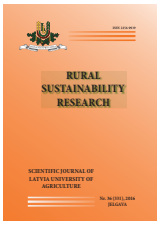
DE GRUYTER OPEN

RURAL SUSTAINABILITY RESEARCH 36(331), 2016
ISSN – 2256-0939

http://www.llu.lv/en/



Application of Statistical Methods in the Analysis of Sentence Structure

*Karolina Piaseckiene
Šiauliai University, Faculty of Technology, Physical and Biomedical Sciences,
Vilniaus st. 141, Šiauliai, LT-76353, Lithuania

Abstract. The goal of this research is to explore sentence structures expressed by parts of speech. Due to a small amount of data, a problem of *sparse* data has arisen, which was solved by recording the annotated sentences and considering a "framework" of a sentence made up from a verb and a noun, which was conditionally called a code. The code of a sentence is created by changing each word of a sentence by a symbol (letter or number) that *encodes* one or other property of that word as a constituent of the sentence. Zipf's law describes sentences, encoded like that, rather well. If we 'learn' well to identify and analyze (annotate, translate, etc.) sentences of the simplest structure, we can automatically process quite a large part of text sentences. It is possible to identify at least 17% of sentences consisting of the simplest structure.
Key words: sentence structure, Zipf's law, coding.

Introduction

In recent years, the processes of language computerization have been rapidly developing all over the world including Lithuania as well. The methods used in foreign countries are not always applicable in the Lithuanian language due to its specificity. The Lithuanian language is a complex inflected language distinguishing itself by a variety of grammar forms, morphological ambiguity, grand inflexion, free word order in a sentence, and so on; therefore, it cannot directly use the software, already created in other countries, e.g. for automatic analysis of syntax, which causes much trouble in developing efficient algorithms for automatic processing of Lithuanian texts.

With the development of structural linguistics, language modeling questions are of particular importance as well. Two kinds of models are used in linguistics: nonstatistical (compiled on the basis of mathematical logics, graph theory) and statistical (created applying the methods of probability theory, information theory and mathematical statistics) (Mauzienė, 2009).

Statistical methods are frequently used in the quantitative linguistic analysis in foreign countries (Abney, 1996; Baayen, 2001; Smith, 2011).

The field of linguistics, based on empirical and statistical methods, is usually called quantitative linguistics. One can single out three methodologies of quantitative linguistics: probabilistic models, statistical linguistics, and computational linguistics. The statistical linguistics is rarely used, the other two methodologies are prevailing.

In Lithuania, the computational linguistics, applied in natural language processing, text mining, information retrieval, is based on *n*-grams (usually trigrams) and hidden Markov models (Rimkutė & Daudaravičius, 2007; Vaičiūnas, 2006) and concrete practical problems are solved: automatic textual annotation, language recognition, correction of mistakes, translation, etc. Thus, statistical methods are rarely applied in scientific research of the language itself in Lithuania. Therefore, the research aims at broader application of statistical methods and performance of a different than regular analysis of sentences of the Lithuanian language.

The goal of this research is to explore sentence structures expressed by parts of speech.

Materials and Methods

Zipf's law

The American linguist George Kingsley Zipf was the first who described the characteristics of frequent

* Corresponding Author's email:
k.piaseckiene@gmail.com

DOI:10.1515/plua-2016-0008

words and their forms (Zipf, 1935). Well-known Zipf's law is an empirical law formulated using mathematical statistics:

$$f_r = \frac{K}{r^{\gamma}}; \quad (1)$$

here $r = 1, 2, \ldots, N$ are the ranks of words arranged in decreasing order of their observed frequencies, f_r is the frequency of words with the rank r, γ is the index of word diversity, $\gamma > 0$, and K is a normalizing constant.

(1) mathematical dependence is a separate case of Zipf-Mandelbrot law.

$$f_r = \frac{K}{(r+B)^{\gamma}};$$

here B is a parameter indicating a deviance from Zipf's law, $B \geq 0$. When $B = 0$, we obtain a classical Zipf's law (Baayen, 1993).

We can find analogies of the law (1) with the second Zipf's law:

$$\log\left(\hat{V}_m\right) \approx \log\left(\hat{V}_+\right) - \log\frac{\Gamma(1-\tau)}{\tau} - (1+\tau)\log(m) \quad (2)$$

which relates the amount of word forms $\hat{V}_m$ found in a text for m times to m because in this expression ranges of word forms r used in (1) have been changed by the frequency m observed for these forms (Kornai, 2002). For the derivation of the formula (2) and more extensive explanations see Piaseckienė & Radavičius, 2014; Piaseckienė, 2014.

