

# **Glottometrics 50 2021**

**RAM-Verlag**

**ISSN 1617-8351  
e-ISSN 2625-8226**

# Glottometrics

(Open Access Journal)

**Indexed in ESCI by Clarivate Analytics and SCOPUS by Elsevier**

**Glottometrics** ist eine regelmäßig erscheinende Zeitschrift (2 Ausgaben pro Jahr) und ist der quantitativen Analyse von Sprache und Texten gewidmet.

**Beiträge** in Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden. Glottometrics ist ein **Open Access Journal**.

**Glottometrics** is a scientific journal for the quantitative research of language and text published 2 times per year.

**Contributions** in English written in a common text processing system (preferably WORD) should be sent to one of the editors. Glottometrics is an **Open Access Journal**.

## Editorial Board

<b>G. Altmann</b> (†)	Univ. Bochum (Germany)	ram-verlag@t-online.de
<b>S. Andreev</b>	Univ. Smolensk (Russia)	smol.an@mail.ru
<b>K.-H. Best</b>	Univ. Göttingen (Germany)	kbest@gwdg.de
<b>R. Čech</b>	Univ. Ostrava (Czech Republic)	cechradek@gmail.com
<b>E. Kelih</b>	Univ. Vienna (Austria)	emmerich.kelih@univie.ac.at
<b>R. Köhler</b>	Univ. Trier (Germany)	koehler@uni-trier.de
<b>H. Liu</b>	Univ. Zhejiang (China)	lhtzju@gmail.com
<b>J. Mačutek</b>	Univ. Bratislava (Slovakia)	jmacutek@yahoo.com
<b>A. Mehler</b>	Univ. Frankfurt (Germany)	amehler@em.uni-frankfurt.de
<b>M. Místecký</b>	Univ. Ostrava (Czech Republic)	MMistecky@seznam.cz
<b>G. Wimmer</b>	Univ. Bratislava (Slovakia)	wimmer@mat.savba.sk
<b>P. Zörnig</b>	Univ. Brasilia (Brasilia)	peter@unb.br

## Editors of this issue:

**E. Kelih, J. Mačutek and R. Čech**

## Editorial and Peer Review Process

Glottometrics is a peer-reviewed scientific journal with an editorial prescreening and assessment (process of publication: 1. first Check, 2. editorial review, 3. peer Review, 4. final decision. For detailed information please see: <https://www.ram-verlag.eu/journals-e-journals/glottometrics/editorial-and-peer-review-process/>)

<https://www.ram-verlag.eu/journals-e-journals/glottometrics/>

Die Deutsche Bibliothek – CIP-Einheitsaufnahme  
Glottometrics. 50 (2021), Lüdenscheid: RAM-Verlag, 2021. Erscheint regelmäßig.  
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse  
<https://www.ram-verlag.eu/journals-e-journals/glottometrics/> verfügbar.  
Bibliographische Deskription nach 50 (2021)  
**online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)**

# Contents

## **Emmerich Kelih, Radek Čech, Jan Mačutek**

Happy Birthday Glottometrics – On the Occasion of the 50th Issue and 20th Anniversary 1–3

## **Sergey Andreev, Gejza Wimmer, Emmerich Kelih**

Gabriel Altmann (1931–2020) 4–8

## **Tayebeh Mosavi Miangah, Relja Vulcanović**

The Ambiguity of the Relations between Graphemes and Phonemes in the Persian Orthographic System 9–26

## **Minna Bao, Saheya Brintag, Dabhurbayar Huang**

English Loanwords in Mongolian Usage 27–41

## **Lu Wang, Yahui Guo, Chengcheng Ren**

A Quantitative Study on English Polyfunctional Words 42–56

## **Alfiya Galieva, Zhanna Vavilova**

Initial and Final Syllables in Tatar: from Phonotactics to Morphology 57–75

## **Alexandr Osochkin, Xenia Piotrowska, Vladimir Fomin**

Automatic Identification of Authors' Stylistics and Gender on the Basis of the Corpus of Russian Fiction Using Extended Set-theoretic Model with Collocation Extraction 76–89

## **Emmerich Kelih, Peter Grzybek**

Glottometrics, 31–50: Bibliography 90–96

## Editorial Note:

# Happy Birthday *Glottometrics* – On the Occasion of the 50<sup>th</sup> Issue and 20<sup>th</sup> Anniversary

Exactly 20 years ago the journal *Glottometrics* was founded by Gabriel Altmann, the well-known nester of quantitative linguistics. He was not only the founding editor, but he also continued to be the editor-in-chief of this journal (published by RAM-Verlag) for the next 20 years. The title of the journal can be seen in close relation to *Glottometrika*, one of the subseries of the *Quantitative Linguistics* book series (published since 1978, in the beginning by Brockmeyer in Bochum, then by WVT Trier and now by de Gruyter). The initial idea of *Glottometrics* was to be a platform for quantitative linguistics and related topics. Indeed, quantitative and statistical methods are becoming more and more popular in linguistics and literary studies. Thus, it is no surprise that since its launch, nearly 400 research articles (including reviews, project descriptions, software presentations etc.) have been published in *Glottometrics*. This is without any exaggeration an occasion for celebrating, but also an occasion for reflecting on the past and planning for the future.

The unique profile of *Glottometrics* is, without any doubt, Gabriel Altmann's merit. It is one of the journals (in addition to the *Journal of Quantitative Linguistics*, and in some respects also *Glottology*) focused on the publication of research articles on quantitative linguistics in a narrow sense. At this point it also has to be remembered that this comfortable situation for scholars is by no means self-evident. We are sure that many colleagues will remember the times when the application and the use of statistical methods in linguistics was not really accepted, or at least it was considered something exotic or a more or less interesting endeavour. Historically seen there have been many journals (cf. the impressive list given by Rostin 2016) open-minded towards statistical methods in linguistics, but many of them were only published for a short period (cf. for instance *SMIL Quarterly – Statistical Methods in Linguistics*, being published only for a short time in the 1960s by Scandinavian colleagues), or published quantitative linguistics topics only sporadically (cf. for instance *Prague Studies in Mathematical Linguistics*, a publication platform for quantitative and algebraic approaches). In this respect, the possibilities for the publication of quantitative linguistic topics significantly increased with the establishment of *Glottometrics*. Since its beginning, it has been a compact and open place for linguistic analyses which used mathematical and statistical methods, while also bearing theoretical perspectives in mind.

Taking a brief look at the articles published in *Glottometrics* in the past years, some clear focuses can be observed, which also give in some respect information about the present shape of quantitative linguistics. One remarkable focus is (in a narrow sense) on three kinds of linguistic laws:

- (1) Functional laws (among them the most prominent one being the Menzerath law, “The longer the constituents, the shorter its constituents),
- (2) Distributional laws (among them the most prominent one being the Zipf law, which is in fact an umbrella term for different statistical interrelations between frequency and rank of (not only linguistic) units, word length and frequency, frequency and polysemy etc.) and
- (3) Developmental laws (such as the Piotrowski law, which contains information about the regular behaviour of language dynamics and change processes (cf. Köhler/Altmann 2005)).

In this respect, mathematical modelling of linguistic phenomena and related problems, as well as theoretical and methodological challenges, has been of outstanding interest for recent quantitative linguistics. Beyond this, there are of course many other areas and branches of

linguistics where quantitative and statistical methods can and should be applied. Based on the publication profile<sup>1</sup> of *Glottometrics*, quantitative and author-specific stylistics (stylometrics), metrical studies, phonological and phonosemantic issues (syllabic structure of languages), vocabulary studies, problems of morphosyntax (e.g. studies about adnominals, compound studies) and quantitative analysis of semantic characteristics (polysemy, synonymy) have been some of the core areas of *Glottometrics* of the past years. One further focus is the history of quantitative linguistics, where many interesting bibliographical sketches and background information of people working “quantitatively” from many different countries have been given.

An empirically based taxonomy of objects studied in *Glottometrics* is given by Lin/Liu (2017: 17), which provides a much more representative overview than the one above:

- (1) System: laws in language systems, properties of a system like economy or symmetry, and relations of levels or elements within a system;
- (2) Phonology and phonetics: phonemes, prosody in literary works, sound symbolism;
- (3) Morphology, lexicology and lexicography: word class, word frequency, word length, type–token relation, entropy, polysemy and synonymy; affix, borrowing and compounding;
- (4) Sentence and syntax: sentence length, syntactic complexity, syntactic network;
- (5) Semantics and pragmatics: lexical semantics, information content in communication;
- (6) Text: text genre and style, translation, text processing;
- (7) Dialectology, typology, diachrony, psycholinguistics, language learning, computational linguistics;
- (8) Script: script complexity, grapheme–phoneme relationship, letters;

As regards the language of publication for *Glottometrics*, a clear trend towards English (cf. Lin/Liu (2017: 6) can be observed, and in the last issues almost all articles appeared in English. Being initially a Germany-based journal, English played an important role from the beginning, thus reflecting the ongoing internationalization of linguistics in general. This is also related to the fact that authors publishing in *Glottometrics* are located all over the world, with a remarkable increase in quantitative linguistics activities in Asia (in particular in China).

Quantitative linguistics is, like every other scientific discipline, subject to ongoing dynamics and changes, and as the scope and publication profile of *Glottometrics* shows, it is upon the quantitative linguistics community to shape the future development of the journal and its contents. It is one of the duties of the editorial team to take note of recent trends in academic publishing, like the required inclusion in relevant citation indexes (*Glottometrics* has been indexed in Emerging Sources Citation Index (ESCI) since 2015, and in Scopus since 2017), an attractive digital form of publishing, etc. The most significant change expected for the near future seems to be a diligent restyling of the layout and the transformation of *Glottometrics* into a fully fledged open-access journal (and listing in <https://doaj.org/>). This is all aimed at facilitating the continuity of *Glottometrics* as a high-quality peer-reviewed platform for research in quantitative linguistics. One also has to remember at this point that *Glottometrics* was initially (also) founded to be a publication platform for young and early career scientists, an idea which will continue to be supported by the future editorial team.

*Glottometrics – Ad multos annos!*

---

<sup>1</sup> Based on the bibliography of issues 31–50, cf. Kelih (2021), and the bibliography of issues 1–30, cf. Grzybek/Kelih (2015).

## References

- Grzybek, P.; Kelih, E.** (2015). Glottometrics 1–30: Bibliography. *Glottometrics* 31, 89–102.
- Kelih, E.** (2021). Bibliography 31–50. *Glottometrics* 50, 90-96.
- Köhler, R.; Altmann, G.** (2005). Aims and Methods of Quantitative Linguistics. In: Gabriel Altmann, Viktor Levickij and Valentyna Perebyjnis (eds.): *Problemy kvantytatyvnoi lingvistyky. Problems of Quantitative Linguistics*. Černivci: Ruta, 12–41.
- Lin, Y., Liu, H.** (2017). Bibliometric Analysis of Glottometrics. *Glottometrics* 39, 1–37.
- Rostin, T.** (2016). List of Journals Containing Contributions to Quantitative Linguistics. In: *Glottometrics* 33, 73–100.

Written by the editors of this issue:

*Emmerich Kelih*<sup>2</sup>


*Radek Čech*<sup>3</sup>

*Jan Mačutek*<sup>4</sup>

---

<sup>2</sup> University of Vienna, [emmerich.kelih@univie.ac.at](mailto:emmerich.kelih@univie.ac.at),  <http://orcid.org/0000-0002-8315-8916>

<sup>3</sup> University of Ostrava, [cechradek@gmail.com](mailto:cechradek@gmail.com),  <http://orcid.org/0000-0002-4412-4588>

<sup>4</sup> Slovak Academy of Sciences, Bratislava, and Constantine the Philosopher University in Nitra, [jmacutek@yahoo.com](mailto:jmacutek@yahoo.com),  <http://orcid.org/0000-0003-1712-4395>

## Gabriel Altmann (1931–2020)

It is with great sadness that we announce that on 2 March 2020 in Lüdenscheid, Germany, peacefully passed Gabriel Altmann, world-renowned linguist and mathematician.

Over the previous year Gabriel had been seriously ill, but his friends and colleagues, though understanding the seriousness of the situation, still continued to hope that he would recover. The news of his death came as a shock and the thought that he is no longer with us is hard to accept.

Gabriel was born on 24 May 1931 in the Slovak village of Poltár, where his father worked as a general practitioner. After basic school, he visited the grammar school in Lučenec and passed his final examinations in 1951. In the times of ‘real socialism’ it was, in particular for people coming from an academic family, not always easy to realize one’s professional wishes and desires, but Gabriel got the opportunity to study Indonesian linguistics and Japanese philology at the Charles University in Prague from 1953 to 1958. Here he came into contact with Vladimír Skalička, a well-known general linguist and typologist, who certainly influenced the way of linguistic thinking of Gabriel. After his PhD, in 1964 he received the state doctorate at the Czechoslovakian Academy of Sciences with his habilitation *Kvantitatívne štúdie indonezistiky* (Quantitative Studies in Indonesian Philology). A look at the list of publications he wrote in the 1960s reveals his primary scientific interests, namely the quantitative analysis of languages and text, in particular phonetic/phonological issues (partly with a typological perspective in cooperation with his younger colleague Viktor Krupa) and some quantitative studies of poetry and rhyme structures. In the years from 1960 until 1968 he worked at the Institute of Oriental Studies of the Slovakian Academy of Sciences (Ústav orientalistiky SAV) in Bratislava (former Czechoslovakia, now Slovakia).

A grant from the Alexander von Humboldt Foundation enabled him to visit the Institute of Phonetics at the University of Cologne from 1968 to 1969. This coincided with the Prague Spring, which ended with the invasion of troops of the Soviet Union and other Warsaw Pact members, radically interrupting all processes of political liberalization in Czechoslovakia. Gabriel decided (along with his family) to start a new life in the Western world and already<sup>5</sup> in 1970 he accepted the position of a researcher in the project Automatic Syntax Analysis of German at the Institut für Deutsche Sprache in Mannheim. Then, again supported by the Alexander von Humboldt Foundation, he was appointed visiting professor for quantitative linguistics at the Department of Linguistics (Ruhr University Bochum). In 1971, he received his German *venia legendi*, this time with his *Habilitationsschrift* entitled *Introduction to Quantitative Phonology*. After that time, he worked as a full professor for mathematical linguistics at this institute until he retired in 1996.

Answering to the demand to overcome the purely descriptive phase in philology, Gabriel introduced exact mathematical methods into linguistic analysis, thus becoming one of the founders of the new stage of quantitative linguistics. This laid the basis for explication of fundamental scientific terms such as ‘theory’, ‘law’, ‘hypothesis’ or ‘explanation’ within a linguistic framework, where these concepts have become blurred and misused over the decades. The construction of a linguistic theory – in the strict sense of the philosophy of science – was for Gabriel the ultimate aim of the study of text and language.

Due to his profound mathematical and statistical background, which is reflected by his publications in mathematical journals, he carried out a number of projects whose results are highly important for both the theoretical development and practical needs of linguistics. Among

---

<sup>5</sup> The next paragraphs are taken from the bibliographical sketch, written by Peter Grzybek and Reinhard Köhler, which was published in the Festschrift dedicated to Gabriel Altmann on the occasion of his 75<sup>th</sup> birthday, to which over 60 colleagues from all over the world contributed (Grzybek/Köhler 2005). Where necessary the original sketch is modified stylistically and in the last third of the obituary some more information about Gabriel’s scientific engagements is given. Gabriel’s first Festschrift was published in 1991 on the occasion of his 60<sup>th</sup> birthday (cf. Grotjahn 1991 et al.) under the title *Viribus Unitis*.

his major achievements is the comprehensive *Thesaurus of Univariate Discrete Probability Distributions*, which he published together with Gejza Wimmer and which contains the mathematical description of some 750 discrete distributions and families (cf. Wimmer/Altmann 1999). It also includes quite a number of distributions which were derived and created by Gabriel.

Directly related to this deep engagement with probability distributions is the development of the Altmann-Fitter. This unique software package is used for the iterative fitting of approximately 200 discrete probability distributions to empirical data, including parameter estimation and goodness-of-fit tests. The program is in use by many researchers from various countries in different disciplines. It allows users to quickly and effectively find proper statistical models for (linguistic) frequency data, and brought about a modelling boom in quantitative studies in general.

Gabriel was against a blind transfer of ‘standard methods’ of mathematical statistics (many of them based on the ‘law of large numbers’ and the ‘normal distribution’) because he was perfectly aware of the specific characteristics of linguistic data, where such methods cannot be properly applied. He formulated this highly important conclusion in his most prominent programmatic papers (Altmann 1972, Altmann 1973, Altmann 1978, Altmann 1985a, 1985b, 1987, 1990, 1993, 1996, 1997, 2006, Köhler/Altmann 1996, Köhler/Altmann 2005a). Gabriel devoted most of his life to developing new statistical methods for linguistic investigations specifically.

Gabriel combined the talent of an outstanding scientist with a very wide range of scientific interests and the gift of a brilliant organizer. He launched numerous national and international research projects, either managed by himself or at least with his ongoing support. Gabriel personally made contacts and helped in establishing contacts with research groups from all over the world, for example in the late seventies with the group *Statistika reči* in the former Soviet Union (led by R.G. Piotrovskij), and with many other groups in Europe, Japan, China and Canada. One can in fact conclude that Gabriel Altmann is not only the founder of quantitative linguistics in Germany, but also the nestor of modern quantitative linguistics in general. In 2005 the handbook *Quantitative Linguistics* was published by de Gruyter in Berlin, edited by Reinhard Köhler, R.G. Piotrovskij and Gabriel Altmann himself. This handbook gives a comprehensive overview of quantitative linguistics and related linguistic and philological disciplines.

In 1978, Gabriel founded, after years of preparation, the book series *Quantitative Linguistics*, with the two sub-series *Glottometrika* and *Musikometrika* (Gabriel was not only an ingenious linguist and mathematician, but also a gifted musician). Within the first ten years, under his supervision thirty volumes by authors from all five continents were published in this series, which prevailed until volume 60 in an almost unchanged form. In 1993, the *Journal of Quantitative Linguistics* was founded with Gabriel as an associate editor. Then, in 1994 the International Quantitative Linguistics Association was founded, where he was since 2005 the Honorary President. In 1995, the comprehensive *Bibliography of Quantitative Linguistics* was published (cf. Köhler 1995), which would not have been possible without Gabriel’s help. In 2001, finally, he started another journal on quantitative linguistics, *Glottometrics*, which he continued to edit until his passing away. He was also the founding editor of the book series *Studies in Quantitative Linguistics*, published by RAM-Verlag, of which 30 issues have been published since 2008. In 2008 he helped to establish the journal *Glottology* (founding editor was Gabriel Altmann’s Slovak colleague Emíla Nemcová), which primarily was meant as an interdisciplinary forum of quantitative and qualitative approaches in linguistics and text analysis (first it was published by the University of Saints Cyril and Methodius in Trnava; it then moved to Akademie-Verlag and finally to de Gruyter in Berlin).

His role in integrating the efforts of the quantitative linguistics community cannot be overestimated. By inviting researchers coming from different academic backgrounds and



cultures to share their knowledge, he made a number of excellent translations, as can be seen, for example, by his translation from Russian into German of the book *Problems of Quantitative-Systemic Lexicology* by Juhan Tuldava (cf. Tuldava 1996), which he published in 1998 as volume 59 of the series *Quantitative Linguistics*.

It is barely possible to name one specific focus of Gabriel's wide scientific horizon. Without any doubt it is a tremendous field of scientific interests and philosophical concerns: from phonetics and phonology to grammar and semantics, including typology, geo-linguistics, dialectology, text analysis, lexicology etc. Gabriel himself classified his collected (for the time being unpublished) works in four volumes (1961–1999) as follows: vol. 1 (General: symmetry, systems, synergetics), vol. 2 (phonology, grammar, structure of units), vol. 3 (semantics, lexicon, dialectology, historical linguistic, areal Linguistics) and vol. 4 (typology, text analysis, probability distributions).

Gabriel was open-minded, always ready to share his ideas and projects, without any academic vanity. He paid a great deal of attention to younger colleagues by giving his advice and by motivating them to follow their own path. Here one has to refer to selected volumes of the book series *Studies in Quantitative Linguistics*, where Gabriel and his colleagues published six books named *Problems of Quantitative Linguistics* (cf. Strauß/Fengxiang/Altmann 2008, Köhler/Altmann 2009, Čech/Altmann 2011, Köhler/Altmann 2014, Altmann 2015, Kelih/Altmann 2018). In this volume selected problems of quantitative linguistics are given in the form of a research hypothesis. By providing the broader background and related references for the interested researcher the 'ingredients' for an empirical analysis are given. The 'problems' can be understood in some respects as the manifold desiderata of quantitative linguistics in general.

To his closer friends, Gabriel was known not only as a scientific genius, but also as the author of a huge collection of humorous short stories (most of them still unpublished), which display his coruscating sense of humour.

We could continue with this description of further examples of Gabriel's eminent competencies, but we should not forget to mention his outstanding personal characteristics. Everyone who knew Gabriel from personal contact, either directly or via online cooperation, has experienced his exceptional, unselfish helpfulness.

Together with an increasing number of his colleagues, Gabriel and his scholars formed an international and interdisciplinary scientific network. The discipline of quantitative linguistics became more and more established. The number of publications in quantitative linguistics and the participants of the conferences devoted to different issues of quantitative studies is growing and growing: Gabriel's scientific life's work has been a great success.

This is now the moment to thank him most cordially, also in the name of an indeterminable number of students, colleagues and friends, whom he supported by giving advice and practical help, with unbelievable patience and good humour, with encouragement and direct engagement.

In May 2020 his ashes were scattered in the Danube, his most beloved river.

May he rest in peace.

