Quantitative Analysis of Bibliographic Corpora: Statistical Features, Semantic Profiles, Word Spectra*

Adam Pawłowski[1], Krzysztof Topolski[2], Elżbieta Herden[1]

1) University of Wrocław, Institute of Information and Library Science
2) University of Wrocław, Institute of Mathematics

Abstract:

The subject of this article is the analysis of the bibliographic corpus, derived from the Polish national bibliography. The research allowed us to discover and compare quantitative characteristics of the bibliographic corpus and of the reference corpus of general language. It was shown that the two corpora differ significantly. In particular, differences in the share of particular parts of speech and of the frequency distribution of lexemes were demonstrated. The statistical distributions of word spectra were also studied. The best fit was obtained for generalized inverse Gauss-Poisson and Zipf-Mandelbrot distributions. The analysis of parameters of both distributions for bibliographic and reference corpora also revealed differences between them. The best perspective for future research on bibliographic corpora is, apart from quantitative linguistics, semantic analysis and text-mining.

Key Words:

quantitative linguistics, corpus linguistics, word spectra, statistical distributions, MARC, bibliography, Polish, book titles

1. Large-scale bibliographies as text corpora

Originally created as registers of literary output, large-scale bibliographies have so far served more practical purposes, in particular for publication searches, their archivization, and/or evaluation of scientific output of institutions, disciplines, and individuals. However, over time, the cognitive potential they have developed has been recognized, allowing them to be treated not as auxiliary tools but as fully-fledged research objects. The conditions for undertaking research on large-scale bibliographies were numerous and, it seems, all of them have now been met. First of all, the critical mass, i.e., the volume of data needed to draw valuable conclusions and generalizations of a scientific nature, has long since been exceeded. Secondly, this data has become available to researchers in digital form. Thirdly, methods of automatic natural language processing have emerged which allow quick analysis of fields

* This research was supported by the Polish National Science Center (NCN) under grant 2016/23/B/HS2/01323 "Methods and tools of corpus linguistics in the research of a bibliography of Polish book publications from 1997 to 2017".

encoded in text format, moving away from the bibliography as storage place. In particular, the processing of large collections consisting of text fields of bibliographic records has been made possible through the use of corpus and quantitative linguistics.

It is noteworthy that the application of quantitative methods in "library science" is not a completely new phenomenon. Such an approach would have unjustifiably depreciated the achievements of researchers of the past, who were not inferior to contemporary researchers, and often put forward great ideas, but who did not have computer tools to implement innovative ideas. One of the first methodical applications of statistics in the study of bibliographic inventories is the work of Gustav Schwetschke (Schwetschke 1850, Schwetschke 1877), who used the catalogues of the bookfairs in Frankfurt am Main and Leipzig from 1564-1846 (the German national bibliography for that period did not exist at the time) to present the geographical and numerical distribution of the German bookselling industry from the 16th to 19th century.

Another example of the use of traditionally understood statistics in bibliography research is research conducted in France the 1950s by Lucien Febvre & Henri-Jean Martin (1958), founders of the *Annales* school. Their groundbreaking work *L'apparition du livre* argued that books were not only carriers of ideas, but also material objects and as such commodities subject to the laws of economics. From more recent studies it is worth mentioning the monumental *The Cambridge History of the Book in Britain* (CHBB 1999-2019), covering the history of the book in the British Isles from 400 to the end of the 20th century. The compendium uses rich bibliographic material to recreate the processes of the creation, production, distribution and reception of the book in the British Isles.

Newer studies apply more advanced methods of statistics and methodology developed within the framework of digital humanities. This type of work consists in analyzing old, digitized bibliographies and comparing them with the resources of modern incunabula or old print databases to determine the state of preservation of old printing or publishing production (Green et al. 2011). In this context, it is worth noting the interdisciplinary research conducted by the Helsinki Centre for Digital Humanities. Researchers associated with the Centre document the history of book production in Finland and Sweden based on the analysis of large collections of bibliographic metadata (databases of retrospective national bibliographies)

(Tolonen et al. 2019a; Lahti et al. 2019). They introduced the term 'bibliographic data science' for this new research paradigm (BDS)[1] (Tolonen et al. 2019b).

