Savoy, J. (2015b). Text Clustering: An Application with the *State of the Union* Addresses. *Journal of the American Society for Information Science & Technology*, 66, 8, 1645-1654.

Savoy, J. (2016). Text Representation Strategies: An Example with the *State of the Union* Addresses. *Journal of the American Society for Information Science & Technology*, 67(8), 1858-1870.

Sherrill, R. (1967). *The Accidental President*. New York: Grossman.

Shogan, C.J., & Neale, T.H. (2012). The President's *State of the Union* Address: Tradition, Function, and Policy Implications. *Congressional Research Service*, no 7-5700.

Slatcher, R.B., Chung, C.K., Pennebaker, J.W. & Stone, L.D. (2007). Winning Words: Individual Differences in Linguistic Style among U.S. Presidential and Vice Presidential Candidates. *Journal of Research in Personality*, 41, 63-75.

Sproat, R. (1992). *Morphology and Computation*. Cambridge: The MIT Press.

Tausczik, Y.R., & Pennebaker, J.W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29, 1, 24-54.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclid Dependency Network. In *Proceedings of NAACL 2003, ACL*, Edmonton (AL), 252-259.

Tulis, J. (1987). *The Rhetorical Presidency*. Princeton: Princeton University Press.

Appendix

Table A.1. List of 45 US Presidents with their number of Inaugural and SOTU speeches together with their political affiliation

#	Name	Inaugural	SOTU	From	To	Party
1	George Washington	2	8	1789	1797	Ind.
2	John Adams	1	4	1797	1801	F
3	Thomas Jefferson	2	8	1801	1809	D–R
4	James Madison	2	8	1809	1817	D–R
5	James Monroe	2	8	1817	1825	D–R
6	John Quincy Adams	1	4	1825	1829	N–R
7	Andrew Jackson	2	8	1829	1837	D
8	Martin Van Buren	1	4	1837	1841	D
9	William H. Harrison	1		1841	1841	Whig
10	John Tyler		4	1841	1845	D
11	James Polk	1	4	1845	1849	D
12	Zachary Taylor	1	1	1849	1850	Whig
13	Millard Fillmore		3	1850	1853	Whig
14	Franklin Pierce	1	4	1853	1857	D
15	James Buchanan	1	4	1857	1861	D
16	Abraham Lincoln	2	4	1861	1865	R
17	Andrew Johnson		4	1865	1869	D
18	Ulysses S. Grant	2	8	1869	1877	R
19	Rutherford B. Hayes	1	4	1877	1881	R
20	James A. Garfield	1		1881	1881	R
21	Chester A. Arthur		4	1881	1885	R
22	Grover Cleveland	1	4	1885	1889	D
23	Benjamin Harrison	1	4	1889	1893	R
24	Grover Cleveland	1	4	1893	1897	D
25	William McKinley	2	4	1897	1901	R
26	Theodore Roosevelt	1	8	1901	1909	R
27	William H. Taft	1	4	1909	1913	R
28	Woodrow Wilson	2	8	1913	1921	D
29	Warren Harding	1	2	1921	1923	R
30	Calvin Coolidge	1	6	1923	1929	R
31	Herbert Hoover	1	4	1929	1933	R
32	Franklin D.	4	12	1933	1945	D
33	Harry S. Truman	1	7	1945	1953	D
34	Dwight D.	2	9	1953	1961	R
35	John F. Kennedy	1	3	1961	1963	D
36	Lyndon B. Johnson	1	6	1963	1969	D
37	Richard Nixon	2	5	1969	1974	R
38	Gerald R. Ford		3	1974	1977	R
39	Jimmy Carter	1	3	1977	1981	D
40	Ronald Reagan	2	7	1981	1989	R
41	George H. Bush	1	4	1989	1993	R
42	Bill Clinton	2	8	1993	2001	D
43	George W. Bush	2	8	2001	2009	R
44	Barack Obama	2	8	2009	2017	D
45	Donald Trump	1	1	2017	-	R

Ind.: Independent D-R: Democratic-Republican N-R: National-Republican
D: Democratic R: Republican

Table A.2. Overall stylistic measurements for some selected presidencies

	Wash.	Madis.	Wilson	Eisen.	Reagan	H. Bush	Obama	Trump	**Mean**
MSL	42.2	**48.3**	33.6	23.4	22.6	*18.8*	21.1	20.5	33.3
BW	32.9%	32.9%	27.9%	**36.1**%	28.7%	*25.6*%	26.3%	29.4%	31.5%
LD	42.4%	43.1%	*41.5*%	**50.3**%	47.3%	45.7%	46.4%	47.6%	44.8%
MATTR	**41.2**%	40.0%	*36.8*%	41.1%	39.7%	37.7%	38.9%	40.3%	38.3%

MSL: mean sentence length
BW: percentage of Big Words
LD: Lexical Density
MATTR: Moving Average Type-Token Ratio

Glottometrics 38, 2017, 77-106

Some Properties of Adnominals in Russian Texts

Sergej Andreev, Ioan-Iovitz Popescu, Gabriel Altmann

Abstract. In the present article the Russian adnominals are described, their distribution is studied and motifs that can be constructed are examined. It is shown that there are some regularities behind their use even if the writers of texts write spontaneously. The authors hope to have found the way to a kind of a new law.

Keywords: *Russian, adnominals, motifs*

Introduction

Our aim is the study of motifs constructed from the sequence of adnominals in Russian. Using motifs, one evidently enters a hierarchically more abstract level. They are constructed of adnominal types whose list for Russian can be found at the end of the Introduction. While moving at the classical levels (phoneme, syllable, morpheme, word, phrase, clause, sentence, etc.) there are for every language ready-made segmentation rules, description of forms, etc., for a slightly different look many entities must be redefined. It is to be remarked that definitions are not "given" (or true) but set up by us and they are formulated in such a way that a kind of segmentation of the text is possible. This is nothing else but a way onto a hierarchically higher level just as in mathematics where nothing exists in reality, everything is constructed by us. One still remembers that there are many definitions of word and sentence and one of them is adequate (not true!) for the given examination, another one for other examination, etc. A definition is appropriate/fruitful if it allows us to set up hypotheses, to find some regularities, to insert the results in a theory, etc.

Up to now, several researchers could show that Köhlerian motifs can be used as "legal" entities because a number of hypotheses have already been corroborated. The motifs of adnominals can be used to compare texts, authors, languages, their behavior, their place in the Köhlerian circuits, i.e. their relation to other entities can be studied, hypotheses about their behavior can be tested, etc. Motifs create a new level in which both qualitative and quantitative entities may exist. Unfortunately, each examination is merely a drop in the sea because nobody can investigate all texts in all languages. Hence, the only success may be a stepwise corroboration of a hypothesis.

Here we shall restrict ourselves to modern Russian texts and search in them for adnominals. The data-base includes 42 texts which belong to modern Russian authors who have been very popular in the 1990s – 2000s and are still popular now. Their popularity is reflected by very large circulations of their works which were reprinted many times, numerous films based on their works and by numerous readers' and critics' ratings. All the authors with one exception continue their creative work. The exception is Sergej Dovlatov who died in 1990 in the USA in emigration; his works became known among the general public in Russia in the 90s and thus can be attributed to this period.

The authors are very different in their creative manner and style, their works belong to different genres. From the point of view of the form these works include novels, novellas ('short novels' or 'long stories') and short stories. Novella (*povest'* – a very popular form in Russian literary tradition), being an intermediate form between the novel and the story both in size and in contents, creates certain difficulties in the classification of some of the works. It

should be also mentioned that the authors themselves sometimes use these terms interchangeably. That is why in borderline cases we shall use the term 'novel'.

The analyzed works belong to different genres, and are often characterized by a certain combination of the features of different genres. One group of works consists of detective stories – classic criminal (Koretskij, Marinina), historical (Akunin), action (Bushkov) and ironic (Dontsova, Ustinova) detective. It should be emphasized that women authors, especially Dontsova and Ustinova combine detective contents with the genre of the love-story. This love-story genre is also present in the works of Demidova and Tret'jakova, in the latter in combination with historical novel genre. Several works by Pelevin and Prilepin may be characterized to some extent as postmodern literature. The genre of the rest of the authors may be attributed to the so-called mainstream trend of belles-lettres, focusing on the moral evolution of the characters.

All the texts were taken from the beginning of the works or else comprise a whole work (some short stories). In the abstract from *Patologii* by Prilepin the part 'Afterward' is included into our analysis because it actually comes before the first chapter.

The average size of the extracts is about 4000 words, in total over 160 thousand. The information about the titles of the works and other information are given in the Tables and in the Appendix.

The adnominals and their abbreviations are presented in the list below. Considering motifs of adnominals, we do not distinguish their position (left or right), only their presence in the sequence.

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 headword, 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 pronoun (*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).

2. Frequencies

Every writer works with special means. Though (s)he does not classify, (s)he has a certain pattern of possible means which are applied according to necessity or possibility. A personal style must be clearly distinguishable from that of other writers. The simplest way to ascertain the existence of a tendency is the comparison of respective adnominals in two or more texts. This can be done e.g. by using the chi-square test for homogeneity of two texts, or by ranking the adnominals in two texts and compare the ranking. This is, at the same time, also a possibility to form a classification of texts or to study the development in one language.

Here we shall simply compute the numbers of individual adnominals in all texts and first present them in a table (cf. Table 2.1)

Table 2.1
Frequencies of adnominals in individual texts

Adnom	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**T11**	**T12**	**T13**	**T14**	**T15**	**T16**
A	229	198	213	219	197	103	198	350	267	349	194	220	209	192	252	358
ADV	3	2	2	1	1	1	0	3	5	2	1	4	4	2	2	1
AO	1	4	6	3	9	0	3	10	0	6	4	3	0	0	2	1
AP	6	18	16	9	41	17	13	18	19	52	17	34	68	23	26	37
APAJ	8	10	9	13	6	6	2	9	5	22	8	24	14	2	8	10
APX	1	8	1	12	8	2	8	9	2	13	9	9	10	33	28	16
AY	6	13	3	13	2	2	4	8	10	6	2	2	9	8	10	7
CN	5	8	2	8	2	4	3	6	2	8	5	4	1	6	8	4
DAT	0	0	0	2	0	0	0	0	2	0	0	1	0	2	7	2
DETF	32	26	28	17	43	22	30	36	26	24	27	38	22	20	22	45
DETH	4	6	7	1	5	8	4	11	12	15	2	4	2	3	4	5
DETN	3	3	3	5	2	2	5	4	1	1	2	3	4	0	0	4

DETQ	36	30	13	16	39	17	30	22	17	29	20	44	35	19	22	42
DETS	59	43	45	46	44	35	68	55	41	23	54	77	65	61	49	45
DETV	1	2	1	2	7	1	8	2	2	1	1	1	0	2	0	2
DETW	0	2	7	2	0	0	0	3	0	0	3	1	3	0	1	0
G	83	107	107	108	90	54	113	137	121	121	58	92	89	87	138	106
I	4	3	2	7	9	2	7	6	0	1	1	0	14	7	8	0
INSTR	0	0	0	2	0	0	0	0	0	1	0	0	0	0	2	1
PR	59	50	45	56	40	46	69	71	43	94	42	50	46	43	53	72
PT	24	21	16	19	25	5	12	35	29	35	19	14	29	13	23	34
PTY	29	24	15	22	15	14	20	24	29	32	18	5	7	28	32	43
RC	21	34	15	17	31	29	40	29	20	23	21	22	13	21	37	49
Sum	614	612	556	600	616	370	637	848	653	858	508	652	644	572	734	884

