Content

[1. Excellence 2](#_Toc145595489)

[1.1 Project objectives 2](#_Toc145595490)

[1.2 RELEVANCE, QUALITY AND NOVELTY OF THE PROJECT 4](#_Toc145595491)

[1.3 METHODOLOGY 5](#_Toc145595492)

[1.4 EXCELLENCE OF THE RESEARCHER 8](#_Toc145595493)

[1.5 EXCELLENCE OF THE APPLICANT/HOST ORGANISATION 15](#_Toc145595494)

[2. Impact 15](#_Toc145595495)

[2.1 THE WIDER IMPACT OF THE PROJECT 15](#_Toc145595496)

[2.2 MEASURES TO MAXIMISE IMPACT – DISEMINATION AND COMMUNICATION, EXPLOITATION OF RESULTS 16](#_Toc145595497)

[3. Implementation 16](#_Toc145595498)

[3.1 PROJECT PLAN AND DELIVERABLES 16](#_Toc145595499)

[3.2 IMPLEMENTATION RISKS AND PROPOSED MEASURES 22](#_Toc145595501)

[3.3 OPERATIONAL CAPACITY OF THE APPLICANT/HOST ORGANISATION 22](#_Toc145595504)

Title of the project: Mathematical Models of Linguistic Laws

Short title of the project/Acronym: MAMOLILA

Category of researcher: R3

Researcher’s job type: full-time researcher

Type of research: basic

Identification of the entity involved in the implementation of the project:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Official name of the entity | Abbreviated name of the entity | Role in the project |
| 1 | Mathematical Institute, Slovak Academy of Sciences | MI SAS | Host organisation |

## 1. Excellence

### 1.1 Project objectives

**Objective 1**: The Menzerath-Altmann law (MAL) - the state-of-the-art analysis

This objective is a necessary first step. While the recent scientific literature is well known to the applicant, new papers with new results appear quite often. Moreover, it will be necessary to obtain older publications which are not available online (especially papers from volumes which were published in minor publishing houses in Germany). A list of mathematical models of the MAL which appeared in the scientific literature will be prepared. Data from older scientific literature will be digitalized. A bibliography will be created, which will be made available on a webpage (with links to the papers which can be find online).

**Objective 2**: A general mathematical model for the MAL.

One can find several mathematical models for the MAL in the scientific literature (and the number of models can grow when older papers are obtained). Mostly, they were suggested as models for relartions between length of particular language units (e.g. one model for the relation between word length and syllable length, another one for the relation between sentence length and clause length, etc). A general mathematical model will be chosen, or a new one will be constructed if none of the ones from the literature will fit all language units from syllable to sentence (or possibly even supra-sentence structures). The interaction between the MAL and the Zipf law (which predicts that shorter units occur more often in texts) will have to be considered (the MAL describes properties of the vocabulary, while the Zipf law takes into account frequencies of language units as they are used in texts). Parameters of the general model of the MAL will be estimated. Results will be published in two scientific papers in peer-reviewed journals.

**Objective 3**: A monograph on the MAL

A monograph on the MAL will be prepared in cooperation with two or three coauthors with different educational backgrounds (Radek Čech, a linguist from the University of Ostrava, Czech Republic; Andrij Rovenchak, a physicist from the Ivan Franko National University in Lviv, Ukraine; possibly another researcher with a background in cognitive science from the quantitative linguistics research group at the Catalonian Polytechnical University in Barcelona, Spain). Such a diverse team of coauthors will guarantee that the monograph will be both formally correct and at the same time relevant not only from the mathematical point of view, but also for linguistic research. The applicant will be the first author of the book.

The monograph will contain a historical overview of the research on the MAL, special cases of the MAL for different language units, a general model of the law, fitting it to diverse datasets (diverse with respect to both languages and language levels, such as syllables, words, clauses, sentences etc), and an attempt to find a linguistically relevant interpretation of its estimated parameters.

The aim is to publish the monograph at de Gruyter publishing house within its Quantitative Linguistics book series (<https://www.degruyter.com/serial/ql-b/html>).

**Objective 4**: A mathematical model for phoneme and grapheme rank frequency distributions - the state-of-the-art analysis

Phonemes and graphemes are exceptional among linguistic units with respect to the mathematical model for their ranked frequencies. Frequencies of syllables, words and units which are even higher in the language unit hierarchy can be modelled by Zipf-like distributions (usually by the Zipf-Mandelbrot distribution). However, the inventory of phonemes and graphemes is much lower, typically several tens. Consequently, their ranked frequencies cannot display the “Zipfian behaviour” for which a long tail (e.g. many words occur only once in a text) is so characteristic. The negative hypergeometric distribution (NHGD) is a model which achieves a very good fit when applied to data from many different languages. The applicant was involved in a project focusing on the mathematical modelling of grapheme rank frequency distributions at the University of Graz (Austria) 15 years ago, however, after the death of the head to the research team, the project remained without a significant output. Although this topic is not so much “alive” as the MAL, some time will have to be dedicated to summarizing the results published in the last roughly 15 years. Similarly to Objective 1, also here we will prepare a bibliography and make it available online.

**Objective 5**: The NHGD, its parameters, and its goodness-of-fit

The NHGD achieves a good fit, but goodness-of-fit is only one of several criteria which a good scientific model should satisfy. Another one is interpretability of the parameters of the model. In this aspect, the NHGD is problematic, as it is derived from a binary urn considerations (balls of two colours are drawn from an urn). For phonemes and graphemes, however, the binarity is unrealistic (perhaps the only reasonable two categories would be vowels and consonants, but their proportions do not correspond to parameter values). One way how to “rescue” the NHGD could be its reinterpretation as a special case (with integer parameter values) of the beta-binomial distribution. Still other possibilities remain open, such as e.g. using some characteristics of the NHGD with a clear interpretation (e.g. its entropy) and to find their relation to the parameter values.

But even the approach to the goodness-of-fit evaluation must be reexamined. If the sample size is small, usually the classical chi-square test is performed, and the p-value is used to decide whether the model for the data well enough. However, in terms of the p-values, almost any null hypothesis is rejected for very large samples (and language corpora are sources of very large samples). Therefore, for large samples, it became standard in linguistic research to compute the value of the chi-square test statistic and to divide the value by the sample size. There are some rules of thumb for the result (if it is below 0.02, the fit is considered very good).