The above literature review indicates that the development of bibliography research in recent years is a fact, but also reveals the one-sidedness of the approach taken so far. In particular, there is a noticeable lack of application of quantitative linguistics methods that is likely to reveal new, cognitively valuable aspects of bibliographies understood as a specific informative text genre. The advantages of bibliographic corpora, important from the perspective of quantitative linguistics, include size (they consist of hundreds of thousands and even millions of records), careful preparation (data are "clean", because they are input by competent employees and not by unprepared users), systematic approach (there are exact dates of each entry, and missing dates can be easily reproduced) and repeatability of structures (each record is assumed to have the same structure). The weakness of bibliographies, when compared to real text corpora, is the lack of longer discursive fragments that convey more complex information than just titles, keywords or generic qualifiers. This fact, however, is not an obstacle to effective quantitative and text-mining tasks. The change in approach to large-scale bibliographies is due to the fact that computer systems allow for automatic extraction of complete sets of data from selected fields (e.g. title or keywords) and their comprehensive analysis by NLP methods. Thanks to this, these large data resources, initially created as databases, also become machine-readable text corpora, effectively used in corpus and quantitative linguistics.

2. Data and hypotheses

The subject of this analysis is a collection of 553,000 records extracted from the resources of the Polish National Library. They represent contemporary texts from the years 1997-2017 and a small number of re-editions of works written earlier (mainly in the 19th century). Records are stored in MARC21 format, used worldwide in database systems of libraries. This format contains a large number of fields and is difficult to process automatically because it is highly redundant (the same information may be repeated in different fields).

From a formal point of view (irrespective of repetitions due to redundancy), the list of fields suitable for linguistic research is as follows:

[1] NB, by introducing the concept of statistical analysis based on bibliographic data, Finnish researchers are actually using a bibliographic method that is well established in the 19th century. Its transfer to the digital sphere obviously creates new research opportunities.

– author (field 100, subfield ‘a’, if subfield ‘e’=‘autor’ or field 700, subfield ‘a’, if subfield ‘e’=‘redakcja’ or ‘e’= ‘autor’);
– title (field 245),
– subject according to the list of subject headings list of the Polish National Library (various fields starting with the digit six – 6xx);
– genre (field 655).

In our case, the title field 245 was selected for the study because it contains the most complete text data in the form of sentence equivalents or full sentences. Although it is difficult to speak of a corpus analysis in the full sense of the word here because in a bibliographic corpus there is no category of a text as a larger, compact whole, the titles fulfill basic communication functions and have such a complex structure that the use of corpus tools is fully justified. Other fields in text format (‘subject’ or ‘genre’) intended for machine-processing, on the other hand, contain only one-word terms that can be machine-processed with corpus or information retrieval tools.

Since the large corpora of titles have not been yet analyzed from the perspective of statistical linguistics, the research conducted was primarily exploratory in nature. Basic quantitative parameters of the corpus were calculated (average lengths of words, sentences, titles, as well as histograms of distributions of these parameters). Frequency lists of words in titles were also prepared and analyzed, as well as the distribution of POS (part of speech) shares in the corpus, which made it possible to create a morphological and a semantic profile of this type of textual resource and to present it against the background of the general language. The nature of these tasks required lemmatization of the text, which was carried out with the use of the WCRF Tagger[2].

The research hypothesis was the assumption that the corpus of titles will be significantly different from the corpus of the general language, which should be reflected, among others, in the values of its quantitative parameters. In order to verify this hypothesis, statistical distributions of vocabulary (so-called word spectra) were generated for the corpus and for the data from the National Corpus of Polish Language. Their parameters were estimated and their values were compared. Apart from that, the comparisons of the POS in the two above mentioned corpora were made, and the relationship between the length of the title and the mean length of the word was also analyzed.

[2] http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki/

3. Research method

Data for research were obtained through the programming interface (API) of the National Library BN Data[3]. It allows one to extract large sets of records or download files with databases containing BN resources representing bibliographical records from longer periods of time. Excerption of the fields containing textual data from the MARC records was performed with our own programming tools. Pre-processing of data for the study required lemmatization, which was performed with the use of WCRFT2 morphosyntactic analyzer, available in the CLARIN-PL infrastructure[4]. This tool also enabled automatic recognition and annotation of POS in the text. Estimation of probability distributions and other statistical tests were conducted with the use of package R. The general methodological principle of corpus linguistics was adopted, which requires that special corpora be compared with reference ones. Following this principle, results generated on the basis of the bibliographic corpus were compared with the National Corpus of Polish (NKJP)[5]. Since not only single words but also sentences were analyzed in the reference corpus, it was necessary to use the corpus manually segmented into sentences. Language is a complex phenomenon and automatic segmentation of Polish texts into sentences gives a high percentage of errors; there are furthermore different ways of calculating sentence length (numbers, units of measurement, auxiliary symbols, etc. can be noted differently). For this reason, we did not use the entire NKJP resource, but only the manually annotated 1-million-word subcorpus[6]. This was not an obstacle to conclusions, because in this case the increase in the volume of the corpus did not significantly increase the quality of estimation of the parameters examined.