## References

- Altmann, G.** (1972): Status und Ziele der quantitativen Sprachwissenschaft. In: Siegfried Jäger (ed.): *Linguistik und Statistik*. Braunschweig: Vieweg (Schriften zur Linguistik, 6), 1–9.
- Altmann, G.** (1973): Mathematische Linguistik. In: Walter A. Koch (ed.): *Perspektiven der Linguistik*. Stuttgart: Kröner (Kröners Taschenausgabe, 446), 208–232.
- Altmann, G.** (1978): Towards a theory of language. In: Gabriel Altmann (ed.): *Glottometrika I*. Bochum: Brockmeyer. (Quantitative Linguistics, 1), 1–25.
- Altmann, G.** (1985a): On the dynamic approach to language. In: Th T. Ballmer (ed.): *Linguistic Dynamics. Discourses, Procedures and Evolution*. Berlin: de Gruyter, 181–189.
- Altmann, G.** (1985b): Sprachtheorie und mathematische Modelle. *SAIS Arbeitsberichte aus dem Seminar für Allgemeine und Indogermanische Sprachwissenschaft* 8, 1–13.
- Altmann, G.** (1987): The levels of linguistic investigation. *Theoretical Linguistics* 14, 1987, 227–239.
- Altmann, G.** (1990): Bühler or Zipf? A re-interpretation. In: Walter A. Koch (ed.): *Aspekte einer Kultursemiotik*. Bochum: Brockmeyer, 1–6.
- Altmann, G.** (1993): Science and Linguistics. In: Reinhard Köhler and Burghard Rieger (eds.): *Contributions to Quantitative Linguistics. Proceedings of the First International Conference of Quantitative Linguistics, QUALICO, Trier, 1991*. Dordrecht, Boston, London: Kluwer, 3–10.
- Altmann, G.** (1996): The nature of linguistic units. *Journal of Quantitative Linguistics* 3, 1, 1–8.
- Altmann, G.** (1997): The art of quantitative linguistics. *Journal of Quantitative Linguistics* 4, 1–3, 13–22.
- Altmann, G.** (2006): Fundamentals of quantitative linguistics. In: J. Genzor and M. Bucková (eds.): *Favete linguis*. Bratislava: Slovak Academic Press, 25–27.
- Grotjahn, R., Kempgen, S., Köhler, R., Lehfeldt, W.** (eds.) (1991): *Viribus unitis. Festschrift für Gabriel Altmann zum 60. Geburtstag*. Trier: WVT.
- Grzybek, P., Köhler, R.** (eds.) (2007): *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin, New York: de Gruyter (Quantitative Linguistics, 62).
- Köhler, R., Altmann, G.** (1996): "Language forces" and synergetic modelling of language phenomena. In: Peter Schmidt (ed.): *Glottometrika 15. Issues in general linguistic theory and the theory of word length*. Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 57), 62–76.
- Köhler, R., Altmann, G.** (2005): Aims and Methods of Quantitative Linguistics. In: Gabriel Altmann, Viktor Levickij and Valentyna Perebyjnis (eds.): *Problemy kvantytatynnoi lingvistyky. Problems of Quantitative Linguistics*. Černivci: Ruta, 12–41.
- Köhler, R.** (1995): *Bibliography of quantitative linguistics*. (with the assistance of Christiane Hoffmann). Amsterdam, Philadelphia: John Benjamins (Amsterdam studies in the theory and history of linguistic science. Series V, Library and information sources in linguistics, 25).
- Köhler, R., Altmann, G., Piotrowski, R.G.** (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

**Problems in Quantitative Linguistics (in chronological order)**

**Strauß, U., Fan, F., Altmann, G.** (2008): *Problems in Quantitative Linguistics 1. Second edition*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 1).

**Köhler, R., Altmann, G.** (2009): *Problems in Quantitative linguistics 2*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 4).

**Čech, R., Altmann, G.** (2011): *Problems in Quantitative Linguistics 3. Dedicated to Reinhard Köhler on the occasion of his 60th birthday*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 12).

**Köhler, R., Altmann, G.** (2014): *Problems in Quantitative Linguistics 4*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 14).

**Altmann, G.** (2015): *Problems in Quantitative Linguistics 5*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 22).

**Kelih, E., Altmann, G.** (2018): *Problems in Quantitative Linguistics 6*. Lüdenscheid: RAM (Studies in Quantitative Linguistics, 28).

Written and compiled by

*Sergey Andreev*<sup>6</sup>,


*Gejza Wimmer*<sup>7</sup>,

*Emmerich Kelih*<sup>8</sup>

---

<sup>6</sup> Smolensk State University (Russia), [smol.an@mail.ru](mailto:smol.an@mail.ru).

<sup>7</sup> Matematický ústav, Slovenská akadémia vied (Bratislava, Slovakia)

<sup>8</sup> University of Vienna, [emmerich.kelih@univie.ac.at](mailto:emmerich.kelih@univie.ac.at).  <http://orcid.org/0000-0002-8315-8916>

# The Ambiguity of the Relations between Graphemes and Phonemes in the Persian Orthographic System

Tayebeh Mosavi Miangah<sup>1</sup>

Relja Vulcanović<sup>2</sup>

## Abstract

In this paper, the degree to which Persian orthography deviates from transparency is quantified and evaluated. We investigate the relations between graphemes and phonemes in Persian, in which the writing system is not fully representative of the spoken language, mostly due to the omission of the short-vowel graphemes. We measure the degree of the Persian orthographic system transparency using a heuristic mathematical model. We apply the same measures to orthographic systems of other languages and compare the results to those obtained for Persian. The results show a relatively high degree of transparency in Persian when it comes to writing, but a low degree of transparency when it comes to reading. We also consider models that avoid the problems related to the short vowels in Persian and these models demonstrate a considerable decrease of the uncertainty in the Persian orthographic system.


**Keywords:** *Persian language, orthographic system, orthographic uncertainty, phonemic uncertainty.*


## 1. Introduction

There is a principle in linguistics known as the one-meaning-one-form principle, based on which forms and meanings in any language system tend to correspond one to one (Anttila 1972, p. 181). The more a language obeys this principle, the more transparent, or the less opaque/uncertain/ambiguous it is. The concepts of transparency and opaqueness/uncertainty/ambiguity are not the same as the simplicity and complexity of languages. Hengeveld and Leufkens point out that “[l]anguages may be complex yet transparent, or simple, yet opaque.” (Hengeveld & Leufkens 2018, 141). Yet, it is argued in Vulcanović (2007) that language complexity should be measured relative to language transparency (if an increase in complexity is not accompanied by a decrease in transparency, the relative complexity of the language is even greater).

The notions of transparency and opaqueness can be extended to any pair of related linguistic units that do not necessarily involve meaning. In this paper, we consider the transparency of the relationship between phonemes and graphemes. More specifically, this paper aims to investigate the transparency of the Persian orthographic system using mathematical models. Our approach is similar to that found in Best & Altmann (2005) and the contributions to the volume *Analyses of script: Properties of characters and writing systems* (Altmann & Fan 2008). However, these works are mainly concerned with the *phoneme-to-grapheme map* and provide measures of the orthographic uncertainty of all phonemes in various languages. As opposed to this, we examine both directions of the relationship between graphemes and phonemes in the written and spoken forms of modern Persian and measure the degree of the overall Persian orthographic system transparency. In addition to the information-theory measure used in Altmann & Fan (2008), we also propose and use a new, slightly simpler one, which is akin to the measures of the degree of violation of the one-meaning-one-form principle (Vulanović & Ruff 2018, Vulcanović & Mosavi Miangah 2020).

---

<sup>1</sup> Department of Linguistics, Payame Noor University, Tehran, Iran. E-mail: [mosavit@pnu.ac.ir](mailto:mosavit@pnu.ac.ir). Work done while visiting Kent State University at Stark.  <http://orcid.org/0000-0002-6528-2876>

<sup>2</sup> Department of Mathematical Sciences, Kent State University at Stark, 6000 Frank Ave NW, North Canton, Ohio 44720, USA. E-mail: [rvulanov@kent.edu](mailto:rvulanov@kent.edu). Corresponding author.  <http://orcid.org/0000-0003-2189-5133>

If each grapheme corresponds to one and only one phoneme, and vice versa, we can say that the orthography of the given language is transparent. On the other hand, when such one-to-one correspondence does not exist, we deal with phonological and/or orthographic uncertainties. As languages differ from one another by the extent to which they are transparent or opaque, we compare the degree of transparency of the orthographic systems in different languages to locate the relative position of the Persian language on a scale.

In the next section, we provide an overview of some fundamental concepts and describe Persian orthography and phonology in general and to the extent that is needed for the present study. Then, in section 3, we survey the relevant literature. This is followed by the introduction of the mathematical models in section 4 and their application in section 5, where the results of the calculations are presented. Concluding remarks are given in section 6.

## 2. Preliminaries

### 2.1 Some fundamental concepts

The specific terms we use are as follows.

- Phoneme: a mental representation of a speech sound made by the mouth, which distinguishes one word from another in a particular language. We indicate phonemes by placing them between two forward slashes, //.
- Letter: a visual building block of written words (the way a word looks, not the way it sounds).
- Grapheme: an individual letter or groups of letters that represent a single phoneme. A grapheme is a written symbol expressing a sound. When we need to emphasize that we are dealing with graphemes, we place them inside angle brackets, < >.

Consider, for example, the word “elephant” in which the grapheme <ph> consists of two letters, <p> and <h>, representing the phoneme /f/.

### 2.2. Persian orthography and phonology

Persian is the official language of Iran. The writing system of this language has been adopted from the Arabic script with some modifications, although the spoken form is very different from Arabic. Persian is written from right to left and most letters have to be joined to their adjacent letters according to their position in the word. Although there is no distinction between capital and small letters, most letters have more than one shape depending on their position in the word, known as initial, medial, and final shapes. There are 36 letters in Persian, out of which 10 letters can only be written joined to the preceding letter (ا - آ - اء - و - ژ - ئ - ؤ - اء - اء), not to the following one.

The Persian phonemic system has 24 consonants and 6 vowels, three long and three short ones.<sup>3</sup> The three long vowels (i, u, A) are realized in written form and the three short vowels (e, o, a) are usually not written, except in two special cases when the phonemes /o/ and /e/ are indicated using the letters و and ة, respectively. The letter و is mainly used to represent a consonant phoneme, as well as a long-vowel phoneme, and ة is mainly used to represent a consonant phoneme. Moreover, superscript and subscript diacritics exist that can be used along with the consonant letters to indicate the short vowels. When added to a letter, the diacritics make the pronunciation of that letter different from the original one. Table 1<sup>4</sup> illustrates the role

---

<sup>3</sup> The issue of whether Persian distinguish between the vowel lengths or the distinction between the vowels is solely based on their quality is of no relevance to this paper. Some general references to Persian phonology are Majidi (1986/1990) and Windfuhr (1997).

<sup>4</sup> In Table 1 and throughout the paper, we use the SAMPA (Speech Assessment Methods Phonetic Alphabet) phonetics for Arabic (Wells 2002) with some modifications adjusted to Persian. The complete list of Persian characters along with their nearest English equivalents is presented in Appendix I.

of the diacritical marks in Persian and the way they make words different in terms of pronunciation and meaning.

Table 1.  
Possible pronunciations of the written word کرم.

No.	Graphemes	Corresponding Phonemes	Pronunciation	Meaning	Graphemes with diacritics
1	کرم	/krm/	[karam]	generosity	کَرَم
2	کرم	/krm/	[kerm]	worm	کِرم
3	کرم	/krm/	[korom]	chrome	کُرَم
4	کرم	/krm/	[kerem]	cream	کِرَم

As Table 1 shows, a string of several consonant graphemes may have several different pronunciations and meanings depending on the type and the location of the short vowels, which are normally not written, in the graphemic string. As a result, phonological and semantic ambiguities arise, specifically for children and others learning Persian for the first time. Such ambiguities partly result from the existence of many Arabic and western loanwords in Persian (e.g., Nos. 1, 3, and 4 in Table 1).

Although short vowels are not normally realized in writing, they are written in two special circumstances. The first one is texts used for and by the beginner learners of Persian (including native Persian-speaking children in the early grades of school). The second case is the religious texts which are almost all borrowed from Arabic. In both cases, it is much easier to read the texts when the short-vowel diacritics are written. However, a major problem arises when the diacritics are omitted and inexperienced Persian readers are required to read the text correctly and understand its meaning. This problem also frequently happens to foreigners learning Persian as their second or foreign language when they try to pronounce written Persian words correctly and get the appropriate meaning. Still, the question is what happens to adult native or matured Persian speakers to be able to read the texts without diacritics and understand their meanings despite the phonological and semantic ambiguities created by the homographs. The answer to this question lies in the fact that after being exposed to texts with and without diacritics in the early years of Persian learning, Persian speakers reach a kind of cognitive maturity based on which reading texts without diacritics is largely possible. In other words, the early year textual materials act as a training corpus with the help of which the readers can develop their reading abilities through the strategies they have previously used. In most cases, their visual ability gained while learning Persian helps them to understand new occurrences of the words using patterns similar to those they have been exposed to earlier. It goes without saying that for the cases which have more than one correct pronunciation with different meanings, adult readers refer to the context to which the given word belongs to decide on the appropriate pronunciation and meaning.

There are also some additional discrepancies between phonemes and graphemes in Persian. An individual grapheme can represent several different phonemes and, similarly, an individual phoneme can be represented by several different graphemes. Tables 2 and 3 depict all possible relationships that exist between these two linguistic units in Persian. For complete lists of the graphemes and phonemes of Persian, see Appendices II and III.

Table 2.

The possible number of phonemes for individual graphemes in Persian (with diacritics)

Graphemes	No. of phonemes	Examples
1- آ (A) 2- او ('u) 3- ؤ (?) 4- ئ (?) 5- ا (?) 6- ای (i) 7- ای (ei)	1	1- آبادان (AbAdAn = a city name) 2- اوست ('ust = s/he is) 3- مسؤول (mas?ul = responsible) 4- رئیس (re?is = boss) 5- هیأت (hei?at = committee) 6- ایران ('irAn = Iran) 7- ای خدا (ei xodA = Oh God)
وا	2	واکسن (vAksan = vaccine) خواهر (xAhar = sister)
ا, ب, د, د, م, س, ت, ر, ر, ن, ز, ش, ک, پ, گ, ف, خ, ق, ل, ج, ح, چ, ژ, ص, ع, ث, ض, ط, غ, ظ, ذ,	4	سال (sAl = year) سفیر (safir = ambassador) سلطان (soltAn = king) سفید (sefid = white) (Example only provided for the grapheme س.)
1- ی (i, y, ya, ye, yo) 2- ه (h, e, he, ha, ho)	5	1- سینی (sini = tray), یاقوت (yAqut = ruby), یزد (yazd = a city name), یک (yek = one), یُد (yod = iodine) 2- مهتاب (mahtAb = moonlight), خانه (xAne = home), هل (hel = cardamom), همیشه (hamiSe = always), هجوم (hojum = rush)

Apart from the short vowels, the long vowels also show some inconsistencies in the Persian orthography. In Table 4, which is a selected extraction from Appendix II, we list the graphemes for the three long vowels of Persian, as those occurring exclusively in initial positions and those appearing in both initial and non-initial positions (each long vowel has a different form when appearing in the initial position). As Table 4 shows, there are some inconsistencies between long-vowel graphemes and phonemes, specifically when the long vowels appear in middle or final positions in the word. In the initial position, the different form of the long-vowel grapheme almost always helps to identify the corresponding phoneme unambiguously. To be concrete, the grapheme <آ>, the other variation of <ا> appearing only in the initial position in the word, has just one possible pronunciation, /A/. The grapheme <ای>, the other variation of <ی> appearing only in the initial position in the word, has two possible pronunciations as /i/ and /ei/; the latter one can only be seen in the word ای (an interjection word which means “Oh!”). And lastly, the grapheme <او>, the other variation of <و> appearing only in the initial position in the word, has a very limited possible set of pronunciations which can be distinguished in a small number of words such as اوستا, اورست and a couple of other words that can be easily memorized. Therefore, it is not very realistic to mention these initial forms as ambiguous graphemes.

Table 3.

The possible number of graphemes for individual phonemes in Persian (without diacritics)

Phonemes	No. of graphemes	Examples
Phonemes other than the ones listed below	1	نمک (namak=salt)
1- /o/ 2- /i/ 3- /t/ 4- /h/ 5- /u/	2	1- استخوان (ostexAn=bone), خوردن (xordan=eating) 2- ایشان (iSAAn = they), بینی (bini = nose) 3- طوفان (tufAn = storm), توپ (tup = ball) 4- هندوانه (hendevAne = watermelon), حیوان (heivAn = animal) 5- اورژانس (urZAns = emergent), سوخت (suxt = fuel)
1- /s/ 2- /A/	3	1- صورت (surat = face), سفید (sefid = white), ثمن (saman = price) 2- مادر (mAdar = mother), آفتاب (AftAb = sunlight), خواندن (xAndan = reading)
/z/	4	زیبا (zibA = nice), ذرت (zorat = corn), ظهر (zohr = noon), ضمن (zemn = while)
/ʔ/	5	سوال (soʔAl = question), مسئله (masʔale = problem), قلعه (qalʔe = castle), امضاء (emzAʔ = signature), مأوا (maʔvA = residence)

Table 4.

Possible corresponding pronunciations of Persian long vowels

Long-vowel graphemes (initial position)	Possible phonemes	Examples
ی (ای)	/i/, /y/, /ya/, /ye/, /yo/	مینا, دستیار, یواش, یگانه, یمن ایستاد,
و (او)	/u/, /o/, /v/, /va/, /ve/, /vo/	جوراب, خورشید, دشوار, دعوت, وصال, وضو, اوست
ا (آ)	/A/, /a/, /e/, /o/	سال, اسیر, امتحان, اسوه, آدم

So, in this paper, we want to calculate the degree of the Persian orthographic ambiguity from the standpoint of both matured and non-matured Persian users. The ambiguity is mainly due to the absence of the short-vowel diacritics from the Persian writing system, which makes the conversion of graphemes to phonemes, that is, reading, difficult. If we look at the Persian orthography as already mastered by adult native speakers, we find the grapheme-to-phoneme relation much more transparent. The same happens if we determine the degree of ambiguity of the Persian orthography by considering texts that use all the diacritics overtly. The other direction, the conversion of phonemes to graphemes, that is, writing, is not a big problem in Persian and we find the Persian orthography more transparent than the orthography of any other language in our sample. These results can be found in section 5 below.

### 3. Related works

There is a related line of work that has grown around the Persian orthographic issues. Of note is a thread of works in this vein by Baluch (Baluch & Besner 1991; Baluch & Shahidi 1991; Baluch 1993, 2005). We would like to mention two separate investigations on the reading of individual Persian words (naming) by children and adult learners. In the first research, Baluch and Shahidi studied the naming of Persian words with consonantal spelling, as opposed to those with vowel letters, by children with the mean age of 8.4 years. Their findings revealed that children made significantly more errors when dealing with opaque words (like /bCh/, meaning *child*) than with transparent words (like /bAzi/, meaning *play*). Consequently, the time taken to



name a list of words with consonantal spelling was shown to be longer than the time taken to name a list of words with vowel-letter spelling (Baluch and Shahidi 1991). Baluch also reported similar findings while performing the same experiment on skilled adult Persian readers. When consonantal words had multiple meanings, their naming times were significantly longer than in the case of consonantal words with a unique meaning. He concluded from there that there was significant difficulty in naming words with consonantal spelling, caused by phonological processes (Baluch 1993).

In an investigation dealing with the Persian spoken in present-day Iran and the relationship between Persian orthography and literacy, Baluch has attempted to emphasize how literacy acquisition by Persian beginner or skilled readers may be affected by peculiarities of Persian orthography. He claimed it was the first time the issue of cognitive processes involved in literacy acquisition of Persian was reviewed. After an extensive elaboration on orthographic and phonological inconsistencies of Persian, he put forward that the main orthographic and phonological factors possibly affecting Persian literacy are the grapheme-phoneme regularity, the phoneme-grapheme ambiguity, and the absence of short vowels in written text. Finally, he points to the fact that perhaps some changes should be introduced into Persian orthography (Baluch 2005).

Another related work is the research reported by Kaveh Ashourinia in his Master's Dissertation. He attempted to quantify and visualize the ambiguities of the semi-consonantal Persian writing system with a glance at its consequences. He introduced an analytic approach as a tool to examine the difficulties resulting from the lack of short vowels in written Persian and the inconsistency of some long vowels with the written form. He concluded that these analytical data support the idea that the Persian writing system is not well-suited for the Persian language (Ashourinia 2019).

However, some other researchers completely ignored the problem of diacritics in Persian and categorized this language as a very transparent language in comparison to other languages. Gholamain & Geva, for instance, use the term "orthographic depth," previously used by Baluch (1993), Baluch & Besner (1991), and Frost, Katz & Bentin (1987), to categorize orthographic systems on a continuum ranging from shallow to deep. They claim that Persian (like some other languages such as Turkish, Hebrew, and Arabic) can be labeled as "shallow" because it has a simple grapheme-phoneme relationship in comparison to other scripts, e.g., the English script, which they labeled as "deep" for its more complicated grapheme-phoneme relations. They argue that learning Persian and mastering to decode Persian words and reading Persian texts accurately for children in the early grades is much easier than it might be the case for reading English. They implicitly conclude that such regularity may have an impact on learning to read, referring to other investigations (Gholamain & Geva 1999). As we have mentioned in subsection 2.2, when we consider Persian texts with all short vowels overt in the script (vowelized script), the grapheme-to-phoneme relationship becomes sufficiently regular.

The only point that is absent in the above literature on Persian orthography is how to measure the degree of orthographic uncertainty in Persian so that it can be compared to other languages. Our work is set to measure for the first time the orthographic ambiguity of Persian by heuristic and robust mathematical formulas. In this way, we can concretely calculate the degree of deviation of the Persian orthographic system from the ideal transparency and compare it to orthographic systems in other languages through an objective criterion. Although this has not been done for Persian before, there are many investigations on quantifying the relationship between phonemes and graphemes in other languages. In addition to Altmann & Fan (2008), which we have already discussed in the introduction, we can mention Best & Altmann (2005). A review of ways to measure orthographic transparency/uncertainty, and of other related issues, is given in Borleffs, Maassen, Lyytinen & Zwarts (2017).

#### 4. Mathematical description

Let us first introduce some mathematical notation. For any set  $A$ , let  $|A|$  denote the number of elements in the set and let  $A^*$  be the set of all strings of elements in  $A$  including the empty string denoted by  $\lambda$ . For two nonempty finite sets  $X$  and  $Y$ , we define  $X \times Y$  as the set of ordered pairs,  $X \times Y = \{(x, y) : x \in X, y \in Y\}$ . A relation  $\Phi$  between  $X$  and  $Y$  is a subset of  $X \times Y$ . The corresponding inverse relation is denoted by  $\Phi^{-1}$ ,  $\Phi^{-1} = \{(y, x) : (x, y) \in \Phi\}$ . Let  $n_x$  indicate how many elements of  $Y$  are paired up in the relation  $\Phi$  with a particular  $x \in X$ ,

$$n_x = n_x(\Phi) = |\{y \in Y : (x, y) \in \Phi\}|.$$

We assume of any relation  $\Phi$  that it involves every element of both  $X$  and  $Y$ , so that for each  $x \in X$  there exists an element  $y \in Y$  such that  $(x, y) \in \Phi$ , and vice versa. This implies that  $n_x \geq 1$  for each  $x \in X$ .

