VILNIUS UNIVERSITY

Karolina Piaseckienė

THE STATISTICAL METHODS IN THE ANALYSIS OF THE LITHUANIAN LANGUAGE COMPLEXITY

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01 P)

Vilnius, 2014

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2008–2013.

Scientific Supervisor

Prof. Dr. Marijus Radavičius (Vilnius University, Physical Sciences, Mathematics – 01 P).

The thesis is defended in the Council of Mathematics at the Vilnius University:

Chairman

Prof. Dr. Habil. Kęstutis Kubilius (Vilnius University, Physical Sciences, Mathematics – 01 P).

Members:

Prof. Dr. Habil. Vilijandas Bagdonavičius (Vilnius University, Physical Sciences, Mathematics – 01 P),

Prof. Dr. Habil. Algimantas Jonas Bikelis (Vytautas Magnus University, Physical Sciences, Mathematics – 01 P),

Prof. Dr. Habil. Jonas Kazys Sunklodas (Vilnius University, Physical Sciences, Mathematics – 01 P),

Prof. Dr. Darius Šiaučiūnas (Šiauliai University, Physical Sciences, Mathematics – 01 P).

Opponents:

Prof. Dr. Kęstutis Dučinskas (Klaipėda University, Physical Sciences, Mathematics – 01 P),

Assoc. Prof. Dr. Rimantas Eidukevičius (Vilnius University, Physical Sciences, Mathematics – 01 P).

The dissertation will be defended at the public meeting of the Council of Mathematics in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p. m. on the 17th of September, 2014.

Address: Akademijos st. 4, LT-08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed on the 18th of August 2014.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University.

VILNIAUS UNIVERSITETAS

Karolina Piaseckienė

STATISTINIAI METODAI LIETUVIŲ KALBOS SUDĖTINGUMO ANALIZĖJE

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01 P)

Vilnius, 2014

Disertacija parengta 2008–2013 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas

prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P).

Disertacija ginama Vilniaus universiteto Matematikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Kęstutis Kubilius (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P).

Nariai:

prof. habil. dr. Vilijandas Bagdonavičius (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P),

prof. habil. dr. Algimantas Jonas Bikelis (Vytauto Didžiojo universitetas, fiziniai mokslai, matematika – 01 P),

prof. habil. dr. Jonas Kazys Sunklodas (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P),

prof. dr. Darius Šiaučiūnas (Šiaulių universitetas, fiziniai mokslai, matematika – 01 P).

Oponentai:

prof. dr. Kęstutis Dučinskas (Klaipėdos universitetas, fiziniai mokslai, matematika – 01 P),

doc. dr. Rimantas Eidukevičius (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2014 m. rugsėjo 17 d. 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2014 m. rugpjūčio 18 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

Introduction

Research area and topicality of the work

In recent years, the processes of language computerization have been rapidly developing all over the world and 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 use directly 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).

Statistical methods are frequently used in the quantitative linguistic analysis (see [1], [4], [18]). The field of linguistics, based on empirical and statistical methods, usually is called a quantitative linguistics. One can single out three methodologies of quantitative linguistics: probabilistic models, statistical linguistics, and computational linguistics. The statistical linguistics is used rarely, the other two methodologies are prevailing. The computational linguistics, applied in natural language processing, text mining, information retrieval, is based on n-grams (usually trigrams) and hidden Markov models 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.

Another problem we face in linguistics is how to define what the research population is. One of the solutions proposed in corpus linguistics is creation of artificial populations – corpora. The basic presumption is corpus randomness which is treated as a *equiprobability* of linguistics objects in the corpus. However, in a large corpus, the data are very heterogeneous, because the corpus consists of different genre texts, meant for different purposes, and different readers. In order to single out the peculiarities, typical of the language itself, from the authors' preferences, an accurate definition of the (finite) population under research is of great importance, for the most part it predetermines the results of statistical analysis.

In researches of the Lithuanian language, a descriptive statistic is used rather often. Much attention is paid to solution of practical problems: a corpus has been compiled (see *http://tekstynas.vdu.lt/tekstynas*), and an automatic programme of morphological analysis *Lemuoklis*, meant for the corpus annotation (see [16]), much attention is also paid to limit morphological ambiguity to automatic syntactic analysis (see [19], [20]), machine translation (translation tools are created, e.g. *http://www.tilde.lt*), to automatic recognition of a language / speech (see [23]) as well as to the investigation of the frequency of phonemes, letters, words and their forms.

R. Merkytė was engaged in the application of mathematical statistics in linguistics ([12], [13]). In the analysis of the Lithuanian language she applied the model, proposed

by W. Fuchs, meant for modeling the number of syllables in a word and developed a method of moments to evaluate its unknown parameters. She has also estimate the literal entropy and information content in Lithuanian texts.

The object of research

The object of research in the dissertation is structural (graphical) models of interdependence of categorical variables and various structures of the Lithuanian language.

The mathematical object of research is the Zipf-Mandelbrot law and structural distribution, related with it, as well as tables of observed frequencies of linguistic objects and models of their distributions.

Linguistic objects of research are sounds, letters, words, meta-comments, sentences (from the aspect of morphology and syntax) and their interrelations; functional styles.

Real data were investigated from the "Corpus of the contemporary Lithuanian language", children's literature present in the library in Šiauliai University, and from a public digital library. In the process of research the objects – corpora – were varying.

The aim and tasks of the work

The target of the work is to apply mathematical and statistical methods in the analysis of the Lithuanian language by identifying and taking into account peculiarities of the Lithuanian language, its heterogeneity, complexity and variability.

To this end, the following problems were solved:

- To conduct a survey of statistical applications in the investigation of the Lithuanian language.
- To explore the properties of the Lithuanian language and their complexity by applying the (graphical) loglinear analysis and other statistical and graphical methods.
- To describe and evaluate heterogeneity and variability of the language stipulated by its author's selection. The aim is to identify properties, connections, and structures relatively little dependent on the author, consequently, potentially typical of the language itself. To prepare an appropriate methodology and apply it to concrete investigations of the Lithuanian language.
- On the basis of the empirical Bayes method, to construct structural distribution estimators of word types in the text that would take into consideration textual non-homogeneity and influence of unseen word types.

Methods of the research

The following research methods are applied in the work: analysis of scientific literature on the topic of the dissertation; survey sampling; elements of the graph theory; mathematical modeling and statistical data analysis, using loglinear and graphical loglinear models as well as logistic regression and zero-inflated negative binomial regression;

Bayes methodology and the empirical Bayes method; asymptotic methods; fundamentals of linguistic science. Statistical software used are SPSS, SAS, and R.

Novelty and practical value of the work

The results of the work done supplement and extend the results of other researches performed in this and related areas.

As usual, in linguistic investigations, based on corpora, the initial research element is a word, a combination of words, and sometimes a sentence. In this work, a primary research element is the author, his choice is the source of heterogeneity and variability. This standpoint, together with survey sampling and statistical analysis methods, makes the basis for a new methodology proposed, that allows us to identify the properties and structures, but little dependent on the author, thus potentially typical of the language. This methodology is applied to two concrete problems of the Lithuanian language analysis.

In solving automatic language processing problems (e.g. automatic syntactic annotation or translation), it is of great importance to evaluate the complexity of the linguistic structures under consideration, because it can predetermine not only selection of the methods used, but also the principles for creating the basic language model and the analysis methodology. In this work, the initial statistical analysis of complexity of the (graphical) syntactic structure of a sentence has been made. Although simple complexity characteristics of the texts in the Lithuanian language were obviously calculated more than once, the works aimed at a more consistent analysis of complexity are unknown to the author.

On the base of the zero-inflated negative binomial regression model and the empirical Bayes method, the structural distribution estimator of word types in the text has been constructed. It employs available auxiliary information on authors of the text and word types, thus allowing us to take into account the non-homogeneity of texts and the effect of unseen word types. The structural distribution is a much more subtle tool of language research than the methods based on parametric models of the Zipf-Mandelbrot type.

This work shows the importance and possibilities of application of the sampling methods and points out to text authors as a basic elements of textual corpora in statistical analysis.

Statements presented for defence

1. Statistical methods are widely applied in the analysis of the Lithuanian language. However, recently the descriptive statistics and procedures, created by computer scientists, have been prevailing, applied more to the English language and oriented to solution of concrete practical problems.

2. In general, Herdan and Zipf laws approximate the amount and distribution of word types in the text of the Lithuanian language rather precisely. However, the values of parameters of those laws described are significantly different between the authors, even more different between the Lithuanian and foreign authors.

3. In the Lithuanian language, directly related words can be considerably distant one from another, therefore models and automatic rules, created on the basis of trigram statistics, have rather limited abilities to properly model and predict the structures of the Lithuanian language.

4. In order that more complex statistical language studies could be performed and interpreted correctly, constantly maintained corpora should render conditions to apply survey sampling methods to data compiling and to select the texts according to various features as well as according to different authors. The corpora that do not allow the control of sampling rules have very limited abilities for statistical analysis.