4. Results – an overview

The results obtained gave an interesting picture of the corpus of titles, confirming the advanced hypotheses. For general statistics on the two corpora, see Table 1. As one can see, the statistical profile of the title corpus differs significantly from that of the reference corpus. Words in titles are on the average longer (there are most probably fewer function words), while titles are on the average shorter than sentences in the general language.

[3] http://data.bn.org.pl/
[4] http://ws.clarin-pl.eu/tager.shtml
[5] http://www.nkjp.uni.lodz.pl/
[6] http://www.nkjp.pl/index.php?page=14&lang=1

Table 1. Basic statistical parameters of the bibliographic and general corpus

Unit	Bibliographic corpus		General language	
	mean length	stand. deviation	mean length	stand. deviation
Word (letters)	*m*=6.49	*d*=3.81	*m*=5.78	*d*=3.36
Sentence / clause (words)	–	–	*m*=13.06	*d*=11.57
Title (words)	*m*=7.32	*d*=5.53	–	–

The analysis of the histograms of the length of titles and their deviations in both corpora gave quite an interesting picture, unheard of in the case of the general language. As can be seen in Figure 1, short titles (from 2 to 4 words) dominate in contemporary publications, while the decrease in the number of longer titles is (surprisingly) almost linear. In order to verify this result, a similar analysis was carried out on the reference corpus. It was assumed that the equivalent of the titles in the bibliographic corpus would be sentences in the general language (represented here by the National Corpus of Polish). Sentences are not the same as titles, just as bibliographic corpora do not show all communication functions of language. In both cases, however, these units are basic carriers of meaning, they are semantically and syntactically closed, and also their length is, at least seemingly, similar. The result of the comparison was in this case difficult to predict, so the approach used was necessarily purely exploratory. Due to the specificity of the data, titles of up to 10 words and sentences of up to 20 words were considered.

Figure 2 shows that titles and sentences have generally similar distributions but differ due to the more complex structure of the general language. What makes them similar is the existence of an extremum to which the values increase and then decrease. This extremum indicates the most common, and therefore communicatively optimal, length of respective units. As one can see, this optimum is in a different place for both corpora. Communication by means of titles takes place through book covers. It requires great conciseness and simplicity: a good title should be captured in the blink of an eye. Normal communication by means of texts or speech is not subject to such pressure: nothing has to be understood at a glance. The curve's shape in Figure 2 corresponds to the natural process of perception, which is determined by Zipf's forces minimizing the effort required to transfer / acquire a given amount of information. This natural sentence length would be close to 7-8 units (thus much longer than a title length).

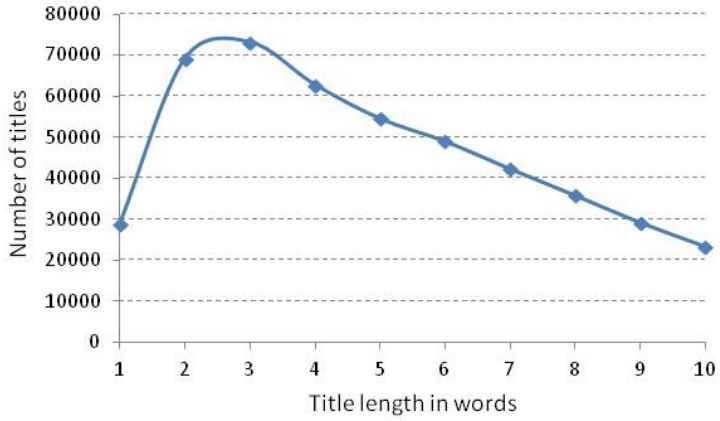

Figure 1. Histogram of title lengths in the bibliographic corpus

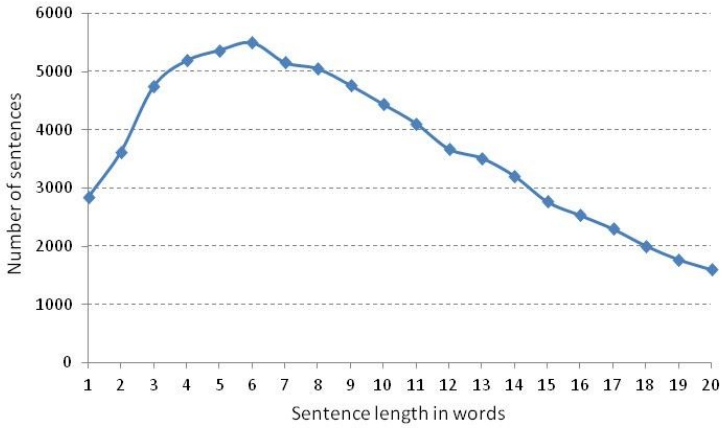

Figure 2. Histogram of sentence lengths in the general language (National Corpus of Polish)