Zipf's law is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency).

Data

Texts that compose the population under consideration are prose books for children (the volume of which is no less than 44 pages) of Lithuanian writers, published in the period 1995–2011. They are meant for children and stored at the library of the Šiauliai university. There are 36 authors in total from the books of each of which an approximate simple random sample without replacement of 20 sentences has been selected. Thus, the sample consists of 720 sentences that were annotated morphologically in a manual way, i.e. the part of speech of each word with the respective properties is pointed out.

Results and Discussion

Due to a small amount of data, a problem of *sparse* data has arisen, which was solved by recording the annotated sentences and considering a "framework" of a sentence made up from a verb and a noun, which was conditionally called a code. The code describes a reduced (simplified) sentence structure. The level of reduction (simplification) depends on what properties and how detailed they are encoded.

The code of a sentence is created by changing each word of a sentence by a symbol (letter or number) that *encodes* one or other property of that word as a constituent of the sentence.

Thus, a sentence becomes as if 'a word' whose 'alphabet' consists of symbols encoding the properties analyzed.

The codes of sentence structures of the following types have been constructed:

I – by keeping the order of the annotated sentence, only nouns *D* and verbs *V* are left, and all the other parts of speech are replaced by a symbol "–", several successive symbols "–" following successively are joined;

Ia – obtained from the code of type I, by joining several successive nouns or verbs;

II – formed just like type I, saving only the information on the case of a noun, i.e. instead of a noun, the case number is written (nominative – 1, genitive – 2, etc.);

IIa – derived from the code of type II, by joining several successive equal symbols.

In Table 1, examples of coding are presented. Here *D* is a noun, *V* is a verb, and *n* is another part of speech.

In Table 2, counts of codes with various observed frequencies, i.e. *frequencies of frequencies* of codes, are presented. Hence we see that 257 sentence codes of type I occur only once, and of type II, regarding

Table 1

Examples of codes of sentence structures

	Without cases	I	Ia	With cases	II	IIa
1	*nVDnD*	*-VD-D*	*-VD-D*	*nV*5*n*5	*-V*5-5	*-V*5-5
2	*DDVnnnnD*	*DDV-D*	*DV-D*	21*Vnnnn*1	21*V*-1	21*V*-1
3	*nnDDnnDVVnDnn*	*-DD-DVV-D-*	*-D-DV-D-*	*nn*11*nn*1*VVn*2*nn*	-11-1*VV*-2-	-1-1*V*-2-

Table 2

Counts of structure codes of sentences with various observed frequencies

Observed frequencies of codes	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Counts for codes of type I	257	31	9	11	3	6	4	1		1	1	2	1	1
Counts for codes of type Ia	120	36	8	6	5	3	3	2	2	1	3	1		1
Counts for codes of type II	407	30	9	6	2	2	5		2		1			
Counts for codes of type IIa	355	37	12	4	4	3	3		4	1	1			
Observed frequencies of codes	15	16	17	18	19	21	23	26	27	30	32	34	38	40
Counts for codes of type I		1		2			1			1	1	1		
Counts for codes of type Ia	1		3		1	1	2	1	1	1			1	1
Counts for codes of type II		1		2						1		1		
Counts for codes of type IIa			1		2					1			1	

the case of noun, even 407 structures are found once (more than a half of all the sentences). In all cases, there are structures met by 30 or even more times.

If Zipf-Mandelbrot law is valid for the "words" formed by the earlier described method, it means that a big part of sentences in a text has almost a standard structure (from the point of view under discussion, defining the coding law used); on the other hand, many sentence structures are (almost) unique, used in texts only once (1 or 2 times).

The parameters of Zipf's law (see Table 3) are calculated for sentence structure codes according to the formula (1). It is simpler to interpret the law, expressed by (1) formula, in the log-log scale. Then we can apply the method of the least squares to assess the parameters.

Table 3

Parameters of Zipf's law

Code	$\log K$	γ
I	1.795	–1.405
Ia	1.585	–1.166
II	1.845	–1.457
IIa	1.863	–1.453

In Figure, the graphs of sentence structures of types I, Ia, II and IIa are presented, where x is the logarithm of code frequency ($x = \lg r$) and y is the logarithm of code reiteration frequency ($y = \lg f_r$).