5. Properly collected data, applying the survey sampling methods, (graphical) loglinear models, as well as the empirical Bayes method enable us to make use of auxiliary information and to model complex language structures, their heterogeneity and individual variability thus forming the basis to define and investigate the properties and relationship typical of the language itself.

Approbation of the work results

The results of the dissertation are published in 3 papers in reviewed scientific periodicals as well as presented at 8 scientific conferences two of which are international.

Structure of the dissertation

The dissertation consists of introduction, three chapters, conclusions, the list of references, the list of the author's publications on the dissertation topic, and one appendix. In the first chapter, a short survey of application of statistical methods in linguistics (in Lithuanian and abroad) is presented. The second chapter contains theoretical material: the basic conceptions, models, and methods, applied in the dissertation, are described. In the third chapter, analysis of application of statistical methods in the Lithuanian language is made.

The general volume of the dissertation is 124 pages, in which there are 45 numbered formulas, 17 figures, and 21 table. The dissertation is based on 99 sources of literature.

Loglinear analysis of Lithuanian texts

In order to take into consideration and estimate language heterogeneity and variability which stem from authors' preferences and choices, the following methodology, based on survey sampling methods and (graphical) loglinear analysis, is proposed.

1. Suppose that a corpus under consideration has tools for text selection by its author. Then a simple random sample without replacement of authors is taken.

2. From each author in the sample, a simple random sample (with replacement) of the linguistic elements or structures (units) of study is drawn. The samples of authors are mutually independent and of equal sizes.

3. Contingency tables of features of the study units are calculated for the sampled dataset of each author represented by the variable (identifier) "author". Loglinear model is fitted to the contingency tables. Initially, all the effects included in the model also have their counterparts containing an interaction with the variable "author".

4. We try to eliminate the effects that have an interaction with the variable "author" or to replace the variable "author" in these effects by a variable representing some general characteristic of the authors. If reduction of this type does not violate the adequacy of the model, the effects in the model without interaction with the variable "author" represent the relations between the linguistic units and their properties that relatively slightly depend on individual choices of the authors and hence express universal relationships and potentially inherent features of the language itself.

The same methodology can be applied in the analysis and generalization not only of the feature "author" but also of other relevant features of texts and linguistic units. We *partly* apply it in the study of meta-comments, of their usage and functions.

Elements of loglinear analysis

Since a loglinear analysis is not widely used in linguistic research, we outline some basic concepts.

Suppose that data of m categorical variables is summarized in the contingency table $\mathcal{T}$. The variables are identified by their numbers, the ith variable $a_i \in A_i := \{1, \dots, k_i\}$, and usually its last category k_i is taken as a reference value ($i \in [m] := \{1, \dots, m\}$). Let $\underline{y} := (y_a, a \in \mathcal{A})$ be a collection of observed frequencies in $\mathcal{T}$, $\mathcal{A} := A_1 \times \dots \times A_m$. Then the expected frequencies $\underline{\mu} := \mathbf{E}\underline{y}$ in a *saturated loglinear model* have the following representation:

$$\ln(\mu_a) = \sum_{J \subset [m]} u^J(a), \quad a \in \mathcal{A}. \tag{1}$$

Here the sum is taken over all subsets J of $[m]$, the functions $u^J(a)$ depend only on components $\{a_j, j \in J\}$ of a and vanish if any of these components takes the reference value, $u^\circ \equiv const$. Thus, the term $u^J(a)$ in the sum has $d = d(J) := \prod_{j \in J}(k_j - 1)$ degrees of freedom, i.e. the number of scalar parameters that identify the function $u^J, J \subset [m]$.

In the loglinear analysis the terms u^J are called *effects* of order $|J| := \mathrm{card}(J)$. Usually high order effects, say of order $|J| > 3$, are statistically insignificant and can be dropped out from the model. Denote by $\mathcal{S}$ a collection of subsets J of $[m]$ corresponding to the statistically significant effects u^J. That yields a loglinear submodel of (1)

$$\ln(\mu_a) = \sum_{J \in \mathcal{S}} u^J(a), \quad a \in \mathcal{A}.$$

The loglinear model with many high order effects is difficult to interpret. Loglinear models, satisfying certain conditions, have a unique graphical representation and are called graphical loglinear models. Each categorical variable in the model is represented

by a vertex of the graph and two variables, say, A and B, are (statistically) dependent if and only if there exists a path in the graph that joins the vertices corresponding to A and B. Thus, the graphical loglinear models have a convenient and clear interpretation in terms of independence and conditional independence.

Example: three-way contingency table. Consider a saturated loglinear model of categorical variables $A \in \{1, \ldots, k_A\}$, $B \in \{1, \ldots, k_B\}$ and $C \in \{1, \ldots, k_C\}$

$$\ln \mu_{ijs} = u^{\circ} + u^{A}(i) + u^{B}(j) + u^{C}(s) + u^{AB}(i,j) + u^{AC}(i,s) + u^{BC}(j,s) + u^{ABC}(i,j,s).$$

The symbolic form of this model is $A+B+C+A*B+A*C+B*C+A*B*C$ with obvious notation. The symbolic form of the loglinear model with mutually independent variables is $A + B + C$. The model $A + B + C + A * B + B * C$ implies that the variables A and C are conditionally independent given values of variable B. The model $A + B + C + A * B + A * C + B * C$ is not graphical.

The standard reference on loglinear (graphical) models, their analysis and interpretation is [2].

Comparison of Lithuanian texts by literal and phonetic composition of words

The proposed methodology is applied to compare Lithuanian texts by the literal and phonetic composition of words, used in these texts. The aim is to study the impact of phonetic composition and length of words on their expected frequency and how this impact depends on the author (text document) and its style, scientific versus fiction.

The initial textual dataset consists of 17 text documents available in digital form and attributed to different authors. A simple random sample of size 3000 words is drawn from each text document and the length as well as 7 binary phonetic features of the words are evaluated, see Table 1. The following categorical variables are used in the analysis: the variable *st* indicates *text style* (scientific versus fiction), the variable *nr* refers to the number of the text document, the variable *i5* stands for 5 word length groups defined by partition into the 5 sets $\{1,2,3\}, \{4,5\}, \{6\}, \{7,8\}, \{9,\ldots\}$, $\{6\}$ being the reference group, the binary variable b takes values $\{0,1\}$, where 1 indicates that the word includes a vowel, similarly the binary variables *by, bn, bl, bp, prd, psb* indicate *long vowels, nasal vowels, front vowels, voiceless consonants*, and*semivowels*, respectively.

Table 1: Coding of vowels and consonats

Vowels, $b=1$					Consonants, $b=0$		
		$by=1$	$by=0$		$prd=1$	$prd=0$	
			$bn=1$	$bn=0$		$psb=1$	$psb=0$
$bl=1$, $bp=0$		ū	ų	o, u	c, č, f, k, p, s, š, t	j, l, m, n, r, v	b, d, g, h, z, ž
$bl=0$	$bp=1$	y	ę, į	e, ė, i			
	$bp=0$		ą	a			

Significant effects included in the fitted model and its goodness-of-fit statistics are presented in Table 2.

Table 2: Significances of effects included in the fitted model

Maximum Likelihood Analysis of Variance					
Source	**DF**	**Pr>ChiSq**	**Source**	**DF**	**Pr>ChiSq**
bl	1	0.0035	psb*nr	16	0.0128
bn	1	<.0001	i5*psb*nr	64	<.0001
bl*bn	1	0.0508	prd	1	<.0001
i5	4	<.0001	i5*prd	4	0.0003
i5*bn	4	<.0001	i5*prd*nr	64	<.0001
b	1	<.0001	i5*b*nr	64	<.0001
i5*b	4	0.0107	bl*nr	16	0.0002
bp	1	<.0001	i5*bl*nr	64	0.0005
by	1	<.0001	bp*nr	16	0.0328
i5*bl	4	0.0176	bp*bn*nr	16	<.0001
i5*by	4	<.0001	i5*bp*nr	64	<.0001
i5*bp*bn	4	0.0144	bl*bn*nr	16	<.0001
i5*bp*by	4	<.0001	st*i5*by	4	<.0001
nr	16	<.0001	by*nr	16	<.0001
i5*nr	64	<.0001	bp*by	1	0.0314
psb	1	<.0001	**Likelihood Ratio**	**361**	**0.0685**

The composition of the fitted loglinear model is as follows:

$$\begin{aligned}\ln \mu_j \;=\;& u^{\circ} + u^{nr}(j_2) + u^{i5}(j_3) + u^{b}(j_4) + u^{by}(j_5) + u^{bn}(j_6) + u^{bl}(j_7) + u^{bp}(j_8) + \\ &+ u^{prd}(j_9) + u^{psb}(j_{10}) + \\ &+ u^{nr,i5}(j_2, j_3) + u^{nr,by}(j_2, j_5) + u^{nr,bl}(j_2, j_7) + u^{nr,bp}(j_2, j_8) + u^{nr,psb}(j_2, j_{10}) + \\ &+ u^{i5,b}(j_3, j_4) + u^{i5,by}(j_3, j_5) + u^{i5,bn}(j_3, j_6) + u^{i5,bl}(j_3, j_7) + u^{i5,prd}(j_3, j_9) + \\ &+ u^{by,bp}(j_5, j_8) + u^{bn,bl}(j_6, j_7) + \\ &+ u^{st,i5,by}(j_1, j_3, j_5) + u^{nr,i5,b}(j_2, j_3, j_4) + u^{nr,i5,bl}(j_2, j_3, j_7) + u^{nr,i5,bp}(j_2, j_3, j_8) + \\ &+ u^{nr,i5,prd}(j_2, j_3, j_9) + u^{nr,i5,psb}(j_2, j_3, j_{10}) + u^{nr,bn,bl}(j_2, j_6, j_7) + \\ &+ u^{nr,bn,bp}(j_2, j_6, j_8) + u^{i5,by,bp}(j_3, j_5, j_8) + u^{i5,bn,bp}(j_3, j_6, j_8).\end{aligned}$$

Here $j = (j_1, \ldots, j_{10}) \in \mathcal{A}$, $\mathcal{A} := \{0,1\} \times [17] \times [5] \times \{0,1\} \times \ldots \times \{0,1\}$. The model has 542 unknown parameters which were estimated from the data. Links between the variables and their mathematical notation are given in Table 3.