The problem is that there is no criterion which could be used to decide which sample are small enough to trust the p-values, and which are already too large. Therefore, a new criterion for evaluating models’ goodness-of-fit will be developed, which will be applicable to any sample regardless of its size. The criterion will be an adjustment of the formula currently used for large samples, with some theoretical properties of the chi-square distribution taken into account.

**Objective 6**: Monitoring the progress of the project and dissemination of its results

The development of the project and the timeline of achieving the results will be continuously monitored. If problems will be anticipated, necessary steps will be undertaken to avoid significant delays in delivering project outputs.

Results obtained will be disseminated in form of four scientific papers in peer-reviewed journals or as chapters in special volumes published by prestigious publishing houses (two papers containing results related to Objective 2, and other two with results related to Objective 5). The monograph on the MAL (Objective 3) will contain results from the past as well as our new findings.

Results achieved will be presented also at several conferences and seminar talks. As minor outputs, two bibliographies (one containing scientific literature in the MAL, the other on mathematical models of phoneme and grapheme frequencies) will be created and made available.

The interim and final report will be presented, and, before the end of the project duration, an application for a new research project will be submitted.

1.2 RELEVANCE, QUALITY AND NOVELTY OF THE PROJECT

Mathematics, statistics, and information technology are quickly becoming integral parts of linguistic research. While not so long ago a linguist cooperating with a mathematician and applying mathematical and statistical methods in the analysis of linguistic data was an exception, a rapid development of computers is mirrored in changing methodological approaches – linguistics is becoming an empirical science, it analyses data, statistical methods are commonly used, and linguistic models are expressed by mathematical formulas.

One of the consequences is a change in understanding of what is a linguistic law. Again, not so long ago, a law in linguistics was a prescription how language should be used, with exceptions having been listed and taught at schools. Nowadays, a linguistic law is more and more often understood as a statistical tendency, and a few data points which do not conform to it are considered random fluctuations rather than exceptions.

This project focuses on two linguistic laws.

The first of them is the MAL, according to which longer units tend to consist of shorter parts (e.g. longer words contain syllables that are on average shorter than those in shorter words). The MAL was originally formulated verbally for the above-mentioned relation between length of words and syllables. Later, it was substantially generalized, it was shown that it is valid for many different language units and parts they consist of. It was also formalized and expressed mathematically. Although the power law is the most frequently used as its mathematical model, several other formulas were suggested as well, mostly within relatively narrow contexts (for specific languages and for specific language units). And although there are several mathematical models, none of them can claim that the role of all its parameters would be understood. This project aims at choosing one of the models (or deriving a new one), which will be general (in the sense that it fits well all languages and all languages units) and its parameters will be interpreted (i.e. their relation to other known language property would be revealed).

The other one is the NHGD. This discrete distribution serves as a model for ranked frequencies of phonemes and graphemes (and no other discrete distribution achieves an acceptable fit). Once again, it is the question of the interpretation of its parameters (and its genesis from an urn scheme with balls of only two colours) which remains unanswered.

The applicant, being a mathematician (although with a long-term experience in work with linguistic data), will try to avoid the risk of deriving mathematically correct, but linguistically nonsensical models. Such a risk is real if one works alone, however, it will become negligible because of the applicant’s cooperation with linguists (Radek Čech, Masaryk University, Brno; Emmerich Kelih, University of Vienna; Miroslav Kubát, University of Ostrava). In all three cases, the cooperation has a long history, the researchers often publish as coauthors. One can see that in spite of different educational background, they managed to find a common ground.

This project will thus promote international collaboration (not only with the three above-mentioned researchers, but also with others). It will pursue scientific excellence and it will thus contribute to a better, more competitive position of the applicant, his collaborators, and in a more general sense also of his institute. The project has some potential to contribute to Europe’s scientific competitiveness and to open also a field for “a typical European applications” – some parameters of the MAL can be used to assess text difficulty, which could be a useful tool in the multilingual EU environment (then one could e.g. compare if the source text and its translation are equally easy to understand for an average reader in different countries).

The research outcomes will be disseminated through peer-reviewed publications in scientific journals and special volumes, lectures on conferences and seminars, and via knowledge exchange with European research institutions (both personally and online).

1.3 METHODOLOGY

A successful completion of the project requires expert knowledge in probability theory and mathematical statistics, experience in programing and working with statistical software, and experience in data analysis. The applicant satisfies these conditions, as can be see also from his publication record (he publishes both papers of a theoretical character on probability theory, and papers focusing on modelling and statistical analysis of linguistic data). A background in linguistics, also necessary as a guarantee that mathematical models and statistical procedures will be relevant for linguistic research, will be provided by a close cooperation with linguists, especially with experts in Slavic languages from Czech Republic and Austria.

The methodological approaches chosen to achieve objectives from Section 1.1 are the following:

**Objective 1**: The Menzerath-Altmann law (MAL) - the state-of-the-art analysis

Task 1.1: Collecting publications on the MAL which the applicant does not have yet.

Task 1.2: Study of publications related to the MAL that are unknown to the applicant so far.

Task 1.3: Creating a bibliography of papers on the MAL and a database of the data used in the literature.

Task 1.4: Creating a list of mathematical models of the MAL.

In Task 1.1, we will read recent issues of journals dedicated to quantitative linguistics (Journal of Quantitative Linguistics, Glottotheory, Glottometrics) and proceedings from conferences in which such presentations appeared. We will also try to obtain older papers from 1980s from book series published especially by minor German publishing houses (e.g. Brockmeyer in Bochum or WVT in Trier), in the ideal case as interlibrary loans. If that would not be possible, a part of resources for research will be used to visit universities which used to buy the production of those publishing houses (at least the universities in Trier, Bochum, and Göttingen) and to scan the relevant papers. Tasks 1.2-1.4 will be started partially simultaneously with Task 1.1, and older papers which are not available online will be processed after they are obtained.

**Objective 2**: A general mathematical model for the MAL

Task 2.1: Analysis of models from Task 1.4.

Task 2.2: Choosing the model(s) with a good fit and desirable properties (which correspond to the nature of data they are supposed to model).