An even more unusual result is the histogram of the average word length in the title plotted against the title length measured by the number of words (Fig. 3). Here, in the case of three- and four-word titles (the most frequent), an anomaly in the form of a sudden drop appears. This anomaly results from the morphosyntactic structure of the Polish language and of other flectional languages. Three- and four-word titles are almost obligatorily expressed by two (three) meaningful lexemes and a function word which determines the relation between them. This is illustrated by the structure of titles such as *Podręcznik dla ośmioklasistów* (Textbook for Eighth Graders), *Droga do Emaus* (The Road to Emmaus), or *Ułan i dziewczyna* (The Lancer and the Girl). With longer phrases, the number of function words no longer deforms the result, because the number of lexical segments also increases. It can also be expected that in analytical languages using articles this anomaly will appear in quite frequent 3- and 4-word titles (e.g., *Histoire de la terre* or *History of the Earth*). However, a histogram prepared according to the same principles in the case of agglutinative languages would certainly look different.

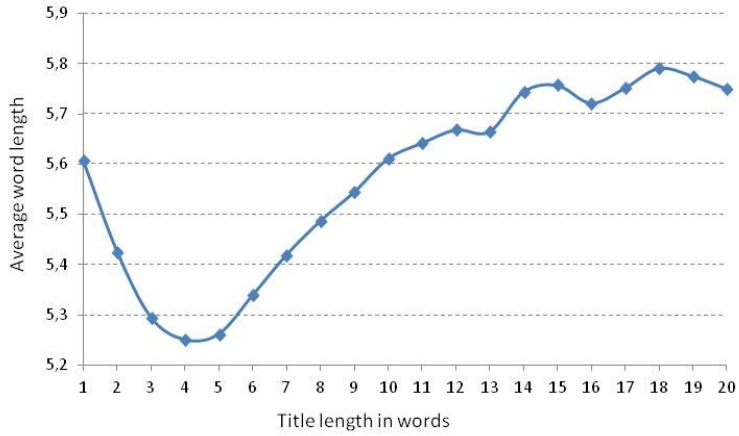

Figure 3. Average word length in a title vs. title length in the bibliographic corpus

As in previous cases, the corpus of the titles was compared with the corpus of the general language. Figures 3 and 4 display a seemingly similar pattern: a sharp decline in average word length followed by an increase. However, differences are noticeable between both corpora. The average word length in titles is lower than the similar parameter in the general language. As we have already said, this is probably due to the fact that titles should be simple to read, thus composed of shorter words.

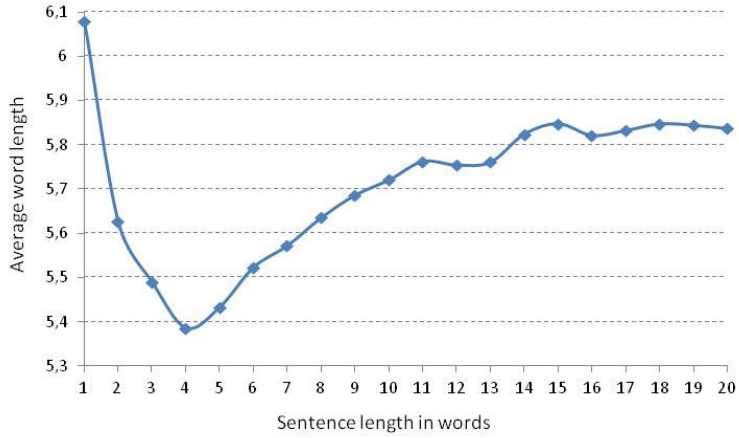

Figure 4. Average word length in a sentence vs. sentence length in the general language (National Corpus of Polish)

The analysis of the relationship between the length of the titles, as well as sentences in the corpus of the general language, and the average length of the words they contain might suggest that Menzerath-Altmann's law would be applicable here. This law states that 'the increase of the size of a linguistic construct results in a decrease of the size of its

constituents'. However, such a relation is valid only for the immediate constituents – and in the case of a sentence such a constituent is a clause, not a word. The same applies to titles which additionally do not have a clear syntactic status. It should come as no surprise, therefore, that the curves in Figures 3 and 4 do not show any law-like tendency, at least from the point of view of quantitative linguistics.

Significant results were obtained in the study of the POS distribution in the corpus of titles as opposed to the general language. The numbers in Table 2 indicate significant differences between the Polish language used in normal communication and the specific language of the title. The language of titles is generally highly nominalized (the proportion of nouns in the title corpus is 57.15%, and in the general language 43.45%), while the proportion of verbs in the bibliographic corpus is significantly lower. Clear differences are also visible in the pronoun group (0.43% vs 1.97%). This is due to the fact that titles are most often impersonal phrases (e.g. *Basics of cellular biology*), while in the general language, the frequency of pronouns increases with the presence of dialogue. While the above differences are due to purely genological specificities, the very high frequency of numbers is due to the fact that titles (and subtitles) often contain indications of edition numbers, parts or dates.