Table 3: Links between the variables and their mathematical notation

Variables	st	nr	$i5$	b	by	bn	bl	bp	prd	psb
Number of categories	2	17	5	2	2	2	2	2	2	2
j	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9	j_{10}

The graphical extension of the model (the least graphical model containing the fitted model) is presented in Fig. 1.

13 out of 31 effects in the model contain the interaction with the variable *nr*. Replacement *nr* by *st* in any of these effects violates the goodness-of-fit of the model. It means that the effects represent author-specific or text-specific characteristics of the literal and phonetic composition of words. Nevertheless, some relationships that are relatively independent of the text document can be stated. The odds ratios of *bn* (the indicator of nasal vowels in a word) with respect to *i5* (the word length group) and

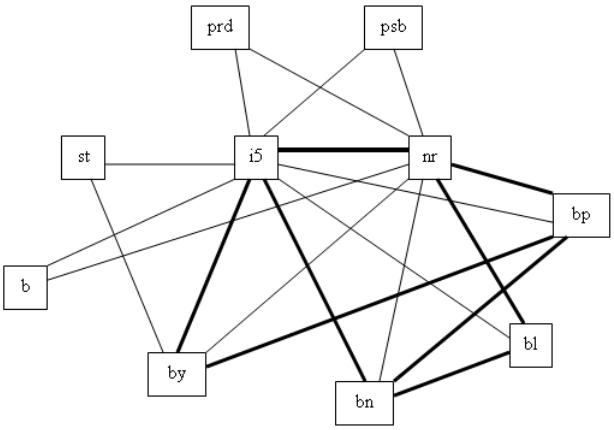

Figure 1: Relationship graph of literal and phonetic composition of words

by (the indicator of long vowels in a word) with respect to *bp* (the indicator of front vowels in a word) are independent of the variable *nr*. For each author and word length group, the variables *prd*, *psb*, and any vowel variable are conditionally independent (see graphical representation of the model Fig. 1). There is only one effect in the model, $st * i5 * by$, containing the interaction with variable *st*, thus representing relationship that could be valid for all text documents (authors) of the same style. Unfortunately, both *i5* and *by* are connected with *nr*.

These findings are, thus far, only preliminary and a further analysis of the phonetic composition of words is necessary taking into account their syllable structure as well.

Statistical analysis of meta-comments and their functions

Meta-comments (MCs) are one of the means to convey imagery of a language, in particular, the Lithuanian language[1]. We investigate effects that may influence the occurrence probabilities of MCs from a collection of simple MCs of the form *participle + adverb*. The adverb in MC is considered to be the basic word that determines the sense and function of MC. The dataset (5778 MC application cases) is taken from the corpus of the contemporary Lithuanian language[2], maintained by the Centre of Computational Linguistics of Vytautas Magnus University. The corpus has no option to select a sample by authors. Thus, we try to express the heterogeneity and variability of probabilities of MC occurrence in written texts in terms of the text style (4 styles are available: administrative, fiction, nonfiction, publicistic) and the basic word of MCs. The initial aggregated data set of occurrence frequencies of MCs in the corpus texts of each style and the corresponding frequencies of their basic words are supplemented by several explanatory variables, representing individual properties of the basic words.

It is assumed that a probability of a basic word, met in the corpus, to be a part of MC from the MC collection is constant in texts of the same style and the observed frequency of this random event has the binomial distribution. A binomial logistic regression model is fitted to the data.

[1]Meta-comment is a comment on the current sentence expressing an attitude of the author to its content (information). For example, "roughly speaking, it is ...".

[2]*http://tekstynas.vdu.lt/*, looked over on the 10th of June, 2012.

K. Piasekienė and M. Radavičius [14] have established that the probability of a basic word to be a part of MC statistically significantly depends on the basic word and the text style. Later the authors attempt to explain the observed MC differences between basic words and styles by their different demand for functions the MC are applied to. It is supposed that MC are used to express the following attitudes to information in a current sentence:

(f1) particularity and concreteness versus generality;
(f2) precision versus imaginativeness;
(f3) validity;
(f4) importance (due to personal values or emotions);
(f5) rotundity;
(f6) criticism, negative attitude.

The corresponding categorical explanatory variables take values from subsets of $\{-2, -1, 0, 1, 2\}$ (validity, importance, and criticism take only nonnegative values) with categories which evaluate direction and intensity of the attitude. The category 0 stands for the neutral attitude. Each MC may have several functions.

The fitted binomial logistic regression model shows that the functions f3, f4, f5 and the interaction f1*f2 of functions f1 and f2 can explain differences in probabilities of basic words to be a part of MC for a major part of the MC collection. However, this explanation fails in 5 (out of 47) rather frequently used MCs. The results obtained are preliminary. They are restricted by available options of the corpus of the contemporary Lithuanian language, which does not allow us] to control the sampling process and hence to apply the methods of survey sampling. Improvements of the list of MC functions as well as of the very methodology of MC analysis is desirable.

Application of statistical methods in the analysis of sentence structure and its complexity

Zipf's law for sentence structures

Assessment of the sentence or text complexity is a very relevant problem in linguistics. Various complexity indexes are proposed that reflect the percentage of more complex grammatical and syntactic combinations in the text, but frequently the simplest ones are used, namely, average length of a word or sentence, percentage of commas, etc.

The complexity characteristics of more interest are the average syntactic depth of a sentence and the average size of syntactic relations as well as the average number of intermediate words between a syntactically related pair [21]. As far as it is known, there was no quantitative analysis of the complexity of the Lithuanian language sentences, though separate characteristics, related with that, were undoubtedly calculated more than once.

The goal of this research is to explore sentence structures expressed by parts of speech, as well as the complexity of sentence structures, expressed by parts of sentence.

Texts that compose the population under consideration are prose books (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 have been selected. Thus, the sample ('corpus') consists of 720 sentences that were annotated morphologically and syntactically in a manual way, i.e. the part of speech of each word with the respective properties is pointed out as well as subjects and predicates (in simple sentences or in components of the simple sentence) with other parts of sentence subordinate to them.

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. 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 and verbs 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 4, examples of coding are presented. Here D is a noun, V is a verb, B is an adjective, I is a pronoun, and n is another part of speech.

Table 4: Examples of codes of sentence structures

	Without cases	I	Ia	With cases	II	IIa
1	nVDnD	-VD-D	-VD-D	nV5n5	-V5-5	-V5-5
2	DDVBnnBD	DDV-D	DV-D	21VBnnB1	21V-1	21V-1
3	IVVnV	-VV-V	-V-V	IVVnV	-VV-V	-V-V
4	DVD	DVD	DVD	1V4	1V4	1V4
5	IDDDnDVnDnD	-DDD-DV-D-D	-D-DV-D-D	I411n1Vn4n4	-411-1V-4-4	-41-1V-4-4

In table 5, 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 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.

The Zipf exponent γ in Zipf's law $f_r = Kr^{-\gamma}$ serves as an index of word diversity. Here $r = 1, 2, \ldots, R$, are the ranks of words arranged in decreasing order of their observed frequencies, f_r is the frequency of words with the rank r, K is a normalizing constant. It is simpler to interpret the law, expressed by this formula, in the log-log

Table 5: 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	

scale. For the codes analyzed, the estimators of the Zipf exponent γ, obtained by the least-squares method, are equal to -1.405, -1.166, -1.457 and -1.453, respectively.

In Fig. 2, the graphs of sentence structures of types I, Ia, II ir IIa are presented, where $x = \lg r$, $y = \lg f_r$.