Task 2.3: Deriving a new model if necessary.

Task 2.4: Attempt to the interpret parameters of the model.

Task 2.1 will focus on mathematical properties of the models, which should mirror properties of real data. E.g., the power law, the simplest and the most often used model, converges to zero, but this property contradicts the property of the data (each syllable contains at least one phoneme, which means that the minimum possible mean syllable length is 1, not 0). Tasks 2.2-2.4 will be performed in a close cooperation with linguists.

**Objective 3**: A monograph on the MAL

Task 3.1: Summarization of results on the MAL published so far and our results from Objective 2.

Task 3.2: Preparing the structure of the book.

Task 3.3: Writing the monograph, closely cooperating with the coauthors.

In Task 3.1, suitable datasets will be chosen. The book will have roughly the following structure: 1) a short historical overview, 2) the MAL for different language units, 3) fitting our model from Objective 2 to several datasets, 4) summarizing papers on the applications of the MAL (comparing authors, genres, etc) and choosing material which will be included in the book. The applicant, being the first author, will act as a coordinator among all coauthors. He will also communicate with the publisher (he has experience of doing it and also knows people at the publishing house from online communicartion, as he was a co-editor of two volumes published there several years ago).

**Objective 4**: A mathematical model for phoneme and grapheme rank frequency distributions - the state-of-the-art analysis

Task 4.1: Collecting the relevant literature published after 2011 (the applicant has the relevant older publications, they were collected during his postdoc stay the University of Graz).

Task 4.2: Study of the publications from Task 4.1.

Task 4.2: Creating a bibliography.

Task 4.3: Creating a list of models for phoneme and grapheme frequencies.

The methodology is basically the same the one for Objective 1, but here the situation is much easier, as the publications from the pre-internet times are ready. As the Mathematical Institute has access to a very broad range of scientific journals, no serious problems are expected here. Most probably, all relevant papers will be downloaded from online version of journals.

**Objective 5**: The NHGD, its parameters, and its goodness-of-fit

Task 5.1: Attempt to obtain an interpretation of parameters using the beta-binomial distribution.

Task 5.2: If the attempt from Task 5.1 is not successful, other characteristics of the NHGD will be applied, e.g. its entropy.

Task 5.3: Developing a rule-of-thumb criterion which is general for all sample sizes

Task 5.4: Fitting the NHGD to various datasets, evaluating its goodness-of-fit

In Task 5.1, the beta-binomial distribution will be studied. Under certain conditions, the NHGD is a special case of the beta-binomial distribution, which can be helpful in getting rid of the binary background of the NHGD. If this approach does not bring the desired result, we will turn our attention to other properties of the NHGD, such as its entropy, moments and cumulants.

The current rule-of-thumb criterion used (not only) in quantitative linguistics will be modified so that it will become appropriate also for small samples. The idea is to develop such a criterion which would have a limit identical to the currently used criterion. Thus, the differences between the new criterion and the “old one” will be negligible for very large samples typical for data from language corpora.

In Task 5.4, data from different languages (from different language families) wil be used. We will thus verify our assumption that the NHGD can be considered another language law (i.e. it is generally valid).

**Objective 6**: Monitoring the progress of the project and dissemination of its results

Task 6.1: To monitor the progress of the project

Task 6.2: If there is a risk of a significant delay, appropriate steps will be undertaken.

Task 6.3: Dissemination of results

Task 6.4: Research proposal

In Task 6.1 we will monitor the development of the project. Potential problems will be addressed, if needed, in Task 6.2. In Task 6.3 we will disseminate the results of the project (a monograph, planned four papers in journals, conference presentations, etc. An application for another project will be written within Task 6.4.

The suggested methodology corresponds closely to the project objectives. The applicant has rich experience with this type of research. In addition, he will regularly consult some aspects of this project with experts with background in both probability and statistics and in linguistics. Attending scientific conferences, where one will meet people working on projects of a similar character, can provide new impulses for research.

Data will be obtained on two steps (both for the analysis of the MAL and for the application of the NHGD as a model for phoneme and grapheme ranked frequencies). In the first step, the Slavic language will be used as the source of data. The reason is that the applicant has established and well-functioning contacts with linguists -specialists in this group of languages. When the models are checked on this language material, other languages from different language families will follow, in order to guarantee the general (as opposed to language specific) validity of the models under study.

Possible challenges:

1. It is too difficult to find a linguistically reasonable interpretation of the parameters of the NHGD.

While such a situation cannot be a priori excluded, we believe we can find at least significant correlations with entropy, moments, cumulants, or other characteristics of the NHGD.

1. Andrij Rovenchak, one of planned coauthors of the monograph on the MAL, will not be able to participate because of the war in Ukraine.

This is a factor which we are not able to influence at all. We sincerely hope it will not happen, but we must consider it a real risk. In the very worst case, the collective of coauthors will have to be smaller. In such a case, writing the monograph will be more difficult, but not impossible, and some delays can be expected.

We will do our best to keep the project as open as possible. Data will be shared unless copyright will forbid us from publishing them on webpage or storing them in a feely accessible repository (we plan to add new data with which no problems of this type are expected, but data taken from older books can be tricky in this aspect). Inasmuch the project budget and journal fees will allow it, we plan to publish papers with open access.

The project is obviously of an interdisciplinary character. It has a potential to attract new researchers to the study of these and other closely related topics. It is quite probable that new contacts will be established, which can lead to an international cooperation in a greater extent.

The applicant will follow principles of equity, personnel decisions will be based strictly on aspects relevant to this project.

1.4 EXCELLENCE OF THE RESEARCHER

The applicant works as a research fellow at the Mathematical Institute of Slovak Academy of Sciences (Bratislava, Slovakia) and as an associate professor at the Department of Mathematics of the Constantine the Philosopher University in Nitra (Slovakia). Until recently, he worked as an an associate professor at the Department of Applied Mathematics and Statistics of the Comenius University in Bratislava (Slovakia). He also taught one academic year at the Higher College of Technology in Muscat (Oman) and worked as a researcher at the University of Graz (Austria) and at the Palacky University (Olomouc, Czech Republic).

His research can be divided into two main areas.