We next define the frequency  $f_k$  of elements in  $X$  that have  $n_x = k$ ,

$$f_k = f_k(\Phi) = |\{x \in X : n_x(\Phi) = k\}|, \quad k = 1, 2, 3, \dots$$

Note that  $0 \leq f_1 \leq |X|$ . If  $f_1 = |X|$ , that is, if  $f_k = 0$  for all  $k \geq 2$ , then the relation  $\Phi$  is a function from  $X$  onto  $Y$ . To measure how far a relation  $\Phi$  is from a function, we can use the formula

$$m(\Phi) = \frac{S(\Phi)}{|X|}, \quad S(\Phi) = \sum_{k \geq 2} f_k(\Phi)(k - 1). \quad (1)$$

In a function, each  $x \in X$  is paired with exactly one  $y \in Y$ , so the above sum  $S(\Phi)$  indicates the number of pairs in  $\Phi$  that goes over the count which would be present in a function. If this count is 0, that is, if  $m(\Phi) = 0$ , then (and only then)  $\Phi$  is a function. The sum  $S(\Phi)$  is divided by the total number of elements in  $X$ , which makes  $m(\Phi)$  a relative measure. The reason for this is illustrated by the following simple abstract example.

**Example.** Consider  $X = \{1\}$ ,  $Y = \{a_1, a_2\}$ , and  $\Phi = \{(1, a_1), (1, a_2)\}$ . We have  $S(\Phi) = 1$  and  $m(\Phi) = 1$ . Now, add 98 more pairs  $(2, a_3), (3, a_4), \dots, (99, a_{100})$  into  $\Phi$  to create a new relation  $\Phi'$ . This new relation is not a function because of only one of the 100 pairs in it, whereas  $\Phi$  is not a function because of one of the two pairs. Nevertheless, the sum  $S(\Phi)$  shows no difference between  $\Phi$  and  $\Phi'$  since  $S(\Phi')$  is still equal to 1. On the other hand,  $m(\Phi') = \frac{1}{99} = 0.0101$ , thus the relative measure  $m(\Phi)$  places  $\Phi$  and  $\Phi'$  in more appropriate positions on a scale indicating for any relation how far it is from a function.

If  $\Phi^{-1}$  is also a function, then  $\Phi$  is a bijection (one-to-one correspondence) between  $X$  and  $Y$ . A measure of how far a relation is from a bijection is introduced in (Vulanović & Ruff 2018) and applied to linguistics. Further modifications and applications of this measure can be found in Vulanović & Mosavi Miangah (2020). As we are about to see, bijections are not suited for the analysis of phoneme and grapheme systems, so we cannot apply the formulas from Vulanović & Ruff (2018) or Vulanović & Mosavi Miangah (2020) in this paper. Nevertheless, there is some similarity between the formulas used in these two works and the formula in (1).

Let  $P$  be the set of all phonemes and  $G$  the set of all graphemes of a language. As Appendix III indicates in the case of Persian, the relation between the phonemes and the graphemes cannot be represented as a subset of  $P \times G$ , but rather as a subset of  $P \times \tilde{G}$ , where  $\tilde{G} \subset G^*$ . This is because  $\tilde{G}$  includes the empty grapheme  $\lambda$ , which otherwise would not be considered an element of  $G$ . Kelih (2008) describes the situation in the Slovene language in a

similar way. He uses the grapheme  $\lambda$  to present the phoneme-grapheme relation, and, although he does not consider the inverse relation, in some other analyses of the grapheme system, he does not include  $\lambda$  as a grapheme. Similarly, Appendix II shows that it is easier to describe the grapheme-phoneme relation in Persian not by referring to a subset of  $G \times P$ , but to a subset of  $G \times \tilde{P}$ , where  $\tilde{P} \subset P^*$ . This is because strings of phonemes need to be used.

Let the phoneme-grapheme relation (also called the *phoneme-to-grapheme map*), as a subset of  $P \times \tilde{G}$ , be denoted by  $\Phi_1$  and let  $\Phi_2$  stand for the grapheme-phoneme relation (or the *grapheme-to-phoneme map*), which is a subset of  $G \times \tilde{P}$ . Then, the measure  $m_1 := m(\Phi_1)$  evaluates the orthographic uncertainty of all phonemes, while  $m_2 := m(\Phi_2)$  measures the phonemic uncertainty of all graphemes. Let also  $p_k = f_k(\Phi_1)$  and  $g_k = f_k(\Phi_2)$ . Then, according to the formulas in (1),

$$m_1 = \frac{1}{|P|} \sum_{k \geq 2} p_k (k - 1), \quad m_2 = \frac{1}{|G|} \sum_{k \geq 2} g_k (k - 1). \quad (2)$$

The greater the value of  $m_1$ , the harder it is to write. If  $m_1 = 0$ , this indicates the easiest writing system in terms of knowing what grapheme to use for any given phoneme. This is so because each phoneme has exactly one graphemic representation (although one grapheme may be used to represent more than one phoneme). Similarly, the greater the value of  $m_2$ , the harder it is to read. An orthography that enables the easiest reading has  $m_2 = 0$  because each grapheme represents exactly one phoneme (although several different graphemes may be used for the same phoneme). When the values of  $m_1$  and  $m_2$  are close, this indicates an orthography in which it is approximately equally easy to write and to read. If  $m_1$  is considerably greater than  $m_2$ , writing is harder than reading, and the other way around.

Typically,  $\Phi_1^{-1} \neq \Phi_2$ . Therefore, it is not appropriate in the present context to ask how far a relation is from a bijection and the measures from Vulcanović & Ruff (2018) and Vulcanović & Mosavi Miangah (2020) cannot be used. Nevertheless, we can still measure the uncertainty of the whole system of phonemes and graphemes by averaging the two measures given in (2),

$$m = \frac{1}{2} (m_1 + m_2). \quad (3)$$

Most of the works on the systems of phonemes and graphemes in various languages are focused on the relation  $\Phi_1$  and do not consider  $\Phi_2$  (Best & Altmann 2005, Altmann & Fan 2008). The orthographic uncertainty of phonemes is measured in these works not by  $m_1$  but by the quantity  $U_1$ ,

$$U_1 = \frac{1}{|P|} \sum_{k \geq 1} f_k(\Phi_1) \log_2 k = \frac{1}{|P|} \sum_{k \geq 2} p_k \log_2 k. \quad (4)$$

By the way, this is the correct mathematical rendering of the formula in Best & Altmann (2005) and Altmann & Fan (2008), which instead of the above sum uses  $\sum_{x \in P} f_x \log_2 n_x$  although the frequency  $f_x$  does not depend on a single phoneme  $x$ .

The measure  $U_1$  is based on information theory. The corresponding weighted measure of the orthographic uncertainty of a single phoneme reduces to entropy (Best & Altmann 2005, Borleffs et al. 2017). We shall apply both measures  $m_1$  in (2) and  $U_1$  in (4) to all the languages in our sample and show that there is a strong correlation between them. Therefore, either measure can be used to arrive at the same general conclusions and  $m_1$  is a little simpler because it does not require the calculations of binary logarithms.

At the same time, it is possible to measure the phonemic uncertainty of graphemes by a quantity  $U_2$  which is defined analogously to  $U_1$  in (4),

$$U_2 = \frac{1}{|G|} \sum_{k \geq 2} g_k \log_2 k. \quad (5)$$

Then, the uncertainty of the whole system of phonemes and graphemes can also be measured by

$$U = \frac{1}{2}(U_1 + U_2). \quad (6)$$

We shall only be able to compare the corresponding measure  $m_2$  in (2) and  $U_2$  in (5), as well as  $m$  in (1) and  $U$  in (6), in the case of Persian and Greek.

## 5. Results

### 5.1. Persian

The table in Appendix III shows the 30 phonemes in Persian. We present in Table 5 the number  $f_k$  of phonemes that have  $k$  graphemic representations.

Table 5.  
The number  $p_k$  of Persian phonemes  
that have  $k$  graphemic representations

<b><math>k</math></b>	1	2	3	4	5
<b><math>p_k</math></b>	19	5	4	1	1

Based on this table, we calculate the orthographic uncertainty of phonemes in Persian using the measure  $m_1$  given in (2),

$$m_1 = \frac{5 \cdot 1 + 4 \cdot 2 + 1 \cdot 3 + 1 \cdot 4}{30} = \frac{20}{30} = 0.6667.$$

The other formula, (4), yields

$$U_1 = \frac{5 \cdot 1 + 4 \log_2 3 + 1 \cdot 2 + 1 \cdot \log_2 5}{30} = 0.5221.$$

Because  $U_1$  uses logarithms, this measure produces values that are less than those of  $m_1$ .

We now consider the phonemic uncertainty of Persian graphemes. Table 6 is derived from Appendix II.

Table 6.  
The number  $g_k$  of Persian graphemes  
that have  $k$  phonemic representations

<b><math>k</math></b>	1	2	3	4	5	6
<b><math>g_k</math></b>	7	1	0	29	2	1

We calculate the value of  $m_2$  using the formula in (2),

$$m_2 = \frac{1 \cdot 1 + 29 \cdot 3 + 2 \cdot 4 + 1 \cdot 5}{40} = \frac{101}{40} = 2.525.$$

Also, from formula (5), we get

$$U_2 = \frac{1 \cdot 1 + 29 \cdot 2 + 2 \log_2 5 + 1 \cdot \log_2 6}{40} = 1.6557.$$

Comparing these values to the above-calculated  $m_1$  and  $U_1$ , we can see that it is much more difficult to read than to write Persian. This is so because the three short-vowel phonemes, /e/, /a/, and /o/, are not written after the consonants. It should be mentioned that we only consider the basic phoneme-to-grapheme and grapheme-to-phoneme maps in Persian. Otherwise, going into details, like in the discussion related to Table 4, can make the maps much more complicated, see also Mohseni Behbahani, Babaali & Turdalyuly (2016).

We can also calculate the overall measures  $m$  and  $U$  using the formulas (3) and (6), respectively,

$$m = \frac{0.6667 + 2.525}{2} = 1.5959, \quad U = \frac{0.5221 + 1.6557}{2} = 1.0889.$$

Let us now consider the situation when three different diacritics are used to indicate the three short vowels after the consonants. This increases the readability of Persian drastically. Table 5 does not change because the empty grapheme  $\lambda$  in Appendix III is replaced with the corresponding short-vowel diacritic. Therefore, the values of  $m_1$  and  $U_1$  remain the same. However,  $m_2$  and  $U_2$  become much smaller. This is because Table 6 changes to Table 7.

Table 7.

The number  $g_k$  of Persian graphemes, including the diacritics to indicate the three short vowels, that have  $k$  phonemic representations

$k$	1	2	3	4
$g_k$	38	3	1	1

This gives

$$m_2 = \frac{3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{43} = 0.186, \quad U_2 = \frac{3 \cdot 1 + 1 \cdot \log_2 3 + 1 \cdot 2}{43} = 0.1531.$$

and

$$m = \frac{0.6333 + 0.186}{2} = 0.4097, \quad U = \frac{0.3732 + 0.1531}{2} = 0.2632.$$

Obviously, the inclusion of the short-vowel diacritics turns the situation in Persian completely around, making its orthography easier for reading than for writing. The same would happen if we treated each consonant grapheme in Appendix II as only representing the corresponding consonant phoneme, thus ignoring the short-vowel phonemes after the consonants. We mention this approach to the analysis of Persian orthography because it is followed in Frost et al. (1987), but short vowels after the consonants should not be ignored because they are phonemes. In this case,  $m_1$  and  $U_1$  would still be the same as above, but the values of  $m_2$  and  $U_2$  would become

$$m_2 = \frac{3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{40} = 0.2, \quad U_2 = \frac{3 \cdot 1 + 1 \cdot \log_2 3 + 1 \cdot 2}{40} = 0.1646,$$

which would give

$$m = \frac{0.6333 + 0.2}{2} = 0.4167, \quad U = \frac{0.3732 + 0.1646}{2} = 0.2689.$$

Therefore, when the short vowels are ignored in this way, the ambiguity of the Persian orthography is reduced.

Each of the two cases considered above, viz., the inclusion of the short-vowel diacritics and the ignoring of the short-vowel phonemes after consonants, simulates a matured Persian user who has no or very little difficulty reading Persian.

## 5.2. A comparison of Persian and Greek

Another paper that analyzes both phoneme-to-grapheme and grapheme-to-phoneme maps is Protopapas & Vlahou (2009). The analysis is applied to the Greek language. In this subsection, we do all the calculations like in 4.1 using the Greek data from (ibid.). Table 8 summarizes the counts.

Table 8.  
The numbers  $p_k$  and  $g_k$  in the Greek orthography

$k$	1	2	3	4	5	6	7	8	9	10	11
$p_k$	7	14	4	4	3	0	2	2	0	0	1
$g_k$	63	14	4	1	1	1	0	0	0	0	0

When counting the Greek phonemes to find the values of  $p_k$ , we made one slight modification of the original data. Protopapas & Vlahou (2009) consider the phoneme strings /k//s/ and /p//s/ as single phonemes /ks/ and /ps/ because they have unique graphemic representations <ξ> and <ψ>, respectively. We did not follow this approach since orthography is not a factor in determining phonemes in a language. Therefore, we did not treat /ks/ and /ps/ as single phonemes. Instead, we added <ξ> to the list of graphemes representing /k/ and also to those representing /s/. Similarly, <ψ> was added to both list of graphemes representing /p/ and /s/. Nemcová & Altmann (2008) did the same with the Slovak grapheme <x>, which they included in each list of graphemes representing the phonemes /k/, /g/, /s/, and /z/.

Based on Table 8, we have for Greek that

$$m_1 = 2.2162, \quad U_1 = 1.3616, \quad m_2 = 0.4048, \quad U_2 = 0.3244.$$

Since  $m_1$  is considerably greater than  $m_2$  (the same relation holds between  $U_1$  and  $U_2$ ), we see that Greek is much easier to read than to write, as opposed to the situation we find in Persian. Using the above values, we can also calculate the overall measures  $m$  and  $U$ . They are given in Table 9 together with the results for the three cases of Persian analysis done in the previous subsection. Although Table 9 only contains four pairs of values of  $m$  and  $U$ , it can be used to verify the correlation between the two measures. The coefficient of correlation is  $R = 0.9988$ , so the correlation between  $m$  and  $U$  is nearly perfect.

Table 9.  
The overall uncertainty of the Persian and Greek  
Orthographies

Language	$m$	$U$
Persian (without the short-vowel diacritics)	1.5833	1.0764
Persian with the short-vowel diacritics	0.4097	0.2632
Persian with short vowels after consonants ignored	0.4167	0.2689
Greek	1.3105	0.8430

Table 10.  
The orthographic uncertainty of phonemes across languages

Language	$m_1$	$U_1$
Greek	2.2162	1.3616
German	1.1795	0.9661
Swedish	1.0556	0.7983
Slovak	0.9318	0.7599
Slovene	0.9310	0.7847
Oriya	0.9167	0.8475
Italian	0.6949	0.5648
Persian	0.6667	0.5221

### 5.3. Other languages

Six other languages, in addition to Persian and Greek, are considered here. They are German and Swedish (Best & Altmann 2005), Slovak (Nemcová & Altmann 2008), Slovene (Kelih 2008), Oriya (Mohanty & Altmann 2008), and Italian (Bernhard & Altmann 2008). However, for these six languages, we only have phoneme-grapheme relations, and because of this, we only calculate  $m_1$  and  $U_1$ . Without changing any of the original data, we got the values presented in Table 10. The results for  $U_1$  were also calculated in the original works. The values we show are essentially the same, but we carried out the calculations with more decimal places. The languages are listed in Table 10 in the decreasing order of the  $m_1$  values. We see that, of the eight languages, Greek is the hardest one to write, whereas Italian and Persian are the easiest ones. Comparisons like this should be taken with a degree of reservation because the eight phoneme-to-grapheme maps are not necessarily given with the same amount of detail. As mentioned before, we here only consider the basic map for Persian. The correlation between the  $m_1$  and  $U_1$  values is very strong,  $R = 0.9620$ .

## 6. Conclusion and implications

In this paper, we have quantitatively described the extent to which the Persian orthography deviates from an ideally transparent orthography. Many languages across the world have different types of inconsistencies in their phonological and orthographic systems due to various factors, but the Persian language has its peculiar problems that make the written form of the language rather hard to read especially for beginner learners. The Persian graphemes are basically taken from Arabic resulting in inconsistencies with the phonological system of Persian. As we have demonstrated, the main problem of Persian orthography is that the short vowels are absent in writing, which leads to ambiguities when reading Persian texts. We have also shown that the ambiguity of the Persian orthographic system becomes much lower when the short vowels are ignored in the pronunciation of Persian words, or when the diacritical marks indicating the short vowels are fully implemented.

We have not investigated the distribution of phonemes or graphemes here, but this also plays a role in the transparency of Persian orthography. The more restricted the distribution, the easier it is to disambiguate phonemes or graphemes in context. This can be analyzed in a continuation of the present study.

Keeping the Persian orthographic system with such ambiguities may have not only national, but also global consequences. As Hengeveld and Leufkens suggest, transparency is an important factor in the learnability of languages, and transparent features of a language are the first to be mastered by young children acquiring their mother tongue (Hengeveld & Leufkens 2018). Thus, regarding literacy acquisition, the inconsistencies of Persian orthography give rise to many problems for both Iranian children learning to read and write Persian in early grades of school and for foreigners trying to learn Persian as a foreign or second language. From the computational linguistics point of view, such ambiguities have serious effects on the quality of Text-To-Speech (TTS) systems, spell checking systems (especially in the second and third phases—generating and ranking candidates—of the whole process, Mosavi Miangah 2014), and the like. These are some arguments indicating that a major reform of the Persian script may be needed, a reform that would render the Persian script much easier to read, not only by the inclusion of the short-vowel graphemes, but also by making the graphemes more uniform (Appendix III shows different written forms for single-consonant phonemes). Nickjoo also points out that the peculiarities of written Persian have implications for literacy. He argues for the abolition of the Persian alphabet and the creation of a Latinized version of Persian (Nickjoo 1979).

The existing semi-Arabic script of Persian cannot answer all the needs of modern life, especially in the digital environment. It is hoped that this paper will stimulate further investigations in the field and motivate appropriate measures towards forward-looking decisions regarding the Persian language and its future challenges.



## References

- Anttila, R.** (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.
- Altmann, G., Fan F. (eds.)** (2008). *Analyses of script: Properties of characters and writing systems*, Berlin/New York: Mouton de Gruyter.
- Ashourinia, K.** (2019). *Quantifying and visualizing the ambiguities of the semi-consonantal Persian writing system, and its consequences*, Master's Dissertation, OCAD University, Toronto, Ontario, Canada.
- Baluch, B., Besner, D.** (1991). Visual word recognition: Evidence for strategic control of lexical and non-lexical routines in oral reading. *Journal of Experimental Psychology, Learning, Memory and Cognition* 17, 644–651.
- Baluch, B., Shahidi, S.** (1991). Visual word recognition in beginning readers of Persian. *Perceptual and Motor Skills* 72, 1327–1331.
- Baluch, B.** (1993). Lexical decisions in Persian: A test of the orthographic depth hypothesis. *International Journal of Psychology* 28, 19–27.
- Baluch, B.** (2005). *Handbook of orthography and literacy*. Routledge Handbooks Online. Cutting edge scholarship from Routledge and CRC Press.
- Bernhard, G., Altmann, G.** (2008). The phoneme-grapheme relationship in Italian. In Altmann, G., Fan F. (eds.), 13–24.
- Best, K.-H., Altmann, G.** (2005). Some properties of graphemic systems. *Glottometrics* 9, 29–39.
- Borleffs, E., Ben A. M. Maassen, Ben A. M., Lyytinen, H., Frans Z.** (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing* 30, 1617–1638.
- Frost, R., Katz, L., Bentin, S.** (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance* 13, 104–115.
- Gholamain, M., Geva, E.** (1999). Orthographic and cognitive factors in the concurrent development of basic reading skills in English and Persian. *Language Learning* 49, 183–217.
- Kelih, E.** (2008). The phoneme-grapheme relationship in Slovene. In Altmann, G. & Fan F. (eds.), 61–74.
- Hengeveld, K., Leufkens, St.** (2018). Transparent and non-transparent languages. *Folia Linguistica* 52, 139–175.
- Majidi, M.-R.** (1986/1990). *Strukturelle Grammatik des Neupersischen (Fārsi)*. Bd. I: *Phonologie*; Bd. II: *Morphologie*. *Forum Phonetikum* 34, 1–2. Hamburg: Buske.
- Mohanty, P., Altmann, G.** (2008). On graphemic representation of the Oriya phonemes. In Altmann, G., Fan F. (eds.), 121–140.
- Mohseni Behbahani, Y., Babaali, B., Turdalyuly, M.** (2016). Persian sentences to phoneme sequences conversion based on recurrent neural networks. *Open Comput. Sci.* 6, 219–225.
- Mosavi Miangah, T.** (2014). FarsiSpell: A spell-checking system for Persian using a large monolingual corpus. *Literary and Linguistic Computing* 29, 56–73.
- Nemcová, E., Altmann, G.** (2008). The phoneme-grapheme relation in Slovak. In Altmann, G., Fan F. (eds.), 79–90.
- Nickjoo, M.** (1979). A century of struggle for the reform of the Persian script. *The Reading Teacher*, 926–929.
- Protopapas, A., Vlahou, E. L.** (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods* 41, 991–1008.
- Vulanović, R.** (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics* 20, 399–427.