Table 2. Distribution of the main POS in the bibliographic corpus (titles) and in the general language

POS	Bibliographic corpus	General language
Noun	57.15%	43.45%
Verb	3.00%	15.24%
Adj.	11.21%	10.75%
Adv.	1.12%	3.91%
Pron.	0.43%	1.97%
Prep	9.14%	10.90%
Conj.	4.97%	3.96%
Num.	5.44%	0,62%

5. Results – statistical distributions

The estimation of statistical distributions of spectral lists of vocabulary in the bibliographic and reference corpus also gave interesting results. To compare bibliographical data with data representing the general language we will use the grouped frequency distribution or the

frequency spectrum *V(m,N)* with $m \geq 1$, which is defined as the number of types with frequency *m* in a sample of *N* tokens. Formally:

$$V(m,N) = \sum_{i=1}^{V(N)} I\{f(i,N) = m\},$$

where $V(N) = \sum_m V(m,N)$ and the indicator function $I\{x\}$ is equal 1, if expression *x* is true, and zero otherwise. There is a natural connection between rank-frequency distribution and frequency spectrum:

$$V(m,N) = \sum_i I\{f(i,N) \geq m\} - \sum_i I\{f(i,N \geq m+1)\}$$

We considered as potentially relevant statistical distributions typically used in the research on lexical data (Baayen 2001). After some preliminary tests it appeared that the best fit of word frequencies was obtained for the generalized Zipf distribution in the case of the general language and the Sichel model for the corpus of titles. The rank frequency distribution described by the Zipf-Mandelbrot law is of the form:

$$f(i,N) = \frac{K}{(i+b)^a},$$

where $a > 1$ and $b \geq 1$ are parameters and K is the normalizing constant (Mandelbrot 1962).

The Sichel model uses the generalized inverse Gauss-Poisson distribution as a description of word probabilities (Sichel 1975). The probability density function for the generalized inverse Gauss-Poisson distribution has the following form:

$$g(x) = Mx^{a-1}\exp(-\frac{x}{c} - \frac{b^2 c}{4x}).$$

The normalizing constant *M* is of the form $M = \frac{(2/bc)^{a+1}}{K_{a+1}(b)}$, where K_a denotes the modified Bessel function of the second kind of order *a*. The maximum likelihood estimators of the generalized inverse Gauss-Poisson distribution parameters are described in Sichel (1982).

After some preliminary tests it appeared that the best fit of word frequencies was obtained for the generalized Zipf distribution in the case of the general language and the Sichel model for the corpus of titles. The results of the estimation are presented in Table 3, where ZM and GIGP denote the Zipf-Mandelbrot distribution and the generalized inverse

Gauss-Poisson distribution respectively. They prove that bibliographical data (titles) differ statistically from the general language (if one assumes that it is represented by the National Corpus)[7].

Table 3. Statistical distributions of word spectra in the general language and in the bibliographical corpus

	Distribution	Par. a	Par. *b*	Par. *c*
Bibliographical corpus (titles)	GIGP	-0.5277	0.00113	0.0014
Reference corpus (general language)	ZM	0.5877	0.00288	-

As a comparison, Table 3a presents the values of the parameters of the fitted Zipf-Mandelbrot distribution for the bibliographical corpus data and the generalized inverse Gauss-Poisson distribution for the general language data.

Table 3a. Statistical distributions of word spectra in the general language and in bibliographical corpus.

	Distribution	Par. a	Par. *b*	Par. *c*
Bibliographical corpus (titles)	ZM	0.5330	0.00133	–
Reference corpus (general language)	GIGP	-0.5995	0.00006	0.0047

This may come as a surprise as the visual inspection of the fits presented in figures 5 and 6 suggests that the observed and the expected data match almost perfectly. However, human perception is usually misleading in the case of big amounts of data (cf. Grotjahn & Altmann 1993). When the chi-squared tests are applied and threshold values are respected, the opposite is proven: both of the fits fail and the advanced hypotheses of similarity should be rejected. This means that both corpora have different lexicostatistical characteristics. Indirectly one should conclude that publication titles, when assembled in a relatively large corpus, do not have the properties of "normal" language as it is used for communication by humans.