First, the applicant studies probability distributions, their transformations, limit properties, systemization (distribution families), and statistical inference (estimations of their parameters).

Second, the applicant focuses on mathematical modelling of linguistic data, both at a theoretical level (language laws, properties common for all languages, specifics of particular languages or language families, diachronic development of language properties) and in applications (text classification. authorship attribution).

These two areas are not so far away from each other, as many properties of linguistic units can be modelled by discrete probability distributions.

The applicant is a co-author of three monographs, and co-editor of six edited volumes. He published 84 peer-reviewed scientific papers (47 of them are included in Web of Science and/or Scopus) as well as nine book reviews. His publications were cited 602 times (402 of the citations can be found in Web of Science and/or Scopus).

Curriculum Vitae

**Personal information**

First and last name: Ján Mačutek

Identifier: 0000-0003-1712-4395 (ORCID)

[26632109800](https://www.scopus.com/authid/detail.uri?authorId=26632109800) (SCOPUS Author ID)

AAH-9490-2020 (Web of Science Researcher ID)

Date of birth: February 15, 1976

Nationality: Slovakia

**Education**

10/2003 – PhD. in probability and mathematical statistics

Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

06/2002 – Mgr. (equivalent of MSc.) in mathematics

Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

**Current position/positions**

since 01/2020 – Research Fellow

Mathematical Institute, Slovak Academy of Sciences, Slovakia

since 09/2019 – Associate Professor

Department of Mathematics, Constantine the Philosopher University in Nitra, Slovakia

**Previous positions**

06/2013 – 08/2021 – Associate Professor

Department of Applied Mathematics and Statistics, Comenius University in Bratislava, Slovakia

01/2013 – 02/2014 – Research Fellow

Department of General Linguistics, Palacky University, Olomouc, Czech Republic

09/2011 – 05/2013 – Assistant Professor

Department of Applied Mathematics and Statistics, Comenius University in Bratislava, Slovakia

01/2009 – 6/2011 – Senior Postdoc

Institute of for Slavic Studies, University of Graz, Austria

09/2006 – 12/2008 – Assistant Professor

Department of Applied Mathematics and Statistics, Comenius University in Bratislava, Slovakia

10/2005 – 07/2006 – Lecturer

Higher College of Technology, Muscat, Oman

11/2001 – 09/2005 – Assistant Professor

Department of Applied Mathematics and Statistics, Comenius University in Bratislava, Slovakia

**Scholarships and awards**

2010 – **OEAD-SAIA Ernst Mach Grant** (Austria – Slovakia)

2009 – 2010 – **FWF Lise Meitner Scholarship** (Austria)

Student and post-docs supervision (if applicable)

since 2022 – one PhD student (currently in the 2nd year)

Mathematical Institute, Slovak Academy of Sciences

2014 – 2021 two PhD students (both successfully defended their dissertation theses)

Department of Applied Mathematics and Statistics, Comenius University in Bratislava

since 2020 – supervisor of 3 diploma students and 1 bachelor student

Department of Mathematics, Constantine the Philosopher University in Nitra, Slovakia

2004 – 2020 supervisor of 21 diploma students and 24 bachelor students

Department of Applied Mathematics and Statistics, Comenius University in Bratislava

**Teaching activities (if applicable)**

since 2001 many courses and practicals on probability theory, mathematical statistics, applied statistics, computer statistics, and foundations of mathematics (Constantine the Philosopher University in Nitra; Comenius University in Bratislava; Higher College of Technology, Muscat, Oman)

**Organisation of scientific meetings (if applicable)**

* **Member of the program committee:**

QUALICO (2023, Lausanne, Switzerland)

SyntaxFest (2021, Sofia, Bulgaria, online)

SyntaxFest (2019, Paris, France)

QUALICO (2014, Olomouc, Czech Republic)

QUALICO (2012, Belgrade, Serbia)

* **Member of the organizing committee:**

ProbaStat 2002

ProbaStat 2015

**Institutional responsibilities (if applicable)**

since 2016 – member of the Council for PhD study program (Faculty of Arts, University of Ostrava,

Czech Republic)

1997 – 1999, 2002 – 2005, 2014 – 2019

member of the Academic Senate of the Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

deputy of the Senate Chair in 2016 – 2019

**Reviewing activities (if applicable)**

* **editor-in-chief**

since 2021 – Glottometrics (<https://glottometrics.iqla.org/>)

indexed in Web of Science and SCOPUS

* **member of editorial board**

since 2008 – Journal of Quantitative Linguistics

(<https://www.tandfonline.com/journals/njql20>)

indexed in Web of Science and SCOPUS

since 2008 – Glottotheory (<https://www.degruyter.com/journal/key/glot/html>)

indexed in SCOPUS

since 2012 – Journal of Language Modelling (<https://jlm.ipipan.waw.pl/index.php/JLM>)

indexed in SCOPUS

* **occasional reviews**

Acta Mathematica Universitatis Comenianae

Acta Universitatis Palackianae Olomucensis Facultas Rerum Naturalium Mathematica

Applications of Mathematics

Communications in Statistics - Theory and Methods

Complexity

Computer Speech and Language

Corpus Linguistics and Linguistic Theory

Entropy

Journal of Physical Studies

Language Sciences

Linguistica Brunensia

Linguistics: An Interdisciplinary Journal of the Language Sciences

Linguistics Vanguard

Mathematica Slovaca

Mathematics (MDPI)

Physica A: Statistical Mechanics and its Applications

PLOS ONE

Quality & Quantity

Russian Linguistics

Scientometrics

Slovo a slovesnost

Statistical Papers

Tatra Mountains Mathematical Publications

Text & Talk

* **Science Fund of the Republic of Serbia**

reviews of two project applications

**Memberships of scientific societies (if applicable)**

International Quantitative Linguistics Asociation

since 2007 – member

2009 – 2016 – treasurer

since 2018 – member of the council of the association

**Major collaborations (if applicable)**

* Radek Čech

Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech

Republic

* Emmerich Kelih

Department of Slavonic Studies, University of Vienna, Austria

* Michaela Koščová

Mathematical Institute, Slovak Academy of Sciences

* Miroslav Kubát

Department of Czech Language, Faculty of Arts, University of Ostrava, Czech

Republic

* George Mikros

College of Humanities oand Social Sciences, Hamad bin Khalifa University, Doha, Qatar