- Vulanović, R., Ruff, O.** (2018). Measuring the degree of violation of the One-Meaning–One-Form Principle. In Wang, L., Köhler, R., Tuzzi, A. (Eds.), *Structure, function and process in texts*, 67–77. Lüdenscheid: RAM.
- Vulanović, R., Mosavi Miangah, T.** (2020). The flexibility of parts-of-speech systems and their grammar efficiency. In Kelih, E., Köhler, R. (eds.), *Words and Numbers (In Memory of Peter Grzybek)*, 129–147. Lüdenscheid: RAM.
- Wells, J. C.** (2002). SAMPA for Arabic. <http://www.phon.ucl.ac.uk/home/sampa/arabic.htm>, accessed 01.02.2021.
- Windfuhr, G.** (1997). Persian phonology. In Kaye, A. S. (ed.) *Phonologies of Asia and Africa (Including the Caucasus)*. Vol 2, 675–689. Winona Lake, Indiana: Eisenbrauns.

Appendix I  
Transcription standards of Persian used in this paper.

Persian character	SAMPA character	Persian example	SAMPA transliteration	English translation
آ	A	آسیب	Asib	damage
ا ل	A	گرما	garmA	warmth
ا	'a	آدیان	'adyAn	religions
ا	'e	ارتباطاتی	'ertebAtAti	communicative
ا	'o	آردک	'ordak	duck
ب ب	B	باادب	bAadab	polite
پ پ	P	پتو	patu	blanket
ت ت	T	تابستان	tAbestAn	summer
ث ث	s	ثابت	sAbet	fixed
ج ج	j	جمله	jomle	sentence
چ چ	C	چرا	CerA	why
ح ح	h	حیاط	hayAt	yard
خ خ	X	خانواده	xAnevAde	family
د د	D	دوست	dust	friend
ذ ذ	Z	ذرت	zorrat	corn
ر ر	R	روز	ruz	day
ز ز	Z	زانو	zAnu	knee
ژ ژ	Z	ژرف	Zarf	profound
س س	S	سفید	sefid	white
ش ش	S	شناگر	SenAgar	swimmer
ص ص	s	صیاد	sayyAd	hunter
ض ض	z	ضروری	zaruri	necessary
ط ط	t	طاووس	tAvus	peacock
ظ ظ	z	ظهر	zohr	noon
ع ع ع	'	عجیب	'ajib	strange
غ غ غ	Q	غایب	QAyeb	absent
ف ف	f	فردی	fardi	personal
ق ق	q	قدیمی	qadimi	ancient
ک ک	k	کثیف	kasif	dirty
گ گ	g	گزارشگر	gozAreSgar	reporter
ل ل	l	لیوان	livAn	glass
م م	m	مرطوب	martub	wet
ن ن	n	نکته	nokte	point
ه ه ه	h	همپایه	hampAye	coordinate
و و	v	وظیفه	vazife	task
ی ی	y	یخچال	yaxCAI	refrigerator
و و	u	مربوط	marbut	related
و و	o	خود	xod	Self

اَ	o	مُدام	modAm	forever
اَ	a	مَدْرَك	madrak	document
اَ	e	مِهْرَبَان	mehrabAn	Kind
اِی یِ	i	شَهْرِی	Sahri	urban
اَؤ	'	مُؤَسَس	mo'asses	founder
اَ اَی	'	اَرَاثَه	'erA'e	presentation

## Appendix II

Persian graphemes with their possible corresponding phonemes.

	Grapheme	Phonemes (how the grapheme can be read)
1	آ	/A/
2	ا	/A/, /e/, /a/, /o/
3	ب	/b/, /be/, /ba/, /bo/
4	د	/d/, /de/, /da/, /do/
5	م	/m/, /me/, /ma/, /mo/
6	س	/s/, /se/, /sa/, /so/
7	او	/u/
8	ت	/t/, /te/, /ta/, /to/
9	ر	/r/, /re/, /ra/, /ro/
10	ن	/n/, /ne/, /na/, /no/
11	ای	/i/
12	ی	/i/, /y/, /ye/, /ya/, /yo/
13	ز	/z/, /ze/, /za/, /zo/
14	ه	/e/, /h/, /he/, /ha/, /ho/
15	ش	/S/, /Se/, /Sa/, /So/
16	ک	/k/, /ke/, /ka/, /ko/
17	و	/o/, /u/, /v/, /ve/, /va/, /vo/
18	پ	/p/, /pe/, /pa/, /po/
19	گ	/g/, /ge/, /ga/, /go/
20	ف	/f/, /fe/, /fa/, /fo/
21	خ	/x/, /xe/, /xa/, /xo/
22	ق	/q/, /qe/, /qa/, /qo/
23	ل	/l/, /le/, /la/, /lo/
24	ج	/j/, /je/, /ja/, /jo/
25	ح	/h/, /he/, /ha/, /ho/
26	چ	/C/, /Ce/, /Ca/, /Co/
27	ژ	/Z/, /Ze/, /Za/, /Zo/
28	ص	/s/, /se/, /sa/, /so/
29	ع	/ʔ/, /ʔe/, /ʔa/, /ʔo/
30	ث	/s/, /se/, /sa/, /so/
31	ض	/z/, /ze/, /za/, /zo/

32	ط	/t/, /te/, /ta/, /to/
33	غ	/Q/, /Qe/, /Qa/, /Qo/
34	ظ	/z/, /ze/, /za/, /zo/
35	ذ	/z/, /ze/, /za/, /zo/
36	وا	/A/, /vA/
37	ؤ	/?/
38	ئ	/?/
39	أ	/?/
40	ء	/?/

### Appendix III

Persian phonemes with their possible correspondent graphemes  
(λ denotes an empty grapheme).

	Phoneme	Graphemes
1	/A/	آ, وا, ا
2	/b/	ب
3	/d/	د
4	/m/	م
5	/s/	ص, ث, س
6	/u/	او, و
7	/t/	ت, ط
8	/r/	ر
9	/n/	ن
10	/i/	ای, ی
11	/y/	ی
12	/z/	ز, ض, ذ, ظ
13	/h/	ه, ح
14	/S/	ش
15	/k/	ک
16	/v/	و
17	/p/	پ
18	/g/	گ
19	/f/	ف
20	/x/	خ
21	/q/	ق
22	/l/	ل
23	/j/	ج
24	/C/	چ
25	/Z/	ز
26	/?/	ع, ء, و, ئ, أ

*Tayebeh Mosavi Miangah, Relja Vulcanović*

27	/q/	ġ
28	/a/	l, λ
29	/e/	o, l, λ
30	/o/	o, l, λ

# English Loanwords in Mongolian Usage

Minna Bao<sup>1</sup>,  
Saheya Brintag<sup>2</sup>,  
Dabhurbayar Huang<sup>3</sup>

## Abstract

Many authors have examined the influence of loanwords in languages using statistical methods. However, English loanwords in Mongolian are rarely studied in quantitative linguistics. The results of the present study show that English loanwords in Mongolian share the universal feature of other tested languages, as their frequency distribution abides by Zipf's Law. In addition, we define and test nine English loanword models depending on borrowing method and parts of speech, and find that the results can be described using a power function.

**Keywords:** *Mongolian, English, loanwords, quantitative linguistics, modelling.*

## 1. Introduction


The Mongolian language is the official language of Mongolia, and the number of speakers across all its dialects may be 10 million, including the vast majority of the residents of Mongolia and many of the Mongolian residents of the Inner Mongolia Autonomous Region in China. Mongolian belongs to the Mongolic family and is a typical agglutinative language that relies on suffix chains in the verbal and nominal domains, and manifests the subject–object–verb (SOV) basic order.

Mongolian includes many words borrowed from other languages, coming from a variety of cultural, trade, political, and military influences. A loanword can be defined as a word that is transferred from a donor language to a recipient language and is used in the recipient language (Joshi & Rajarshi 2017). In the history of the language – in the course of the last nearly eight hundred years –, the Mongols have used no fewer than 4,000 loanwords (Muren 1984), borrowed from about 30 languages (Tumurtogoo 2018). First, Mongolian adopted loanwords from Old Turkic, Sanskrit, Persian, Arabic, Greek, Sogd, Tibetan, Tungusic, and Chinese. However, recent loanwords come from Russian, English, and Mandarin Chinese (mainly in Inner Mongolia). Despite phonetic differences, Mongolian dialects often share common loanwords borrowed from other languages and keep using them in daily life.


As a result of recent socio-political changes, Mongolian has also borrowed many words from English. The English words for new objects or for new concepts are examples of technical borrowings. Words like *kompiüter*<sup>4</sup> (“computer”), *layiser* (“laser”), *radar* (“radar”), or *disk* (“disk”) abound in technical Mongolian. In recent times, numerous loanwords of English origin concerning daily life have also become more common. A considerable number of words in this category have acquired a very wide circulation – for example, *radio* (“radio”), *feyil* (“file”), or *imel* (“e-mail”). As such, the number of English loanwords in all spheres (innovation technology, media, economy, fashion, etc.) is constantly increasing to satisfy the communication demands of society.

---

<sup>1</sup> School of Mongolian Studies, Inner Mongolia Normal University, Inner Mongolia, China.

 <http://orcid.org/0000-0001-9988-4704>

<sup>2</sup> Mathematics Science College, Inner Mongolia Normal University, Inner Mongolia, China.

 <http://orcid.org/0000-0002-0398-1452>

<sup>3</sup> School of Mongolian Studies, Inner Mongolia University, Inner Mongolia, China. Correspondence address: Dabhurbayar Huang, e-mail address: [dabhurbayar@163.com](mailto:dabhurbayar@163.com).  <http://orcid.org/0000-0001-9475-8354>

<sup>4</sup> The Latin transcription of Mongolian follows the Latin transcription by Poppe (1954).

In quantitative linguistics, loanwords have been analysed in many ways – for example, from the perspectives of adoptions, structures, processes, influences, and their interrelations. Many authors have examined loanwords in German (Best 2001; Körner 2004; Ternes 2011; Liu 2013), English (Best 2006), and Russian (Stachowski 2010, 2018); their methodology is characterized by the use of quantitative methods and tools ranging from (simple) quantitative description to simulation and mathematical modelling. Among the dozens of investigations, the most important hypothesis is known as Piotrowski Law, capturing the law-like process of the incorporation of loanwords in many languages. The model has been revised by Altmann (1983, 1992), and obtained a form which has been positively tested in almost all respective research.

Best (2006) treated the process of transferring German words to English and the spectrum of fields, finding that the frequency of borrowings follows a regular rank-frequency distribution. The same author (Best 2005, 2013, 2014) also focused on the development of borrowings in German and demonstrated that this process abides by Piotrowski Law.

Liu (2013) dealt with German words of Chinese origin and analysed them using the methods of quantitative linguistics. The investigation showed that more than 160 words of the Chinese origin are in active use in the modern German language. Many Chinese loanwords were probably brought from China to the West by migrants. Many of the first words borrowed from Chinese migrated from the Cantonese dialect via English to German and other European languages. However, the investigation has yet to determine whether the reception process of German words of the Chinese origin abides by Piotrowski Law, or not.

Stachowski (2010) carried out research of loanword adaptation in a different way. His new method of preparing data for a quantification of loanword adaptation was illustrated with the example of Russian loanwords in Dolgan. The result is an attempt to measure the commonness and meaningfulness of adaptations, and an index of loanword nativization. Stachowski (2018) analysed the distribution of counts of phonetic renderings in 25 adaptations, together with the specific results that they yield during or after borrowing. Using loanwords from three methodologically different datasets, which contained the Arabic loanwords in sixteenth-century Ottoman Turkish, compound words borrowed from German to Polish, and the Russian loanwords in contemporary Dolgan, he confirmed the hypothesis that the distribution of counts of renderings in loanword adaptation is consistent with the Zipf-Alekseev distribution.

Furthermore, special attention has been paid to the investigation of the structure and cohesion of borrowings. In order to study the development of English-origin expressions, Gnatchuk (2015) analysed English–German as well as German–English hybrid compounds used in the newspaper *Kleine Zeitung* from 1995 until 2015. This analysis shows that the tendency of usage of new English borrowings is modelled using statistical methods, such as Piotrowski Law. The rank-frequency distribution of the English–German (German–English) hybrid compounds can be fitted using the power function. Furthermore, the investigation of cohesion for English–German and German–English compounds shows that the total values of English–German hybrid compound cohesion and their rank-frequency distribution can be fitted using the Zipf-Alekseev function.

The English loanwords in Mongolian have been researched in many ways (Muren 1984; Tumurtseren 2004; Norjin 2007; Bao 2017). However, a quantitative analysis of the distribution of probabilities in such tendencies has not yet been conducted. Given that the previous research on loanword adaptation has typically focused on historical investigation and transcription differences between source and loanword sounds in morphologically simplex words, few studies on English loanwords in Mongolian have considered their dynamic change, since their frequency distribution has been less studied and less understood.

In order to examine the adaptation of English loanwords in Mongolian dynamically, the present study investigates the following questions:

1. In the present-day usage, what is the rank-frequency distribution of rank for loanwords of English origin in Mongolian?
2. Does the frequency of English loanwords occurring in a text follow some general frequency distribution of rank?
3. Are there any morphological mechanisms in borrowings?

In the view of the above, we concentrate on the frequency distribution of English loanwords and their morphological characteristics. It is assumed that the rank of loanwords is arranged according to decreasing frequencies, and that the frequency of the structural patterns of English loanwords is related to complexity formed by grammatical rules in Mongolian. The rest of this paper is organized as follows: Section 2 introduces the material and the method used in this study; in section 3, the analysis results of English loanwords are illustrated and discussed; the final section contains some concluding remarks.

## **2. Data and Method**

### **2.1 Corpus**

A growing number of English loanwords have become current in official media and publications in Mongolian. Since the task of our analysis is to presuppose mechanisms that are responsible for borrowings in usage, we illustrate the performance of our proposed method by collecting data from the TV news of Inner Mongolia News and the newspaper Inner Mongolia Daily. As a result, we use a 299,027-word news corpus, consisting of 562 news sources broadcasted or published from 2012 to 2016. The collection of data from each year is given in Table 1.

Table 1.  
Text collection and proportion of each year

<b>Year</b>	<b>Words</b>	<b>Sample collection</b>	<b>Percentage of tokens (%)</b>
2012	13.901	Inner Mongolia News	4.65
2013	81.270	Inner Mongolia News	27.18
2014	22.521	Inner Mongolia News	7.53
2015	81.387	Inner Mongolia Daily	27.22
2016	99.948	Inner Mongolia Daily	33.42
<b>Total</b>	299.027		100

## **3. Results and Discussion**

### **3.1 Frequency distribution and the result of computation**

Word frequencies are central to lexicology investigation – they are always used to illustrate the relation between quantitative and qualitative methodologies; the studies of the phenomenon have focused mainly on the distribution of counts of phonetic (Stachowski 2010, 2018) and structural patterns (Gnatchuk 2015) of the loanwords, and on the estimation of the actual proportion of loanwords in a language (Joshi & Rajarshi 2017). In order to obtain the frequencies of English loanwords in Mongolian, we calculated occurrences of these words in the 299,027 tokens corpus. As a result, we have 118 distinct words together with their frequencies. Table 2 gives the frequencies and the ranks of selected English loanwords (the full list is shown in the Appendix 1).



Table 2.  
Frequencies of English loanwords

Loanwords	Part of speech	English	Frequency	Rank
<i>mašin</i>	noun	machine	226	1
<i>kompani</i>	noun	company	219	2
<i>inženering</i>	noun	engineering	216	3
<i>šistem</i>	noun	system	193	4
<i>kadri</i>	noun	cadre	192	5
<i>materijal</i>	noun	material	158	6
<i>energi</i>	noun	energy	119	7
<i>kilometr</i>	noun	km	88	8
<i>telvis</i>	noun	television	65	9
<i>ton</i>	noun	ton	59	10
...	...	...	...	...
<i>kod</i>	noun	code	1	118

On the basis of this list, it is observed that the absolute frequency of English loanwords is 2,348, so that the average frequency of English loanwords in our corpus is 0.79%. In addition, the English-origin noun is 2,303 make up 98.08% of the English loanwords in corpus.

Fundamental laws in quantitative linguistics describe proportionality phenomena related to frequencies of units or of classes made up of features of the units (Andersen 2002). Here, we used the Altmann-Fitter software to fit the data, and the result of the computation shows that the frequency distribution of English loanwords follows the Zipf-Mandelbrot Law.

Zipf's Law is as mathematicised as follows –

$$f_r = cr^{-1}; \tag{2}$$

based on Zipf's theory, Joos and Mandelbrot (Joos 1936, Mandelbrot 1953) proposed an improved distribution model by treating the exponent in Zipf's formula as a parameter, shown in the following formula:

$$f_r = cr^{-b}, \tag{3}$$

where  $b > 0, c > 0$ .

When we choose the parameter  $b = 1$ , then the formula of Joos is reduced to the traditional formula of Zipf's law. On the basis of the above two formulas, Mandelbrot (Mandelbrot 1953, 1962) studied the frequency distribution of words by using methods of information theory and probability theory, and proposed a three-parameter model:

$$f_r = c(r + a)^{-b}, \tag{4}$$

where  $a \geq 0, b > 0, c > 0$ .

This formula was called the Mandelbrot formula by later generations. When we choose the parameter  $a = 0$ , then the formula is reduced to Joos's formula, and when the parameters  $a = 0, b = 1$ , the formula is reduced to the formula of Zipf's Law. This means that the formula of Zipf's Law and the formula of Joos are special cases of the Mandelbrot formula.

The data of 118 Mongolian loanwords in the attached appendix are fitted by Altmann-Fitter, and the automatic fitting property tells us that the Zipf-Mandelbrot model is the best of the models available in the Fitter. In order to get more detailed information of Zipf-Mandelbrot fitting, we use the Mathematica software to compute the fitting results. Take the logarithm of both sides of the equation (4) and obtain

$$\ln(f_r) = \ln(c) - b * \ln(r + a). \quad (5)$$

By using the “Nonlinear ModelFit” function in Mathematica on the data set of Mongolian loanwords, we get the parameters table as following:

	Estimate	Standard Error	t-Statistic	P-Value
$a$	8.88575	1.00941	8.80292	$1.582 \times 10^{-14}$
$b$	2.37427	0.06926	34.28274	$3.279 \times 10^{-62}$
$\ln(c)$	11.182701	0.33143	33.74042	$1.741 \times 10^{-61}$

The value of R Squared is  $R^2 = 0.9661$ ; from the testing parameters, we can see that the Zipf-Mandelbrot model fits very well. The data curve and fitting curve of Zipf-Mandelbrot are drawn in Figure 1, which visually shows that the fitting effect of the model is excellent.

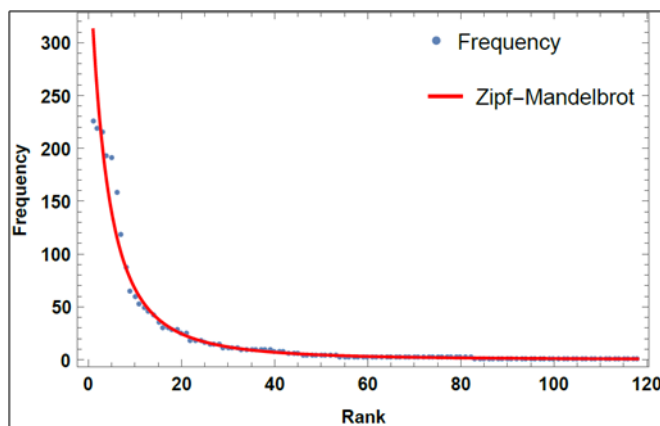


Figure 1. Original data and fitted Zipf-Mandelbrot curve

### 3.2 Borrowing methods and the results of computation

#### 3.2.1 Borrowing methods

The aim of this analysis is to research the borrowing methods of English loanwords in Mongolian and to investigate their frequency distributions. Careful attention should be drawn to the fact that the structural patterns of borrowing methods can be of three types: transliteration, agglutination, and acronymization.

*Transliteration* is a mapping from the letters of the English script to letters pronounced similarly in the Mongolian script. As the relationship between letters and sounds is similar in both, transliteration is very close to transcription. In practice, there are some mixed transliteration/transcription systems that transliterate part of the original script and transcribe the rest. Therefore, there is more than one standard transliteration system. However, unsystematic transliteration is common. For the sake of illustration, let us take Mongolian words as examples – *injenering* is a transliteration of “engineering”, *minüt* is a transcription of “minute”, and *bangqi* is a mixture of transliteration/transcription of “bank”.

*Agglutination* means that words are derived from other words by adding suffixes to invariable primary stems (Poppe 1954). From the morphological point of view, all words can be divided into two classes – those with primary stems and those with secondary stems. The

agglutinative words are derived from primary stems by means of suffixes; e.g., *autočilaysan* (“automotive”) is derived from *auto* (“auto”).

*Acronymization* is when words are borrowed into Mongolian without any transcription into Mongolian scripts, for example, “CPI” (“Consumer Price Index”), “GDP” (“Gross domestic product”), or “TV” (“television”).

As far as the procedure of our study is concerned, we have analysed the individual English loanwords according to their borrowing methods and morphological features. The results are given in Table 3 (the full list is shown in the Appendix 2).

Table 3.  
The English loanwords in terms of borrowing methods and morphological features

Borrowing methods	Morphological features	Examples	Total
Transliteration	Noun < <sup>5</sup> Noun	<i>ķompani</i> (“company”)	89
	Adjective < Adjective	<i>oryaniy</i> (“organic”)	2
Agglutination	Noun < Noun + Suffix -či/-čín	<i>boķsčín</i> (“boxer”)	3
	Noun < Noun + Suffix -la-/-le-(-ra-/-re-) + Suffix -yči/-gči	<i>ķomandalayči</i> (“commander”)	1
	Noun < Noun + Suffix -čila-/-čile- + Suffix -l	<i>autočilal</i> (“automation”)	2
	Adjective < Noun + Suffix -tu/-tü	<i>motortu</i> (“motor-assisted”)	3
	Adjective < Adjective + Suffix -tu/-tü	<i>oryaniytu</i> (“organic”)	1
	Verb < Noun + Suffix -čila-/-čile- + Nomen Perfecti -ysan-/ -gsen	<i>autočilaysan</i> (“automotive”)	3
Acronymization	Noun < Noun	CPI	14
<b>Total</b>			<b>118</b>

Table 4.  
Function of each suffix

Suffix	Function
-či/-čín	Nouns designating names of vocations
-la-/-le- (-ra-/-re-)	Acquirement of a quality
-yči/-gči	Nomen actoris designates the person acting and sometimes the process of an action; it is used as subject, object, attribute, and with a copula, as predicate
-čila-/-čile	Indication of the fact the object is rendered into, made into, or made like the thing or quality designated by the primary word
-l	Nouns designating abstract ideas
-tu/-tü	Adjectives designating possession of or containment in something

<sup>5</sup> “<” means “developed from”.