Adnom	**T17**	**T18**	**T19**	**T20**	**T21**	**T22**	**T23**	**T24**	**T25**	**T26**	**T27**	**T28**	**T29**	**T30**	**T31**	**T32**
A	405	211	168	272	259	320	368	139	297	231	144	531	316	213	174	201
ADV	4	1	4	2	1	1	1	0	3	2	0	1	1	0	0	0
AO	6	10	7	3	1	2	2	1	5	2	3	3	16	1	2	12
AP	28	15	17	8	16	23	35	2	22	45	13	41	13	11	8	34
APAJ	6	3	3	5	4	2	4	1	2	6	2	6	5	8	9	8
APX	10	4	2	6	10	21	10	2	9	16	3	8	4	20	8	20
AY	11	6	10	7	5	6	8	1	5	8	2	9	5	2	3	4
CN	5	1	0	2	3	3	3	2	4	12	4	9	10	3	1	2
DAT	0	0	0	0	0	0	1	0	1	1	3	3	3	1	0	0
DETF	40	26	32	20	32	22	22	8	43	50	34	69	25	15	18	20
DETH	4	21	20	8	9	10	6	0	14	7	1	6	2	5	6	6
DETN	2	13	5	8	3	3	0	2	4	3	0	0	0	0	0	0
DETQ	30	35	27	19	16	34	17	19	28	28	19	42	17	12	8	19
DETS	72	47	35	29	30	38	31	7	25	36	23	95	45	44	26	38
DETV	2	1	3	3	2	2	4	0	2	2	3	7	1	0	0	1
DETW	0	5	4	3	1	1	4	1	1	9	3	2	2	0	2	2
G	140	60	50	51	114	42	89	60	84	113	67	196	165	71	57	80
I	5	2	0	2	2	0	0	2	2	3	1	7	4	1	0	0
INSTR	0	0	1	2	0	0	0	3	0	1	0	0	4	1	0	0
PR	69	41	41	49	26	63	70	36	27	87	15	120	67	25	40	35
PT	45	24	18	23	19	27	51	31	34	50	18	74	31	14	7	15
PTY	45	30	6	27	23	17	29	30	37	30	12	33	14	8	4	3
RC	28	31	14	18	32	12	22	7	24	25	5	30	12	6	13	6
Sum	957	587	467	567	608	649	777	354	673	767	375	1292	762	461	386	506

Adnom	**T33**	**T34**	**T35**	**T36**	**T37**	**T38**	**T39**	**T40**	**T41**	**T42**
A	216	156	204	152	256	156	207	230	226	173
ADV	1	0	0	1	1	2	0	3	4	0
AO	1	5	8	2	7	5	0	7	2	1
AP	12	9	3	20	8	1	7	23	13	19
APAJ	2	5	10	5	8	3	7	3	2	3
APX	6	9	1	6	1	5	8	7	0	1
AY	9	4	5	1	14	5	7	5	11	5
CN	0	0	9	1	2	0	1	1	2	3
DAT	0	0	0	2	1	0	0	0	1	0

DETF	17	8	34	21	28	31	24	28	11	32
DETH	5	6	7	6	8	2	10	7	4	2
DETN	1	0	4	1	1	0	0	4	1	1
DETQ	23	16	29	19	20	23	19	33	27	30
DETS	58	20	28	18	65	57	37	56	78	56
DETV	4	0	1	0	0	1	0	0	2	3
DETW	0	3	0	1	0	1	0	1	0	1
G	79	97	77	77	96	69	146	63	109	37
I	2	0	0	0	2	0	1	0	4	0
INSTR	0	0	0	2	0	1	0	1	0	0
PR	39	30	65	17	59	36	57	53	45	32
PT	19	23	19	17	28	15	26	28	49	24
PTY	36	24	10	18	37	36	45	42	56	12
RC	47	7	14	27	34	36	14	23	17	22
Sum	577	422	528	414	676	487	616	618	664	457

Though there are many cells with zeros, the chi-square test for groups can be performed if we add all frequencies of a special adnominal for female and male texts respectively. In this way we obtain the result presented in Table 2.2.

Table 2.2
Sums of individual adnominal for female and male texts

Adnominal	**Female**	**Male**	**Sum**
A	4804	5169	9973
ADV	45	22	67
AO	78	88	166
AP	482	378	860
APAJ	173	105	278
APX	191	175	366
AY	139	124	263
CN	84	75	159
DAT	16	17	33
DETF	576	592	1168
DETH	146	129	275
DETN	70	28	98
DETQ	542	498	1040
DETS	993	911	1904
DETV	42	35	77
DETW	34	35	69
G	1912	1988	3900
I	80	31	111
INSTR	9	13	22
PR	1079	1044	2123
PT	463	619	1082
PTY	465	556	1021
RC	513	435	948
SUM	12936	13067	26003

Performing the chi-square test with DF = 22 we obtain X^2 = 139.25463 which is highly significant. That means, female texts differ from male texts in using the types of adnominals. However, the greatest differences are represented by the adnominals of the type AP, APAJ, DETN and PT. The same test could be performed also using ranks and taking their means. We may omit this procedure and accept the difference though we know that the chi-square increases with increase of the sample size. A more precise result can be presented only after one has analyzed the same difference in several languages.

Taking the texts individually and ranking the frequencies (as is usual with classification) we obtain for each the common Zipf-Mandelbrot distribution/function. Here we shall use the function with or without added 1 because we omit all cells containing zero that is

$$y = (1) + \frac{a}{(b+x)^c}.$$

Sometimes, it is more adequate to fit merely the power function (P) defined as

$$y = (1) + ax^b$$

as can be seen e.g. in T 16. Since in many cases the greatest ranks yield frequently very great expected values, we used also the Menzerathian (Menz) function defined as

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

The data and the results of fitting are presented in Table 2.3.

Table 2.3
Fitting the given functions to ranked frequencies of adnominal types

	T 1 (M+1)		**T 2 (M+1)**		**T 3**		**T 4 (M+1)**		**T 5 (M+1)**	
1	229	227.49	198	198.52	213	213.75	219	219.54	197	196.30
2	83	95.66	107	100.42	107	100.11	108	104.06	90	90.69
3	59	58.27	50	63.02	45	58.55	56	61.50	44	58.11
4	59	41.16	43	44.26	45	38.65	46	41.03	43	42.51
5	36	31.53	34	33.33	28	27.53	22	29.57	41	33.42
6	32	25.41	30	26.32	16	20.68	19	22.48	40	27.50
7	29	21.21	26	21.51	16	16.14	17	17.78	39	23.35
8	24	18.16	24	18.04	15	12.97	17	14.49	31	20.28
9	21	15.86	21	15.45	15	10.66	16	12.11	25	17.93
10	8	14.07	18	13.45	13	8.94	13	10.31	15	16.07
11	6	12.63	13	11.87	9	7.60	13	8.93	9	14.57
12	6	11.46	10	10.59	7	6.56	12	7.84	9	13.33
13	5	10.48	8	9.55	7	5.71	9	6.96	8	12.29
14	4	9.66	8	8.68	6	5.03	8	6.25	7	11.40
15	4	8.96	6	7.95	3	4.46	7	5.66	6	10.64
16	3	8.36	4	7.33	3	3.99	5	5.16	5	9.98
17	3	7.84	3	6.79	2	3.59	3	4.75	2	9.40
18	1	7.38	3	6.33	2	3.24	2	4.39	2	8.88
19	1	6.97	2	5.92	2	2.95	2	4.08	2	8.43

20	1	6.61	2	5.56	1	2.69	2	3.82	1	8.02
21			2	5.25	1	2.47	2	3.59		
22							1	3.38		
23							1	3.20		
	a = 213.6973 b = -0.0467 c = 1.2161 R^2 = 0.9803		a = 469.9164 b = 0.7633 c = 1.5281 R^2 = 0.9900		a = 752.1128 b = 0.9736 c = 1.8505 R^2 = 0.9997		a = 780.4354 b = 0.9917 c = 1.8474 R^2 = 0.9953		a = 189.4385 b = -0.0273 c = 1.1005 R^2 = 0.9702	

	T 6		**T 7**		**T 8**		**T 9**		**T 10**	
1	103	101.63	198	196.16	350	350.01	267	267.86	349	348.24
2	54	60.56	113	117.29	137	135.24	121	109.99	121	131.45
3	46	41.33	69	77.05	71	77.74	43	63.14	94	77.04
4	35	30.52	68	54.01	55	52.52	41	42.13	52	53.21
5	29	23.75	40	39.71	36	38.76	29	30.65	35	40.08
6	22	19.16	30	30.27	35	30.24	29	23.57	32	31.86
7	17	15.89	30	23.74	29	24.52	26	18.85	29	26.25
8	17	13.46	20	19.06	24	20.46	20	15.52	24	22.22
9	14	11.60	13	15.60	22	17.42	19	13.07	23	19.19
10	8	10.13	12	12.97	18	15.10	17	11.20	23	16.84
11	6	8.95	8	10.94	11	13.26	12	9.73	22	14.96
12	5	7.98	8	9.33	10	11.78	10	8.56	15	13.43
13	4	7.18	7	8.04	9	10.57	5	7.61	13	12.17
14	2	6.50	5	6.99	9	9.56	5	6.82	8	11.10
15	2	5.93	4	6.13	8	8.70	2	6.16	6	10.20
16	2	5.43	4	5.41	6	7.97	2	5.60	6	9.42
17	2	5.00	3	4.81	6	7.34	2	5.12	2	8.74
18	1	4.63	3	4.30	4	6.79	2	4.70	1	8.15
19	1	4.30	2	3.86	3	6.31	1	4.34	1	7.62
20					3	5.89			1	7.16
21					2	5.51			1	6.74
	a = 360.9558 b = 1.3685 c = 1.4699 R^2 = 0.9786		a = 4777.6786 b = 2.9982 c = 2.3039 R^2 = 0.9919		a = 343.0462 b = -0.0147 c = 1.3573 R^2 = 0.9983		a = 368.7042 b = 0.2370 c = 1.5023 R^2 = 0.9890		a = 271.8881 b = -0.1839 c = 1.2180 R^2 = 0.9933	

	T 11		**T 12**		**T 13**		**T 14 (M+1)**		**T 15**	
1	194	193.17	220	217.29	209	206.28	192	190.98	252	255.29
2	58	70.76	92	108.02	89	104.46	87	93.01	138	116.30
3	54	43.52	77	68.02	68	66.99	61	58.32	53	72.01
4	42	31.49	50	48.11	65	48.15	43	41.32	49	50.96
5	27	24.69	44	36.48	46	37.02	33	31.47	37	38.86
6	21	20.33	38	28.96	35	29.75	28	25.14	32	31.10
7	20	17.28	34	23.76	29	24.68	23	20.78	28	25.75
8	19	15.03	24	19.99	22	20.96	21	17.62	26	21.85
9	18	13.31	22	17.14	14	18.13	20	15.24	23	18.90
10	17	11.94	14	14.92	14	15.91	19	13.39	22	16.59
11	9	10.83	9	13.16	13	14.13	13	11.92	22	14.75
12	8	9.91	5	11.72	10	12.68	8	10.72	10	13.24

13	5	9.13	4	10.54	9	11.47	7	9.74	8	11.99
14	4	8.47	4	9.55	7	10.45	6	8.91	8	10.94
15	3	7.90	4	8.70	4	9.58	3	8.21	8	10.04
16	2	7.40	3	7.98	4	8.83	2	7.62	7	9.27
17	2	6.96	3	7.36	3	8.18	2	7.10	4	8.60
18	2	6.57	2	6.81	2	7.61	2	6.65	2	8.01
19	1	6.22	1	6.33	1	7.11	2	6.25	2	7.48
20	1	5.91	1	5.91					2	7.02
21	1	5.63	1	5.53					1	6.61
	a = 109.6192 b = -0.4384 c = 0.9820 R^2 = 0.9813		a = 390.6485 b = 0.5264 c = 1.3870 R^2 = 0.9818		a = 333.9856 b = 0.4498 c = 1.2973 R^2 = 0.9806		a = 317.5661 b = 0.4501 c = 1.3824 R^2 = 0.9925		a = 302.9980 b = 0.1464 c = 1.2536 R^2 = 0.9829	