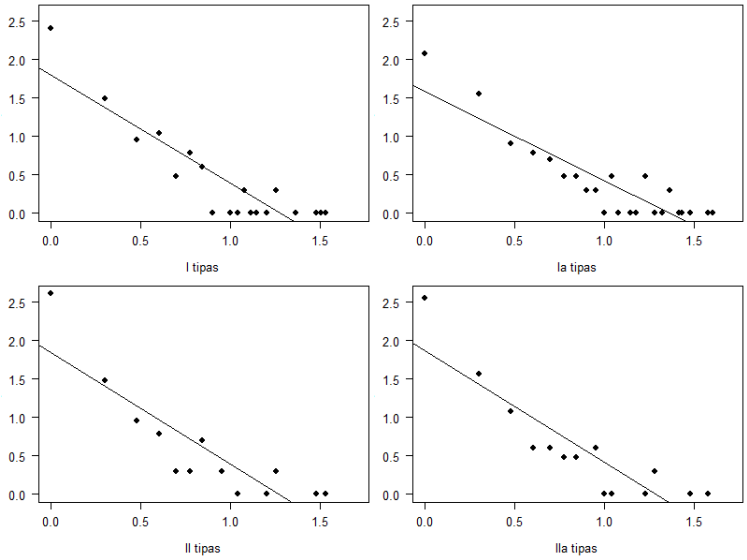

Figure 2: Log-log graphs of sentence structure code frequency

Note that the fitted lines describe the data of pairs (r, f_r), in the log-log scale rather well. Thus, 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. If we treat structures, that occurred, say, no less than 10 times, as simple structures, we can identify 17.64% of sentences by code II, and even 33.75% of sentences by code I.

Some statistics of syntactic complexity

The tree of subordination, used for the syntactic structure of other languages, cannot reflect all the syntactic relations, present in a Lithuanian sentence. As it is stated by D. Šveikauskienė [19, 20], only the graph of dependencies can correctly represent syntactic structures of the Lithuanian language.

On the basis of foreign experience, when modeling a language, in Lithuania as well, the Hidden Markov model of the 2nd order remains the most popular language model, based on trigrams, i.e. on the statistic of three successive words in the text (see [16], [23]). Due to the specificity and complexity of the Lithuanian language, algorithms based on trigrams seems to be not very promising in the research of the Lithuanian language.

The complexity of sentence structure can be conceived as *depth* and *width* of a sentence according to the words subordinate to the main parts of sentence (the subject and predicate). The *depth* of a sentence is the maximal length of a word sequence made up from words directly subordinate one to another in a sentence. The *width* of a sentence is comprehended as the number of words in an appropriate structure. For instance, if a word has no words subordinate to it, the width of such a group is assumed to be 1.

So, in order to find out whether the algorithms, based on trigrams, are applicable to the Lithuanian language, it is important to evaluate what proportion is of the sentences in which all the directly syntactically related words are no more distant one from another than over one intermediate word in Lithuanian texts (in this case, in the literature devoted to children).

In this work, only the sentences the structure of which can be represented by a graph-tree, i.e. 327 out of 720 sentence samples available are annotated simple sentences. The depth and width of these sentences has been estimated. The maximal depth of a sentence in the data under consideration was equal to 9. In Fig. 3, frequencies of the subject and predicate structure depth are illustrated. We see that the majority of predicate structures consist of 3–4 words, and the subject structures of 1–2 words. This fact shows that predicates are prone to attach longer sequences of words subordinate to them.

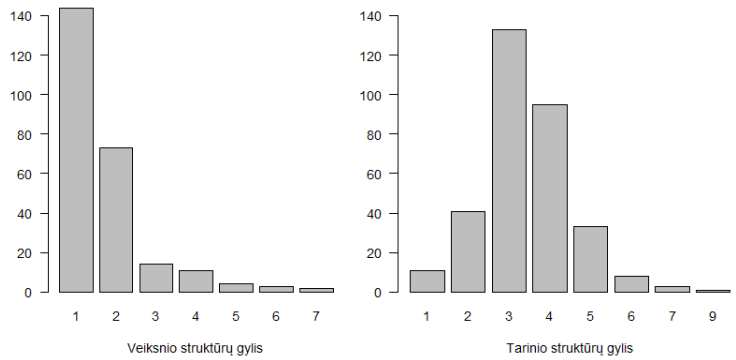

Figure 3: Depth of the subject and predicate structures

In Fig. 4 the width of the subject and predicate structures is presented. Note that

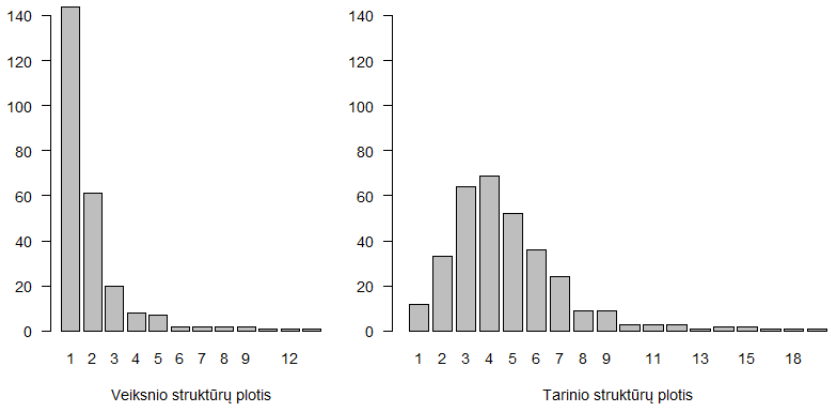

Figure 4: Width of the subject and predicate structures

the words belonging to the subject structure can be distant from the subject even over 14 words, though the majority of them is side-by-side with the subject (the width is equal to 1), while the words, belonging to the predicate structure can be distant from the predicate even by 21 word and the width of most predicate structures is 2–4 words (it is worth recalling that the depth of most predicate structures was 3–4 words as well).

In the analysis of the general width of a sentence, we have established that proportion of the sentences the width of which is no less than 4 reaches up to 74.62% (the approximate 95% confidence interval of the proportion is 70% to 79.3%) among all the annotated sentences analyzed. This is especially typical for the predicate structures the width and depth of which in many cases is larger than it is needed to express them by trigrams.

Analysis of structural distributions of Lithuanian texts

A *structural distribution* is one of the main objects of study in statistical linguistics closely related to probabilistic approach, in particular to Zipf-Mandelbrot law ([25], [11]), Yule-Simon law ([24], [17]), etc. A. Utka ([22]) presents the structural distribution of word frequency counts for a Lithuanian corpus of about 102 million word tokens.

We construct an estimator of structural distributions of words by making use of a simple statistical model and the empirical Bayes approach (see [15]). B. P. Carlin and T. A. Louis (2000) provide an overview with extensive reference list on the empirical Bayes methods.

Structural distribution

Let S be a fixed population of subjects or sources of textual information that are assumed to be statistically independent. Let W_s denote a set of word types in (the vocabulary of) a source $s \in S$ and let $V_s := |W_s|$ be its total number of types (the vocabulary size). It is supposed that all vocabularies W_s are subsets of the general vocabulary $\mathcal{W}$. Thus, the data we deal with is $\{(y_w(s), x_w(s)), w \in \mathcal{W}, s \in S\}$, where $y_w(s)$ is the frequency of a word type $w \in \mathcal{W}$ in the source $s \in S$, $x_w(s)$ is a vector

of the corresponding explanatory variables and represents some auxiliary information about both the word type $w \in \mathcal{W}$ and its source $s \in S$.

The fundamental assertion in quantitative linguistics states that, in principle, the vocabulary $\mathcal{W}$ is unbounded (see, e.g., [4], [10]). To put it formally, let us introduce an asymptotic parameter $M \to \infty$ that represents the overall size of text documents under consideration. Thus $\mathcal{W} = \mathcal{W}_M$ and in theoretical considerations we require that

$$V = V^{(M)} := V(\mathcal{W}_M) \to \infty, \quad M \to \infty.$$

Let word types in the general vocabulary $\mathcal{W}$ of the size $V = V^{(M)}$ be arranged in a certain order r to get $\underline{w} = \underline{w}_V(r) := (w_1, \ldots, w_V)$. The observed and expected word type frequencies, y_w and $\mu_w := \mathbb{E}y_w$, and the explanatory variables x_w (as well as other related objects) are accordingly arranged giving $\underline{y} := (y_1, \ldots, y_V)$ $\underline{\mu} := (\mu_1, \ldots, \mu_V)$ and $\underline{x} := (x_1, \ldots, x_V)$, respectively.

When dealing with word count data it is sometimes reasonable to assume that the word type identifiers themselves are irrelevant, they are not interesting for a researcher. Thus, in this case, the expected frequencies of word types w in $\mathcal{W}$ are completely represented by their empirical distribution function (edf)

$$\hat{F}(u) = \frac{1}{V} \sum_{i=1}^{V} \mathbb{I}\{\mu_i \le u\}, \quad u \ge 0. \tag{2}$$

Edf $\hat{F}$ is referred to as an *empirical structural distribution.*

Definition (cf. [9], [7])**.** *Suppose that edf $\hat{F}(\rho t)$ with a scaling factor $\rho = \rho(M)$ weakly converges to the distribution function F, as $M \to \infty$. Then F is called a structural distribution of the expected frequencies $\underline{\mu}$ with the scaling factor ρ.*

Remark. E. V. Khmaladze (1988, [9]) and B. van Es et al. (2003, [7]) considered the multinomial sampling scheme $\underline{y} \sim Multinomial_n(N, \underline{p})$, where $\underline{p}$ is a given vector of probabilities. They defined (empirical) structural distribution by

$$\hat{F}(u) = \frac{1}{V} \sum_{i=1}^{V} \mathbb{I}\{np_i \le u\}. \tag{3}$$

This corresponds to (2) with $\underline{\mu} = N\underline{p}$ and the scaling factor $\rho = N/n$.