On the other hand it is known that the Pearson chi-square goodness-of-fit test rejects all null hypotheses if the sample size is sufficiently large and for this reason it creates a problem with correct interpretation of the test prediction. In Mačutek & Wimmer (2013), it was suggested that it is possible to solve this problem by taking into account not only significance level, but also the so-called test resistance. One of the simplest possible tests of this type is based on *C*, the discrepancy coefficient defined as:

[7] Calculations have been performed with the *ZipfR* package (Evert & Baroni 2007; http://zipfr.r-forge.r-project.org/).

$$C = \frac{\chi^2}{N},$$

where χ^2 is the value of the Pearson test statistics for the data considered and N is the size of the sample.

When the differences between theoretical and empirical relative frequencies are fixed, the chi-square statistic increases linearly with the sample size. For this reason, Cressie & Read (1984) propose the use of discrepancy to evaluate the quality of fit. For the linguistic data, as a rule of thumb, a fit is considered to be satisfactory if $C < 0.02$. Several more advanced possible solutions of this problem have been mentioned and discussed in Mačutek & Wimmer (2013). In Table 4 we present the values of the chi-square statistics and the corresponding values of the discrepancy coefficient C, for the best fit distributions presented in Table 3 and for the alternative fit from Table 3a.

Table 4. The result of the Pearson goodness of fit test and corresponding discrepancy coefficient for the distributions of word spectra in the general language and in the bibliographical corpus

	Distribution	χ^2	C
Bibliographical corpus	GIGP	108.733	0.000029
Reference corpus	GIGP	5007.673	0.000021
Bibliographical corpus	ZM	1402.910	0.000369
Reference corpus	ZM	3514.091	0.000015

The values of coefficient C are consistent with results presented in figures 5-6. The fit for the bibliographical corpus is better than for the general language data, and the value of C suggests that the distribution from the classes usually considered in the literature gives a reasonable fit. Of course, it seems to be interesting to examine the distributions from the broader classes of possible distributions and check the obtained fit using characteristics other than the discrepancy coefficient. However, this would require using the raw data instead of frequency spectrum data on which the analysis presented is based.

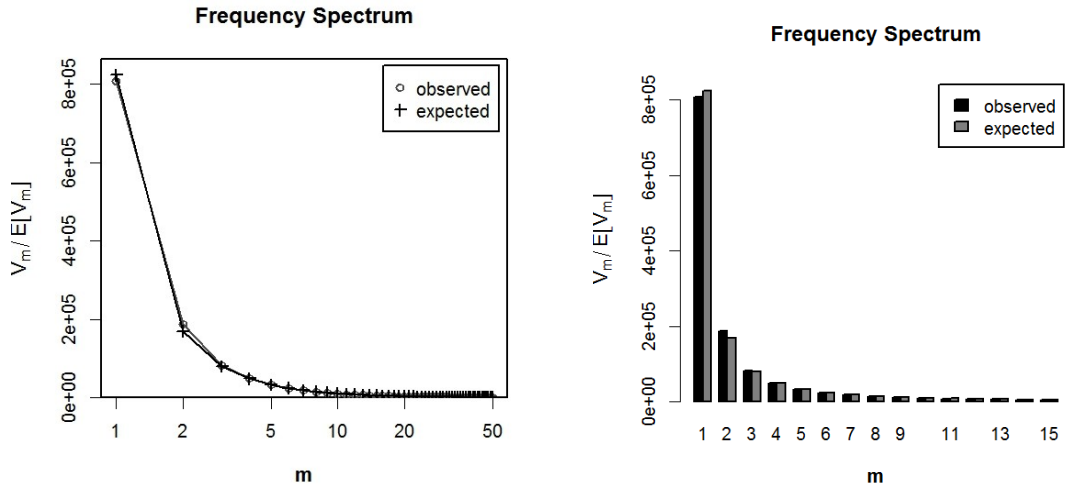

Figure 5. Left panel: the frequency spectrum for the general language data (circles), the Zipf-Mandelbrot fit (solid line and crosses). Right panel: bar plot for first 15 spectrum elements