	T 16 (P+1)		**T 17 (M+1)**		**T 18**		**T 19**		**T 20**	
1	358	350.60	405	404.58	211	210.51	168	167.47	272	271.83
2	106	134.68	140	142.51	60	70.09	50	59.35	51	60.84
3	72	77.18	72	82.85	47	45.36	41	38.03	49	36.73
4	49	52.12	69	57.36	41	34.41	35	28.53	29	26.97
5	45	38.51	45	43.45	35	28.09	32	23.06	27	21.59
6	45	30.13	45	34.78	31	23.93	27	19.48	23	18.16
7	43	24.52	40	28.89	30	20.95	20	16.93	20	15.77
8	42	20.55	30	24.65	26	18.71	18	15.02	19	14.00
9	37	17.60	28	21.46	24	16.96	17	13.53	18	12.64
10	34	15.34	28	18.99	21	15.54	14	12.33	8	11.55
11	16	13.57	11	17.01	15	14.36	10	11.34	8	10.66
12	10	12.14	10	15.40	13	13.38	7	10.52	8	9.92
13	7	10.99	6	14.07	10	12.53	6	9.81	7	9.29
14	5	9.99	6	12.95	6	11.80	5	9.20	6	8.75
15	4	9.17	5	11.99	5	11.16	4	8.67	5	8.28
16	4	8.47	5	11.17	4	10.60	4	8.21	3	7.87
17	2	7.87	4	10.45	3	10.09	3	7.79	3	7.51
18	2	7.35	4	9.82	2	9.64	3	7.42	3	7.18
19	1	6.89	2	9.26	1	9.23	2	7.09	2	6.89
20	1	6.48	2	8.77	1	8.86	1	6.78	2	6.63
21	1	6.13			1	8.53			2	6.39
22									2	6.17
	a = 349.6021 b = -1.3869 R^2 = 0.9755		a = 266.2500 b = -0.2959 c = 1.1858 R^2 = 0.9931		a = 86.5280 b = -0.6851 c = 0.7695 R^2 = 0.9786		a = 78.1813 b = -0.6031 c = 0.8245 R^2 = 0.9809		a = 70.2035 b = -0.7943 c = 0.8538 R^2 = 0.9917	

	T 21		**T 22 (M+1)**		**T 23 (M+1)**		**T 24 (M+1)**		**T 25**	
1	259	259.98	320	319.87	368	367.56	139	137.9	297	296.86
2	114	100.89	63	69.43	89	100.86	60	65.13	84	85.02
3	32	56.02	42	42.11	70	58.49	36	38.31	43	50.43
4	32	36.61	38	31.03	51	41.28	31	25.48	37	36.08

5	30	26.27	34	24.92	35	31.97	30	18.33	34	28.17
6	26	20.05	27	21.01	31	26.13	19	13.94	28	23.15
7	23	15.97	23	18.27	29	22.14	8	11.05	27	19.68
8	19	13.14	22	16.24	22	19.23	7	9.03	25	17.13
9	16	11.08	21	14.67	22	17.03	7	7.57	24	15.18
10	16	9.53	17	13.42	17	15.29	3	6.48	22	13.63
11	10	8.34	12	12.39	10	13.90	2	5.65	14	12.38
12	9	7.39	10	11.54	8	12.75	2	4.99	9	11.34
13	5	6.63	6	10.81	6	11.78	2	4.47	5	10.47
14	4	6.00	3	10.19	4	10.97	2	4.04	5	9.72
15	3	5.48	3	9.64	4	10.26	2	3.69	4	9.08
16	3	5.04	2	9.17	4	9.65	1	3.39	4	8.51
17	2	4.67	2	8.74	3	9.11	1	3.15	3	8.02
18	2	4.35	2	8.37	2	8.64	1	2.94	2	7.58
19	1	4.07	1	8.03	1	8.22	1	2.76	2	7.19
20	1	3.83	1	7.72	1	7.84			2	6.83
21	1	3.61							1	6.51
22									1	6.22
	a = 373.1744 b = 0.2524 c = 1.6232 R^2 = 0.9848		a = 78.9890 b = -0.8123 c = 9.8341 R^2 = 0.9928		a = 139.0540 b = -0.6148 c = 1.0160 R^2 = 0.9932		a = 541.2512 b = 1.0624 c = 1.9111 R^2 = 0.9849		a = 114.4684 b = -0.6328 c = 0.9511 R^2 = 0.9922	

	T 26		**T 27 (P+1)**		**T 28 (Menz)**		**T 29 (P)**		**T 30**	
1	231	228.57	144	145.94	531	526.66	316	324.68	213	212.85
2	113	123.82	67	59.86	196	214.99	165	128.25	71	73.04
3	87	80.63	34	35.74	120	125.70	67	74.49	44	40.47
4	50	57.96	23	24.90	95	85.14	45	50.66	25	26.85
5	50	44.31	19	18.88	74	62.50	31	37.56	20	19.61
6	45	35.34	18	15.11	69	48.29	25	29.42	15	15.19
7	36	29.06	15	12.55	42	38.64	17	23.93	14	12.26
8	30	24.46	13	10.71	41	31.73	16	20.01	12	10.18
9	28	20.97	12	9.33	33	26.57	14	17.09	11	8.65
10	25	18.25	5	8.26	30	22.60	13	14.84	8	7.48
11	16	16.08	4	7.42	9	19.46	12	13.06	8	6.56
12	12	14.31	3	6.73	9	16.94	10	11.62	6	5.82
13	9	12.85	3	6.16	8	14.87	5	10.44	5	5.21
14	8	11.62	3	5.69	7	13.15	5	9.45	3	4.71
15	7	10.58	3	5.29	7	11.70	4	8.62	2	4.28
16	6	9.69	3	4.94	6	10.47	4	7.90	1	3.92
17	3	8.92	2	4.64	6	9.42	4	7.29	1	3.61
18	3	8.24	2	4.38	3	8.51	3	6.75	1	3.34
19	2	7.65	1	4.15	3	7.71	2	6.28	1	3.10
20	2	7.13	1	3.95	2	7.02	2	5.86		
21	2	6.66			1	6.41	1	5.49		
22	1	6.25					1	5.16		
23	1	5.87								
	a = 572.2487		a = 144.9367		a = 542.8982		a = 324.6839		a = 166.2514	

	b = 0.8890 c = 1.4429 R^2 = 0.9869	b = -1.3002 R^2 = 0.9914	b = -1.2488 c = 0.0304 R^2 = 0.9940	b = -1.3401 R^2 = 0.9833	b = -0.1665 c = 1.3567 R^2 = 0.9984

	T 31		**T 32**		**T 33**		**T 34**		**T 35**	
1	174	173.74	201	200.89	216	214.73	156	160.20	204	203.03
2	57	61.01	80	78.94	79	90.09	97	73.33	77	86.75
3	40	34.58	38	47.03	58	55.17	30	44.57	65	52.02
4	26	23.36	35	32.80	47	39.14	24	30.90	34	36.04
5	18	17.31	34	24.88	39	30.04	23	23.13	29	27.07
6	13	13.58	20	19.88	36	24.23	20	18.20	28	21.40
7	9	1.08	20	16.45	23	20.21	16	14.83	19	17.54
8	8	9.29	19	13.98	19	17.28	9	2.41	14	14.75
9	8	7.96	15	12.11	17	15.07	9	0.59	10	12.66
10	8	6.94	12	10.65	12	13.31	8	9.19	10	11.05
11	7	6.12	8	9.49	9	11.91	7	8.08	9	9.76
12	6	5.47	6	8.54	6	10.76	6	7.18	8	8.72
13	4	4.93	6	7.75	5	9.80	5	6.44	7	7.85
14	3	4.47	4	7.09	4	8.99	5	5.82	5	7.13
15	2	4.09	3	6.52	2	8.29	4	5.30	4	6.52
16	2	3.76	2	6.03	2	7.69	3	4.85	3	5.99
17	1	3.48	2	5.60	1	7.17			1	5.54
18			1	5.23	1	6.71			1	5.14
19					1	6.30				
	a = 128.5939 b = -0.2095 c = 1.2799 R^2 = 0.9972		a = 164.1989 b = -0.1552 c = 1.1960 R^2 = 0.9921		a = 188.5258 b = -0.1064 c = 1.1566 R^2 = 0.9850		a = 238.8790 b = 0.3316 c = 1.3950 R^2 = 0.9662		a = 228.6866 b = 0.0951 c = 1.3107 R^2 = 0.9904	

	T 36		**T 37**		**T 38**		**T 39 (P+1)**		**T 40**	
1	152	153.31	256	254.75	156	153.37	207	219.99	23	229.16
2	77	65.40	96	106.34	69	81.51	146	101.58	0	79.01
3	27	39.45	65	64.42	57	52.71	57	64.81	63	50.30
4	21	27.50	59	45.27	36	37.77	45	47.20	56	37.61
5	20	20.77	37	34.46	36	28.84	37	36.97	53	30.33
6	19	16.51	34	27.60	36	22.99	26	30.31	42	25.58
7	18	13.59	28	22.88	31	18.91	24	25.65	33	22.20
8	18	11.48	28	19.45	23	15.93	19	22.22	28	19.68
9	17	9.89	20	16.86	15	13.67	14	19.59	28	17.71
10	17	8.66	14	14.84	5	11.90	10	17.52	23	16.13
11	6	7.68	8	13.22	5	10.49	8	15.84	23	14.82
12	6	6.88	8	11.90	5	9.35	7	14.46	7	13.74
13	5	6.21	8	10.80	3	8.40	7	13.31	7	12.81
14	2	5.66	7	9.87	2	7.60	7	12.32	7	12.01
15	2	5.18	2	9.08	2	6.93	1	11.48	5	11.31
16	2	4.78	2	8.40	1	6.35	1	10.75	4	10.70
17	1	4.42	1	7.80	1	5.85			3	10.15
18	1	4.11	1	7.28	1	5.41			3	9.66
19	1	3.84	1	6.82	1	5.03			1	9.23

20	1	3.60	1	6.41					1	8.83
21	1	3.38							1	
	a = 162.5025 b = 0.0469 c = 1.2707 R^2 = 0.9751		a = 236.2628 b = -0.0606 c = 1.2052 R^2 = 0.9983		a = 349.3789 b = 0.7850 c = 0.3379 R^2 = 0.9660		a = 218.9926 b = -1.1224 R^2 = 0.9448		a = 103.4700 b = -0.6162 c = 0.8302 R^2 = 0.9706	

	T 41 (Menz)		**T 42**	
1	226	223.39	173	172.44
2	109	118.75	56	62.66
3	78	77.75	37	39.00
4	56	55.45	32	28.59
5	49	41.43	32	22.70
6	45	31.89	30	18.92
7	27	25.05	24	16.27
8	17	19.97	22	14.32
9	13	16.11	19	12.81
10	11	13.11	12	11.62
11	11	10.75	5	10.65
12	4	8.86	3	9.84
13	4	7.35	3	9.16
14	4	6.12	3	8.58
15	2	5.11	2	8.08
16	2	4.28	1	7.64
17	2	3.60	1	7.25
18	2	3.03	1	6.91
19	1	2.56	1	6.60
20	1	2.17		
	a = 254.3231 b = -0.7245 c = 0.1297 R^2 = 0.9925		a = 92.5876 b = -0.4734 c = 0.9607 R^2 = 0.9728	

As can be seen, the usual ranking functions excellently express the trends. Whatever the author strives, (s)he adheres to a subconscious regularity. Not even later corrections of the text can change it. The fact that we used three different functions (Power, Mandelbrot, Menzerath) merely shows that there are some boundary conditions leading to a slightly different regime. The three functions belong to the unified theory (cf. Wimmer, Altmann 2005) and can be expressed by the differential equations

$$\frac{dy}{y(-1)} = \frac{b}{x} \quad (Power\ \ ft.)$$

$$= \frac{-c}{b+x} \ (Mandelbrot\ ft.)$$

$$= \frac{b}{x} - c \ (Menzerath\ ft.)$$

On the left hand side, (-1) in the denominator is constant and may, but need not be used. It can be shown that its use or omitting may lead to better results. In any case, we omit the classes having frequencies smaller than 1.