* Andrij Rovenchak

Department of Theoretical Physics, Faculty of Physics, Ivan Franko National University, Lviv, Ukraine

* Gejza Wimmer

Mathematical Institute, Slovak Academy of Sciences

**Overview of the researcher’s most important projects in the last 5 years** (max. 5)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project name/identification** | **Source of funding** | **Budget (EUR)** | **Project period** | **The role of the researcher in the project** |
| APVV-21-0216: Advanced mathematical and statistical methods for measurement and metrology | Slovak Research and Development Agency | 248 482 | 2022-2025 | Research team member |
| SK-AT-20-0003: Frequency and declensional morphology in Slavic languages  (Russian, Slovak, Slovene) | Slovak Research and Development Agency | 1 862 | 2021-2023 | Principal investigator for SAS |
| Operational Programme  Integrated Infrastructure (OPII) for the project 313011BWH2: “InoCHF – Research and development in the field of innovative technologies in the management of patients  with CHF”, co-financed by the European Regional Development Fund | European Regional Development Fund | 1 695 449 | 2022-2023 | Research team member |
| VEGA 2/0096/21: Probability distributions and their applications in modelling and testing | Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences | 18 950 | 2021-2023 | Principal investigator |
| VEGA 2/0054/18: New statistical methods for special classes of probability distributions and their applications | Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences | 38 539 | 2018-2021 | Research team member |

**Overview of the researcher’s most important outputs** (max. 5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Output name/identification** | **Type of output** *(e.g., publication, dataset, software, patent, service, product, etc.)* | **Short description** | **The role of the researcher** |
| Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20(3), 227‑240. | publication | It is a well-known fact that statistical tests tend to reject almost any hypothesis if the sample size is very large. Linguistic corpora often contain millions of items, this using p-values to test hypotheses is very problematic. Some alternative approaches (at least rules of thumb) are suggested for such situations. | All parts, together with the co-author |
| Mačutek, J., Wimmer, G., Koščová, M. (2022). On a parametrization of partial-sums discrete probability distributions. *Mathematics* 10(14), 2476. | publication | Partial summations are special transformations of discrete distributions. In previous research it was proved that for each partial summation there is one and only one distribution which is invariant with respect to the summation. In this paper, summations are parametrized. It is shown that parametrized partial summations divide discrete distributions into two groups. Distribution from one group remain invariant, but those from the other “lose” their invariancy and are transformed into two-parametric distributions. | All parts, together with the co-authors |
| Koščová, M., Harman, R., Mačutek, J. (2020). Iterated partial summations applied to finite-support discrete distributions. *Mathematica Slovaca* 70(2), 489-496. | publication | Partial summations can be repeatedly applied to a discrete distribution. This paper proves that, under certain conditions, iterative partial summations have their limit distributions. | All parts, together with the co-authors |
| Mačutek, J. (2008). A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Reports in Physics* 60(3), 501-509. | publication | In linguistic research, the negative hypergeometric distribution is used as a model for ranked frequencies of phonemes and graphemes. However, this distribution is derived form a binary urn model (balls of two colours are drawn form an urn), which makes the distribution difficult, if not impossible, to interpret in the linguistic context (where binarity is out of the question). In this paper an alternative model, which avoids these problems, is presented. . | All parts |
| Mačutek, J. (2022). Why do parameter values in the Zipf-Mandelbrot distribution sometimes explode? *Journal of Quantitative Linguistics* 29(4), 413-424. | publication | The Zipf-Mandelbrot distribution is a ubiquitous model for rank-frequency distributions in many branches of science (linguistics, biology, geography, etc). While it almost always fits the data very well, its parameters are virtually uninterpretable in some cases (they attain very high values, and in addition they are extremely sensitive to the estimation method chosen). The paper shows that this happens if both parameters diverge to infinity, but their ratio remains constant. Under these conditions, the Zipf-Mandelbrot distribution converges to the geometric distribution. | All parts |

Research results achieved by the applicant include:

* Study of partial-sums discrete probability distributions. Partial summations take one discrete distribution (parent) and transform it into another one (descendant). Relations between characteristics (probability generating functions, moments) of parent and descendant distributions were derived. It was shown that for each partial summation there is one and only one distribution which is invariant with respect to the summation, and that every two discrete distributions are connected by a partial summation. A parametrization of partial summations can lead to a different behaviour of parent and descendant distributions, which was also studied. Some conditions for the existence of limit distributions for iterated partial summations were found. A generalization for multivariate discrete distributions was presented.
* A proof that the Zipf-Mandelbrot distribution converges to the geometric distribution if its parameters diverge to infinity, but their ratio remains constant. This convergence explains a problematic behaviour of the Zipf-Mandelbrot distribution parameters in some cases.
* A simple parameter estimation in the Schröter family of discrete distributions. These distributions are used in insurance mathematics, and previously suggested estimations are not explicit - they are results of iterative numerical algorithms which can be computationally problematic, as they can involve e.g. inversions of sparse matrices. The suggested simple estimation have themselves their share of problems (based on a simulation studies, it seems that the estimation of one of the parameters in not consistent), but it is very easy to evaluate, and it can serve as a reasonable initial value for more sophisticated numerical procedures.
* A suggestion how to evaluate goodness-of-fit of models when sample sizes are huge (which is often the case in linguistics, where language corpora contain more and more texts). The classical statistical tests reject practically any null hypothesis if the sample size is large enough. Thus, it was suggested to use the p-values for relatively small samples, and to use a different approach to decision-making for large samples.
* Mathematical models for some linguistic laws were studied, where linguistic laws are understood as statistical tendencies as opposed to prescriptions. The applicant focused mainly on the Menzerath-Altmann law (which says that longer language units consist, on average, of shorter parts, e.g. longer words tend to contain shorter syllables than short words) and on the model for phoneme and grapheme frequencies (the negative hypergeometric distribution is a candidate for a general model, but it has its share of problems, mainly the unclear interpretation of its parameters). Statistical problems related to this area which were at least partially solved were e.g. parameter estimation in the models, statistical tests for differences of parameter values in different languages, text types, etc.,
* New statistical methods for classification of languages and texts were suggested and applied. In the context of the Czechoslovak and Czech history, these methods can e.g. capture stylistic differences between speeches of democratic and communist presidents.