Under the Poisson sampling model, values of y chosen at random with equal probabilities from the observed frequencies $\{y_1, \ldots, y_V\}$, satisfies

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} \hat{F} \quad (\stackrel{\mathcal{L}}{=} \text{ defines a distribution law}), \tag{4}$$

where $Poisson(\lambda)$ denotes the Poisson distribution law with the mean $\lambda > 0$. If the sequence $\rho^{-1}\underline{\mu}_V := (\rho^{-1}\mu_1, \ldots, \rho^{-1}\mu_V)$ is a sequence of iid random variables with a common distribution F, then F obviously is the structural distribution of $\mathcal{W}$ with the scaling factor ρ. Thus, the word count distribution, determined by (4), can be approximated by the Poisson mixture model

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\rho\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} F.$$

Empirical studies (see, e.g., [4], [6]) show that a better fit to real data is obtained with improper structural distributions. S. Evert [6] considers Zipf's law in the LNRE

model with the parameters $(\tau, b) \in (0,1) \times (0,\infty)$. The improper structural density f of this distribution law is

$$f(z) = z^{-\tau-1}, \quad 0 < z \le b < \infty,\ \tau \in (0,1).$$

The parameter τ determines the *Zipf's exponent* $1/\tau$.

In order to avoid dealing with improper distributions, we can approximate the empirical structural distribution by the *finite Zipf-Mandelbrot* (or *truncated Pareto*) distribution $F_\varrho(\cdot \mid \tau, \delta)$ with the density

$$\begin{aligned} f_\varrho(z \mid \tau, \delta) &:= \varrho^{-1} f_1(\varrho^{-1} z \mid \tau, \delta), \quad 0 < \delta < 1,\ \tau \in (0,\infty), \\ f_1(z) &:= c_1 z^{-\tau-1} \mathbb{I}\{\delta < z < 1\}, \end{aligned}$$

where the scale parameter ϱ, the lower bound δ of the support of f_1 and hence the normalizing constant c_1 may depend on the asymptotic parameter M. Here we suppose that $\tau > 0$ is fixed,

$$\lim_{M\to\infty} \varrho(M) = b_0 \in (0,\infty], \quad \lim_{M\to\infty} \delta(M)\varrho(M) = a_0 \in [0, b_0).$$

Under these assumptions one can derive the following approximations. The first one describes the relation between the observed number of word types $\widehat{V}_+$ and the observed number of word tokens N in a large corpus

$$\log\left(\widehat{V}_+\right) \approx \text{const} + \alpha \log N \tag{5}$$

and it is known as Herdan-Heaps law. Here $\alpha \in (0,1]$ is some constant usually close to 1. The second one is given by

$$\log\left(\widehat{V}_m\right) \approx \log(\widehat{V}_+) - \alpha_1 - (1+\tau)\log(m), \tag{6}$$

and is called *Zipf's Second Law*. Here $\widehat{V}_m$ denotes the number of word types observed exactly m times and $\alpha_1 := \lim_{M\to\infty} \log(\widehat{V}_+/\widehat{V}_1) \in (0,\infty)$.

Other parametric models of structural distributions are discussed in [3], [4], [6], [5]. The Zipf-Mandelbrot distribution is one of the best-fitting models in real applications.

E. V. Khmaladze ([9]) has that a straightforward estimator of G obtained by substituting y_j/N for p_j $(j = 1, \ldots, n)$ in (3) generally yields an inconsistent estimator. The consistency of structural distribution estimators based on grouping as well as kernel estimators is proved by B. van Es, C. A. J. Klaassen and R. M. Mnatsakanov [7] under certain smoothness conditions.

In this work, the empirical Bayes method is applied in estimating the structural distribution. The main idea of the empirical Bayes approach is to estimate unknown hyperparameters of the prior distribution from the data under consideration. It is also assumed that some auxiliary information in terms of explanatory variables x is available.

Empirical Bayes estimator of structural distribution

E. V. Khmaladze [9] has pointed out that the structural distribution can be treated as a latent mixing distribution in the empirical Bayes approach. Here we present a simple and convenient for computations, yet rather informative, Bayes statistical model and parametric empirical Bayes approach for estimating expected frequencies of word types and their structural distributions.

The Bayesian model incorporates that the conditional distribution of frequency of $y_w(s)$ of the word type w in the source s, given a value x of $x_w(s)$, depends on x through scalar functions $p = p(x)$, $\mu = \mu(x)$ and $\kappa = \kappa(x)$. To be precise,

$$\begin{aligned}
[y_w(s) \mid z_w(s) = 0] &= 0, && (7)\\
[z_w(s) \mid x_w(s) = x] &\stackrel{\mathcal{L}}{=} Binomial(1, 1 - p(x)), && (8)\\
[y_w(s) \mid z_w(s) = 1, \lambda_w(s) = \lambda] &\stackrel{\mathcal{L}}{=} Poisson(\lambda), && (9)\\
[\lambda_w(s) \mid x_w(s) = x] &\stackrel{\mathcal{L}}{=} Gamma(\kappa(x), \mu(x)). && (10)
\end{aligned}$$

Here $\{z_w(s), w \in \mathcal{W}, s \in S\}$ are latent binary random variables (mutually) conditionally independent, when the values of the explanatory variables $\{x_w(s), w \in \mathcal{W}, s \in S\}$ are given. In turn, $\{y_w(s), w \in \mathcal{W}, s \in S\}$ are random variables (mutually) conditionally independent, when the values of latent positive random variables $\{\lambda_{sw}, w \in \mathcal{W}, s \in S\}$ and that of the explanatory variables are kept fixed. $Gamma(\kappa, \mu)$ denotes the Gamma distribution law with the mean $\mu > 0$ and variance $\kappa\mu^2$. The value 0 of the binary latent variable $z_w(s)$ indicates that a word type w is irrelevant (not expected) to the source s, $p(x)$ is the irrelevance probability among cases with $x_w(s) = x$. The latent variable $\lambda_w(s)$ is the expected frequency of the relevant word type w in the source s.

The marginal (and conditional for given x's) distribution of y's is obtained by integrating out the unobservable random variables z's and λ's

$$\begin{aligned}
Q_k(x) &:= \mathbb{P}(y_w(s) = k \mid x_w(s) = x)\\
&= p(x)\mathbb{I}\{k = 0\} + (1 - p(x)) \int_0^\infty \Pi_k(u)\, g\left(u \mid \kappa(x), \mu(x)\right) \mathrm{d}u, \qquad (11)
\end{aligned}$$

which actually is a mixture, respectively with the prior probabilities $p(x)$ and $1-p(x)$, of the degenerate at 0 distribution and the negative binomial distribution q with the mean parameter $\mu = \mu(x)$ and the dispersion parameter $\kappa = \kappa(x)$. Equations (7)–(10) define a conjugate hierarchical Bayesian model with the product of two-component Gamma-Poisson mixtures as the prior distribution. The prior distribution is determined by the mutually independent pairs $(z_w(s), \lambda_w(s))$, $w \in \mathcal{W}, s \in S$, of unknown parameters with distributions

$$\begin{aligned}
[z_w(s) \mid p_{ws} = p] &\stackrel{\mathcal{L}}{=} Binomial(1, 1 - p),\\
[\lambda_w(s) \mid z_w(s) = 1, \mu_{ws} = \mu,\, \kappa_{ws} = \kappa] &\stackrel{\mathcal{L}}{=} Gamma(\kappa, \mu),\\
[\lambda_w(s) \mid z_w(s) = 0, \mu_{ws} = \mu,\, \kappa_{ws} = \kappa] &= 0
\end{aligned}$$

dependent on the *hyperparameters* $p_{ws} := p(x_w(s))$, $\mu_{sw} := \mu(x_w(s))$, $\kappa_{sw} := \kappa(x_w(s))$, $w \in \mathcal{W}, s \in S$. Hence the posterior distribution of the unknown parameters, based

on a sample $y\{D\} := \{y_w(s), w \in \mathcal{W}, s \in D\}$, $D \subset S$, is again the product of two-component Gamma-Poisson mixture with the updated hyperparameters

$$\hat{p}_{sw} = \hat{p}_{sw}(y\{D\}) := \frac{p_{sw}\mathbb{I}\{y_w(s) = 0\}}{p_{sw}\mathbb{I}\{y_w(s) = 0\} + (1 - p_{sw})\, q(y_w(s) \mid \mu_{sw}, \kappa_{sw})}, \quad (12)$$

$$\hat{\mu}_{sw} = \hat{\mu}_{sw}(y\{D\}) := \frac{\mu_{sw}(1 + \kappa_{sw} y_w(s))}{1 + \kappa_{sw}\mu_{sw}}, \quad (13)$$

$$\hat{\kappa}_{sw} = \hat{\kappa}_{sw}(y\{D\}) := \frac{\kappa_{sw}}{1 + \kappa_{sw} y_w(s)}, \quad s \in D. \quad (14)$$

The main problem in Bayesian statistics is specification of the prior distribution. In our setting, it means a specification of the hyperparameters $p_{sw}, \mu_{sw}, \kappa_{sw}$, $w \in \mathcal{W}, s \in S$. According to the *empirical Bayes approach*, the hyperparameters are estimated by fitting the marginal distributions (11) of y's to the available data $y\{D\}$. Assuming a special parametric form of the functions $p(\cdot), \kappa(\cdot)$ and $\mu(\cdot)$ we can solve this task efficiently. For instance, if $p(\cdot)$ and $\mu(\cdot)$ depend on linear predictors with the logit and logarithmic link functions, respectively, and $\kappa(\cdot)$ is a constant then equations (7)–(10) yield *zero-inflated negative binomial* regression model ([8]). The standard statistical software (R, SAS, STATA) can be applied to fit the model.