As a matter of fact, we merely confirmed the existence of some regulation and, at the same time, the reasonability of ranking in linguistically defined classes. Analysis of other languages may lead to other results but we expect that ranking will be everywhere reasonable. The fact that ranking is sufficient also at this high abstraction level is a good strengthening of the idea of ranking at all.

3. Motifs

The description of motifs may be found especially in the works by Köhler (see the References). The problems connected with them and their philosophy can be found in Altmann (2016). It must be remarked here that they are "legal" linguistic entities positioned at a higher abstraction level than the well known units (phonemes, syllables, morphemes, phrases, clauses, sentences, etc.). They behave according to some laws which are not yet known because the research proceeds very slowly. Here we shall try to show their behavior if they are constructed from adnominals in modern Russian.

A question which has not been touched in the history of motifs is the existence of a possible hierarchy, i.e. the existence of some relations between motifs constructed from lower and higher elementary entities, e.g. there are some regularities holding true for syllabic motifs and other ones holding true for parts of speech. Is there some relation between them, e.g. model generalization, adding new parameters, considering different stochastic processes, do they belong to Köhlerian control circuits, etc.? Here we shall restrict ourselves to adnominals and shall not study possible hierarchies, a task for great teams.

3.1. Types

If we consider adnominal motifs, we can omit the placing in front of or behind the noun and obtain a sequence in which the distinction between -R and -L disappears. The qualitative Köhlerian motifs are defined here as follows: the motif consists of not repeated symbols; the next motif mostly begins with a symbol occurring in the previous motif but does not contain two or more symbols of the preceding one. A simple example from Text 1 (Demidova) having the form:
[A,G,PTY,P,G,C,G,PTY,DETQ,DETF,P,A,PT,G,A,G,A,A,A,A,DETS,A,G,…..]

can be dissected in motifs as follows:

[A,G,PTY,P],[G,C],[G,PTY,DETQ,DETF,P,A,PT],[G],[A,G],[A],[A],[A],[A,DETS],[A,G],
….

Here one can see that the fifth motif, [A,G], does not begin with a symbol occurring in the previous one but "A" cannot stay in the previous motif because it cannot contain two symbol from the third motif.

Now qualitative motifs have at least the following properties:

1. A distribution of types considered either in ranked or in spectrum form.

2. Length consisting of the number of symbols in individual motifs, yielding again a distribution/function.

In the first case one obtains most probably a very long and flat function which depends on the extent of the list of different adnominals. The more adnominal types there are, the flatter will be the function. On the contrary, the spectrum will have a great excess. Nevertheless, we conjecture that all this may be expressed by reasonable formulas. In the second case, again, the more types of adnominals exist, the longer will be the individual motifs.

However, the text is subdivided in sentences hence one can study also the Belza-chains of motifs, i.e. the distribution of the number of subsequent sentences containing the same motif. Or, if one subdivides the texts in Frumkina passages, containing, say 20 sentences each, one can perform the same investigation. Here we shall consider merely the variant considering the whole text as a unit.

Let us consider first the motif types. If there are many different types then the text is rich in motifs, it displays great variability, it is stylistically rich. The richness can be measured in various ways. A simple expression is the mean: if the mean is small, there is a tendency to use many different motifs, that is, the smaller is the mean the richer is the text. A slightly more complex measure is that of Ord: the steeper is the function or distribution of types, the richer is the text. In order to study the possible development of motif-richness, one can use Ord's criterion (I,S) and study the course of $S = f(I)$ where $I = m_2/m_1'$ and $S = m_3/m_2$ and m_r are the r-th central moments. The mean, m_1' is the first moment about zero.

Let us illustrate these concepts using Text 1 by Demidova. The spectrum of types is presented in Table 3.1. The numbers in the first two columns are to be read as "there are 98 motif-types each occurring exactly once", etc. There are N = 285 motifs out of which 125 are different, hence the mean $m_1' = \bar{x} = \frac{1}{N}\sum_x xf_x$ = 285/125 = 2.28. The other values (m2, m3, I, S) are displayed in the fourth column of Table 3.1. Here we treat the sequence as a distribution.

Now, the spectrum can be given either by a discrete distribution or simply, by a function. It must be remarked that mathematical models do not represent the truth but are merely exact means of our concept formation. Here we shall use also functions because there are many x-values having frequency zero. They should be chosen in such a way that the determination coefficient R^2 is greater than 0.8 and the number of parameters is minimal. The exponential function is defined as $y = 1 + aexp(-bx)$; we add everywhere 1, otherwise the values in the tail of function are too small. The fitted power function is given as $y = 1 + ax^{-b}$. In both, $y = f_x$. In the power function, the parameter a yields a more realistic value than that in the exponential one.

An example of constructing motifs is presented below using Text 8 by Rubina:

[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].

The individual motifs have a length consisting of the number of elements in them, and one can also study how many times a special motif occurred. The results give a picture of the stylistic richness: the more types, the richer is the text stylistically; the longer are the motifs, the more complex are the sentences stylistically.

Table 3.1
The spectrum of motif-types in Text 1 by Demidova

Spectrum x	**Frequency f_x**	**Computed frequency**		
		Exponential	**Power function**	
1	98	97.96	97.97	$\bar{x} = 2.28$
2	11	11.88	11.84	$m_2 = 47.6256$
3	6	2.22	4.01	$m_3 = 3225.7885$
4	2	1.14	2.21	I = 20.88
6	1	1.00	1.34	S = 67.73
7	1	1.00	1.21	
8	2	1.00	1.14	
11	2	1.00	1.05	
12	1	1.00	1.04	
76	1	1.00	1.00	
		a = 863.8401	a = 96.9656	
		b = 2.1871	b = 3.1614	
		R^2 = 0.9978	R^2 = 0.9992	

The table should be read as follows: There is one type occurring 98 times; there are two types occurring 11 times, etc.

The values of all texts are presented in Table 3.2. The first line contains the numbers x, the second the frequencies and the third the computed values (Power + 1). i.e. $y = ax^b$. As can be seen, there is always a very preferred adnominal type – usually it is an adjective -, the other ones have a much smaller frequency. However, the decrease is not always monotonous but this occurs mostly with high frequency motifs.

Table 3.2a
Fitting and characterizing the frequency of motif-types (female texts)
by the power function (P)

Text	**Frequencies**
T 1	1 2 3 4 6 7 8 11 12 76
f_x	98 11 6 2 1 1 2 2 1 1
P +1	97.97 11.84 4.01 2.21 1.34 1.21 1.14 1.05 1.04 1.00
	a = 96.9656, b = -3.1614, R^2 = 0.9992, I = 20.88, S = 67.73
T 2	1 2 3 4 5 7 12 15 53
f_x	123 11 4 2 2 1 1 1 1
P+1	122.99 11.29 3.42 1.87 1.39 1.12 1.02 1.01 1.00
	a = 121.9900, b = -3.5677, R^2 = 0.9998, I = 11.57, S = 46.76
T 3	1 2 3 4 5 6 15 19 73
f_x	97 9 1 2 1 3 1 1 1
P+1	97.01 8.52 2.70 1.59 1.26 1.13 1.00 1.00 1.00
	a = 96.0118, b = -3.6739, R^2 = 0.9991, I = 22.10, S = 63.93

T 4	1 2 3 4 5 6 17 20 69
f_x	89 16 4 3 2 2 1 1 1
P+1	89.06 14.97 5.76 3.22 2.26 1.75 1.05 1.03 1.00
	a = 88.0619, b = -2.6562, R^2 = 0.9993, I = 18.98, S = 58.95
T 5	1 2 3 4 5 6 7 17 49
f_x	120 10 5 1 2 4 2 1 1
P+1	119.98 10.73 3.25 1.80 1.36 1.18 1.11 1.00 1.00
	a = 118.9751, b = -3.6120, R^2 = 0.9989, I = 9.86, S = 40.37
T 6	1 2 3 4 8 9 25
f_x	71 12 4 1 1 1 1
P+1	71.05 11.33 4.37 2.52 1.22 1.16 1.01
	a = 70.0460, b = -2.7609, R^2 = 0.9983, I = 4.46, S = 19.48
T 7	1 2 3 4 5 7 9 18 24 67
f_x	106 12 3 4 2 2 1 1 1 1
P+1	105.99 12.06 3.97 2.17 1.56 1.19 1.08 1.01 1.00 1.00
	a = 104.9900, b = -3.2463, R^2= 0.9995, I = 17.51, S = 55.05
T 8	1 2 3 4 5 7 8 11 16 18 28 144
f_x	123 11 1 3 4 1 1 2 1 1 1 1
P+1	123.01 10.63 3.18 1.76 1.34 1.10 1.06 1.02 1.0 1.0 1.0 1.0
	a = 122.0052, b = -3.6631, R^2 = 0.9989, I = 5103, S=131.39
T 9	1 2 3 4 5 6 8 11 27 99
f_x	95 12 1 3 1 2 1 2 2 1
P+1	95.03 11.05 3.71 2.07 1.52 1.29 1.11 1.04 1.00 1.00
	a = 93.0347, b = -3.2277, R^2= 0.9985, I = 32.73, S = 86.58
T 10	1 2 3 4 6 7 8 10 11 13 16 18 150
f_x	122 15 5 2 1 3 1 1 1 1 1 1 1
P+1	122 14.94 4.94 2.61 1.45 1.28 1.18 1.09 1.07 1.04 1.02 1.01 1.00
	a = 121.0032, b = -3.1181, R^2 = 0.9997, I = 53.26, S=140.37
T 11	1 2 3 4 5 6 7 8 9 10 76
f_x	85 7 3 2 2 1 1 1 1 1 1
P+1	84.99 7.34 2.40 1.48 1.21 1.11 1.06 1.04 1.02 1.02 1.00
	a = 83.9889, b = -3.7272, R^2 = 0.9998, I = 23.74, S = 69.53
T 12	1 2 3 4 5 6 8 9 10 11 13 20 73
f_x	107 12 3 2 2 2 1 1 1 1 1 1 1
P+1	107.01 11.74 3.81 2.09 1.52 1.29 1.11 1.07 1.05 1.04 1.02 1.01 1.00
	a = 106.0098, b = -3.3031, R^2 = 0.9999, I = 18.69, S = 61.54
T 13	1 2 3 4 5 7 8 9 17 20 60
f_x	107 13 5 2 1 2 1 2 1 1 1
P+1	107.00 13.14 4.42 2.39 1.69 1.24 1.16 1.11 1.02 1.01 1.00
	a = 105.9955, b = -3.1261, R^2 = 0.9998, I = 14.16, S = 48.09

T 14	1 2 3 4 6 8 9 13 60
f_x	102 10 4 3 1 2 1 1 1
P+1	101.98 10.47 3.37 1.89 1.22 1.08 1.06 1.02 1.00
	a = 100.9809, b = -3.4149, R^2 = 0.9997, I = 14.96, S = 52.73
T 15	1 2 3 4 6 7 10 23 26 89
f_x	119 13 6 3 1 1 1 1 1 1
P+1	118.97 13.75 4.47 2.38 1.38 1.23 1.07 1.01 1.00 1.00
	a = 117.9692, b = -3.2096, R^2 = 0.9997, I = 26.24, S = 76.52
T 16	1 2 3 4 5 6 7 8 9 12 13 18 28 129
f_x	116 11 7 3 3 1 1 1 1 3 1 1 1 1
P+1	115.94 12.55 4.01 2.16 1.55 1.30 1.18 1.12 1.08 1.03 1.02 1.01 1.0 1.0
	a = 114.9373, b = -3.3150, R^2 = 0.9985, I = 0.69, S = 1.67
T 17	1 2 3 4 5 6 7 8 9 11 12 26 123
f_x	116 17 11 1 1 2 3 1 1 1 1 1 1
P+1	115.93 18.43 6.78 3.64 2.44 1.88 1.58 1.40 1.29 1.17 1.13 1.02 1.0
	a = 114.9349, b = -2.7213, R^2 = 0.9975, I = 64.37, S =163.26
T 18	1 2 3 4 5 6 8 10 86
f_x	108 8 3 5 1 1 2 1 1
P+1	107.98 8.61 2.62 1.54 1.23 1.12 1.04 1.02 1.00
	a = 106.9793, b = -3.8127, R^2 = 0.9986, I = 21.68, S = 70.29
T 19	1 2 3 4 5 6 7 75
f_x	93 12 1 1 3 1 1 1
P+1	93.95 10.80 3.64 2.04 1.51 1.28 1,17 1.00
	a = 92.0473, b = -3.2318, R^2 = 0.9984, I = 24.21, S = 70.80
T 20	1 2 3 4 5 6 7 11 13 133
f_x	77 13 4 1 2 2 1 1 1 1
P+1	77.05 12.21 4.66 2.65 1.89 1.54 1.35 1.10 1.06 1.00
	a = 76.0511, b = -2.7618, R^2 = 0.9992, I = 57.14, S= 126.24