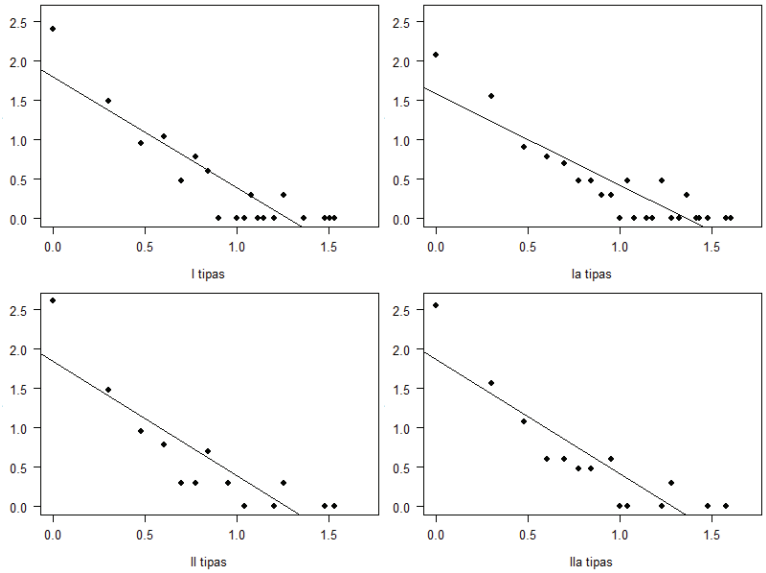


Figure. Log-log graphs of sentence structure code frequency.

Note that the fitted lines describe the data of pairs (r, f_r), in the log-log scale rather well. Here r is a range of a "word" form, and f_r is the amount of "word" forms whose range is r. In all cases, with the exception of the simplest (most reducing) coding Ia, Zipf straight line's parameters are very similar. So, Zipf's law describes the sentences, encoded like that, rather well.

If we treat structures that occurred, say, no less than 10 times, as simple structures, we can identify 17.64% of sentences (according to the data displayed in Table 2) by code II (approximate 95% confidence interval of the proportion is from 14.86% to 20.42%), and even 33.75% of sentences by code I (approximate 95% confidence interval of the proportion is from 30.3% to 37.2%).

Having perfected the coding, probably, it would be possible to achieve even better results.

Conclusions

1. If we 'learn' well to identify and analyze (annotate, translate, etc.) sentences of the simplest structure, we can automatically process quite a large part of text sentences.
2. It is possible to identify at least 17% of sentences consisting of the simplest structure.
3. The sentences encoded by the method discussed above are described by Zipf law quite well.

References

1. Abney, S. (1996). Statistical Methods and Linguistics. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. 1–26.
2. Baayen, R.H. (1993). Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. *Computers and the Humanities*. 26, 347–363.
3. Baayen, R.H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers.
4. Kornai, A. (2002). How many words are there? *Glottometrics*. 4, 61–86.
5. Mauzienė, L. (2009). Lingvistiniai ir psichologiniai lingvodidaktikos pagrindai (teorinė interpretacija) (Linguistic and Psychological Basics of Linguistic Didactics (Theoretical Interpretation)). *Santalka. Filologija. Edukologija*. 17(2), 61–67. (in Lithuanian)
6. Piaseckienė, K., & Radavičius, M. (2014). Empirical Bayes estimators of structural distribution of words in Lithuanian texts. *Nonlinear Analysis: Modelling and Control*. 19(4), 611–625.
7. Piaseckienė, K. (2014). *Statistiniai metodai lietuvių kalbos sudėtingumo analizėje* (*The statistical methods in the analysis of the Lithuanian language complexity*). Doctoral dissertation,Vilnius, 63–66. (in Lithuanian)
8. Rimkutė, E., & Daudaravičius V. (2007). Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas (Morphological Annotation of the Contemporary Lithuanian Language Corpus). *Kalbų studijos*. 11, 30–35. (in Lithuanian)
9. Smith, N.A. (2011). Linguistic Structure Prediction. *Synthesis Lectures on Human Language Technologies*. Morgan&Claypool Publishers.
10. Vaičiūnas, A. (2006). *Lietuvių kalbos statistinių modelių ir jų taikymo* šnekos *atpažinimui tyrimas, kai naudojami labai dideli* žodynai (*Research on the Lithuanian Language Statistical Models and Their Application in Speech Recognition when High-Volume Vocabularies Are Used*). Doctoral dissertation, Kaunas. (in Lithuanian).
11. Zipf, G.K. (1935). *The Psycho-Biology of Language*. New York: Houghton Mifflin.

Acknowledgements

This research was done in the framework of the Nordplus Higher Education 2016 project NPHE-2016/10342 "Raising awareness about the role of math skills in building specialists competence for the sustainable development of society".