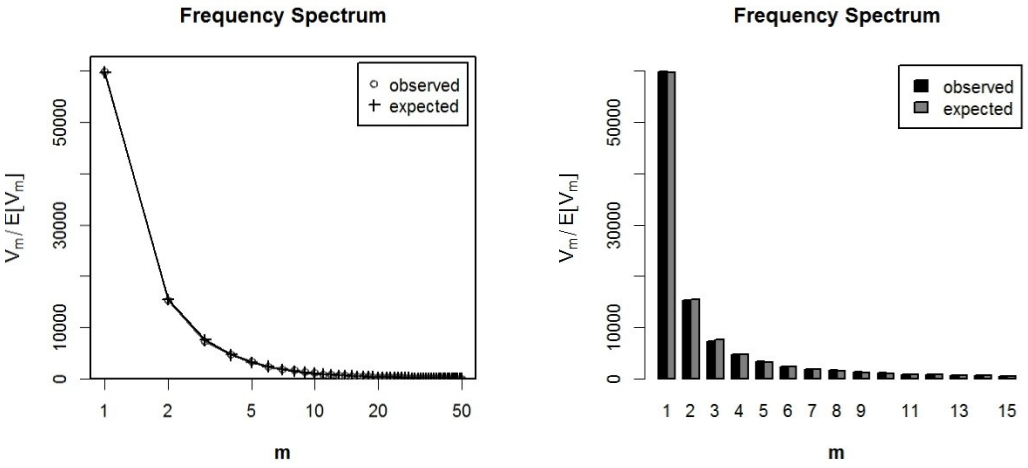

Figure 6. Left panel: the frequency spectrum for the bibliographical corpus (circles), the generalized-Gauss-Poisson fit (solid line and crosses). Right panel: bar plot for first 15 spectrum elements

The above conclusion, which actually confirms the hypothesis put forward at the beginning, is supported by the histogram of the frequency of words in both corpora (Figure 7), which shows significant differences.

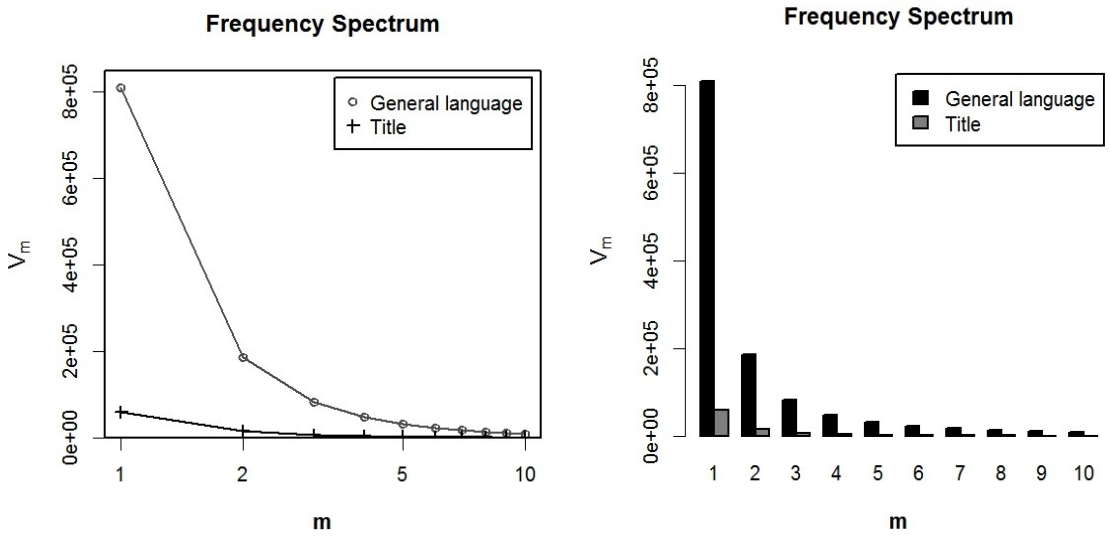

Figure 7. Left panel: the frequency spectrum for the general language data (circles) and the title data (crosses). Right panel: bar plot for first 15 spectrum elements

6. Conclusions

The aim of this study was to present the quantitative characteristics of the bibliographic corpus and to compare it with the reference corpus of the general language. The tests carried out proved that the "language of titles", i.e., a large corpus composed of book titles extracted from the Polish National Library databases, differs in many respects from the language used in normal communication. It was shown that basic quantitative parameters of both corpora are different (Table 1), and that there are striking differences in the distribution of frequency of parts of speech (Table 2). The same tendency was observed when histograms of title and sentence lengths frequencies were compared (Figure 1 and 2): both corpora turned out to have different quantitative characteristics. We also analyzed the relationship between the average word length in titles and the length of a title (Figure 3 and 4). Both corpora were different and no law-like tendency (e.g. resembling Menzerath-Altmann's law) was observed.

Interesting results were obtained while estimating word spectra distribution of both corpora. It turned out that the generalized Zipf-Mandelbrot distribution and the Generalized Inverse Gauss-Poisson distribution give the best fitting levels. The evaluation of the best fit of both distributions proved to be a problem, as the most common chi-square goodness-of-fit test is designed so that with large corpora the result is always negative. Therefore, in order to evaluate the goodness of fit, resistance and discrepancy tests were used (Table 4). As shown in Figure 7, statistical distributions of word spectra in the bibliographical corpus (titles) and in the general language do not follow the same pattern.