•

1.5 EXCELLENCE OF THE APPLICANT/HOST ORGANISATION

Mathematical Institute of the Slovak Academy of Sciences is a scientific institute focused mainly on basic research in mathematics and theoretical informatics. The Institute has a long tradition in several important branches of pure and applied mathematics. It participated in a number of successful projects in both basic and applied research, including projects of Frame Projects of EU, Structural projects of EU, and projects of domestic research agencies APVV and VEGA. The researchers of the Institute belong to the top in their research, in a world-wide context, and are engaged in multiple collaborations with experts from internationally renowned institutions. In collaboration with the Comenius University in Bratislava, the Institute organises a PhD study program, with an increasing proportion of PhD students coming from abroad. Many young scientists and students use the Slovak fellowship program SAIA for short term study stays at the institute.

There are two other researchers with similar research interest at the Institute (both Gejza Wimmer and Michaela Koščová are statisticians working, among others, also in mathematical models in linguistics). Thus, the applicant will have the possibility to consult with colleagues directly at his workplace. In addition, statisticians from the Institute closely cooperate with members of the Department of Applied Mathematics and Statistics of the Comenius University, and thus they have the opportunity to exchange ideas with a larger group of experts.

Mathematical Institute has very good research infrastructure with adequate hardware and software equipment. The library of Mathematical Institute SAS belongs to the best mathematical libraries in Slovakia, with access to many journal nets of the world publishing houses, scientific-information, and technical databases, such as e.g. Mathematical Review, Zentralblatt Mathematik, Web of Knowledge, Scopus, etc.

## Impact

### 2.1 THE WIDER IMPACT OF THE PROJECT

The project focuses on basic research, but the planned results have potential to be used in automatic text analysis, authorship attribution, automatic assessing of text readability, etc.

An analysis of language laws contributes to a better understanding of human language, its use, cognitive processing, and in a broader sense to a better understanding of any information exchange. E.g., the MAL was observed not only in human language, but also in animal communication, in the DNA structure, in music – all these domains, although being quite distant form each other, have an exchange of information as their “common denominator”.

Results of the project will be disseminated in international journals and presented at international conferences. Thus, new contacts can be established, which can provide new ideas to both sides and can lead to a more intense international cooperation with other institutions and their employees. Given that this topic is known to be genuinely interdisciplinary (typically, conferences are attended by linguistics, mathematicians, physicists, computer scientists etc.), there is a relatively high probability of finding people with complementary skills, which can make a cooperation more efficient.

For the applicant, the project will be a huge opportunity for a further personal development and for building a research team which can remain functional also after this project is finished. Also the host institution will benefit from the results (directly and in near future by an increased number of publications, in a longer perspective the project can attract potential PhD students and new researchers).

As major outcomes, it is expected that one monograph will be published at de Gruyter and at least four papers will be published in prestigious journals. Minor outcomes include two bibliographies, which will also the applicant and the host institution more visible.

* Potential obstacles to the planned impact of the project

Obstacles to gains of the applicant, his research team, and the host institution are not expected. As far as potential applications are concerned (authorship attribution and similar), it can happen that AI-based language models like eg. ChatGPT will achieve so good results that they will, at least temporarily, overshadow methods based on theoretical results.

### 2.2 MEASURES TO MAXIMISE IMPACT – DISEMINATION AND COMMUNICATION, EXPLOITATION OF RESULTS

The outcomes of the project will be published as a monograph at de Gruyter and in prestigious journals (with open access inasmuch possible). In addition, the applicant will attend international conferences (both statistical ones as well as those focusing on quantitative linguistics) which will further increase the impact.

Data and two bibliographies will be made freely available (with possible exception of data covered by copyright) also after the project completion:

Results will be also made more known in the scientific community when the applicant will visit research institutions abroad (presentations at local seminars are expected).

## Implementation

3.1 PROJECT PLAN AND DELIVERABLES

The project is divided into six work packages, each corresponding to one objective:

1. MAL – state-of-the-art
2. MAL – general model
3. MAL - monograph
4. Models for phoneme and grapheme frequencies – state-of-the-art
5. NHGD, parameters, goodness-of-fit
6. Dissemination

The project is planned for 24 months. The tasks planned within the project are accordingly included in work packages. The package WP1 and WP4 are necessary prerequisites for continuation of the research. The package WP2 focuses on the analyses of different models for the MAL, and choosing the best one, or deriving a new one if none from the literature has desirable properties. WP3 involves writing a monograph on the MAL, exploiting results from WP2. The package WP5 will concentrate on attempts to interpret the parameters of the NHGD, a model for ranked frequencies of phoneme and grapheme frequencies. Finally, WP6 contains work on dissemination of the results achieved and checking how the project develops, as well as on identifying potential threats and, if needed, on undertaking necessary steps to successfully complete the project.

Timeline of the project:

WP1 – months 1-3

WP2 – months 2-6

WP3 – months 4-16

WP4 – months 12-14

WP5 – months 17-24

WP6 – months 2-24

More details on project work packages can be found below.

3.1.1 Work packages

|  |  |
| --- | --- |
| Work package number | WP1 |
| Title of the work package | MAL – state-of-the-art |
| **Start of implementation of the work package (Mx Month)** | M1 |
| End of implementation of the work package (Mx month) | M3 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 2 person months |
| **Personnel costs (in EUR)** | 13646 EUR  (9646 EUR researcher salary + 4000 Eur salary for 1 R2 person month for a research team, help with literature) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 3705,75 EUR  (2000 EUR research costs + 1705,75 indirect costs) |
| Objectives | |
| Study of the literature on the MAL. Obtaining older papers which are not available online. Creating the bibliography.on the MAL. | |
| Description of the work package | |
| Task 1.1: Collecting publications on the MAL which the applicant does not have yet.  Task 1.2: Study of publications related to the MAL that are unknown to the applicant so far.  Task 1.3: Creating a bibliography of papers on the MAL and a database of the data used in the literature.  Task 1.4: Creating a list of mathematical models of the MAL. | |
| Deliverables | |
| D1: **Bibliography**  A bibliography containing papers on the MAL | |