Given the updated hyperparameters (12)–(14), the structural distribution of word types for a source $s \in S$ can be estimated directly as follows

$$\hat{F}_s(u) = \frac{1}{V} \sum_{w \in \mathcal{W}} \left(\mathbb{I}\{\hat{\mu}_{ws} \le u, y_w(s) > 0\} + (1 - \hat{p}_{sw})\mathbb{I}\{\hat{\mu}_{ws} \le u, y_w(s) = 0\}\right). \quad (15)$$

The second summand in this expression estimates the contribution of *unseen* word types for the source $s \in S$. In order to obtain an estimator of the structural distribution of word types of the general vocabulary $\mathcal{W}$, one can take the weighted average of structural distribution estimators (15)

$$\hat{F}_*(u) := \frac{1}{\omega_+} \sum_{s \in S} \omega_s \hat{F}_s \left(\frac{u N^*}{\widehat{\mu}_{*s}}\right) \quad (16)$$

appropriately scaled to have the same estimated expected text sizes N^*. In the empirical study (see the next subsection), the equal weights and the weights proportional to the source text (vocabulary) size are considered.

Results of empirical study

In the empirical study, the texts are taken from the digital library collection[3] that consists of the recommended school imaginative pieces of Lithuanian and foreign authors, overall 80 text documents of 63 authors with 206453 word types (different words in a text) and 2567290 word tokens (running words in a text) in total.

[3] *http://ebiblioteka.mkp.emokykla.lt*, looked over on the 30th of April, 2013.

The author is grateful to Education Development Center, the Ministry of Education and Science of the Republic of Lithuania and EU Structural Fund Support of project "Development of the Key Competencies in Basic School (grade 5-8)".

It is assumed that the authors are independent and heterogeneous sources of textual data. Therefore different s stand for different authors. The general vocabulary $\mathcal{W}$ is taken as $\cup_{s\in S}W_s$. The vector of explanatory variables $x_w(s)$ consists of two categorical variables, s and ℓ_w, and their interactions. The categorical variable $s \in S$ has $|S| = 63$ categories, the categorical variable $\ell_w \in \{2, \ldots, 10\}$ is the group of length of the word type w. The group with $\ell_w = 2$ consists of word types of length 1 or 2, word types in a group with $\ell_w = 10$ have 10 or more letters, in the rest groups the word type length and group number is coincident. We also use a derivative feature *native* indicating weather an author is native Lithuanian or he/she is foreign.

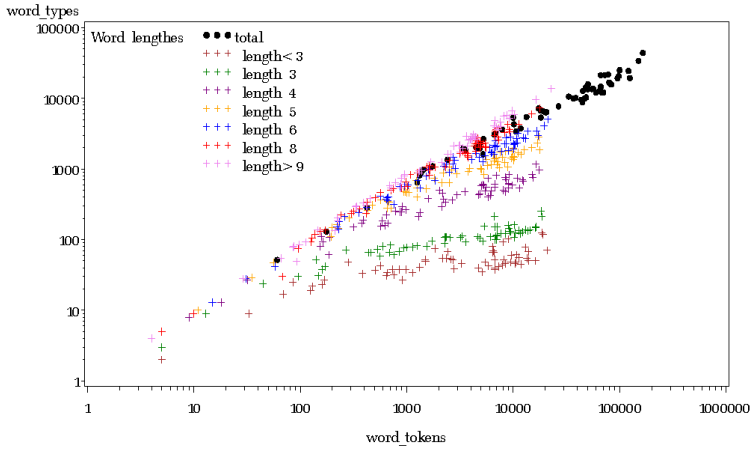

Figure 5: The Herdan-Heaps law for the words of different length and in total (groups of the length 7 and 9 are omitted)

Fig. 5 illustrates the Herdan-Heaps law (5) and shows the distribution of text and vocabulary sizes among the authors. The data in the log-log scale fits the straight line very well even for short text documents. However, it does not hold for shorter word types (as $\ell_w < 6$). Graphical illustrations of Zipf's second law (6) are presented in Fig. 6. Again, one can notice a significant dependence of the slopes in the Zipf's law on the length of word types (more vivid for longer ones with $\ell_w > 7$).

Fig. 7 contains the scatter plot of estimated slopes in approximate linear equations (6) by formal fitting a linear regression model to data of each source (we do not check the model fit and validity). Note a tendency of slopes to be smaller in absolute value for foreign authors as compared with native ones. These observations show that the choice of $x_w(s) = (s, \ell_w)$ (as the initial step in this direction) is quite reasonable.

The empirical Bayes approach applied to the available data enables us to estimate the number of *unseen word types* in each source and hence adjust estimators of the structural distributions, respectively. The effect of the adjustment is apparent in Fig. 8, where histograms of structural distribution estimators, obtained by making use of different methods, are drawn. The structural distribution estimators are respectively scaled to match the total $N^* = 10^6$ of word tokens and their histograms are standardized for a text document with the vocabulary size of 10^6 word types. *Source_Equal* and *Source_Totals* label the estimates obtained by (16) with the equal weights $\{\omega_s\}$ and the weights proportional to the text size N_s of sources $s \in S$, respectively. The

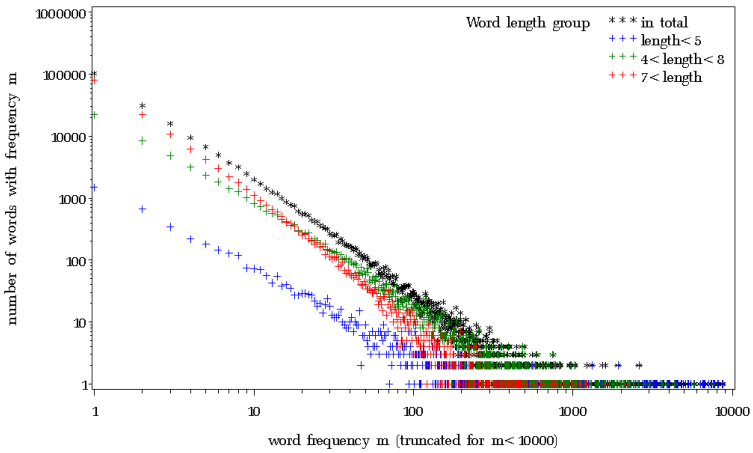

Figure 6: Zipf's second law for the words of different length and in total

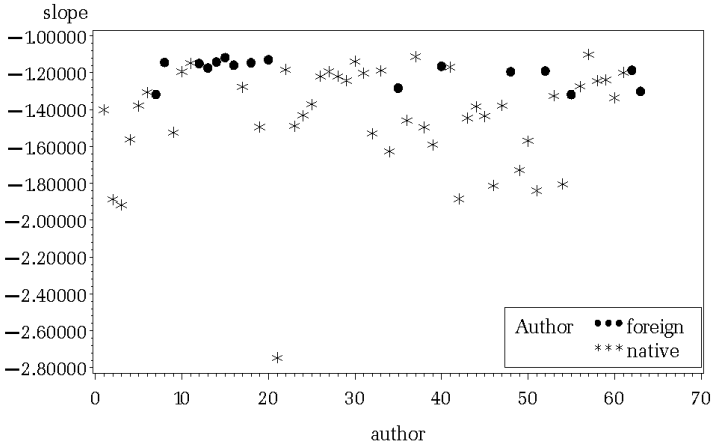

Figure 7: Scatter plot of slope estimates in Zipf's second law

estimates with the weights proportional to the source vocabulary size are very close to that with equal weights and therefore are not presented here. In addition, histograms of analogous structural distribution estimates, but now based on a random subsample $S_9 \subset S$ of size 9, are given. That is indicated by adding the word *Sample* to the labels. For comparison, we include the histogram of respectively scaled observed frequencies.

The weighted structural distribution estimates *Source_Equal_Sample* and *Source_Totals_Sample* yield reasonable predictions of the respective estimates *Source_Equal* and *Source_Totals*, based on the whole sample S, and probably of the true structural distribution F.