Table 3.2b
Fitting and characterizing the frequency of motif-types (male texts)

T 21	1 2 3 4 6 7 8 12 20 32 97
f_x	75 13 4 2 1 1 2 1 1 1 1
P+1	75.05 12.31 4.77 2.73 1.58 1.38 1.26 1.09 1.02 1.01 1.00
	a = 74.0455, b = -2.7108, R^2 = 0.9995, I = 33.56, S = 82.29
T 22	1 2 3 4 5 6 9 11 21 161
f_x	78 7 4 4 5 1 1 1 1 1
P+1	77.94 8.27 2.83 1.69 1.32 1.17 1.04 1.02 1.00 1.00
	a = 76.9434, b = -3.4038, R^2 = 0.9957, I = 72.86, S = 152.30
T 23	1 2 3 4 5 7 8 9 12 15 20 79

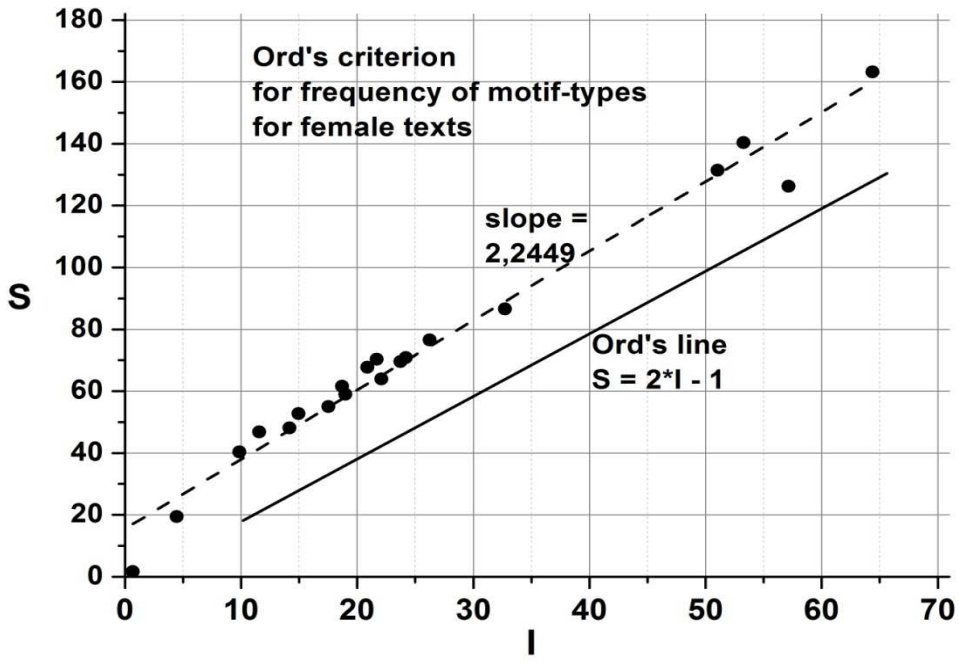
f_x	92 10 3 3 4 1 1 3 1 1 1 1
P+1	91.98 10.34 3.47 1.96 1.46 1.15 1.10 1.07 1.03 1.01 1.00 1.00
	a = 90.9770, b = -3.2836, R^2 = 0.9994, I = 76.03, S = 169.12
T 24	1 2 3 4 5 6 7 8 11 53
f_x	57 6 3 1 1 2 1 1 1 1
P+1	56.99 6.21 2.30 1.49 1.23 1.12 1.07 1.05 1.02 1.00
	a = 55.9931, b = -3.4252, R^2 = 0.9994, I = 15.93, S = 45.93
T 25	1 2 3 4 5 7 8 9 18 139
f_x	102 12 3 3 2 2 1 1 1 1
P+1	102.00 11.86 3.95 2.17 1.57 1.19 1.13 1.09 1.01 1.00
	a = 101.0001, b = -3.2175, R^2 = 0.9997, I = 55.75, S = 131.86
T 26	1 2 3 4 7 10 11 13 17 23 61
f_x	138 9 5 4 1 1 1 1 1 1 1
P+1	137.97 10.01 2.84 1.59 1.07 1.02 1.01 1.01 1.00 1.00 1.00
	a = 136.9727, b = -3.9256, R^2 = 0.9993, I = 13.97, S = 47.94
T 27	1 2 3 4 5 6 10 14 61
f_x	74 3 2 1 1 1 1 1 1
P+1	74.00 3.14 1.27 1.06 1.02 1.01 1.00 1.00 1.00
	a = 72.9985, b = -5.0907, R^2 = 0.9999, I = 20.38, S = 54.21
T 28	1 2 3 4 5 6 7 9 10 11 15 16 22 37 41 203
f_x	145 26 6 3 2 2 1 1 3 2 1 1 1 1 1 1
P+1	145.15 23.78 8.74 4.60 2.99 2.22 1.81 1.42 1.31 1.24 1.11 1.09 1.04 1.01 1.01 1.0
	a = 144.1468, b = -2.6616, R^2 = 0.9989, I = 67.92, S = 182.41
T 29	1 2 3 4 6 8 10 15 37 42 127
f_x	100 10 2 4 2 2 1 1 1 1 1
P+1	99.99 9.99 3.21 1.82 1.20 1.07 1.03 1.01 1.00 1.00 1.00
	a = 98.9939, b = -3.4604, R^2 = 0.9991, I = 46.69, S = 106.56
T 30	1 2 3 4 6 1 14 101
f_x	74 6 2 2 2 2 1 1
P+1	74.00 6.10 2.08 1.36 1.08 1.01 1.00 1.00
	a = 72.9958, b = -3.8380, R^2 = 0.9995, I = 40.78, S = 93.22
T 31	1 2 3 4 5 6 9 10 14 82
f_x	54 6 1 2 1 2 1 1 1 1
P+1	54.01 5.70 2.14 1.42 1.19 1.10 1.02 1.02 1.01 1.00
	a = 53.0079, b = -3.4962, R^2 = 0.9990, I = 32.02, S = 73.93
T 32	1 2 3 4 5 6 8 9 14 18 86
f_x	81 8 3 1 2 1 1 1 1 1 1
P+1	81.00 8.06 2.71 1.62 1.29 1.15 1.05 1.04 1.01 1.00 1.00
	a = 79.9984, b = -3.5023, R^2 = 0.9998, I = 29.23, S = 76.78
T 33	1 2 3 4 5 8 11 12 78

f_x	91 14 7 1 1 2 1 2 1
P+1	91.00 14.19 5.29 2.93 2.04 1.28 1.12 1.09 1.00
	a = 90.0032, b = -2.7706, R^2 = 0.9987, I = 22.40, S = 70.03
T 34	1 2 3 4 5 11 18 58
f_x	69 7 2 2 2 2 1 1
P+1	69.00 6.99 2.45 1.53 1.24 1.02 1.00 1.00
	a = 67.9978, b = -3.5044, R^2 = 0.9995, I = 17.74, S = 48.64
T 35	1 2 3 4 5 6 11 13 82
f_x	83 9 3 1 2 1 2 2 1
P+1	83.00 8.92 3.02 1.76 1.36 1.19 1.03 1.01 1.00
	a = 82.0032, b = -3.3725, R^2 = 0.9995, I = 26.55, S = 73.28
T 36	1 2 3 4 5 13 17 59
f_x	78 10 1 1 1 1 1 1
P+1	78.04 8.91 3.09 1.81 1.39 1.02 1.01 1.00
	a = 77.0444, b = -3.2834, R^2 = 0.9987, I = 18.52, S = 51.16
T 37	1 2 3 4 5 6 7 9 10 16 83
f_x	99 14 2 5 1 1 1 2 2 1 1
P+1	99.03 13.26 4.63 2.53 1.78 1.45 1.29 1.13 1.10 1.02 1.00
	a = 98.0270, b = -2.9997, R^2 = 0.9981, I = 23.16, S = 73.97
T 38	1 2 4 5 7 8 10 49
f_x	88 13 1 4 1 1 1 1
P+1	88.00 12.87 2.62 1.85 1.32 1.22 1.12 1.00
	a = 87.0060, b = -2.8732, R^2 = 0.9988, I = 11.61, S = 42.76
T 39	1 2 3 4 5 6 7 10 24 30 73
f_x	97 11 3 2 1 1 1 1 1 1 1
P+1	97.01 10.70 3.54 1.98 1.47 1.26 1.15 1.05 1.00 1.00 1.00
	a = 96.0127, b = -3.3069, R^2 = 0.9999, I = 22.67, S = 58.88
T 40	1 2 3 4 5 6 7 9 10 13 87
f_x	109 8 2 2 2 2 1 1 1 1 1
P+1	109.00 8.02 2.42 1.46 1.19 1.09 1.05 1.02 1.01 1.00 1.00
	a = 107.9979, b = -3.9440, R^2 = 0.9998, I = 26.64, S = 80.00
T 41	1 2 3 4 5 6 9 10 13 21 75
f_x	102 13 4 3 1 2 3 1 1 1 1
P+1	102.00 12.95 4.43 2.41 1.71 1.41 1.12 1.08 1.04 1.01 1.00
	a = 100.9987, b = -3.0789, R^2 = 0.9995, I = 19.56, S = 63.57
T 42	1 2 3 4 6 9 73
f_x	79 4 6 6 3 1 1
P+1	78.95 5.92 1.98 1.31 1.06 1.01 1.00
	a = 77.9494, b = -3.9853, R^2 = 0.9907, I = 22.88, S= 67.54

The results of fitting are excellent. Even if the data contain some irregularities, (see below), the high value of the determination coefficient is in all cases given.

The relation between I and S is quite concentrated and does not display great deviations. This means that the authors adhering merely to the knowledge of their language subconsciously follow a kind of law that can be expressed by a formula and its parameters, or by the relation of two indicators. Needless to say, there are some "exceptions", e.g. in Text T 42 the value of x = 2 is smaller than that of x = 3 but one can consider them as some boundary conditions, post-writing corrections, etc.

The (I,S) points can be seen in Figures 3.1 and 3.2.