<b>-ysan-/-gsen</b>	Nomen perfecti express a completed past action, e.g., “someone who has died” or “is dead”; this form is used as subject, object, attribute, and predicate
---------------------	---

### 3.2.2 Distribution of borrowing methods and morphological features

The frequencies of models for English loanwords in terms of borrowing methods and morphological features are given in Table 5.

Table 5.  
The frequencies of models for English loanwords in terms of borrowing methods and morphological features

Borrowing method	Part of speech	Morphological features	Name of models	Absolute frequency
Transliteration	Noun	< Noun	M1	2.273
	Adjective	< Adjective	M2	19
Agglutination	Noun	< Noun + Suffix -čila-/-čile- + Suffix -l	M3	3
		< Noun + Suffix -či/-čín	M4	2
		< Noun + Suffix -la-/-le- (-ra-/-re-) + Suffix -γči/-gčín	M5	2
	Adjective	< Noun + Suffix -tu/-tü	M6	11
		< Adjective + Suffix -tu/-tü	M7	8
	Verb	< Noun + Suffix -čila-/-čile- + Nomen Perfecti -ysan-/-gsen	M8	7
Acronymization	Noun	< Noun	M9	23
<b>Total</b>				2.348

It can be seen from Table 5 that transliteration is the primary borrowing method, with the total count of 2.292 tokens representing the proportion of 97.61% of the total number of 2,348 words.

### 3.2.3 The morphological model frequency distribution of rank for English loanwords in Mongolian

In the following section, we deal with two variables in the research – the model of English loanwords and the frequency of loanwords, which may give a better estimation of the manner of loanword adoption. In this analysis, the relationships have been captured by means of a power function –

$$y_r = 1 + a * r^b, \tag{6}$$

where,  $y$  is the morphological model frequency,  $r$  is the rank and  $a, b$  are the parameters. We used the Mathematica software to fit the data; the outcome of the computation is as follows –

$$y_r = 1 + 2271.9915r^{-6.5607}. \tag{7}$$

As a result, the relation between the analysed variables has been positively confirmed. The results are presented in Table 6.

Table 6.  
Numerical results for the frequency of models of English loanwords and ranks

Rank	Name of models	Frequency	Computed
1	M1	2,273	2,272.99
2	M9	23	25.0681
3	M2	19	2.6833
4	M6	11	1.2550
5	M7	8	1.0590
6	M8	7	1.0178
7	M3	3	1.0065
8	M4	2	1.0027
9	M5	2	1.0013
$a = 2271.9915, b = -6.5607, R^2 = 0.9996$			

The fitting curve and the frequency data for the model are shown in Figure 2.

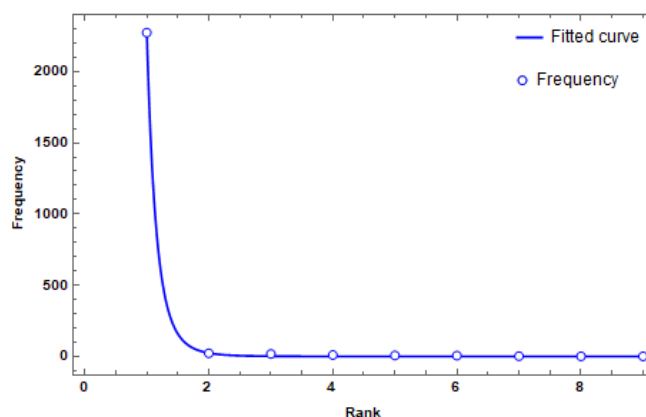


Figure 2. Rank-frequency plot of both of the frequency data and the fitted modelling

Therefore, the above-formulated hypothesis regarding English loanwords in Mongolian has been positively corroborated (see the value of the determination coefficient). Nevertheless, it is necessary to investigate more data of loanword adaptations, both of English origin and coming from other languages, in order to corroborate the aforementioned result and discover possible language laws.

#### 4. Concluding remarks

English-originated loanwords make up an increasingly higher proportion of the words in Mongolian. The number of recent English loanwords in Mongolian is considerable and their influence continues, not only in the domain of sciences and technology, but also in the language of everyday communication. However, a quantitative analysis of the distribution of probabilities of English loanwords in Mongolian has not yet been conducted. In order to obtain the actual proportion of English loanwords in usage, we calculated occurrences of these words in a 299.027-word corpus. As a result, 2.348 English loanwords were discovered, and the main observations being listed here.

First, in order to take a look at the internal dynamics of English loanwords and discover mechanisms that are responsible for borrowings, we collected 118 distinct loanwords together with their frequencies in the real corpus. It was observed that the average frequency of English

loanwords in Mongolian is 0.79% and that English-origin nouns make up a large majority of English loanwords in Mongolian.

Second, regarding the frequency and rank of the loanwords, it is assumed that the list is ordered by decreasing frequencies. We confirmed the hypothesis that the distribution of loanword adaptation in the corpus does exactly satisfy the Zipf-Mandelbrot Law.

Third, loanwords are quickly integrated into the Mongolian language system by transliteration, agglutination, and acronymization. Transliterating and using acronyms are fairly easy ways of accepting foreign terms. On the other hand, there is a strong tendency to create semi-Mongolian equivalents for English words by means of derivation and affixation.

Finally, we developed nine models for English loanword borrowing methods and morphological features, which give a better estimation of the manner of loanword adoption. Nevertheless, it is entirely clear that we will need more data of loanword adaptation both in English and across languages (and fitted by different models, too) before we can move closer to finding a reasonable explanation and a better understanding of the nature of loanword features.

In conclusion, dynamic change of English loanwords has not been studied extensively in the Mongolian vocabulary, probably because of lack of data; nonetheless, this study has found generally consistent patterns and demonstrated the dominant role of the borrowing method and morphological features in the loanword adaptation process. The findings and the analysis of English loanwords in Mongolian can contribute to the theories of loanword adaptation in particular, and to the feature theory in general, and may be helpful for opening a new perspective in the statistical lexical research of the Mongolian language.

### **Acknowledgements**

This work is supported by the National Social Science Foundation of China (Grant No. 17CYY046), and the National Social Science Foundation of China (Grant No. 19AYY018). I would like to express my special thanks to Haitao Liu and Gabriel Altmann, for their advice and warm support during the whole process.

## References

- Altmann, G.** (1983). Das Piotrowski–Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung*. Göttingen: Herodot, 54–90.
- Altmann, G.** (1992). Piotrowski's Law of Language Change. In: Saukkonen, P. (ed.), *What is Language Synergetics?* Oulu: Acta Universitatis Ouluensis, 34–35.
- Altmann, G.** (2002). Zipfian Linguistics. *Glottometrics* 3, 19–26.
- Andersen, S.** (2002). Freedom of Choice and Psychological Interpretation of Word Frequencies in Texts. *Glottometrics* 2, 45–52.
- Bao, M.** (2017). The English Loanword in Mongolian. *Journal of Mongolian Studies in China* 39(3), 14–18.
- Best K.-H.** (2001). Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft* 5, 7–20.
- Best, K.-H.** (2005). Turzismen im Deutschen. *Glottometrics* 11, 56–63.
- Best, K.-H.** (2006). Deutsche Entlehnungen im Englischen. *Glottometrics* 13, 66–72.
- Best, K.-H.** (2013). Iranismen im Deutschen. *Glottometrics* 26, 1–8.
- Best, K.-H.** (2014). Hebraismen im Deutschen. *Glottometrics* 27, 10–17.
- Gnatchuk, H.** (2015). Anglicisms in the Austrian Newspaper *Kleine Zeitung*. *Glottometrics* 31, 38–49.
- Joshi, K., Rajarshi, M. B.** (2017). Estimation of Actual Proportion of Loanwords in a Language. *Sankhyā: Indian Journal of Statistics* 79, 60–75.
- Joos, M.** (1936). Review of Zipf's *The Psycho-Biology of Language*. *Language* 12(3), 196–210.
- Körner, H.** (2004). Zur Entwicklung des deutschen (Lehn-) Wortschatzes. *Glottometrics* 7, 25–49.
- Liu, Y.** (2013). Words of Chinese Origin in German and Their Development Tendency. *Journal of Zhejiang University (Humanities and Social Sciences)* 43(4), 122–134.
- Mandelbrot, B.** (1953). *An Informational Theory of the Statistical Structure of Language*. *Communication Theory*, Washington: Butterworths Scientific Publications, 486–502.
- Mandelbrot, B.** (1962). On the Theory of Word Frequencies and on Related Markovian Models of Discourse. *Structure of Language and Its Mathematical Aspects* 12, 190–219.
- Muren, M.** (1984). The Standardization of Borrowings in Mongolian. *Journal of Inner Mongolia Normal University (Humanities and Social Science)* 39(1), 37–43.
- Norjin, Ts.** (2007). *Monggol kelen-deki angyli jigelege üges-un bičilge-yin dürim ba hamiya büküü ügülel material* [Writing Standards of English Borrowings in Mongolian and Reference], Hohhot: Inner Mongolia Educational Publishing House.
- Poppe, N.** (1954). *Grammar of Written Mongolian*. Wiesbaden: Harrassowitz.
- Prün, C., Zipf, R.** (2002). Biographical notes on G.K. Zipf. *Glottometrics* 3, 1–10.
- Stachowski, K.** (2010). Quantifying Phonetic Adaptations of Russian Loanwords in Dolgan. *Studia Linguistica Universitatis Jagellonicae Cracoviensis* 127, 101–177.
- Stachowski, K.** (2018). A Report on the Distribution of Phonetic Renderings in Loanwords. *Journal of Quantitative Linguistics* 25(1), 38–52.
- Ternes, K.** (2011). Entwicklungen im deutschen Wortschatz. *Glottometrics*, 21, 25–53.
- Tumurtsereen, J.** (2004). *Monggol kelen-ü uge-yin sang-un sudulul* [A Study of Mongolian Lexicology]. Hohhot: Inner Mongolia Educational Publishing House.
- Tumurtoogoo, D.** (2018). *Monggol kelen-dü oroysan gadagadu üge-yin huriyangyui tayilburi toli* [A Concise Explanatory Dictionary of Loan-words in Mongolian]. Ulaanbaatar: Admon Printing Company.
- Zipf, G. K.** (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge: Addison–Wesley.

*English Loanwords in Mongolian Usage*

Appendix I

The frequencies and the ranks of the distinct English loanwords

<b>Latin transliteration</b>	<b>Part of speech</b>	<b>English</b>	<b>Frequency</b>	<b>Rank</b>
<i>mašin</i>	noun	machine	226	1
<i>ķompani</i>	noun	company	219	2
<i>inķenering</i>	noun	engineering	216	3
<i>ķistem</i>	noun	system	193	4
<i>ķadr</i>	noun	cadre	192	5
<i>materiyal</i>	noun	material	158	6
<i>energi</i>	noun	energy	119	7
<i>ķilometr</i>	noun	km	88	8
<i>telvis</i>	noun	television	65	9
<i>ton</i>	noun	ton	59	10
<i>bilet</i>	noun	billet	53	11
<i>radio</i>	noun	radio	49	12
<i>metr</i>	noun	meter	46	13
<i>minüt</i>	noun	minute	43	14
<i>bangqi</i>	noun	bank	36	15
<i>tegnig</i>	noun	technique	31	16
<i>front</i>	noun	front	30	17
<i>elektoron</i>	noun	electron	29	18
<i>professor</i>	noun	professor	29	19
<i>ķiris</i>	noun	series	25	20
<i>ķart</i>	noun	card	25	21
<i>gradüs</i>	noun	grade	19	22
<i>ķilovat</i>	noun	kilowatt	19	23
<i>oryaniy</i>	adjective	organic	18	24
<i>medal</i>	noun	medal	17	25
<i>motor</i>	noun	motor	15	26
<i>element</i>	noun	element	14	27
<i>boķs</i>	noun	box	14	28
<i>baķteri</i>	noun	bacteria	12	29
<i>ķilogram</i>	noun	kilogram	12	30
<i>stok</i>	noun	stock	11	31
<i>doctor</i>	noun	doctor	11	32
<i>diyametr</i>	noun	diameter	10	33
<i>vičat</i>	noun	wechat	10	34
<i>virüs</i>	noun	virus	10	35
<i>cels</i>	noun	celsius	10	36
<i>nomertu</i>	adjective	number	9	37
<i>vidio</i>	noun	video	9	38
<i>ķarton</i>	noun	cartoon	9	39
<i>alkuul</i>	noun	alcohol	8	40
<i>oryaniyту</i>	adjective	organic	8	41



<i>kompiüter</i>	noun	computer	7	42
<i>program</i>	noun	program	6	43
<i>mašinčılal</i>	noun	machine	6	44
<i>sport</i>	noun	sport	6	45
<i>cm</i>	noun	cm	5	46
<i>autočilaysan</i>	verb	auto	5	47
<i>benzen</i>	noun	benzene	5	48
<i>meqaniy</i>	noun	mechanic	5	49
<i>doktorant</i>	noun	doctor student	5	50
<i>radar</i>	noun	radar	5	51
CPI	noun	CPI (consumer price index)	4	52
<i>dollar</i>	noun	dollar	4	53
POS	noun	POS (point of sale)	3	54
<i>autočilal</i>	noun	auto	3	55
<i>alken</i>	noun	alkene	3	56
<i>internet</i>	noun	internet	3	57
<i>net</i>	noun	net	3	58
<i>postdoktor</i>	noun	postdoctor	3	59
<i>qormon</i>	noun	hormone	3	60
<i>milimetr</i>	noun	milimeter	3	61
<i>sentimetr</i>	noun	centimeter	3	62
<i>totem</i>	noun	totem	3	63
<i>deyita</i>	noun	data	3	64
<i>filim</i>	noun	film	3	65
<i>konsül</i>	noun	consul	3	66
<i>klüb</i>	noun	club	3	67
<i>hektar</i>	noun	hectare	3	68
DNA	noun	DNA	2	69
TV	noun	TV	2	70
<i>auto</i>	noun	auto	2	71
<i>etil</i>	noun	ethyl	2	72
<i>eten</i>	noun	ethylene	2	73
<i>inžener</i>	noun	engineer	2	74
<i>olimpiķ</i>	noun	Olympic	2	75
<i>model</i>	noun	model	2	76
<i>spirt</i>	noun	spirit	2	77
<i>traķtor</i>	noun	tractor	2	78
<i>telvisčid</i>	noun	television	2	79
<i>fiziķ</i>	noun	physics	2	80
<i>ķalz</i>	noun	calcium	2	81
<i>ķomandalayči</i>	noun	command	2	82
APEC	noun	APEC (Asia Pacific Economic Cooperation)	1	83
CEO	noun	CEO (chief executive officer)	1	84

English Loanwords in Mongolian Usage

COMT	noun	COMT (catechol-O-methyltransferase)	1	85
GDP	noun	GDP (gross domestic product)	1	86
km	noun	km	1	87
LOGO	noun	LOGO	1	88
PVC	noun	PVC (polyvinyl chloride)	1	89
<i>amper</i>	noun	ampere	1	90
<i>algebra</i>	noun	algebra	1	91
<i>aķademiĉi</i>	noun	academy	1	92
<i>nomer</i>	noun	number	1	93
<i>boķsĉin</i>	noun	box	1	94
QQ	noun	QQ	1	95
<i>pasport</i>	noun	passport	1	96
<i>piza</i>	noun	pizza	1	97
<i>gram</i>	noun	gram	1	98
<i>gen</i>	noun	gene	1	99
<i>general</i>	noun	general	1	100
<i>mašinĉilaysan</i>	verb	machine	1	101
<i>mikrometr</i>	noun	micrometer	1	102
<i>motortu</i>	adjective	motor	1	103
<i>motorĉin</i>	noun	motor	1	104
<i>motorĉilaysan</i>	verb	motor	1	105
<i>liter</i>	noun	liter	1	106
<i>loyiy</i>	noun	logic	1	107
<i>ķistemtū</i>	adjective	system	1	108
<i>ķekūnt</i>	noun	second	1	109
<i>tangķ</i>	noun	tank	1	110
<i>romantiķ</i>	adjective	romantic	1	111
<i>rūbli</i>	noun	ruble	1	112
<i>visa</i>	noun	visa	1	113
<i>vitamin</i>	noun	vitamin	1	114
<i>feudal</i>	noun	feudal	1	115
<i>ķalun</i>	noun	clone	1	116
<i>ķaluri</i>	noun	calorie	1	117
<i>ķod</i>	noun	code	1	118

Appendix II

Borrowing methods and morphological features of the distinct English loanwords

Borrowing methods	Morphological features	Examples	Total
Transliteration	Noun <sup>6</sup>Noun	<p><i>mašin</i> (“machine”), <i>kompani</i> (“company”), <i>inženiring</i> (“engineering”), <i>šistem</i> (“system”), <i>kađr</i> (“cadre”), <i>materiyal</i> (“material”), <i>energi</i> (“energy”), <i>kađometr</i> (“kilometer”), <i>telvis</i> (“television”), <i>ton</i> (“ton”), <i>bilet</i> (“billet”), <i>radio</i> (“radio”), <i>metr</i> (“meter”), <i>minüt</i> (“minute”), <i>bangqi</i> (“bank”), <i>tegnig</i> (“technic”), <i>front</i> (“front”), <i>elektoron</i> (“electron”), <i>professor</i>(„professor), <i>širis</i> (“series”), <i>kađart</i> (“card”), <i>gradüs</i> (“grades”), <i>kađilovat</i> (“kilowatt”), <i>medal</i> (“medal”), <i>motor</i> (“motor”), <i>element</i> (“element”), <i>boks</i> (“box”), <i>bađteri</i> (“bacteria”), <i>kađilogram</i> (“kilogram”), <i>stođ</i> (“stock”), <i>doctor</i> (“doctor”), <i>diyametr</i> (“diameter”), <i>vičat</i> (“WeChat”), <i>virus</i> (“virus”), <i>cels</i> (“Cels”), <i>vidio</i> (“video”), <i>kađarton</i> (“cartoon”), <i>alkuul</i> (“alcohol”), <i>kompiüter</i> (“computer”), <i>program</i> (“programme”), <i>sport</i> (“sport”), <i>benzene</i> (“benzene”), <i>međaniđ</i> (“mechanic”), <i>doktorant</i> (“doctoral student”), <i>radar</i> (“radar”), <i>dollar</i> (“dollar”), <i>alken</i> (“alkene”), <i>internet</i> (“internet”), <i>net</i> (“net”), <i>postdođktor</i> (“post-doctor”), <i>qormon</i> (“hormone”), <i>milimetr</i> (“millimetre”), <i>sentimetr</i> (“centimetre”), <i>totem</i> (“totem”), <i>deđvita</i> (“data”), <i>filim</i> (“film”), <i>kađonsül</i> (“consul”), <i>kađlüb</i> (“club”), <i>heđktar</i> (“hectare”), <i>auto</i> (“auto”), <i>etil</i> (“ethyl”), <i>eten</i> (“ethylene”), <i>inđerener</i> (“engineer”), <i>olimpiđ</i> (“Olympics”), <i>model</i> (“model”), <i>spirt</i> (“spirit”), <i>trađktor</i> (“tractor”), <i>fiziđ</i> („Physics), <i>kađalz</i> („calcium), <i>amper</i> (“ampere”), <i>algebra</i> (“algebra”), <i>nomer</i> (“number”), <i>passport</i> (“passport”), <i>piza</i> (“pizza”), <i>gram</i> (“gram”), <i>gen</i> (“gene”), <i>general</i> (“general”), <i>mikrometr</i> (“micrometer”), <i>liter</i> (“liter”), <i>lođiđ</i> (“logic”), <i>sekünt</i> (“second”), <i>tangđ</i> (“tank”), <i>rübli</i> (“Rouble”), <i>visa</i> (“visa”), <i>vitamin</i> (“vitamin”), <i>feudal</i> (“feudal”), <i>kađalun</i> (“clone”), <i>kađaluri</i> (“calories”), <i>kađod</i> (“code”)</p>	89

6 “<” means “developed from”.

*English Loanwords in Mongolian Usage*

	Adjective < Adjective	<i>oryaniy</i> (“organic”), <i>romantiy</i> (“romantic”)	2
Agglutination	Noun < Noun + Suffix -či/-čin	<i>motorčin</i> (from “motor”), <i>boḵščin</i> (from “box”), <i>akademiči</i> (from “academy”)	3
	Noun < Noun + Suffix -la-/le- (-ra-/re-) + Suffix -γči/-gči	<i>ḵomandalayči</i> (from “command”)	1
	Noun < Noun + Suffix -čila-/čile- + Suffix -l	<i>autočilal</i> (from auto) <i>mašinčilal</i> (from machine)	2
	Adjective < Noun + Suffix -tu/-tü	<i>motortu</i> (from “motor”), <i>nomertu</i> (from “number”), <i>šistemtü</i> (from “system”)	3
	Adjective < Adjective + Suffix -tu/-tü	<i>oryaniytu</i> (from “organic”)	1
	Verb < Noun + Suffix -čila-/čile- + Nomen Perfecti - ysan-/gsen	<i>mašinčilaysan</i> (from “machine”), <i>autočilaysan</i> (from “auto”), <i>motorčilaysan</i> (from “motor”)	3
Acronymization	Noun < Noun	cm, km, CPI, TV, GDP, IP, IT, QQ, COMT, APEC, LOGO, PVC, DNA, POS	14
<b>Total</b>			118

# A Quantitative Study on English Polyfunctional Words

Lu Wang<sup>1</sup>,  
Yahui Guo<sup>2</sup>,  
Chengcheng Ren<sup>3</sup>

## Abstract

This paper reports quantitative research on the parts of speech of English words using the data from British National Corpus. Most of the part-of-speech investigations focus on the rank-frequency distribution. However, in English and many other languages, we can find that part of speech can be ambiguous. For example, *hope* can be a noun and a verb. Such words are called polyfunctional words, while other words, which belong to only one part of speech, are called monofunctional words. The number of parts of speech that a word belongs to is referred to as polyfunctionality. First, we study polyfunctionality distribution of English words and find that the Shenton-Skees-geometric and the Waring distributions capture the data very well. Then, we group words according to their part of speech, e.g., monofunctional nouns, like *Saturday*, and polyfunctional nouns, like *hope* (noun, verb) compose noun group, and try to work out a general model for all the groups. The result is that the extended positive binomial distribution captures all the groups except the article group, because of the sparsity of the data. Last, we study the diversification variants. Since there are polyfunctional words in each group – e.g., in a noun group, a polyfunctional noun may also be a verb, we consider the “verb” function as a diversification variant and try to model the rank-frequency distribution of variants with the Popescu-Altman function, as used in the previous investigation. The results show very good fit for all groups except conjunction group.