To illustrate the textual data heterogeneity, histograms of *Source_Equal* type estimates of structural distributions for two groups of authors (native vs. foreign) are presented in Fig. 9. Note a more subtle word frequency pattern of foreign authors as compared to the scatter plot of slopes in Zipf's second law in Fig. 7. On the one hand, translated texts tend to use more the standard vocabulary (reduction of expected frequencies in the interval $(0, 0.8)$ in the $\log_{10}$ scale), on the other hand, they

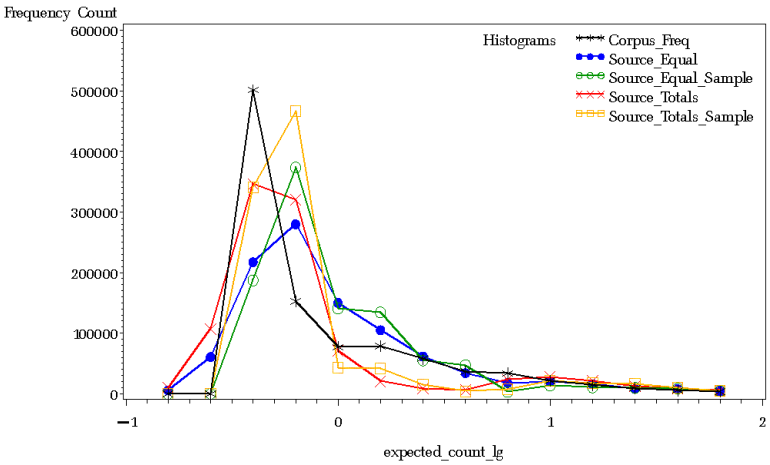

Figure 8: Histograms of structural distribution estimates and empirical histogram: a comparison

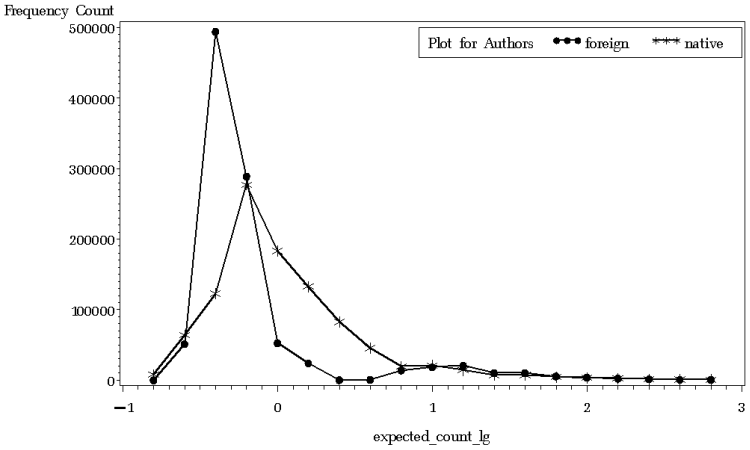

Figure 9: The structural distribution estimates for native and foreign authors

contain words related to culture and specific being of other nations and hence rare in the original Lithuanian texts (the pike at -0.4).

General conclusions

1. When applying statistical methods, it is of great importance to exactly define the research object and population thereby, since the interpretation and reliability of the results obtained depend on that.

2. As shown by analogues of the Zipf-Mandelbrot and Herdan-Heaps laws and by analysis of the structural distribution, the data of corpora are non-homogeneous, they depend not only on style, genre, and the like, but also on the author himself. Considerable differences between styles and the authors' texts are noticed. When making statistical inference in linguistics, non-homogeneity and variability should be taken into consideration.

3. The Bayes model, based on the zero-inflated negative binomial regression, and the empirical Bayes method enabled us to construct estimators of the structural distribution that take into account corpus non-homogeneity and bias that occur due to unseen word types.

4. Loglinear and graphical loglinear models are rather flexible and allow us to describe complex structures, to represent them graphically. Thus, they are a convenient tool to state and test various hypotheses about structures and dependencies of linguistic objects as well as properties that are typical for the language itself.

5. By encoding words in a special manner, we can treat a sentence as a new word and investigate the analogues of the Zipf-Mandelbrot law. The initial analysis shows that significant part (more than 17%) of sentences have the simplest structures.

6. The syntactic structures of sentences of the Lithuanian language are too complicated so as to be described by models, based on the trigram statistics. For instance, the width of sentences of as much as 74.62% is large than 3.

List of author's publications on the topic of dissertation

1. K. Piaseckienė, M. Radavičius, R. Stiklius. Lietuviškų tekstų stilių palyginimas remiantis universalių kiekybinių charakteristikų statistine analize. *Lietuvos matematikos rinkinys. LMD darbai*, 2010, **51**: 307–312.

2. K. Piaseckienė, M. Radavičius. Lietuvių kalbos vaizdingumo raiškos priemonių analizė. *Lietuvos matematikos rinkinys. LMD darbai*, 2011, **52**: 220–224.

3. K. Piaseckienė, M. Radavičius. Empirical Bayes estimators of structural distribution of words in Lithuanian texts. *Nonlinear Analysis: Modelling and Control*, 2014, **19(4)**. (Accepted for publication)

List of literature, referenced in this summary

1. S. Abney. Statistical Methods and Linguistics. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1996, 1–26.

2. A. Agresti. *Categorical Data Analysis*. New York: Wiley & Sons, 2002.

3. R. H. Baayen. Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. *Computers and the Humanities*, 1993, **26**: 347–363.

4. R. H. Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

5. M. Baroni, S. Evert. Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, 2007, 904–911.

6. S. Evert. A simple LNRE model for random character sequences. *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT2004)*, Belgium: Louvain-la-Neuve, 2004, 411–422.

7. B. van Es, C. A. J. Klaassen, R. M. Mnatsakanov. Estimating the structural distribution function of cell probabilities. *Austrian Journal of Statistics*, 2003, **32**: 85–98.

8. M. Jansche. Parametric Models of Linguistic Count Data. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2003, **1**:288–295.

9. E. V. Khmaladze. The statistical analysis of large number of rare events. *CWI Report MS-R8804*, Amsterdam: Dept.Math.Statist., 1988.

10. A. Kornai. How many words are there? *Glottometrics*, 2002, **4**: 61–86.

11. B. Mandelbrot. An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. *Communication Theory*, London: Acad. Press, 1953, 503–512.

12. R. Merkytė, V. Kalinka. Apie W. Fuchso lingvistinių elementų susidarymo dėsnį. *Lietuvos matematikos rinkinys*, 1968, **8(2)**: 279–287. (Rusų kalba)

13. R. Merkytė. Skiemenų ir fonemų skaičiaus lietuvių kalbos žodžiuose savitarpio priklausomybės tyrimas. *Eksperimentinė ir praktinė fonetika*. Vilnius, 1974, 73–84.

14. K. Piaseckienė, M. Radavičius. Lietuvių kalbos vaizdingumo raiškos priemonių analizė. *Lietuvos matematikos rinkinys. LMD darbai*, 2011, **52**: 220–224.

15. K. Piaseckienė, M. Radavičius. Empirical Bayes estimators of structural distribution of words in Lithuanian texts. *Nonlinear Analysis: Modelling and Control*, 2014, **19**.

16. E. Rimkutė, V. Daudaravičius. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos*, 2007, **11**: 30–35. http://donelaitis.vdu.lt/publications/Rimkute_2007.pdf.

17. H. A. Simon. On a class of skew distribution functions. *Biometrika*, 1955, **42**: 425–440.

18. N. A. Smith. Linguistic Structure Prediction. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, 2011.

19. D. Šveikauskienė. *Lietuvių kalbos vientisinių sakinių automatinė sintaksinė analizė*: daktaro disertacija. Vilnius, 2009.

20. D. Šveikauskienė. Lietuvių kalbos sintaksinė analizė. *Lietuvių kalba*, 2013, **7**. http://www.lietuviukalba.lt/index.php?id=231.

21. L. Tanguy, N. Tulechki. Sentence Complexity in French: a Corpus-Based Approach. *Intelligent Information Systems*, 2009, 1—14.

22. A. Utka. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės. *Lituanistica*, 2005, **1(61)**: 48–55.

23. A. Vaičiūnas. *Lietuvių kalbos statistinių modelių ir jų taikymo šnekos atpažinimui tyrimas, kai naudojami labai dideli žodynai*: daktaro disertacija. Kaunas, 2006.

24. G. U. Yule. A mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, 1924, **B 213**: 21–87.

25. G. K. Zipf. *The Psycho-Biology of Language*. New York: Houghton Mifflin, 1935.

Reziumė

Pagrindinis darbo tikslas – pritaikyti matematinius ir statistinius metodus lietuvių kalbos analizėje, identifikuojant ir atsižvelgiant į lietuvių kalbos ypatumus, jos heterogeniškumą, sudėtingumą ir variabilumą.

Siekiant numatyto tikslo, buvo sprendžiami tokie uždaviniai:

• Atlikti statistinių taikymų lietuvių kalbos tyrimuose apžvalgą.

• Taikant (grafinę) logtiesinę analizę bei kitus statistinius ir grafinius metodus ištirti lietuvių kalbos savybes, struktūras ir jų sudėtingumą; statistiniais metodais ištirti, ar kitoms kalboms pastebėti ypatumai tinka lietuvių kalbai.