Fgure 3.1. Ord's criterion for female motif-types

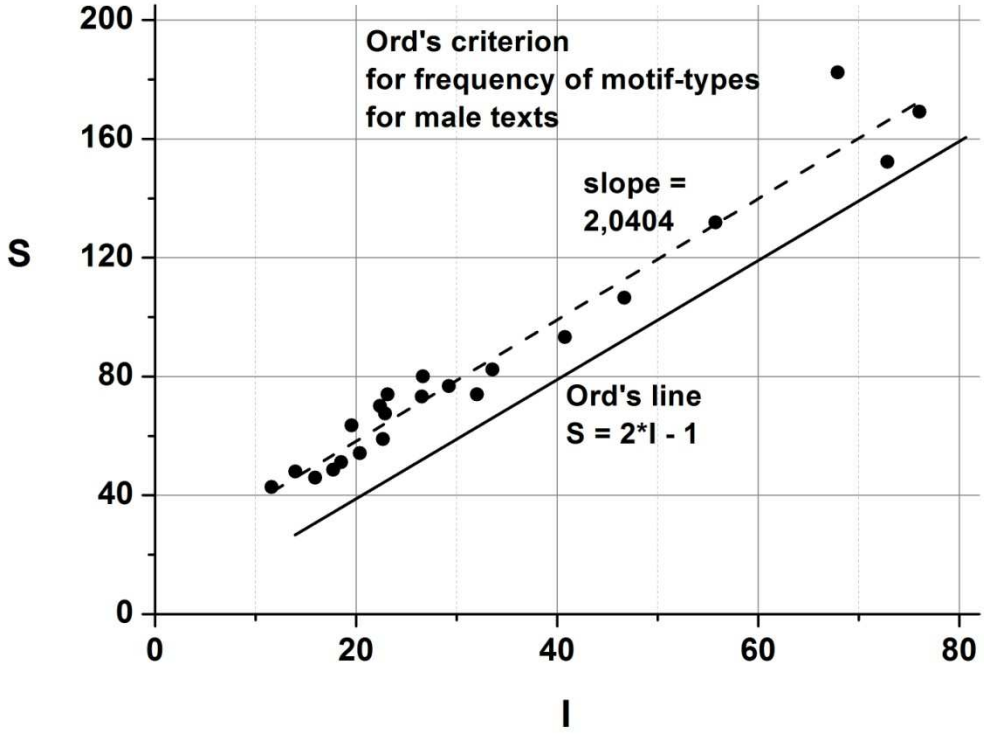


Figure 3.2. Ord's criterion for motif types in male texts

The Figures 3.1. and 3.2 show two very remarkable features. (1) The (I,S) points are positioned on a very emphasized straight line which is parallel to the Ord's line S = 2I – 1 and over the Ord's line. (2) In female texts the straight line has a slightly greater slope than that of male texts. Automatically one may ask whether this is the situation also in other languages, with other text types, present in the evolution of the given language or person, etc. For literary scientists a very extensive research domain opens.

3.2. Length

Another property of motifs is their length computed in terms of the number of symbols in them. Again, the longer are the motifs, the more variegated is the text, and we can speak about its richness. Here we do not speak about vocabulary richness but rather about stylistic richness. Since motifs are segmented by applying some prescriptions (s. above), another aspect of this stylistic property would be the computation of distances between equal symbols in the sense of the Skinner hypothesis.

Consider first the situation in Text 1 by Demidova. The data are presented in Table 3.3. We can, again, choose between a function and a probability distribution. The latter case is more appropriate here than with motif-types because the lengths attain only few values and all values are non-zero. However, it has been shown that lengths of any kind in texts follow the Zipf-Alekseev function (cf. Popescu, Best, Altmann 2014) hence we shall apply it here, too.

Table 3.3
Computing the length of motifs in Text 1 by Demidova

x	**f_x**	**Zipf-Alekseev ft.**
1	106	105.32
2	85	88.51
3	54	47.89
4	27	24.85
5	9	13.36
6	2	7.62
7	2	4.68
		a = 0.5573, b = -1.1698, c = 104.3226 R^2 = 0.9894

Table 3.4a
Fitting the Zipf-Alekseev function to the motif-lengths in all texts (female authors)

Text	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Results**
T1	106	85	54	27	9	2	2				a = 0.5333,b = -1.1341
	105.34	88.41	48.15	24.95	13.17	7.18	4.06				c =105.3428, R^2 = 0.99
T2	91	73	52	29	12	2	3	0	0	1	a = 0.4928, b = -1.0149
	89.99	77.77	45.43	25.34	14.35	8.37	5.03	3.11	1.98	1.29	c =89.990, R^2 = 0.99
T3	102	73	40	25	10	4	2				a = 0.2377, b = -0.9885
	101.84	74.67	40.11	21.19	11.54	6.53	3.83				c = 101.8427,R^2 = 1.00
T 4	99	85	47	24	14	4					a = 0.5963, b = -1.1676
	98.92	85.34	46.53	23.97	12.55	6.78					c = 98.9241.R^2 = 1.00
T 5	95	83	46	34	12	2	0	1			a = 0.6082, b = -1.1239
	94.64	84.07	47.55	25.36	13.70	7.63	4.38	2.60			c = 94.6397, R^2 = 0.99
T 6	53	36	37	17	11	3	3				a = 0.1842, b = -0.6973
	51.87	42.15	27.37	17.53	11.46	7.69	5.30				c = 51.8680, R^2 = 0.93
T 7	114	81	53	20	14	3	2				a = 0.2701,b = -1.0056
	113.46	84.40	45.36	23.89	12.95	7.29	4.26				c =113.4646, R^2 = 0.99
T 8	186	123	68	25	14	6	1				a = 0.2064,b = -1.1098
	185.63	125.67	61.01	29.29	14.61	7.62	4.15				c = 185.6289,R^2 = 1.00
T 9	142	111	47	25	8	0	1				a = 0.6781,b = -1.4973
	142.02	110.67	49.09	20.46	8.75	3.91	1.83				c = 142.0203,R^2 = 1.00
T 10	195	115	68	31	12	5	2				a =-0.0825, b = -0.9025
	194.45	119.04	59.76	30.61	16.44	9.25	5.43				c =194.4488, R^2 = 1.00
T 11	103	72	37	11	11	5	1	1			a = 0.3281,b = -1.2036
	102.96	72.49	34.54	16.06	7.73	3.89	2.05	1.12			c = 102.9551,R^2 = 1.00
T12	125	99	52	22	11	5					a = 0.5581, b = -1.2707
	124.87	99.84	49.74	23.55	11.40	5.74					c = 124.8668,R^2 = 1.00
T 13	123	96	53	21	11	4	1				a = 0.5307,b = -1.2411
	122.72	97.65	49.15	23.58	11.58	5.91	3.14				c =122.7190, R^2 = 1.00
T 14	85	69	56	29	5	4	1	1			a = 0.5983, b = -1.0590
	83.53	76.03	44.90	25.02	14.09	8.15	4.85	2.98			c = 83.5285, R^2 = 0.96
T 15	140	86	58	30	10	8	4				a = -0.0883, b = -0.7703,
	139.34	90.52	49.91	28.05	16.44	10.03	6.35				c = 139.3407, R^2 = 0.99
T 16	185	142	65	29	11	6	2				a = 0.6012, b = -1,4154

	184.97 142.15 64.87 28.04 12.45 5.78 2.80	c = 184.9682, R^2 = 1.00
T 17	195 142 82 30 11 6 3 194.37 146.08 71.82 34.04 16.60 8.44 4.47	a = 0.4327, b = -1.2187, c = 194.3682, R^2 = 0.99
T 18	103 82 34 28 11 5 3 103.34 79.54 41.41 20.84 10.76 5.77 3.22	a = 0.3999, b = -1.1217, c = 103.3407, R^2 = 0.99
T 19	103 42 41 18 11 5 102.30 48.73 28.71 18.90 13.34 9.87	a = -0.9215, b = -0.2139, c = 102.2983, R^2 = 0.96
T 20	122 72 33 19 4 121.96 72.18 33.41 15.72 7.78	a = -0.0358, b = -1.0402, c = 21.9624, R^2 = 1.00

Table 3.4b
Fitting the Zipf-Alekseev function to the motif-lengths in all texts (male authors)

Text	1 2 3 4 5 6 7 8	Results
T 21	131 101 54 15 8 2 130.66 103.20 47.38 20.47 9.06 4.19	a = 0.6565, b = -1.4381 c = 130.6603, R^2 = 0.99
T 22	173 105 47 19 6 2 1 172.88 105.88 44.81 18.84 8.31 3.87 1.90	a = 0.1845, b = -1.2865 c = 172.8810, R^2 = 1.00
T 23	206 125 56 23 7 2 2 206.86 126.04 53.48 22.55 9.97 4.66 2.29	a = 0.1796, b = -1.2803 c = 205.8596, R^2 = 1.00
T 24	73 57 30 13 5 72.90 57.62 28.48 13.41 6.46	a = 0.5425, b = -1.2725 c = 72.9005, R^2 = 1.00
T 25	164 90 57 17 12 5 163.57 93.80 46.59 23.98 13.02 7.43	a = -0.2197, b = -0.8405 c = 163.5717, R^2 = 0.99
T 26	115 101 47 37 19 7 2 0 1 115.31 98.59 55.02 29.19 15.74 8.76 5.04 2.99 1.83	a = 0.5390, b = -1.1038 c = 115.3059, R^2 = 0.97
T 27	83 51 28 15 8 1 82.87 51.84 26.76 14.07 7.74 4.46	a = -0.0744, b = -0.8689 c = 82.8702 , R^2 = 1.00
T 28	281 197 110 38 17 7 0 1 280.20 202.30 96.44 44.57 21.27 10.61 5.53 3.00	a = 0.3863, b = -1.2353 c = 280.2003, R^2 = 1.00
T 29	184 123 60 23 7 3 1 183.74 124.78 55.42 23.98 10.78 5.09 2.52	a = 0.3523, b = -1.3137 c = 183.7432, R^2 = 1.00
T 30	122 72 33 19 4 121.96 72.18 33.41 15.72 7.78	a = -0.0358, b= -1.0402 c = 121.9624, R^2 = 1.00
T 31	107 55 30 9 5 3 106.86 56.42 25.66 12.19 6.15 3.29	a = -0.2767,b = -0.9301 c = 106.8553, R^2 = 1.00
T 32	123 76 33 18 6 5 123.05 75.58 34.52 15.84 7.62 3.86	a = 0.0724, b = -1.1190 c = 123.0471, R^2 = 1.00
T 33	100 91 51 18 9 1 2 99.65 92.90 46.16 20.93 9.57 4.53 2.23	a = 0.9237, b = -1.4783 c = 99.6379, R^2 = 1.00
T 34	87 60 33 14 8 2 0 1 86.81 61.18 30.43 14.87 7.52 3.97 2.18 1.25	a = 0.2632, b = -1.1082 c = 86.8145, R^2 = 1.00
T 35	117 68 40 24 8 2 1 116.62 70.50 36.69 19.65 11.05 6.50 3.98	a = -0.1677, b = -0.8056 c = 116.6213, R^2 = 0.99
T 36	88 46 40 15 6 3 1 87.24 51.83 28.49 16.34 9.85 6.20 4.05	a = -0.2943, b = -0.6594, c = 87.2425, R^2 = 0.97
T 37	107 97 65 21 13 4 1 106.18 101.25 54.84 27.32 13.70 7.07 3.78	a = 0.8422, b = -1.3139 c = 106.1773, R^2 = 0.99
T 38	72 64 39 20 12 5	a = 0.5967, b = -1.0796,

	71.87 64.69 37.61 20.64 11.46 6.54	c = 71.8652, R^2 = 1.00
T 39	120 80 45 34 5 4 1 1 119.53 82.38 44.01 23.49 13.00 7.49 4.48 2.77	a = 0.0996. b = -0.9185, c = 119.5334, R^2 = 0.98
T 40	121 74 44 31 10 6 1 120.57 76.46 42.21 23.92 14.16 8.74 5.59	a = -0.1475, b = -0.7353, c = 120.5731, R^2 = 0.99
T 41	119 108 46 28 9 3 1 1 119.19 106.76 50.54 21.90 9.61 4.38 2.08 1.03	a = 0.9044, b = -1.5340, c = 119.1860, R^2 = 1.00
T 42	102 69 28 20 7 3 102.14 67.79 32.19 15.14 7.41 3.80	a= 0.1942, b = -1.1333, c = 102.1403, R^2 = 0.99

As can be seen, the determination coefficient is in all cases very high, hence the Zipf-Alekseev function is an adequate expression of this phenomenon. The parameter *c* represents always the beginning (first value) of the function, parameter *b* is situated in the vicinity of -1 and parameter *a* in the vicinity of zero. It seems that we, perhaps, approach a law expressed on a very high abstraction level.

In order to characterize the complete field we compute the criterion of Ord and present the results in Table 3.5.