The final conclusion that emerges from the analysis of the results confirms the general principle of statistical linguistics which states that the texts created in natural, spontaneous communication differ significantly from the texts prepared by the researchers as artificial collections, according to arbitrary criteria, such as, for example, the identical context of use, function, or genre of the text. One of the differences between natural texts and "artificial" text corpora is that the latter do not necessarily follow the statistical laws of language, determined in the context of natural communication. These laws are not just mathematical formulas that the researcher "matches to a line", that is, to an empirical histogram of some relationship of observed variables. Rather, they reflect, using the formalized language of mathematics, the possibilities and limitations of the human brain, which in a peculiar but effective way optimizes processing of perceived stimuli, aiming at the best adaptation of the human being to its information environment[8].

A general remark that can be indirectly inferred from the analyses carried out indicates that bibliographies are these days becoming the subject of linguistic and cultural studies, where they are treated as fully fledged text corpora containing valuable, reliable, and easily quantifiable data on culture, society, science, technology, etc. Admittedly, bibliographic corpora are also an emerging object of analysis in quantitative linguistics. Although it is difficult to say whether these new types of data will become a favourite topic of lexicostatistical research, they will certainly be the preferred target of text-mining algorithms – in this case, quantitative research will play an important, albeit auxiliary, role.

References

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Springer-Science+Business Media. doi: 10.1007/978-94-010-0844-0

CHBB.1999-2019. *The Cambridge history of the book in Britain*. Vol. 1-7, Cambridge: Cambridge University Press.

[8] One of the basic rules of human information processing is the principle of least effort. To a large extent, it shapes the form of language (length of units, speech sound sequences, etc.), but the question of whether it works similarly in other areas of language, for example, in semantics or axiology of the world image, remains open.

Cressie, Noel & Timothy R. C. Read. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 46.3. 440–464, https://www.jstor.org/stable/2345686

Evert, Stefan & Marco Baroni. 2007. zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, Prague, Czech Republic, June 25-27, http://www.stefan-evert.de/PUB/EvertBaroni2007.pdf

Febvre, Lucien & Henri-Jean Martin.1958. *L'apparition du livre*, avec le concours de Anne Basanoff, Henri Bernard-Maitre, Moché Catane et al., Paris: A. Michel.

Green, Jonathan, Frank McIntyre & Paul Needham. 2011. The Shape of Incunable Survival and Statistical Estimation of Lost Editions. *The Papers of the Bibliographical Society of America* 105.2. 141-175. doi: 10.1086/680773

Grotjahn, Rüdiger & Gabriel Altmann.1993. Modelling the distribution of word length: some methodological problems. In Reinhard Köhler, Burghard B. Rieger (eds.), *Contributions to quantitative linguistics*, 141-153. Dordrecht: Kluwer. doi: 10.1007/978-94-011-1769-2_9

Lahti, Leo, Jani Marjanen, Hege Roivainen & Mikko Tolonen. 2019. Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloguing & Classification Quarterly* 57.1. 5-23. doi: 10.1080/01639374.2018.1543747

Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20.3. 227-240. doi: 10.1080/09296174.2013.799912

Mandelbrot, Benoît. 1962. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jacobson (ed.), *Structure of Language and its Mathematical Aspects* (Proceedings of Symposia in Applied Mathematics 12), 190–219. Providence, RI: AMS.

Schwetschke, Gustav. 1850. *Codex nundinarius Germaniae literatae bisecularis. Teil: 1564 - 1765*. Halle: G. Schwetschke's Verlags-Handlung und Buchdruckerei, https://reader.digitale-sammlungen.de//resolve/display/bsb11199701.html

Schwetschke, Gustav. 1877. *Codex nundinarius Germaniae literatae bisecularis. Teil: Forts. 1766 bis 1846*, Halle: G. Schwetschke's Verlags-Handlung und Buchdruckerei, https://digital.slub-dresden.de/werkansicht/dlf/102071/1/0/

Sichel, Herbert S. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association* 70. 542–547. doi: 10.2307/2285930
1982. Asymptotic efficiency of the three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika* 69. 467–472. doi: 10.2307/2335423

Tolonen, Mikko, Jani Marjanen, Hege Roivainen & Leo Lahti. 2019a. Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52.1. 57-78. doi: 10.1080/01615440.2018.1526657

Tolonen, Mikko, Jani Marjanen, Hege Roivainen & Leo Lahti. 2019b. Scaling Up Bibliographic Data Science, *DHN*. 450-456, https://cst.dk/DHN2019Pro/papers/40_2019DHNBDS.pdf

Text corpora and software

BN Data: http://data.bn.org.pl/
CLARIN-PL infrastructure: http://clarin-pl.eu
NKJP: http://www.nkjp.uni.lodz.pl/
WCRFT2 morphosyntactic tagger: http://ws.clarin-pl.eu/tager.shtml
ZipfR package: http://zipfr.r-forge.r-project.org/