• Aprašyti ir įvertinti kalbos heterogeniškumą ir variabilumą (kintamumą), kurį sąlygoja jos autoriaus pasirinkimai, siekiant identifikuoti nuo autoriaus santykinai mažai priklausančias, vadinasi, potencialiai pačiai kalbai būdingas savybes, sąryšius ir struktūras. Sudaryti atitinkamą metodiką, ją pritaikyti konkretiems lietuvių kalbos tyrimams.

• Pademonstruoti matematinių metodų galimybes sprendžiant konkrečius lietuvių kalbos uždavinius: skirtingų funkcinių stilių identifikavimas remiantis raidžių / garsų proporcijomis, metakalbinių komentarų konstravimo ypatumai, sakinio struktūros sudėtingumo matavimas (charakteristikos ir jų pasiskirstymas).

• Pritaikyti empirinį Bajeso metodą aprašant kalbos variabilumą ir vertinant struktūrinį skirstinį.

Mokslinis darbo naujumas ir praktinė vertė.

Atlikto darbo rezultatai papildo ir praplečia kitų šioje bei giminiškose srityse atliktų tyrimų rezultatus.

Lingvistiniuose tyrimuose, kurie remiasi tekstynais, pirminis tyrimo elementas yra žodis, žodžių junginys, kartais – sakinys. Šiame darbe pirminis tyrimo elementas yra autorius, jo pasirinkimas yra kalbos heterogeniškumo ir variabilumo šaltinis. Šis požiūris, drauge su atitinkamais imčių bei statistinės analizės metodais, sudaro siūlomos naujos metodologijos, kuri leidžia nustatyti nuo autoriaus santykinai mažai priklausančias, vadinasi, potencialiai pačiai kalbai būdingas savybes, sąryšius ir struktūras, pagrindą.

Sprendžiant automatinio kalbos apdorojimo uždavinius labai svarbu įvertinti tiriamų lingvistinių struktūrų sudėtingumą, nes jis gali nulemti ne tik naudojamų metodų pasirinkimą, bet ir bazinio kalbos modelio sudarymo principus bei analizės metodiką. Šiame darbe atlikta pradinė sakinio (grafinės) sintaksinės struktūros sudėtingumo statistinė analizė. Nors paprastos tekstų lietuvių kalba sudėtingumo charakteristikos, matyt, jau skaičiuotos ne kartą, vis dėlto darbai, skirti nuoseklesnei jų sudėtingumo analizei, autorei nežinomi.

Remiantis neigiamos binominės regresijos su pertekliniu nulių kiekiu modeliu ir empiriniu Bajeso metodu buvo sukonstruotas žodžių formų tekste struktūrinio skirstinio įvertinys, kuris panaudoja turimą papildomą informaciją apie teksto autorius ir žodžio formas ir tokiu būdu leidžia atsižvelgti į tekstų nehomogeniškumą bei nestebėtų žodžio formų efektą. Struktūrinis skirstinys yra žymiai subtilesnis kalbos tyrimo įrankis negu metodai, kurie remiasi parametriniais Zipfo-Mandelbroto tipo modeliais.

Šis darbas parodo imčių metodų taikymo svarbą ir galimybes.

Ginamieji disertacijos teiginiai.

• Statistiniai metodai plačiai taikomi lietuvių kalbos analizėje, bet pastaruoju metu vyrauja aprašomoji statistika ir informatikų sukurtos procedūros, pritaikytos daugiau anglų kalbai ir orientuotos į konkrečių praktinių uždavinių sprendimą.

• Apskritai Herdano ir Zipfo dėsniai gana tiksliai aproksimuoja žodžių formų kiekį ir pasiskirstymą lietuvių kalbos tekstuose. Tačiau tuos dėsnius aprašančių parametrų reikšmės tarp autorių ženkliai skiriasi, dar labiau tarp lietuvių ir užsienio autorių.

• Lietuvių kalboje betarpiškai susiję žodžiai gali būti gerokai nutolę vienas nuo kito, todėl modeliai ir automatinės taisyklės, sudarytos remiantis trigramų statistika, turi gana ribotas galimybes tinkamai modeliuoti ir prognozuoti lietuvių kalbos struktūras.

• Tam, kad sudėtingesni statistiniai kalbos tyrimai būtų atlikti ir interpretuoti korektiškai, pastoviai palaikomi tekstynai turėtų suteikti galimybę tyrimo duomenų sudarymui taikyti imčių metodus ir išrinkti tekstus pagal įvairius požymius, taip pat ir pagal autorius. Tekstynai, kurie neleidžia kontroliuoti imties sudarymo taisyklių, turi labai ribotas statistinės analizės galimybes.

• Tinkamai taikant imčių metodus surinkti duomenys, (grafiniai) logtiesiniai modeliai, taip pat ir empirinis Bajeso metodas leidžia išnaudoti turimą papildomą informaciją ir modeliuoti sudėtingas kalbos struktūras, jų heterogeniškumą bei individualų kintamumą ir tokiu būdu sudaro pagrindą nustatyti ir tyrinėti pačiai kalbai būdingas savybes bei sąryšius.

Išvados.

1. Taikant statistinius metodus labai svarbu tiksliai apibrėžti tyrimo objektą, o tuo pačiu ir tiriamąją populiaciją, kadangi nuo to priklauso gautų rezultatų interpretacija ir patikimumas.

2. Kaip rodo Zipfo-Mandelbroto ir Herdano-Hipso dėsnių analogai bei struktūrinio skirstinio analizė, tekstynų duomenys yra nehomogeniški, priklauso ne tik nuo stiliaus, žanro ir pan., bet ir nuo pačio autoriaus. Stebimi ryškūs skirtumai tarp stilių ir tarp pačių tekstų autorių. Darant išvadas apie statistinius dėsningumus lingvistikoje reikėtų į tą nehomogeniškumą atsižvelgti.

3. Bajeso modelis, kuris remiasi neigiama binomine regresija su pertekliniu nulių kiekiu, ir empirinis Bajeso metodas leido sukonstruoti struktūrinio skirstinio įvertinius, kurie atsižvelgia į tekstynų nehomogeniškumą ir poslinkį, atsirandantį dėl juose nestebimų žodžių formų.

4. Logtiesiniai ir grafiniai logtiesiniai modeliai yra gana lankstūs ir leidžia aprašyti sudėtingas struktūras, jas pavaizduoti grafiškai. Taigi, tie modeliai yra patogus įrankis formuluoti ir tikrinti įvairias hipotezes apie lingvistinių objektų struktūras ir priklausomybes, taip pat ir pačiai kalbai būdingas savybes.

5. Specialiu būdu užkodavus žodžius, į sakinį galima žiūrėti kaip į naują žodį ir traktuojant sakinius kaip žodžius galima tirti Zipfo-Mandelbroto dėsnio analogus. Pirminė analizė rodo, kad nemaža dalis (daugiau negu 17%) sakinių turi paprasčiausią struktūrą.

6. Lietuvių kalbos sintaksinės (sakinių) struktūros yra per daug sudėtingos, kad jas būtų galima aprašyti modeliais, besiremiančiais trigramų statistika. Pavyzdžiui, net 74,62% sakinių plotis yra didesnis negu 3.

Briefly about the author

Education

- 1998 m. Šiauliai University, Faculty of Humanities, bachelor of Lithuanian linguistics and literature.
- 2001 m. Šiauliai University, Faculty of Humanities, master of Lithuanian philology.
- 2005 m. Šiauliai University, Faculty of Mathematics and Informatics, bachelor of mathematics.
- 2007 m. Šiauliai University, Faculty of Mathematics and Informatics, master of mathematics.

Working experience

- 2005–2006 m. Dainai secondary school, Šiauliai, mathematics teacher.
- Since 2007 m. Šiauliai University, Department of Mathematics, assistant.

Trumpai apie autorę

Išsilavinimas ir kvalifikacija

- 1998 m. Šiaulių universiteto Humanitariniame fakultete baigus Lietuvių kalbos ir literatūros studijų programos bakalauro studijas suteiktas humanitarinių mokslų bakalauro ir lietuvių kalbos ir literatūros mokytojo kvalifikacinis laipsnis.
- 2001 m. Šiaulių universiteto Humanitariniame fakultete baigus Lietuvių kalbotyros studijų programos magistro studijas suteiktas humanitarinių mokslų magistro ir gimnazijos mokytojo kvalifikacinis laipsnis.
- 2005 m. Šiaulių universiteto Matematikos ir informatikos fakultete baigus Matematikos studijų programos bakalauro studijas suteiktas matematikos bakalauro kvalifikacinis laipsnis.
- 2007 m. Šiaulių universiteto Matematikos ir informatikos fakultete baigus Matematikos studijų programos magistro studijas suteiktas matematikos magistro ir matematikos mokytojo kvalifikacinis laipsnis.

Darbo patirtis

- 2005–2006 m. Šiaulių m. Dainų vidurinės mokyklos matematikos mokytoja.
- Nuo 2007 m. Šiaulių universiteto Matematikos katedros asistentė.