**Keywords:** *polyfunctionality, polyfunctional words, parts of speech, BNC.*

## 1. Introduction


There is a phenomenon in English, as well as in many other languages, that the same word may have several different grammatical functions. For example, in the sentence *A canner can can a can*, the first *can* is a modal verb which means “be able to”, the second *can* is a verb which means “to preserve food by putting it in a can”, the last *can* is a noun which means “a metal container”. Linguists call this class cleavage (Bloomfield, 1933), multiple class membership (Bloomfield, 1933; Allerton, 1979; Biber, 1999; Hudson & Hudson, 2007; Jackson & Amvela, 2007; Jackson, 1988; Nida, 1948), multifunctionality (Harris, 1946; Braun, 2009), decategorization (Hopper & Thompson, 1984), intercategory polysemy (Zawada, 2005), zero-derivation (Kastovsky, 2005), transcategorization (Halliday & Matthiessen, 2006), heterosemy (Enfield, 2006), conversion (Balteiro, 2007), word class expansion (Fan & Altmann, 2008) and polyfunctionality (Wang, 2016). In this paper, we follow the terminology from Wang (2016): words that have more than one part of speech are called polyfunctional words, while words with only one part of speech are called monofunctional words; the number of parts of speech they have is referred to as polyfunctionality (PF). Polyfunctionality is different from polysemy, because all the meanings of a polysemous word may belong to one part of speech. The word is monofunctional, with  $PF = 1$ . Only when the meanings belong to different parts of speech, the word is also polyfunctional, such as *can* in the abovementioned sentence,  $PF = 2$ . The present

---


<sup>1</sup> Computational Linguistics and Digital Humanities, University of Trier, Germany.

[wanglu-chn@hotmail.com](mailto:wanglu-chn@hotmail.com).  <http://orcid.org/0000-0002-4366-2739>

<sup>2</sup> Handan Vocational Center, Handan, China. [894132316@qq.com](mailto:894132316@qq.com).

 <http://orcid.org/0000-0003-0938-4672>

<sup>3</sup> School of Foreign Languages, Dalian Maritime University, Dalian, China. [1026806608@qq.com](mailto:1026806608@qq.com).

 <http://orcid.org/0000-0002-1414-8124>

paper attempts to study the polyfunctionality distribution of English words.

We can group words according to their parts of speech, e.g., monofunctional nouns, like *Saturday*, and polyfunctional nouns, like *hope* (noun, verb) compose the noun group; monofunctional verbs, like *rely*, and polyfunctional verbs, like *hope* (verb, noun) compose the verb group; monofunctional adjectives, like *fantastic*, and polyfunctional adjectives, like *mean* (adjective, noun, verb) compose the adjective group, etc. Thus, each group forms a polyfunctionality distribution. We try to analyse the polyfunctionality distributions and find a general model to capture all the groups.

If a word group includes polyfunctional words, the part(s) of speech other than the shared one are variants. Consider the above-mentioned five words as a small corpus: the noun group includes *Saturday*, *hope* and *mean*, where the part of speech “verb” from *hope* and *mean* and the part of speech “adjective” from *mean* are variants. Similarly, in the verb group (*rely*, *hope*, *mean*), the parts of speech “noun” and “adjective” are variants; in the adjective group (*fantastic*, *mean*), the part of speech “noun” and “verb” are variants. We focus on the variants and their rank-frequency distribution of each group and test if all the groups abide by the same model.

## 2. Data

Our data is extracted from British National Corpus<sup>4</sup> (BNC), which contains over 100 million words. The whole corpus is tagged with the Constituent Likelihood Automatic Word-tagging System<sup>2</sup> (CLAWS) according to C5 tagset<sup>3</sup>. This tagset contains over 60 tags, in which 53 denote parts of speech, such as NN0 refers to common noun, neutral for number; NN1 refers to singular common noun; NN2 refers to plural common noun; NP0 refers to proper noun. In analogy to these noun tags, most of the 53 tags are subcategories. However, it is not our aim to distinguish sub-classes of a part of speech, so we merge them to their corresponding part of speech (POS) – adjective, article, adverb, conjunction, numerals, pronoun, interjection, noun, preposition and verb – ten classes in total as shown in Table 1.

Within BNC there is 3.3% part-of-speech ambiguity<sup>5</sup>. The word *round* in the following sentence is an example: its tag PRP-AVP means preposition or adverb particle.

*James Rogers is 23 and needs <w PRP-AVP>round the clock medical attention.*

Since those ambiguity tags<sup>6</sup> increase the error rate<sup>7</sup> of BNC part-of-speech tagging, they are excluded from the present study. Thus, the error rate reduces to 0.7%, which we consider acceptable. We also exclude word combinations, words with hyphens, words with Arabic numerals, incorrect spellings and non-English words as shown in the following sentences. Words are not lemmatized, since a previous study (Wang and Guo, 2018) adopts English dictionary data (i.e. lemmas). In this study we focus on word forms in running texts. Finally, we obtain 278,966 word types, case insensitive.

---

<sup>4</sup> <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554>

<sup>5</sup> <http://ucrel.lancs.ac.uk/claws/>

<sup>6</sup> <http://ucrel.lancs.ac.uk/claws5tags.html>

<sup>7</sup> <http://ucrel.lancs.ac.uk/bnc2/bnc2error.htm>

*Ours was an <w AJ0>ad hoc group.*

*Pound was a bohemian figure, despite his Quaker origins, who espoused an <w AJ0>anti-credit economic philosophy which thrust him into <w NN1>anti-Semitism.*

*Times to beat from last year are Kevin Brown's <w CRD>1:07:21 and Chris Buckley's <w CRD>1:24:25 in the women's race.*

*<w AT0>The effort of the natives to be heard by the Greeks was evidently encouraged by the curiosity of the Greeks about the natives and, generally speaking, corresponded to the political situation.*

*From <w AT0>de lift to <w AT0>de balcony.*

Table 1.  
BNC tags and parts of speech

BNC tags	Description	POS
AJ0	Adjective (general or positive) (e.g. <i>good, old</i> )	Adjective (adj.)
AJC	Comparative adjective (e.g. <i>better, older</i> )	
AJS	Superlative adjective (e.g. <i>best, oldest</i> )	
AT0	Article (e.g. <i>the, a, an, no</i> )	Article (art.)
AV0	General adverb: an adverb not subclassified as AVP or AVQ (e.g. <i>often, furthest</i> )	Adverb (adv.)
AVP	Adverb particle (e.g. <i>up, out</i> )	
AVQ	Wh-adverb (e.g. <i>when, how</i> )	
EX0	Existential <i>there</i> , i.e. <i>there</i> occurring in the “ <i>there is ...</i> ” or “ <i>there are ...</i> ” construction	
XX0	The negative particle “ <i>not</i> ” or “ <i>n't</i> ”	
CJC	Coordinating conjunction (e.g. <i>and, or</i> )	Conjunction (conj.)
CJS	Subordinating conjunction (e.g. <i>although, when</i> )	
CJT	The subordinating conjunction “ <i>that</i> ”	
CRD	Cardinal number (e.g. <i>one, 3, fifty-five, 3609</i> )	numerals (num.)
ORD	Ordinal numeral (e.g. <i>first, sixth, 77th, last</i> ).	
DPS	Possessive determiner-pronoun (e.g. <i>your, their, his</i> )	Pronoun (pron.)
DT0	General determiner-pronoun: i.e. a determiner-pronoun which is not a DTQ or an AT0.	
DTQ	Wh-determiner-pronoun (e.g. <i>which, what, whose, whichever</i> )	
PNI	Indefinite pronoun (e.g. <i>none, one [as pronoun]</i> )	
PNP	Personal pronoun (e.g. <i>I, them</i> )	
PNQ	Wh-pronoun (e.g. <i>who, whoever</i> )	

*A Quantitative Study on English Polyfunctional Words*

PNX	Reflexive pronoun (e.g. <i>myself, ourselves</i> )	
ITJ	Interjection or other isolate (e.g. <i>oh, yes, mhm, wow</i> )	Interjection (interj.)
NN0	Common noun, neutral for number (e.g. <i>aircraft, data</i> )	Noun (n.)
NN1	Singular common noun (e.g. <i>pencil, goose</i> )	
NN2	Plural common noun (e.g. <i>pencils, geese</i> )	
NP0	Proper noun (e.g. <i>London, IBM</i> )	
PRF	The preposition “ <i>of</i> ”	Preposition (prep.)
PRP	Preposition (except for “ <i>of</i> ”) (e.g. <i>at, in</i> )	
VBB	The present tense forms of the verb <i>BE</i> , except for <i>is, 's</i> : i.e. <i>am, are, 'm, 're</i> and <i>be</i> [subjunctive or imperative]	Verb (v.)
VBD	The past tense forms of the verb <i>BE</i> : <i>was</i> and <i>were</i>	
VBG	The -ing form of the verb <i>BE</i> : <i>being</i>	
VBI	The infinitive form of the verb <i>BE</i> : <i>be</i>	
VBN	The past participle form of the verb <i>BE</i> : <i>been</i>	
VBZ	The -s form of the verb <i>BE</i> : <i>is, 's</i>	
VDB	The finite base form of the verb <i>BE</i> : <i>do</i>	
VDD	The past tense form of the verb <i>DO</i> : <i>did</i>	
VDG	The -ing form of the verb <i>DO</i> : <i>doing</i>	
VDI	The infinitive form of the verb <i>DO</i> : <i>do</i>	
VDN	The past participle form of the verb <i>DO</i> : <i>done</i>	
VDZ	The -s form of the verb <i>DO</i> : <i>does, 's</i>	
VHB	The finite base form of the verb <i>HAVE</i> : <i>have, 've</i>	
VHD	The past tense form of the verb <i>HAVE</i> : <i>had, 'd</i>	
VHG	The -ing form of the verb <i>HAVE</i> : <i>having</i>	
VHI	The infinitive form of the verb <i>HAVE</i> : <i>have</i>	
VHN	The past participle form of the verb <i>HAVE</i> : <i>had</i>	
VHZ	The -s form of the verb <i>HAVE</i> : <i>has, 's</i>	
VM0	Modal auxiliary verb (e.g. <i>will, could</i> )	
VVB	The finite base form of lexical verbs (e.g. <i>forget, send</i> ) [Including the imperative and present subjunctive]	
VVD	The past tense form of lexical verbs (e.g. <i>forgot, sent</i> )	
VVG	The -ing form of lexical verbs (e.g. <i>forgetting, sending</i> )	



VVI	The infinitive form of lexical verbs (e.g. <i>forget, send</i> )	
VVN	The past participle form of lexical verbs (e.g. <i>forgotten, sent</i> )	
VVZ	The -s form of lexical verbs (e.g. <i>forgets, sends</i> )	

### 3. Discussion

#### 3.1 Polyfunctionality distribution

We find that 27,592 out of the total 278,966 words in the BNC data are polyfunctional accounting for 9.89%, and 251,374 words are monofunctional accounting for 90.11%. Polyfunctionality arranges from 1 to 6 as shown in Table 2. The word *like* extends to six parts of speech as shown in the following sentences.

*Corridors that twisted upwards <prep.>like corkscrews.*

*In the woman's eyes he saw a <adj.>like recognition and knew his senses did not deceive him.*

*Pull out the magazine schemes that appeal most to you and stick them in a file; mark the pages you <v.>like in books.*

*What I'd like is just a few regulars, that'd come by appointment, <adv.>like, so I could stay at home.*

*He knew full well: by Acts of Parliament, voted by landlords to benefit their <n.>like.*

*They'll miss him, <conj.>like they missed many of their first team players who're saving themselves for a vital league game next week.*

Fan & Altmann (2008) found that the Shenton-Skees geometric distribution is an excellent model for their 165-English-word data:

$$P_x = pq^{x-1} \left[ 1 + a \left( x - \frac{1}{p} \right) \right],$$

where  $x$  stands for PF ( $x = 1, 2, 3, \dots$ );  $p, q$  and  $a$  are parameters ( $0 < p < 1, q = 1-p, 0 < a < \frac{1}{q-1}$ ). Later studies on the polyfunctionality distribution model include Wang (2016), which adopts the Modern Chinese Dictionary (the 5th edition), and Wang & Guo (2018), which adopts CELEX dictionary data (Baayen et al., 1995) from German, Dutch, and English. The results show the Waring distribution –

$$P_x = \frac{b}{b+n} * \frac{n^{(x)}}{(b+n+1)^{(x)}}$$

$$x = 0, 1, 2, \dots (b > 0, n \geq 0)$$

where  $x$  stands for PF ( $x = 0, 1, 2, \dots$ );  $b$  and  $n$  are parameters ( $b > 0, n \geq 0$ );  $n^{(x)} = n(n+1)(n+2)\dots(n+x-1)$  – can fit the data of all these four languages better than Shenton-Skees geometric distribution. In the present study, both models are fitted as shown in Table 2. C and R<sup>2</sup> values indicate goodness-of-fit. C is the coefficient of discrepancy. The fit is considered to be acceptable, when  $C \leq 0.02$ . R<sup>2</sup> is the coefficient of determination, originally applied to linear models, but also used to evaluate non-linear models, if they obtain a sufficiently high R<sup>2</sup> value:

$R^2 > 0.9$  is considered to be a good fit and  $R^2 > 0.8$  an acceptable fit. (Macůtek & Wimmer, 2013).

Table 2.  
Fitting Waring and Shenton-Skees geometric distributions to the polyfunctionality data

x[i]	f[i]	Waring	Shenton-Skees-geometric
1	251374	251800.7	251323.8
2	25191	24519.98	25309.55
3	2308	2387.72	2151
4	84	232.51	168.15
5	8	22.64	12.5
6	1	2.44	0.97
		b = 77730601.38 n = 8385897.13	p = 0.9391 a = 0.6265
		C = 0.0005 $R^2 > 0.999$	C = 0.0002 $R^2 > 0.999$

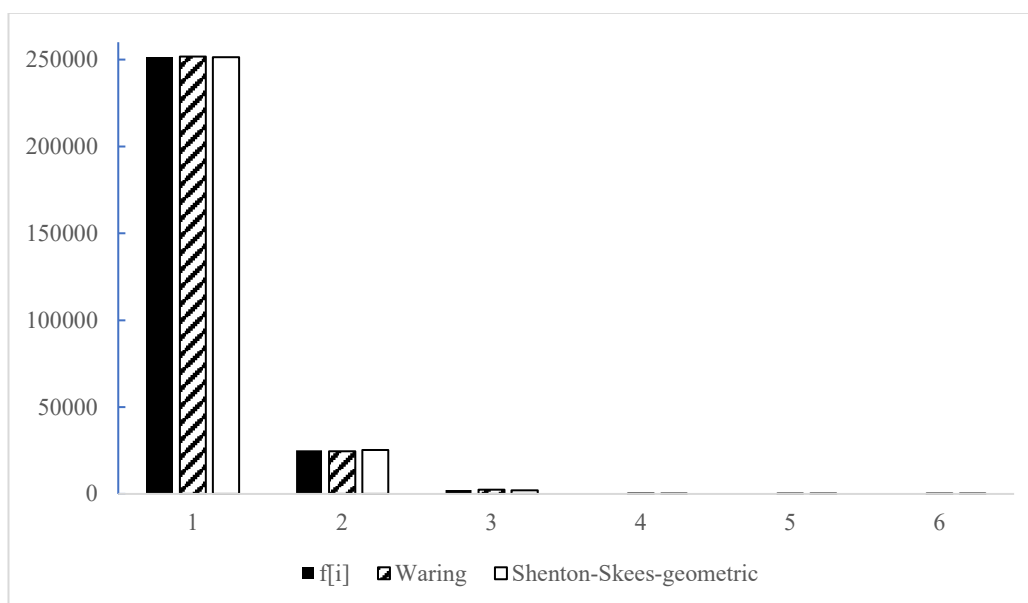


Figure 1. Fitting Waring and Shenton-Skees geometric distributions to polyfunctionality of BNC data

### 3.2 Polyfunctionality distribution of word groups

The polyfunctionality distribution of each group is demonstrated in Table 3. The data for verb, preposition and conjunction groups show bell-shaped forms. The data for noun, adjective, adverb, pronoun, numerals and interjection groups are monotonically decreasing. The data of these nine groups abide by the extended positive binomial distribution

$$P_x = \begin{cases} 1 - \alpha, & x = 0 \\ \frac{\alpha \binom{n}{x} p^x (1 - p)^{n-x}}{1 - (1 - p)^n}, & x = 1, 2, 3, \dots \end{cases}$$

where  $x$  stands for PF;  $n, p$  and  $\alpha$  are parameters,  $n = 1, 2, 3, \dots, 0 \leq p \leq 1, 0 \leq \alpha \leq 1$ . Since the article group consists of only three data points (Table 4), it is insufficient to fit a three-parameter model like extended positive binomial distribution. We tried other models with fewer parameters, such as the binomial distribution, because the extended positive binomial belongs to the binomial family, and Shenton-Skees-geometric and Waring, which are reported to capture polyfunctionality distributions by Fan and Altmann (2008) and Wang (2016) respectively. For such a small data set, we use the  $\chi^2$ -test and its  $p$ -value instead of the discrepancy coefficient  $C$ , which is applied only when the sample is large and  $\chi^2$ -test loses its reliability (Macůtek & Wimmer, 2013). The results show that, the  $p$ -values of binomial, Shenton-Skees-geometric and Waring are 0.4645, 0.4625 and 0.7438 respectively, all greater than 0.05, thus indicating the models are acceptable. Caution is required because the article data set with only five words and three data points is quite small, therefore the results are questionable. However, they are still worth reporting to provide information for further studies.

Table 3.

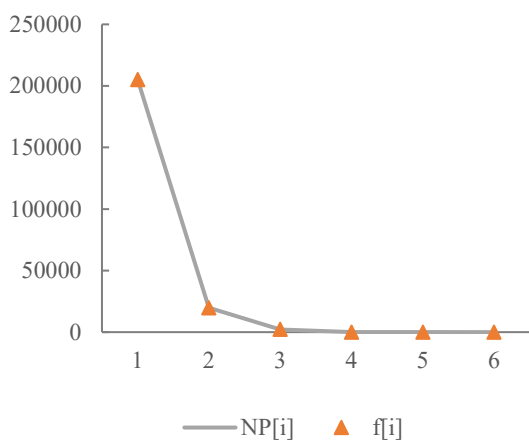
Fitting extended positive binomial distribution to the polyfunctional data of word groups

PF	n.		v.		adj.		adv.	
x[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]	NP[i]
1	205222	205222	16811	16811	23019	23019	5612	5612
2	19904	19910.57	18575	18580.81	11018	11067.13	548	508.74
3	2279	2230.91	2156	2103.7	2221	2050.88	161	224.19
4	78	124.98	67	119.09	71	190.03	71	49.4
5	8	3.5	8	3.37	6	8.8	7	5.44
6	1	0.04	1	0.04	1	0.16	1	0.24
	n = 5 p = 0.0531 $\alpha$ = 0.0979		n = 5 p = 0.0536 $\alpha$ = 0.5531		n = 5 p = 0.0848 $\alpha$ = 0.3665		n = 5 p = 0.1806 $\alpha$ = 0.1231	
	C = 0.0001 $R^2 > 0.999$		C = 0.0009 $R^2 > 0.999$		C = 0.0025 $R^2 = 0.9999$		C = 0.0049 $R^2 = 0.9998$	

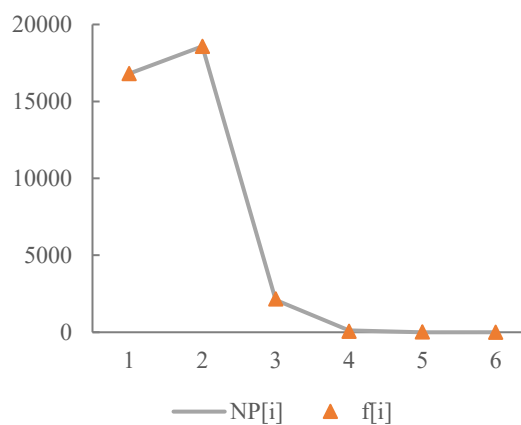
PF	prep.		conj.		pron.		num.		interj.		art.
x[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]	NP[i]	f[i]
1	32	32	18	18	123	123	309	309	226	226	2
2	51	48.75	23	21.75	54	54.51	45	42.78	162	154.08	2
3	39	44.47	14	16.18	22	21.04	12	15.55	20	31.04	0

*A Quantitative Study on English Polyfunctional Words*

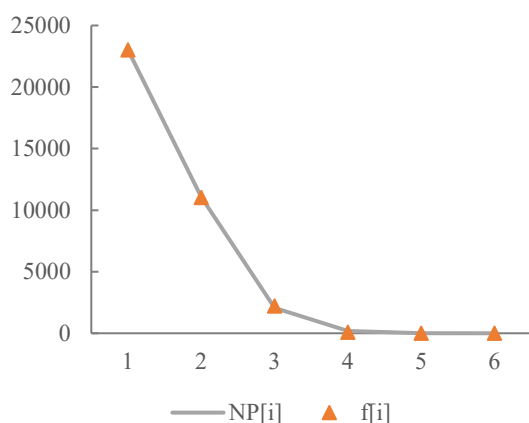
4	27	22.54	8	7.29	5	5.31	3	2.51	5	2.78	1
5	6	6.86	2	2.22	1	1.14	1	0.15	1	0.09	
6	1	1.38	1	0.57							
	n = 7 p = 0.2332 $\alpha$ = 0.7949		n = 12 p = 0.1191 $\alpha$ = 0.7273		n = 51 p = 0.0152 $\alpha$ = 0.4		n = 4 p = 0.1951 $\alpha$ = 0.1649		n = 4 p = 0.1184 $\alpha$ = 0.4541		
	C = 0.012 R <sup>2</sup> = 0.9724		C = 0.0068 R <sup>2</sup> = 0.9833		C = 0.0004 R <sup>2</sup> = 0.9999		C = 0.0043 R <sup>2</sup> = 0.9997		C = 0.0187 R <sup>2</sup> = 0.9958		