Table 3.5
Ord's criterion for motif-length in individual texts

Text	I	S	Text	I	S	Text	I	S
1	0.69	1.36	15	0.86	1.76	29	0.61	1.51
2	0.83	2.01	16	0.69	1.67	30	0.57	1.13
3	0.79	1.57	17	0.70	1.65	31	0.68	1.69
4	0.70	1.24	18	0.84	1.71	32	0.71	1.69
5	0.71	1.28	19	0.84	1.42	33	0.64	1.47
6	0.86	1.36	20	0.57	1.13	34	0.75	1.78
7	0.73	1.54	21	0.57	1.26	35	0.61	1.61
8	0.71	1.59	22	0.61	1.60	36	0.76	1.50
9	0.59	1.30	23	0.63	1.70	37	0.67	1.32
10	0.73	1.62	24	0.56	1.01	38	0.74	1.23
11	0.86	2.09	25	0.61	1.17	39	0.76	1.63
12	0.68	1.40	26	0.86	1.71	40	0.80	1.46
13	0.69	1.49	27	0.69	1.28	41	0.70	1.68
14	0.71	1.43	28	0.65	1.53	42	0.70	1.47

The results are graphically presented in Figure 3.3, 3.4, and 3.5.

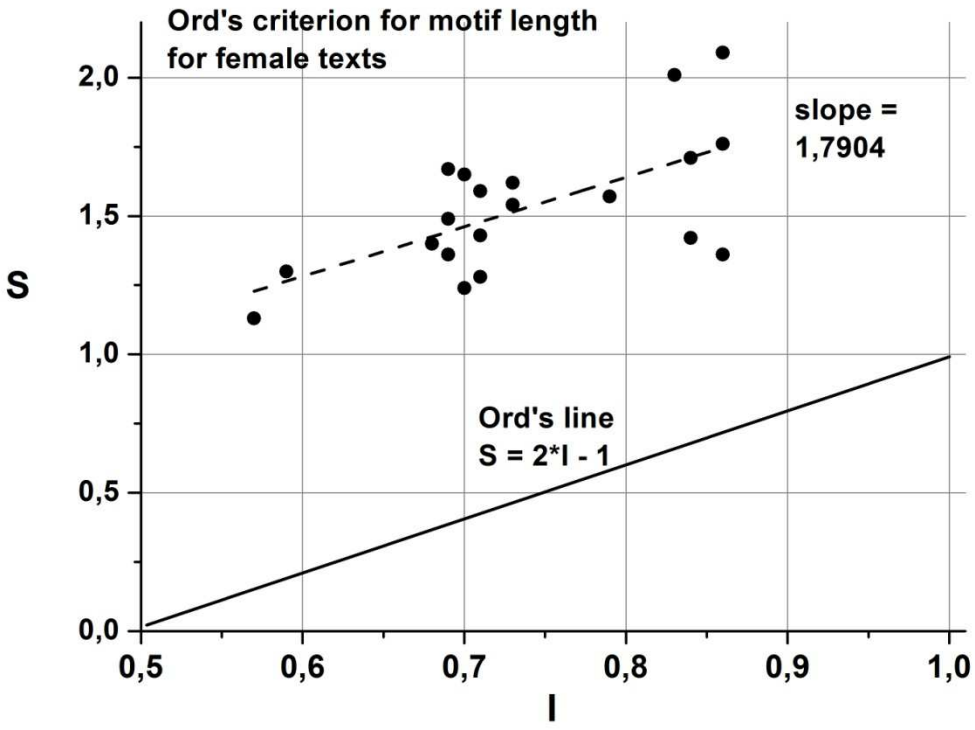


Figure 3.3. Ord's criterion for female texts

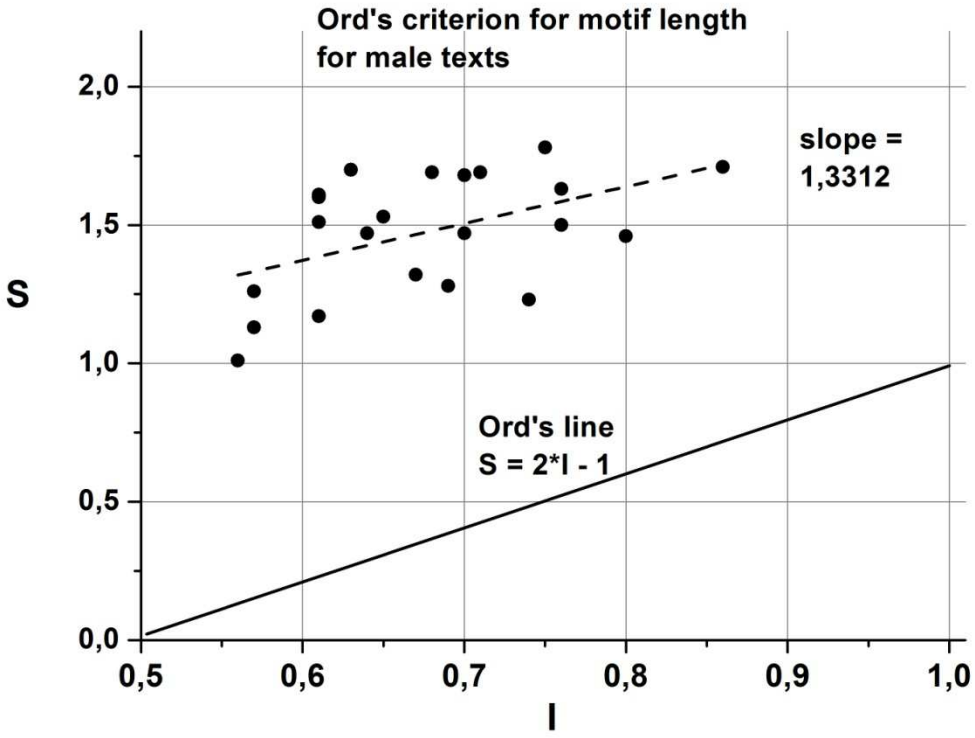


Figure 3.4. Ord's criterion for male texts

As can be seen, the length of motifs is not strictly concentrated as that of types, but both male and female texts are positioned above the Ord's line and lie rather in an ellipse. The I-value of male texts is slightly greater than that of female texts. Presented in a common figure, one can see that there is some background mechanism caring for a strong stylistic concentration.

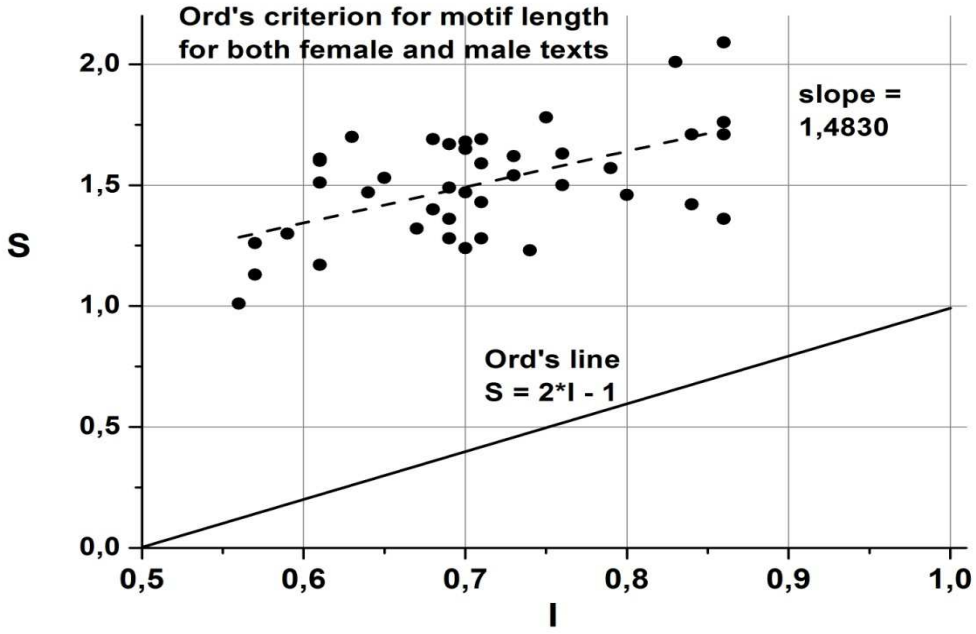


Figure 3.5. Ord's criterion for motif-length in female and male texts

Again, both the homogeneity of, say, texts written by men/women or the difference between men and women can be tested. Here we shall perform merely the same test as above. We add all frequencies of individual lengths with female texts and male text separately and obtain the results presented in Table 3.5.

Table 3.5
Frequencies of motif lengths in male and female texts

Length	1	2	3	4	5	6	7	8	10
Female	2467	1767	1023	494	212	70	31	3	1
Male	2800	1910	1016	458	187	67	14	4	

Performing the usual chi-square test for homogeneity we obtain chi-square = 1.79 with DF = 9 yielding P = 0.99. Hence we may preliminarily accept that motif lengths of adnominals in Russian do not differ between male and female writers.

4. Conclusions

Adnominals are only one of the many phenomena of the text. They can be attained by classifying some words or expressions of the text. However, different linguists may perform the classification in different ways. The approximation to some real phenomenon can be judged only if one sets up hypotheses and tests them using statistical tests. The hypotheses may be (preliminarily) considered as corroborated if the results are positive. But one never knows whether the phenomenon behaves in the same way in all languages. Nevertheless, in the mean time one can make some further steps in such a way that the given entities are

ordered in more abstract classes. Here we tried to analyze the motifs of adnominals. A different way would be the study of runs of adnominals, forming of Belza-chains based on adnominals taking into account also the sentences, classifying "hrebs" and studying their motifs, correlating the above results with other properties of the texts, etc. Evidently, this is a way without end.

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.	1995	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	572
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	***Oligarh s Bol'shoj Medveditsy***	2004	4749	587

T. Ustinova	(*Oligarch from the Big Dipper*). Novel.			
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 B. Akunin	***Pikovyj valet*** (*Jack of spades*). Long story.	1999	4043	650
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 S. Dovlatov	Predislovie k sborniku rasskazov «***Chemodan***» i rasskaz ***Krepovye finskie noski*** (Preface to the collection of stories *Suitcase* and the story *Crepe Finnish socks*).	1986	3100	386
T 32 S. Dovlatov	***Filial*** (*The branch*). Long story.	1987	3079	506
T 33 V. Erofeev	***Russkaja krasavitsa*** (*Russian beauty*).	1990	4206	577
T 34 D. Koretskij	***Antikiller.*** Novel.	1995	2948	422
T 35 D. Koretskij	***Antikiller-5.*** Novel.	2014	3937	528
T 36 D. Koretskij	***Antikiller-6.*** Novel.	2016	2815	414
T 37 V. Pelevin	***Operatsija "Burning Bush"*** (*Operation "Burning Bush"*).	2010	3392	676

T 38 V. Pelevin	***Assasin*.**	2008	3565	487
T 39 V. Pelevin	***Grecheskij variant*** (*Greek variant*). Novel.	1997	2891	616
T 40 Z. Prilepin	***Obitel'*** (*Convent*). Novel.	2014	4523	618
T 41 Z. Prilepin	***Patologii*** (*Pathologies*). Novel.	2005	3968	**664**
T 42 Z. Prilepin	***Sher amin'*** (*Cher amen*). Story.	2016	3774	457

References

Altmann, G. (2016). On Köhlerian motifs. In: Kelih, E., Knight, R., Mačutek, J., Wilson, A. (eds.), *Issues in Quantitative Linguistics V. 4, 2-8*. Lüdenscheid: RAM-Verlag.

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative Properties of Argumentation Motifs. In: Obradović, I.; Kelih, E.; Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012: 33-43.* Belgrade.

Boroda, M.G. (1973). K voprosu o metroritmičeski elementarnoj edinice v muzyke. *Bulletin of the Academy of Sciences of the Georgian SSR 71(3),745-748.*

Boroda, M.G. (1982a). Die melodische Elementareinheit. In: Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. Quantitative Analysen: 205-222.* Bochum: Brockmeyer.