|  |  |
| --- | --- |
| Work package number | WP2 |
| Title of the work package | MAL – general model |
| **Start of implementation of the work package (Mx Month)** | M2 |
| End of implementation of the work package (Mx month) | M6 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 3 person months |
| **Personnel costs (in EUR)** | 21469 EUR  (14469 EUR researcher salary + 7000 EUR for 1.5 person month, R2 – consultation with linguists, writing a paper) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 6183,63 EUR  (3500 EUR research costs + 2863,63EUR indirect costs) |
| Objectives | |
| A general mathematical model for the MAL | |
| Description of the work package | |
| Task 2.1: Analysis of models from Task 1.4.  Task 2.2: Choosing the model(s) with a good fit and desirable properties (which correspond to the nature of data they are supposed to model).  Task 2.3: Deriving a new model if necessary.  Task 2.4: Attempt to the interpret parameters of the model. | |
| Deliverables | |
| D2 2 research papers  Two publications (an introduction of the general model of the MAL, its application to data) | |

|  |  |
| --- | --- |
| Work package number | WP3 |
| Title of the work package | MAL - monograph |
| **Start of implementation of the work package (Mx Month)** | M4 |
| End of implementation of the work package (Mx month) | M16 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 10 person months |
| **Personnel costs (in EUR)** | 70230 EUR  (48230 EUR researcher salary + 22000 EUR for 5 person month R3, consultation with linguistics, data preparation, writing the monograph) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 19278,75 EUR  (10500 research costs + 8778,75 EUR indirect costs) |
| Objectives | |
| Writing a monograph which will be published at de Gruyter.. | |
| Description of the work package | |
| Task 3.1: Summarization of results on the MAL published so far and our results from Objective 2.  Task 3.2: Preparing the structure of the book.  Task 3.3: Writing the monograph, closely cooperating with the coauthors | |
| Deliverables | |
| D3: **conference presentation**  Results obtained will be presented at an international conference.  D4 **monograph**  A monograph will be published at de Gruyter | |

|  |  |
| --- | --- |
| Work package number | WP4 |
| Title of the work package | Models for phoneme and grapheme frequencies – state-of-the-art |
| **Start of implementation of the work package (Mx Month)** | M12 |
| End of implementation of the work package (Mx month) | M14 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 2 person months |
| **Personnel costs (in EUR)** | 13646 EUR  (9646 EUR researcher salary + 4000 EUR FOR 1 R2 person month for a research team, help with literature) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 3705,75 EUR  (2000 EUR research costs + 1705,75 EUR indirect costs) |
| Objectives | |
| A mathematical model for phoneme and grapheme rank frequency distributions - the state-of-the-art analysis | |
| Description of the work package | |
| Task 4.1: Collecting the relevant literature published after 2011 (the applicant has the relevant older publications, they were collected during his postdoc stay the University of Graz).  Task 4.2: Study of the publications from Task 4.1.  Task 4.2: Creating a bibliography.  Task 4.3: Creating a list of models for phoneme and grapheme frequencies. | |
| Deliverables | |
| D5: **bibliography**  Bibliography containing papers on the models for ranked frequencies of phonemes and graphemes | |

|  |  |
| --- | --- |
| Work package number | WP5 |
| Title of the work package | NHGD, parameters, goodness-of-fit |
| **Start of implementation of the work package (Mx Month)** | M17 |
| End of implementation of the work package (Mx month) | M24 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 3 person months |
| **Personnel costs (in EUR)** | 21469 EUR  (14469 EUR researcher salary + 7000 EUR for 1.5 person month, R2 – consultation with linguists, writing a paper) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 6183,62 EUR  (3500 EUR research costs + 2683,62 EUR indirect costs) |
| Objectives | |
| The NHGD, its parameters, and its goodness-of-fit | |
| Description of the work package | |
| Task 5.1: Attempt to obtain an interpretation of parameters using the beta-binomial distribution.  Task 5.2: If the attempt from Task 5.1 is not successful, other characteristics of the NHGD will be applied, e.g. its entropy.  Task 5.3: Developing a rule-of-thumb criterion which is general for all sample sizes  Task 5.4: Fitting the NHGD to various datasets, evaluating its goodness-of-fit | |
| Deliverables | |
| D6: **2 journal publications**  2 papers in scientific journals – one focusing on the criterion for goodnes-of-fit which is applicable to samples regardless of their size, the other on application of the NHGD to data  D7 : **conference presentation**  Results obtained will be presented at an international conference. | |

|  |  |
| --- | --- |
| Work package number | WP6 |
| Title of the work package | Dissemination |
| **Start of implementation of the work package (Mx Month)** | M2 |
| End of implementation of the work package (Mx month) | M24 |
| **Involvement (expressed in Person Months)** | Ján Mačutek - 4 person months |
| **Personnel costs (in EUR)** | 23292 EUR  (19292 EUR researcher salary + 4000 EUR for 1 person month, R2 – consultation with linguists, writing a paper) |
| Other eligible costs, excluding personnel costs (in EUR excluding VAT) | 7411,5 EUR  (4500 EUR research costs + 2911,5 EUR indirect costs) |
| Objectives | |
| Monitoring the progress of the project and dissemination of its results | |
| Description of the work package | |
| Task 6.1: Monitoring the progress of the project  Task 6.2: If there is a risk of a significant delay, appropriate steps will be undertaken.  Task 6.3: Dissemination of results  Task 6.4: Research proposal | |
| Deliverables | |
| D8 **Interim report**  Report summarizing results achieved in the middle of the duration of project, their evaluation, possible delays or deviations to the plan, together with their justification and comparison between obtained and expected results.  D9 **Final report**  Report summarizing all results obtained during implementation of the project and comparison between achieved and expected results.  D10 **New research project (grant) application**  Research proposal for a new project (grant) , which will involve the applicant. | |

3.1.2 List of work packages:

|  |  |  |  |
| --- | --- | --- | --- |
| Work package number | Title of the work package | **Start of activities** | **End of activities** |
| WP1 | MAL – state-of-the-art | M1 | M3 |
| WP2 | MAL – general model | M2 | M6 |
| WP3 | MAL - monograph | M4 | M16 |
| WP4 | Models for phoneme and grapheme frequencies – state-of-the-art | M12 | M14 |
| WP5 | NHGD, parameters, goodness-of-fit | M17 | M24 |
| WP6 | Dissemination | M2 | M24 |