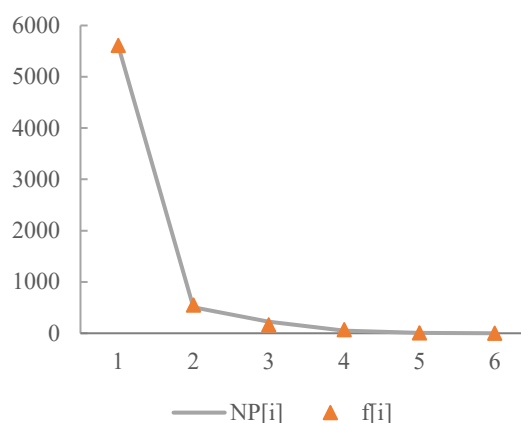
(1) nouns



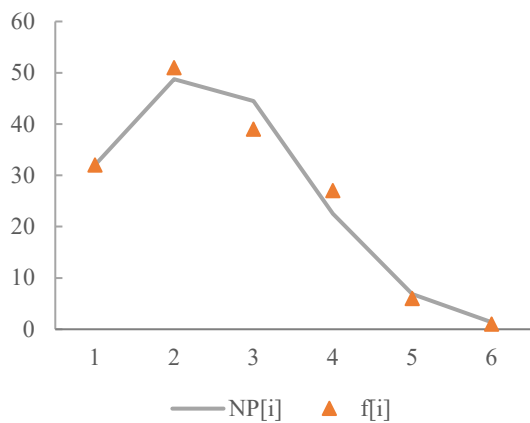
(2) verbs



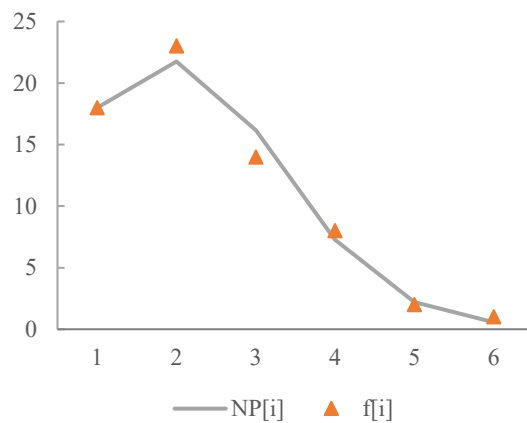
(3) adjectives



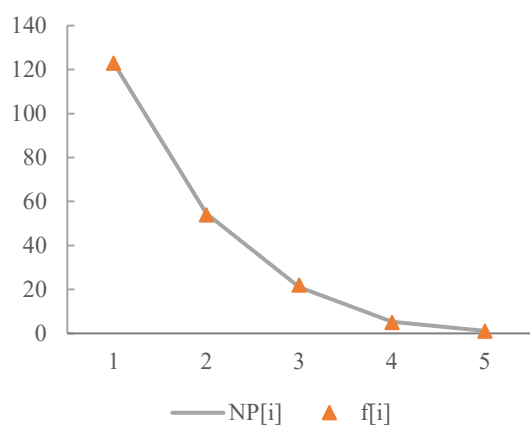
(4) adverbs



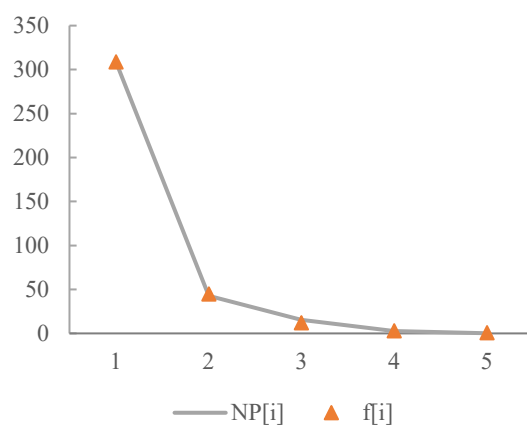
(5) prepositions



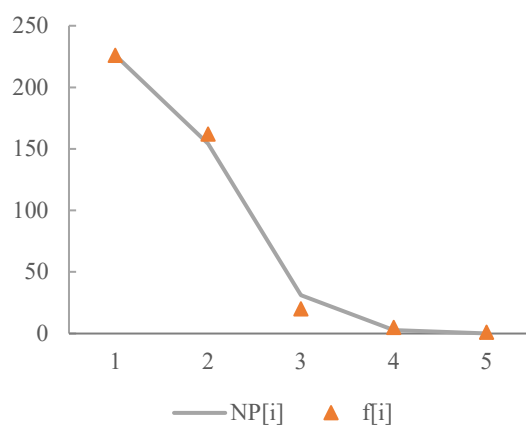
(6) conjunctions



(7) pronouns



(8) numerals



(9) interjections

Table 4.  
Article group

Words	POS	PF	Examples
every	art.	1	<i>It should have been no surprise: Johnson scored two of the goals in last season's play-off final against Brighton and was booed &lt;art.&gt;every time he touched the ball by home fans with good memories.</i>
the	art.	1	<i>But &lt;art.&gt;the Greeks were seldom in a position to check what the natives told them: they did not know the languages.</i>
a	art. n.	2	<i>It is delighted &lt;art.&gt;a college preparatory school headed by Dr Norman &lt;n.&gt;A Palmer.</i>
an	art. n.	2	<i>As early as the 17th minute, Johnson sold McCarthy &lt;art.&gt;an extravagant dummy and hit a 20-yard shot that just cleared Digweed's crossbar.</i>  <i>A mile to the west of Loch Gorm the road to Gruinart cuts through a settlement at &lt;n.&gt;An Sithern of about a dozen round houses which were constructed in the late Bronze Age and which a cursory examination has shown that they were used at least three times.</i>
no	art. adv. interj. n.	4	<i>Nevertheless &lt;art.&gt;no reader takes the passage like that.</i>  <i>But there is &lt;adv.&gt;no doing so unless we accept that the literal writer has an imagination.</i>  <i>Oh &lt;interj.&gt;no!</i>  <i>Fact &lt;n.&gt;no 11, Income Support for residential and nursing homes, has been updated.</i>

### 3.3 Diversification variants

We have observed that each group includes polyfunctional words. Take the article group as an example – the words *a*, *an* and *no* are polyfunctional words, in that all the three have the noun function besides the article function. The part of speech “noun” is considered as a variant with three observations. In the same way, the parts of speech “adverb” and “interjection” from the word *no* are variants, each obtaining one observation. Thus, the rank-frequency distribution is obtained and shown in Table 5. These variants are generated by part-of-speech diversification processes such as conversion. According to diversification studies (Köhler & Altmann, 2009), if an entity diversifies, the frequency of individual elements abides by a regular distribution or a function. Here, we choose Popescu-Altmann’s function (Popescu, Altmann & Köhler, 2009):

$$y = 1 + ae^{-bx} ,$$

which is proved to be the best model for diversification variants in Wang (2016). The fitting results are shown in Table 5 and Figure 3, mostly acceptable except the conjunction group obtaining  $R^2 = 0.6578$ .

Table 5.

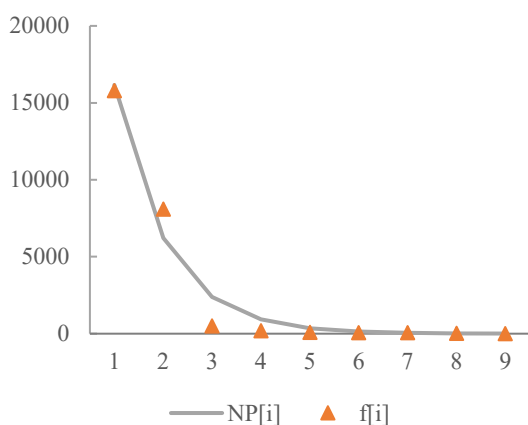
Fitting Popescu-Altmann's function to rank-frequency distribution of part-of-speech variants

x[i]	n.			v.			adv.		
	variants	f[i]	NP[i]	variants	f[i]	NP[i]	variants	f[i]	NP[i]
1	v.	15773	16160.30	n.	15773	16061.20	n.	490	528.38
2	adj.	8083	6216.70	adj.	7147	5588.55	adj.	410	291.37
3	adv.	490	2391.88	adv.	96	1944.98	v.	96	160.88
4	interj.	177	920.66	prep.	42	677.34	prep.	63	89.03
5	prep.	71	354.75	interj.	22	236.31	pron.	25	49.47
6	num.	59	137.07	pron.	19	82.87	conj.	22	27.69
7	pron.	55	53.34	conj.	18	29.48	interj.	7	15.69
8	conj.	22	21.13	num.	8	10.91	num.	2	9.09
9	art.	3	8.74				art.	1	5.45
	a = 42010.1968 b = 0.9554 R <sup>2</sup> = 0.968			a = 46161.3184 b = 1.0558 R <sup>2</sup> = 0.9726			a = 957.8383 b = 0.5967 R <sup>2</sup> = 0.9254		

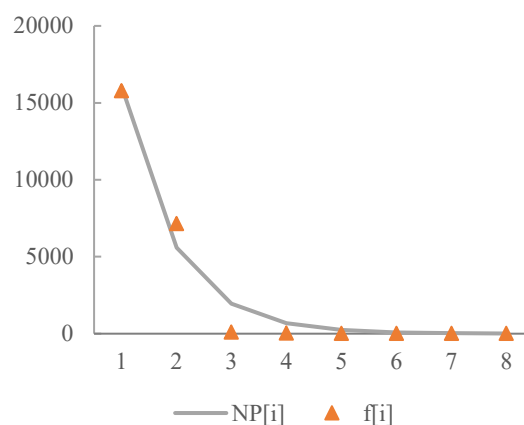
x[i]	adj.			pron.			prep.		
	variants	f[i]	NP[i]	variants	f[i]	NP[i]	variants	f[i]	NP[i]
1	n.	8083	8872.82	n.	55	54.28	n.	71	77.53
2	v.	7147	4362.20	adv.	25	28.18	adv.	63	55.37
3	adv.	410	2144.87	v.	19	14.87	v.	42	39.63
4	prep.	37	1054.88	adj.	7	8.08	adj.	37	28.44
5	interj.	10	519.07	num.	4	4.61	conj.	21	20.49
6	pron.	7	255.67	interj.	3	2.84	num.	3	14.85
7	num.	5	126.19	prep.	2	1.94	pron.	2	10.84
8	conj.	3	62.54	conj.	2	1.48			
	a = 18047.6134 b = 0.7101 R <sup>2</sup> = 0.8512			a = 104.4288 b = 0.6729 R <sup>2</sup> = 0.9876			a = 107.7328 b = 0.3419 R <sup>2</sup> = 0.9102		

*A Quantitative Study on English Polyfunctional Words*

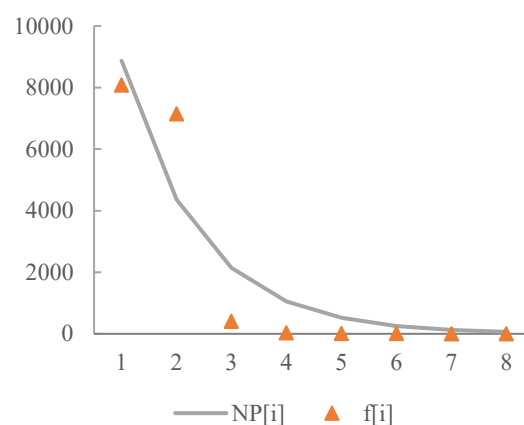
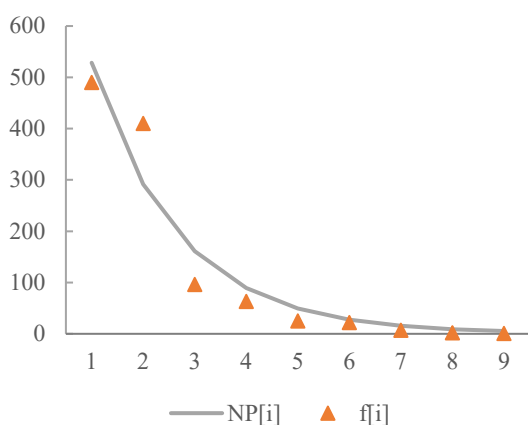
x[i]	num.			interj.			conj.			art.		
	variants	f[i]	NP[i]	variants	f[i]	NP[i]	variants	f[i]	NP[i]	variants	f[i]	NP[i]
1	n.	59	58.93	n.	177	176.87	n.	22	25.80	n.	3	2.99
2	v.	8	8.98	v.	22	23.87	adv.	22	19.95	interj.	1	1
3	adj.	5	2.10	adj.	10	3.97	prep.	21	15.49	adv.	1	1
4	pron.	4	1.15	adv.	7	1.39	v.	18	12.07			
5	prep.	3	1.02	pron.	3	1.05	adj.	3	9.46			
6	adv.	2	1.00	num.	1	1.01	pron.	2	7.47			
7	interj.	1	1.00	art.	1	1.00						
a = 420.2848 b = 1.9817 R <sup>2</sup> = 0.9915			a = 1352.518 b = 2.0399 R <sup>2</sup> = 0.997			a = 32.4507 b = 0.2688 R <sup>2</sup> = 0.6578			a = 18305.8794 b = 9.1218 R <sup>2</sup> > 0.999			



(1) noun group



(2) verb group





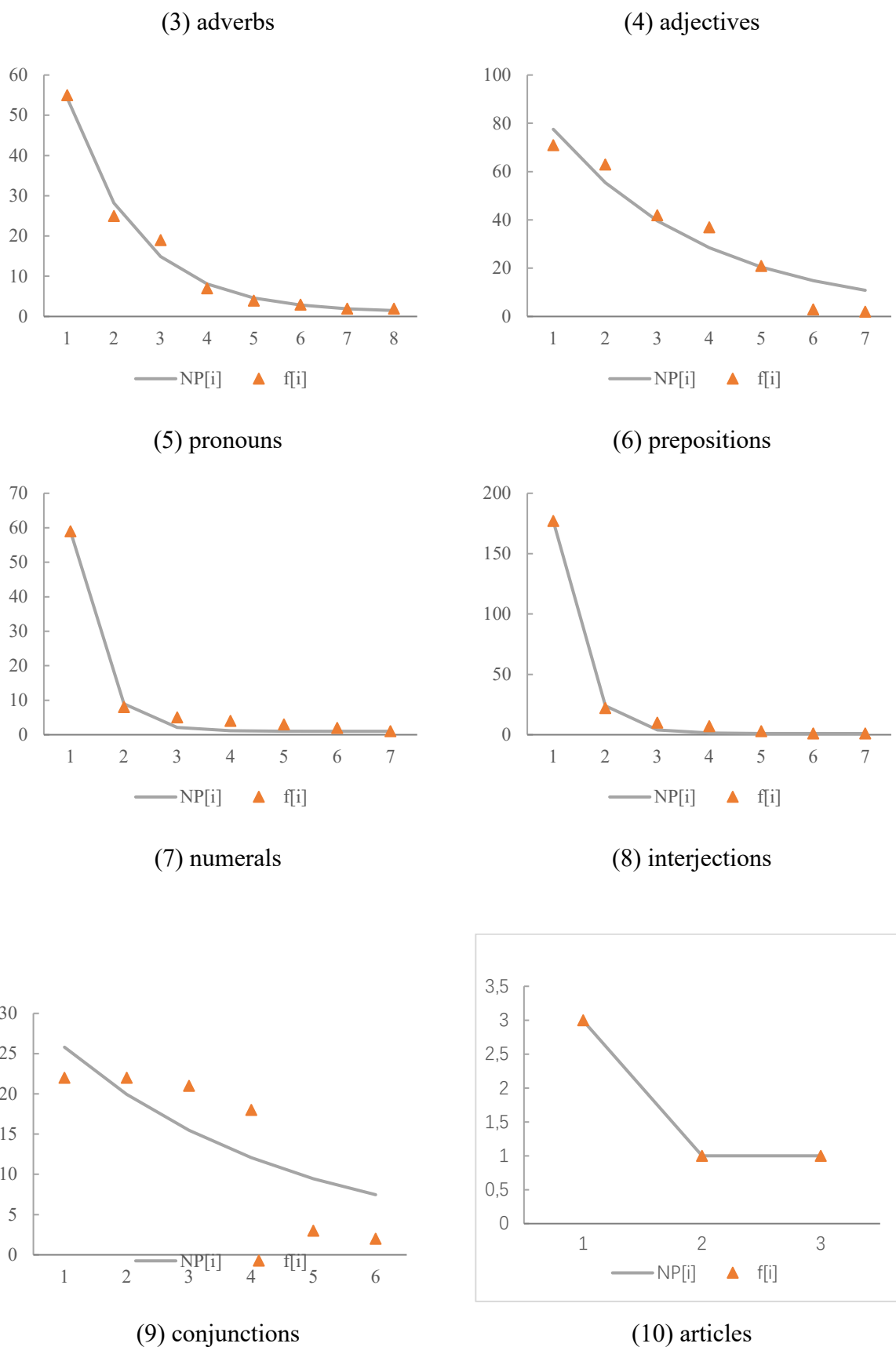


Figure 3. Fitting Popescu-Altmann's function to rank-frequency distribution of part-of-speech

#### **4. Discussion and Conclusion**

The present paper shows the results of investigating polyfunctional English words based on data extracted from the BNC. The polyfunctionality data are shown to abide by Shenton-Skees-geometric and Waring distributions. Further, the polyfunctionality data of the word groups that classified according to part of speech were analysed. Extended positive binomial distribution captures nine groups perfectly. The article group, too small for this model, obtains good fitting results from binomial, Shenton-Skees-geometric and Waring distributions, but it should be noted that, the reliability of fitting such a small data set is doubtful. The rank-frequency distribution of diversification variants of the word groups abide by Popescu-Altmann function.

## References

- Allerton, D. J.** (1979). *Essentials of Grammatical Theory: A Consensus View of Syntax and Morphology*. London: Routledge & Kegan Paul.
- Baayen, R. H., Piepenbrock, R. and Gulikers, L.** (1995). *The CELEX lexical database (CD-ROM)*, Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Balteiro, I.** (2007). *A Contribution to the Study of Conversion in English*. Münster /NewYork /München /Berlin: Waxmann.
- Biber, D., Johansson, S., Leech, G., Conrad S., & Finegan, E.** (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bloomfield, L.** (1933). *Language*. New York: Holt, Rinehart, & Winston.
- Braun, M.** (2009). *Word-Formation and Creolisation: The Case of Early Sranan*. Tübingen: Max Niemeyer.
- Enfield, N. J.** (2006). Heterosemy and the grammar-lexicon trade-off. In: Ameka, F. K., Dench, A. C. & N. Evans. (eds), *Catching Language: The Standing Challenge of Grammar Writing*. Berlin: Walter de Gruyter, 297 - 320.
- Fan, F., Altmann, G.** (2008). On Meaning Diversification in English. *Glottometrics* 17, 69–81.
- Halliday, M. A. K., & Matthiessen, C.** (2006). *Construing Experience through Meaning: A Language-based Approach to Cognition*. London: A&C Black.
- Harris, Z. S.** (1946). From Morpheme to Utterance. *Language*, 22, 161-183.
- Hopper, P. & Thompson, S.** (1984). The Discourse Basis for Lexical Categories in Universal Grammar. *Language*, 60, 703 - 752.
- Hudson, R. A. & Hudson, R.** (2007). *Language Networks: The New Word Grammar*. Oxford University Press.
- Kastovsky, D.** (2005). Conversion and/or Zero: Word-formation Theory, Historical Linguistics, and Typology. In: Bauer L. & Valera S. (eds), *Approaches to Conversion/Zero-derivation*, Münster: Waxmann, 31-49.
- Köhler, R. & Altmann, G.** (2009) *Problems in Quantitative Linguistics 2*. Trier: Ram Verlag.
- Jackson, H.** (1988). *Words and Their Meaning*. London: Longman.
- Jackson, H. & Amvela, E. Z.** (2007). *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. London: Bloomsbury Publishing.
- Nida, E. A.** (1948). The Identification of Morphemes. *Language*, 24, 414 - 441.
- Macůtek, J. & Wimmer, G.** (2013) Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20(3), 227-240.
- Popescu, I., Altmann, G., & Köhler, R.** (2009). Zipf's law—another view. *Quality and Quantity* 44(4), 713–731.
- Wang, L.** (2016). Parts of speech studies in Chinese. *Journal of Quantitative Linguistics* 23(3), 235–255.
- Wang, L. & Guo, Y.** (2018). Polyfunctionality Studies in German, Dutch, English and Chinese. In: L. Wang, R. Köhler & A. Tuzzi (eds.), *Structure, Function and Process in Texts*. Germany: RAM-Verlag, 55–66.
- Zawada, B.** (2005). *Linguistic creativity and mental representation with reference to intercategoryal polysemy*. University of South Africa.

# Initial and Final Syllables in Tatar: from Phonotactics to Morphology

Alfiya Galieva<sup>1</sup>  
Zhanna Vavilova<sup>2</sup>

## Abstract

The paper proposes a methodology for analyzing the syllabic structure of Tatar words using fiction text data. Syllable construction rules are unique for each language as they are determined by the laws that govern its specific internal structure. However, the issue of the syllable finds a rather superficial description in Tatar grammars. Thus, possible correlations of the syllable structure with morphological features of the language will be examined in this paper. We analyze the distribution of syllable types in Tatar texts and represent their ranked frequencies and theoretical values fitted by means of the Zipf-Mandelbrot distribution. The main part of the study is devoted to inquiry into the structure of initial and final syllables. We proceed from the hypothesis that distributions of syllable structures in word-initial and word-final positions should be marked by statistically important differences due to discriminative structural features of stems and affixal chains. The study is based on a selection of obstruent and sonorant consonants. To evaluate statistical significance of these differences, the well-known  $\chi^2$  test is applied.