Boroda, M.G. (1988). Towards a problem of basic structural units of musical texts. *Musikometrika 1, 11-69.* Bochum: Brockmeyer.

Köhler, R. (2000). A study on the informational content of sequences of syntactic units. In: L.A. Kuz'min (ed.), *Jazyk, glagol, predloženie. K 70-letiju G.G. Sil'nitskogo.* Smolensk, 51-61.

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

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

Köhler, R., Naumann, S. (2008c). *Quantitative text analysis using L-, F- and T-segments*. In: Preisach, B.; Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications*: *637-646.* Berlin, Heidelberg: Springer,

Köhler, R., Naumann, S. (2010). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, P.; Kelih, E.; Mačutek, J. (eds.), *Text and Language: 81-89.* Wien: Praesens.

Mačutek, J. (2009). Motif richness. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM-Verlag.

Mačutek, J., Mikros, G. (2015). Menzerath-Altmann law for word length motifs. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 125-131*. Berlin/Boston: de Gryuter.

Milička, J. (2015). Is the distribution of L-motifs inherited from the word length distribution? In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 133-145.* Berlin/Boston: de Gryuter.

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

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives: 183-194.* Wien: Praesens.

Glottometrics 38, 2017, 107-111

Commentary

A commentary on "The Now-Or-Never Bottleneck: A Fundamental Constraint on Language", by Christiansen and Chater (2016)

Ramon Ferrer-i-Cancho[1]

In a recent article, Christiansen and Chater (2016) present a fundamental constraint on language, i.e. a now-or-never bottleneck that arises from our fleeting memory, and explore its implications, e.g., chunk-and-pass processing, outlining a framework that promises to unify different areas of research. Here we explore additional support for this constraint and suggest further connections from quantitative linguistics and information theory.

(a) Further support for the now-or-never bottleneck

Memory limitations are at the core of the now-or-never bottleneck. In section 3 of their article, Christiansen and Chater (2016) review well-established facts from psychological experiments on memory constraints. Further support for this view comes from statistical research showing that, for instance, the probability that two syntactically related words are at a certain distance in a sentence decays exponentially with distance in sentences of the same length (Ferrer-i-Cancho 2004). This decay of probability might be mirroring an exponential decay of memory as a function of time. A power-law decay has been reported when mixing distances from sentences of different length (Liu 2007), but this result could be an artefact of mixing information from sentences of different length (Ferrer-i-Cancho and Liu 2014).

In section 6.1.2, the authors review the statistical evidence of one implication of the now-or-never bottleneck "*The prevalence of local linguistic relations*" (presented in section 1). Further confirmation of this prediction of the now-or-never bottleneck is provided by statistical analyses showing that

- About 50% of adjacent words are linked syntactically (Yuret 2006) and about 50% of dependencies involve adjacent words (Ferrer-i-Cancho 2004; Liu 2008). This might be reflecting the need to recode the input as it comes along (Christiansen and Chater 2016).
- Constraints of syntactic dependencies such as planarity (non-crossing dependencies) could be a side effect of pressure for linguistic relations to be local (Ferrer-i-Cancho and Gómez-Rodríguez 2016b and references therein).

The latter finding is of enormous theoretical importance: it frees grammar, the language faculty,… from the responsibility of the rather low frequency of crossing dependencies in

[1] Complexity and Quantitative Linguistics Lab. LARCA Research Group, Departament of Computer Science, 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.

real languages. Connecting the dots, one could conclude that the now-or-never bottleneck predicts a structural property of syntax. That complements connections between chunk-and-pass and a wide range of linguistic phenomena that the authors indicate at the end of section 6.1.2. From a theoretical perspective, dependency length minimization not only predicts that crossings dependencies should have a low frequency (Ferrer-i-Cancho 2014b, Ferrer-i-Cancho and Gómez-Rodríguez 2016b) but also the consistency of branching and its direction, and the fixed order of certain kinds of syntactic relationships, e.g. adjective-noun (Ferrer-i-Cancho 2015a-b). The now-or-never bottleneck is crucial for the development of a parsimonious theory of syntax (Ferrer-i-Cancho and Gómez-Rodríguez 2016a).

Christiansen and Chater (2016) propose a unified approach across scales, from the sentence scale to the historical/biological scale passing through the individual scale (Fig 3 of their article). Interestingly, the hypothesis of a word order permutation ring that constrains word order evolution (Ferrer-i-Cancho 2015a) can be reinterpreted as another implication of the now-and-never bottleneck at the historical level. Under pressure for dependency length minimization, that ring predicts that the development of SVO order from SOV is more likely than the development of OVS from SOV, although both SVO and OVS are convenient from the perspective of dependency length minimization (Ferrer-i-Cancho 2015a), in full agreement with historical data (Gell-Mann and Ruhlen 2011).

(b) The now-or-never bottleneck as a solution to puzzles

The now-or-never bottleneck and its implications can help us to understand various puzzles. As we mentioned above, the probability that a syntactic dependency involves words at a certain distance decays exponentially with distance but slows down for distances about 4-5 onwards in Czech (Ferrer-i-Cancho 2004; see Fig. 4 B and D). The point is: if the length of a syntactic dependency is a burden, why the decay of that probability slows down at some point? The paradox could be solved hypothesizing a chunk-and-pass strategy and multiple levels of representation to fight against memory limitations at long distances, supporting Christiansen and Chater's (2016) view. Interestingly, a distance of 4-5 could be related with the size of word chunks at some level. Suppose that dependents can go eihter to the left and to the right of their head. Then a crossover at distance d translates into a chunk size of $2d+1$. Applying this theoretical argument, the crossover at distance 4-5 words yields chunk sizes of 9-10 words. In contrast, if dependents can go only at one side of their head (consistent branching), a crossover at distance d translates into a chunk size of $d+1$ and then the crossover at distance 4-5 yields a chunk size of 5-6 words. The relationship between that crossover and chunk size should be investigated with the help of dependency treebanks.

Another puzzling fact comes from analyses of the growth of the sum of dependency lengths as a function of sentence length: this sum is below a random baseline (supporting dependency length minimization) but clearly above the theoretical minimum, the one that is obtained by solving the minimum linear arrangement problem (Ferrer-i-Cancho 2004, Ferrer-i-Cancho and Liu 2014). The question is: why are real languages not getting to that minimum? One answer is that dependency length minimization is in conflict with other principles or that pressure to minimize dependency lengths increases with the length of the sentence (Ferrer-i-Cancho 2014a). A more interesting possibility follows from Christiansen and Chater's now or never bottleneck: an online, incremental, chunk-and-pass language production would lead to suboptimal linear arrangements and the real sum of dependency lengths would reflect it. Real sentences are unlikely to be optimized following the batch processing that algorithms for solving the minimum linear arrangement problem imply (Chung 1984; Shiloach 1979).

(c) Conflicts between implications of the now-or-never bottleneck

At the beginning of section 1, the authors enumerate the implications of the now-or-never bottleneck:

> 2. The prevalence of local linguistic relations
>
> …
>
> 4. The use of prediction in language interpretation and production.

Some aspects of these two implications are incompatible: the word order that maximizes locality is incompatible with the word order that maximizes prediction (Ferrer-i-Cancho 2014a). That conflict could be one of the reasons why the sum of dependency lengths of sentences does not get to the theoretical minimum (as explained in (b)). This kind of conflict is reminiscent of other conflicting constraints in language optimization at the level of the mapping of words into meanings (Ferrer-i-Cancho 2005).

Conflicts between implications are not a problem for the now-or-never bottleneck. They may simple illuminate the wide range of solutions adopted by language and why languages keep evolving.

(d) Further insights from the information theory

In section 6.1.3, the Christiansen and Chater (2016) write "*The problem of encoding and decoding digital signals over an analog serial channel is well-studied in communication theory (Shannon, 1948)—and, interestingly, the solutions typically adopted look very different from those employed by natural language.*"

The authors are missing recent research based on information theory predicting solutions that resemble a lot those of natural language: duality of patterning (Plotkin and Nowak 2000; relevant for section 6.1.4), Zipf's law for word frequencies (Ferrer-i-Cancho 2016a, Prokopenko et al 2010, Ferrer-i-Cancho 2005), Zipf's law of abbreviation (Ferrer-i-Cancho et al 2015), Menzerath's law (Gustison et al 2016), Clark's principle of contrast (Ferrer-i-Cancho 2017), a vocabulary learning bias in children (Ferrer-i-Cancho 2017),…

Information theory can make even more general predictions on language and beyond. Power-law-like distributions such as Zipf's law for word frequencies can be regarded as manifestations of critical-like behaviour (Kello et al 2010). In general, the critical-like behavior that many natural systems such as language exhibit can be explained using information theory: "*criticality turns out to be the evolutionary stable outcome of a community of individuals aimed at communicating with each other to create a collective entity*" (Hidalgo et al 2014).

As the authors say, "*the Now-or-Never bottleneck provides a constant pressure towards reduction and erosion across the different levels of linguistic representation, providing a possible explanation for why grammaticalization tends to be a largely unidirectional process*" but there are even stronger predictions that can be made from the now-or-never bottleneck. The authors argue that "*the brain must compress and recode linguistic input as rapidly as possible*" to deal with the "Now-or-Never" bottleneck. It is precisely a generalized principle of compression (the minimization of the mean energetic cost of units) that predicts three linguistics laws: Zipf's law for word frequencies (Ferrer-i-Cancho 2016a), Zipf's law of abbreviation (Ferrer-i-Cancho et al 2015) and Menzerath's law (Gustison et al 2016). Furthermore, that principle could be related to the principle of dependency length minimization: compression can help to reduce the actual distance between

syntactically related words when measured in syllables or phonemes (Ferrer-i-Cancho 2015b). Thus, memory limitations could promote compression across levels and domains.

We hope that our comments stimulate further research in quantitative linguistics.

Acknowledgements

We are grateful to N. Chater, M. Christiansen and C. Gómez-Rodríguez for helpful comments and stimulating discussions. This work was supported 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

Christiansen, M. and Chater, N. (2016). The now-or-never bottleneck: a fundamental constraint on language. *Brain and Behavioral Sciences 39, e62.*

Chung, F.R.K. (1984). On optimal linear arrangements of trees. *Computers & Mathematics with Applications 10 (1), 43-60.*

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. (2014a). 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. (2014b). A stronger null hypothesis for crossing dependencies. *Europhysics Letters 108 (5), 58003.*

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). Compression and the origins of Zipf's law for word frequencies. *Complexity 21 (S2), 409-411*.

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

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

Ferrer-i-Cancho, R. and Gómez-Rodríguez, C. (2016a). Liberating language research from dogmas of the 20th century. *Glottometrics 33, 33-34*.

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

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

Gell-Mann, M. and Ruhlen, M. (2011). The origin and evolution of word order. *PNAS 108 (42), 17290–17295.*

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

Hidalgo, J., Grilli, J., Muñoz, M. A., Banavar, J. R. and Maritan, A. (2014). Information-based fitness and the emergence of criticality in living systems. *Proceedings of the National Academy of Sciences 111 (28), 10095-10100.*

Kello, C.T., Brown, G.D.A., Ferrer-i-Cancho, R., Holden, G., Linkenkaer-Hansen, K., Rhodes, T. and Van Orden, G.C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences 14 (5), 223-232.*

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

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

Plotkin, J.B. and Nowak, M.A. (2000). Language evolution and information theory. *Journal of Theoretical Biology 205 (1), 147-159.*

Prokopenko, M., Ay, N., Obst, O. and Polani, D. (2010). Phase transitions in least-effort communications. *Journal of Statistical Mechanics, P11025*.

Shiloach, Y. (1979). A minimum linear arrangement algorithm for undirected trees. *SIAM Journal on Computing 8 (1), 15-32.*

Yuret, D. (2006). Dependency parsing as a classification problem. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X).* New York City: Association for Computational Linguistics, pp. 246-250.

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, 172 pp.