3.1.3 List of deliverables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Deliv. num. | Deliverable | WP# | Type | Access and dissemination | Method of verification | Delivery |
| D1 | Bibliography - MAL | WP1 | bibliography | public | Accesibility check | M3 |
| D2 | Research papers | WP2 | publication | public | Peer review | M6 |
| D3 | Conference contribution | WP3 | publication | public | Peer review | M6 |
| D4 | Monograph | WP3 | publication | public | Peer review | M16 |
| D5 | *Bibliography – models for*  *rank-frequency distributions* | WP4 | bibliography | public | Accesibility check | M14 |
| D6 | *Research papers* | WP5 | publication | public | Peer review | M22 |
| D7 | Conference contribution | WP4 | publication | public | Peer review | M21 |
| D8 | Interim report | WP6 | report | non-public | Progress  evaluation | M12 |
| D9 | Final report | WP6 | report | non-public | Project evaluation | M24 |
| D10 | Project proposal | WP6 | proposal | non-public | Proposal  evaluation | M24 |

3.1.4 List of milestones:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone number | Milestone | Work package number | Method of verification | Expected time to reach the milestone (project month) |
| MS1 | Deriving a general model for the MAL | WP2 | peer review | M6 |
| MS2 | Submitting the final version of the monograoh to the publishing house | WP3 | peer review | M16 |
| MS3 | A criterion for goodness-of-fit applicable regardless of sample size | WP5 | peer review | M22 |

3.2 IMPLEMENTATION RISKS AND PROPOSED MEASURES

3.2.1 Risks of implementation:

|  |  |  |
| --- | --- | --- |
| **Description of the risk of implementation (severity)** | **Work package** | Proposed measures for risk mitigation or elimination |
| Failure at parameter interpretation  of NHGD (medium) | W5 | Focusing on correlations between parameter values and other characteristics of the NHGD, eg. entriopy. |
| Data are not reliable (low) | W5 | Looking for data from other languages, establishing collaboration with experts on other language families. |
| A planned coauthor of the monograph unavailable  because of the war in Ukraine (high) | W2, W3 | In the worst-case scenario, writing the monograph with a smaller team – some delays can occur in such a case. |
| Publication delays (medium) | W5 | Diversification of publications, contacting editors. |

3.3 OPERATIONAL CAPACITY OF THE APPLICANT/HOST ORGANISATION

3.3.1 Description of the research/innovation infrastructure of the applicant/host organisation that is necessary for the implementation of the project:

|  |  |
| --- | --- |
| Name of infrastructure or equipment | Short description |
| DEVANA | supercomputer |
| Access to databases | Science Direct, Scopus, WoS, SpringerLink, SpringerNATURE, JSTOR, PLOS, PNAS, Taylor & Francis, Wiley Online Library |
| Standard scientific infrastructure | Computers, internet connection |
| library |  |

3.3.2 List of the five most important projects of the applicant/host organisation and their relevance to the proposed project (in the last 5 years):

|  |  |  |
| --- | --- | --- |
| Project name/identification | Programme/scheme/grant provider | Short description |
| APVV-21-0216: Advanced mathematical and statistical methods for measurement and metrology | Slovak Research and Development Agency | Focus on theoretical methods in probability theory and mathematical statistics, applicant is the principal researcher |
| SK-AT-20-0003: Frequency and declensional morphology in Slavic languages  (Russian, Slovak, Slovene) | Slovak Research and Development Agency | Mathematical modelling of linguistic data, applicant is the principal researcher |
| VEGA 2/0096/21: Probability distributions and their applications in modelling and testing | Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences | Focus on theoretical methods in probability theory and mathematical statistics, applicant is the principal researcher |
| VEGA 2/0054/18: New statistical methods for special classes of probability distributions and their applications | Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences | Focus on theoretical methods in probability theory and mathematical statistics, applicant is a member of a research team |
| , APVV SK-SRB-2016-0021:Quantitative analysis of syllables in Slavic languages (Russian, Slovak, Serbian) | Slovak Research and Development Agency | Mathematical modelling of linguistic data, applicant is the principal researcher |

3.3.3 List of maximum five most important outputs of the applicant/host organisation relevant to the submitted project:

|  |  |  |
| --- | --- | --- |
| Output name/identification | **Type of output** | Short description |
| Wimmer, G., Altmann, G. (1999). Thesaurus of discrete univariate probability distributions. Essen: Stamm. | monograph | A dictionary of discrete distribution, with probability generating functions, relations to other distributions, and a very rich list of references |
| Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In Koehler, R, Altmann, G., Piotrowski, R.G, (eds), Handbook of Quantitative Linguistics (pp. 791-807).. Berlin: de Gruyter | book chapter | A very general model, special cases of which can be considered a mathematically formulated theory of language (language laws are special cases of this formula) |
| Mačutek, J. (2022). Why do parameter values in the Zipf-Mandelbrot distribution sometimes explode? *Journal of Quantitative Linguistics* 29(4), 413-424. | article | The Zipf-Mandelbrot distribution is a ubiquitous model for rank-frequency distributions in many branches of science (linguistics, biology, geography, etc). While it almost always fits the data very well, its parameters are virtually uninterpretable in some cases (they attain very high values, and in addition they are extremely sensitive to the estimation method chosen). The paper shows that this happens if both parameters diverge to infinity, but their ratio remains constant. Under these conditions, the Zipf-Mandelbrot distribution converges to the geometric distribution. |
| Kubát, M., Mačutek, J., Čech, R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech presidential speeches. *Digital Scholarship in the Humanities* 36(1), 138-152. | article | Application of statistical methods. It is shown using quantitative and statistical methods that democratic Czechoslovak and Czech presidents used “other language” if compared with communist presidenrts |
| Koščová, M., Mačutek, J., Kelih, E. (2016). A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies. *Journal of Quantitative Linguistics* 23(2), 177-190. | article | Introduction of a new clustering method, with an application to language classification. |