**Keywords:** *syllable, syllable structure, the Tatar language, phonotactics and morphology, quantitative linguistics.*


## 1. Introduction


Discovering statistically significant connections and consistent patterns between different levels of a language is a task that is successfully solved by means of quantitative linguistics. A discovery of such dependencies can provide new information about the internal structure of the language and the laws that govern it. Linguistic phenomena, despite certain deviations and diversity in their behavior, are characterized by a well-defined regularity and a stable relative frequency. Texts, as manifestations of languages, consist of a large number of elements of different nature whose connections are complex, being influenced by random factors. So text data provide us with information on language regularities within certain variances.

Syllables constitute the most important level between the meaningless (phonemes) and the meaningful (morphemes and words) language units. Many languages of the world have a fixed syllabic structure: possible combinations of phonemes in a syllable are rigidly determined. Relations between the phonotactic organization of words and the morphology of the language is an issue yet poorly described in linguistics. Thus, developing a methodology of such research is a topic of current interest.

The main task of this paper is to propose a method for discovering possible relationships between syllable patterns and morphological features of Tatar, a language with a rich agglutinative morphology. We compare the structure of the initial syllables of polysyllabic words (which are stems or parts of stems) and that of the final syllables (which are mainly affixes or parts of affixal chains) and conduct special tests to determine statistically significant

---

<sup>1</sup> Kazan Federal University, Kazan, Russian Federation, [amgalieva@gmail.com](mailto:amgalieva@gmail.com).  <http://orcid.org/0000-0003-2915-4946>.

<sup>2</sup> Kazan State Power Engineering University, Kazan, Russian Federation, [zhannavavilova@mail.ru](mailto:zhannavavilova@mail.ru).  <http://orcid.org/0000-0002-0247-8257>.

differences between them. Classical and modern texts of Tatar literature serve as an empirical source for the study.

The body of the paper is organized as follows: Section 2 covers research background. Section 3 contains basic information on Tatar as demanded by the study goals. Section 4 outlines the main stages of data preparation. Section 5 represents the quantitative data on syllable structures in the analyzed texts and the results of the  $\chi^2$  test with evaluation of statistical significance of differences between structural features of the initial and final syllables in polysyllabic words; Section 6 concludes and outlines the prospects of future work.

Tatar words, except in cases when it is necessary to represent their original graphical form, are introduced in the extended Latin transcription.

## 2. Research background

In recent decades, a large number of syllable studies of different languages of various types and structures have appeared, providing researchers with multiple perspectives on the topic. A book called *The Notion of Syllable across History, Theories and Analysis* edited by D. Russo (2015) introduces investigations into the syllable from four points of view: historical, descriptive, analytical-instrumental, and theoretical. Discussions on the nature and structure of the syllable bring into question both the status of the minimal unit of language and methods of linguistic analysis.

Distinguishing segments within the syllable depends on the theoretical assumptions of researchers and on the language type, so it differs in various approaches. Harry van der Hulst and Nancy A. Ritter distinguish onset-rhyme models, mora models and hybrid models (1999: 22-38). The most frequently model subdivides the syllable into three constituents: onset, nucleus, and coda (Haugen, 1956; Davis, 1988); see Figure 1. This last approach will be used to analyze the syllabic structure of Tatar words.

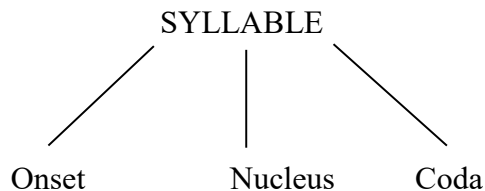


Figure 1. The syllable constituents

The onset is a constituent comprising the syllable-initial consonant or consonant cluster; the nucleus consists of the vowel or syllabic consonant and is considered the peak of the syllable and its obligatory constituent; the coda contains the syllable-final consonant or consonant cluster (Davis, 2006).

Many researchers address the problem of ranking the constraints that determine the syllable structure in a particular language by creating syllable division models within the Optimality Theory (Prince & Smolensky, 1993; Féry & van de Vijver, 2003).

Languages may differ by the degree of distinctness of syllabication; from this perspective, the so-called quantum and wave languages are distinguished. The former has clear-cut syllables with a strictly defined structure, whereas in wave languages, the structures of consonant combinations are vague, with fuzzy syllabic boundaries – so fuzzy that native speakers can mark them in different ways (Kodzasov & Muravyova, 1980). According to this classification, Tatar belongs to quantum languages.

Syllable structure impacts word length (Grzybek, 2007; Antić, Kelih, & Grzybek, 2007), language complexity (Fenk, Fenk-Oczlon, & Fenk, 2006), as well as other linguistic phenomena and has a variety of dimensions (Russo, 2015; Zörnig et al., 2019).

Researchers use different classifications of phonemes when analyzing syllabic structures. Some do not distinguish consonant subclasses (for example, Zörnig et al., 2019). In Russian linguistics, obstruent and sonorant consonants are traditionally opposed (Knyazev, 2006), which is often taken into consideration by researchers of languages of the Russian Federation. For example, Moroz (2019) describes the syllable structure of Adyghe, a Northwestern Caucasian language spoken in Russia and some other countries. Because of unclear syllabification rules in Adyghe, the author uses only word-initial syllable onsets and word-final codas of items taken from an Adyghe-Russian dictionary, thereby disregarding word-medial consonant clusters. It is revealed that the structures of onsets and codas in Adyghe differ, and that nearly all attested consonantal classes can occur in both positions.

An international research team inspired by G. Altmann published a book called *Quantitative Insights into Syllabic Structures* (Zörnig et al., 2019) which generalized data of languages with different morphologies and represented a quantitative analysis of syllable types, syllable length, open and closed syllables, asymmetry of onsets and codas, distances, and syllabic sequences in German, Polish, Slovak, Slovene, Russian, Romani, Chinese, Tatar and some other languages. The book examines a variety of statistical methods applied to the syllable data, and compares theoretical values with the empirical ones provided by languages of different types and origin.

Quantitative studies of linguistic structures, including the types and structure of syllables, based on the evidence of individual languages, raise the question of validity of the existing models (considering a significant number of fluctuations in empirical data, creativity of the text production process, etc.) (Altmann & Gerlach, 2016). Thus, the notion of syllable, seeming intuitively clear, remains at the intersection of discussions, and syllabic phenomena may serve as a good field for discovering alternative solutions for building language models.

As for Turkic languages, where Tatar belongs, there are special tools used to extract syllables. In particular, TASA (Turkish Automatic Spelling Algorithm) was developed for Turkish and was tested over five different corpora (Aşliyan & Günel, 2005). In Tatar linguistics, the concept of syllable remains on its periphery, although modern computer technologies make it easy to develop tools for automatically selecting syllable structures, as well as for their quantitative study. The available Tatar grammars pay very little attention to the problem of syllabification and syllable structure (Zakiev, 1993: 85-87; Khisamova, 2015: 40-41), and the number of special studies is very limited (Galieva, 2020). Therefore, quantitative analysis of syllable structures would be the first and a very important step to building a model of the syllable based on Tatar language data.

### **3. Overview of Tatar**

Tatar, a Turkic language, is spoken in the Volga and Ural regions of Russia and in some regions of Siberia. It is the second most common language of the Russian Federation and is, on a par with Russian, the official language of the Republic of Tatarstan. According to the 2010 Census, the number of Tatar speakers in Russia is 5.31 million people (Vserossijskaya perepis, 2010).

The location of Tatar culture at the intersection of Occidental and Oriental civilizations leads to active language contacts both with the Arab-Muslim and the European cultural areas. A significant part of abstract vocabulary in Tatar is of Oriental (Arabic and Persian) origin, and

many scientific and technical terms come from Europe. Historical contacts with Russian and its current dominant role as the state language of the Russian Federation became a cause of a huge number of words and constructions borrowed and calqued (component-by-component translated) from Russian. Consequently, modern Tatar has a large number of synonymous items of different – Turkic, Russian, European, and Oriental – origin (Galieva, 2018). As a rule, oriental loanwords were borrowed centuries (and even millennia) ago, which resulted in their phonetic assimilation. Unlike those, loanwords of European origin appeared through Russian mediation during the last centuries, maintaining a graphical and phonological shape typical for Russian.

In the 8th – 10th centuries, ancient Turkic peoples used a runic script – the so-called Orkhon-Yenisey script, named after the Orkhon Valley in Mongolia. After Volga Bulgars converted to Islam in 922, thus establishing firm contacts with Arabic and Muslim cultural areas, the ancient runic script was replaced by the Arabic script to be used by the ancestors of modern Tatars for over a millennium. In 1928, the Arabic script was changed to Latin. Since 1938 – 1940, Cyrillic is the official Tatar alphabet which employs all Russian letters and 6 additional ones to designate specific Tatar sounds. Therefore, the total number of letters in present Tatar alphabet is 39. Nevertheless, this script maps pronunciation of Tatar words not consistently enough, allowing for variations in writing and ambiguities in reading.

Modern Tatar has a rich system of phonemes: it includes 9 original vowels and 3 additional ones used in loanwords; the consonant system comprises 25 original and 5 additional consonants used in loanwords. Sonorant consonants in Tatar include glides *j* and *w*, liquids *r* and *l*, and nasals *m*, *n*, *ŋ* (Zakiev, 1993; Khisamova, 2015). In original Tatar, there are no affricates, and obstruents are divided into stops and fricatives.

According to Tatar grammars, original Tatar words are constructed from syllables of six types: V, CV, VC, CVC, VSC, CVSC<sup>3</sup>; some other syllable types can be found in loanwords (Zakiev, 1993: 85; Khisamova, 2015: 40).

The most important phonetic feature of Turkic languages is progressive vowel harmony. In Tatar, vowel harmony is a morphonological assimilatory process involving agreement between vowels within a word form. Original Tatar one-root words contain only back vowels (*a*, *o*, *u*, *ɯ*) or only front vowels (*ä*, *ö*, *ü*, *e*, *i*). Due to progressive (from-beginning-to-end) direction of vowel harmony, the quality of vowels in affixes and in affixal chains is determined by the quality of vowels in stems; the latter do not alternate, leaving it for alternating affixes (derivational and inflectional ones) to follow the vowel harmony rules.

This is how vowel harmony works for a word containing back vowels:

*Bala* ‘child’

*Bala-lar-da*

Child-PL, LOC<sup>4</sup>

‘in children’

This is how vowel harmony works for a word containing front vowels:

*Mäktäp* ‘school’

*Mäktäp-lär-dä*

School-PL, LOC

‘in schools’

---

<sup>3</sup> C – obstruent consonant, S – sonorant consonant, V – vowel.

<sup>4</sup> PL – Plural, LOC – Locative.

Some Tatar particles of Turkic origin also obey the vowel harmony law which governs the co-occurrence of vowels within a span of utterance. For example, particles *da* ‘too, also’ and *gina* ‘only, merely’, like affixes, have phonetic variants depending on the nature of the preceding word:

*Ber bala gına* ‘only one child’

*Ber kön genü* ‘only a day’.

Such items form phonological words according to vowel harmony rules.

Words with vowel harmony violations contain a mixed set of vowels – front and back vowels at the same time. These exceptions are mainly compound words consisting of two or more stems or loanwords (from Arabic, Persian, European languages or Russian), for example:

*Kitap* ‘book’, Arabic loanword;

*Tarih* ‘history’, Arabic loanword;

*Maşina* ‘machine’, Greek loanword.

Tatar is characterized by rich agglutinative morphology. The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and clear-cut monosyllabic derivational and inflectional affixes to the stem. The boundaries between the affixes within the word form are distinct and transparent, so that the affixal joint in many cases coincides with the syllabication (Guzev & Burykin, 2007).

#### **4. Data preparation**

The preparatory stage of the research included the following main steps:

- selection of linguistic data;
- conversion of written text items into phonological form;
- dividing words into syllables.

We did not use data of Tatar dictionaries for several reasons:

- they contain a large number of loanwords with a phonological structure which is not typical for Tatar (for example, items with complex consonant clusters), and a great number of these words are rarely used in real texts;
- they fix words in basic forms with their typical structure (for example, verbs in Tatar are given in the Infinitive or Verbal Noun forms);
- they do not map inflected forms of words, and those are crucial for analyzing Tatar with its agglutinative morphology.

The objective was to study the distribution of syllables in real use, so we consider textual material to be the best source. Therefore, fiction texts of different genres (poetry and prose by Tatar classical and modern writers) were selected as a source of Tatar language patterns (basic information on this selection is given in Appendix; the titles of the texts are given in transliteration and in translation).

The next stage was bringing the written text to the standard form: 1 letter – 1 sound. It is believed that Tatar writing is generally based on exactly this principle (nevertheless, there are some exceptions). For this purpose, it seems to be relevant to mention here the main features of Tatar spelling.

1. In Tatar, there are two letters (*ь* and *ӱ*) that do not denote any sound but determine the pronunciation of adjacent letters.
2. Letters *я* and *ю* denote correspondingly a couple of sounds *ya/ yä* or *yu/ yü* (the choice of *a / ä* and *u / ü* is determined by the vowel structure of the word).



3. *E* may be pronounced as *ye*, *yi* or *e* depending on its position in the word and the word vowel structure.
4. *Y* and *y* after *a* / *ä* are pronounced as *w* sonorant consonant and as *u* or *ü* vowels in any other case.
5. Besides, *ө* may be pronounced as *v* in Russian and European loanwords and as *w* in original Tatar and Oriental (Arabic and Persian) loanwords.

So special rules were set to convert Tatar texts into a phonologically relevant form.

Then phonological structure of words was mapped as frames consisting of vowels, sonorant (*l, r, m, n, ŋ, w, j*) and obstruent consonants. The differentiation between sonorant and obstruent consonants fits well for modeling syllables in Turkic languages.

Next, rules of dividing words into syllables were developed, and syllables were mapped basing on the available grammars of the Tatar language (Zakiev, 1993; Khisamova, 2015).

Table 1.  
Main stages of word analysis

Original Cyrillic word form	Phonological mapping of the word	Syllable structure of the word
урман 'forest, wood'	/urman/	VS-SVS
егет 'young man'	/yeget/	SV-CVC
ямьле 'nice'	/yämle/	SVS-SV
аулай 'to hunt'	/awlaw/	VS-SVS
юл 'road, way'	/yul/	SVS

At the last stage, the data were statistically processed and the results were visualized<sup>5</sup>.

## 5. Syllable structures in Tatar

### 5.1. Main syllable patterns

In the framework of our study, it is important to divide consonants into obstruents and sonorants according to the ratio of voice and obstructing airflow. First, let us see how syllables of different structures are distributed in the Tatar texts. Table 2 represents the distribution of syllable patterns in 10 Tatar texts (only 10 most frequent syllable types are presented). The table shows that syllables of simple structure, composed of an initial consonant (obstruent or sonorant one, CV and SV types together) and a vowel, make up about 40-50% of all syllables. Syllables consisting of one consonant onset, a nucleus vowel and one consonant coda account for 38-48%. Relative frequencies of syllable patterns in individual texts may differ significantly. The CV syllable has rank 1 in all the texts processed, in other words, it is the most frequent type. The SV and CVS syllable patterns have rank 2 or 3, depending on the text. The CVS type has rank 2 or 3, and the SVS type has rank 5 in most of the texts. It is noteworthy that the SVS type is absent among the 6 canonical syllable types presented in Tatar grammars; perhaps this is due to the fact that the SVS pattern occurs mainly at the end of words in affixal chains (see Figure 3 where the distribution of types of syllables in word-initial and word-final positions is presented).

<sup>5</sup> All the stages were implemented in R programming language (R Core Team, 2018); besides, tidyverse (Wickham 2017) and stringr (Wickham 2019) packages were used.

Table 2.  
Frequencies of syllables of different structures in Tatar texts

Type	Eniki			Tukay, <i>Şüräle</i>			Alish		
	Rank	Obs.	Expected	Rank	Obs.	Expected	Rank	Obs.	Expected
CV	1	1531	1482	1	506	545	1	559	574
SV	2	897	1037	3	268	279	3	313	315
CVS	3	776	729	2	392	390	2	444	428
SVS	4	471	510	4	222	200	4	222	225
CVC	5	417	362	5	207	144	5	180	169
V	6	359	255	7	86	74	6	141	124
SVC	7	243	178	6	103	103	7	107	90
VS	8	129	124	8	66	53	8	79	68
VC	9	96	89	9	41	38	9	79	56
CVSC	10	26	59	11	6	19	10	6	34
	s = 125.96, b = 351.96, $\chi^2=225.4$			s = 124.73, b = 370.31 $\chi^2=86.2$			s = 123.45, b = 401.78, $\chi^2=82.8$		

Type	Amirhan			Tukay, <i>Käcä belän sarık</i>			Tukay, <i>Su anası</i>		
	Rank	Obs.	Expected	Rank	Obs.	Expected	Rank	Obs.	Expected
CV	1	421	386	1	417	364	1	246	238
SV	2	216	274	3	157	177	3	115	126
CVS	3	210	195	2	190	252	2	170	173
SVS	4	132	139	5	111	89	5	78	67
CVC	5	113	99	4	127	125	4	87	91
V	7	71	50	7	49	46	6	48	49
SVC	6	89	71	6	76	64	7	43	35
VS	8	29	36	8	37	34	8	37	26
VC	9	23	26	9	31	25	9	19	19
CVSC	10	3	18	10	10	18	12	2	7
	s = 124.82, b = 363.59, $\chi^2=57.9$			s = 12.88, b = 33.53 $\chi^2=47.3$			s = 123.59, b = 385.36, $\chi^2=23.1$		

Type	Amirhan			Tukay, <i>Käcä belün sarık</i>			Tukay, <i>Su anası</i>		
	Rank	Obs.	Expected	Rank	Obs.	Expected	Rank	Obs.	Expected
CV	1	421	386	1	417	364	1	246	238
SV	2	216	274	3	157	177	3	115	126
CVS	3	210	195	2	190	252	2	170	173
SVS	4	132	139	5	111	89	5	78	67
CVC	5	113	99	4	127	125	4	87	91
V	7	71	50	7	49	46	6	48	49
SVC	6	89	71	6	76	64	7	43	35
VS	8	29	36	8	37	34	8	37	26
VC	9	23	26	9	31	25	9	19	19
CVSC	10	3	18	10	10	18	12	2	7
	s = 124.82, b = 363.59, $\chi^2 = 57.9$			s = 12.88, b = 33.53 $\chi^2 = 47.3$			s = 123.59, b = 385.36, $\chi^2 = 23.1$		

Type	Zulfat		
	Rank	Obs.	Expected
CV	1	95	99
SV	3	54	54
CVS	2	73	73
SVS	4	42	43
CVC	5	35	29
V	7	19	16
SVC	6	23	22
VS	9	6	9
VC	8	11	12
CVSC	10	1	6
	s = 123.15, b = 399.17, $\chi^2 = 7.5$		

To compare, Table 2 also puts forward theoretical values calculated on the basis of the Zipf-Mandelbrot distribution which has the following probability mass:

$$(1) p(x) = \frac{(x+b)^{-s}}{\sum_{i=1}^N (i+b)^{-s}}$$

where  $x = 1, 2, \dots, N$ .  $S, b > 0$  are shape parameters,  $x$  is the rank of the data, and  $N$  is the number of ranks. B. Mandelbrot put forward this distribution to estimate word frequencies (Mandelbrot, 1965); the Zipf-Mandelbrot distribution is often used for modeling syllable frequencies (see, for example, Radojičić et al., 2019). In Table 2, we list the computed parameter values and the measure  $\chi^2$  for the goodness of fit (in the table, 10 most frequent syllable types are presented and the fit applied only to them). According to our data,  $s$  parameter in 8 texts lies in the interval from 123 to 126, and  $b$  parameter lies in the interval from 350 to 402. Two texts

(the tale *Käcä belän sarık* by Tukay with parameters  $s = 12.88$  and  $b = 33.53$  and the text by Gilman with parameters  $s = 55.27$  and  $b = 156.04$ ) appear to be outliers.

As was noted above, Tatar grammars represent 6 canonical types of syllables and mention that other types can be found in loanwords (Zakiev, 1993: 85; Khisamova, 2015: 40). We found 22 different syllable types in the examined texts, 7 of which are quite frequent and make up at least 5% each. Syllable patterns with rank lower than 10 characterize rather random features of the texts. Syllable types missing in grammar books come from loanwords; besides, they can be found in morpheme junctions in original Tatar word forms, for example the SVS type in the example below:

*barmıym* (CVS-SVSS)  
go-NEG, PRES, 1SG<sup>6</sup>  
'I do not go'.

The distribution of units with different ranks and frequencies of syllable types found in the text by Eniki is presented in the chart (Figure 2). Theoretical values are fitted by means of the Zipf-Mandelbrot distribution.

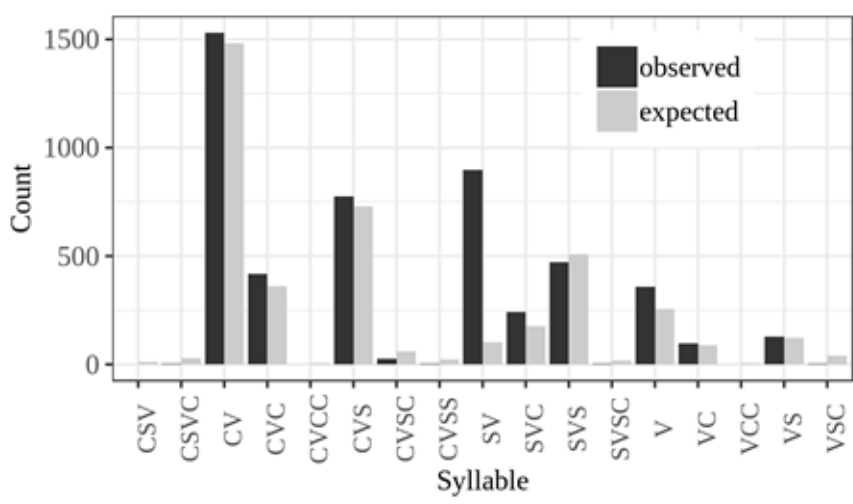


Figure 2. Distribution of syllables of different structures in the text by Eniki

Data represented in Table 2 evidence that complex consonant clusters in syllable onsets and codas occur relatively infrequently. It should be noted that, for example, SC cluster (not syllable type) by itself is quite frequent in Tatar words; however, its elements, when inflected, are often distributed over different syllables. See examples below:

*kayt* (CVSC) 'return (Imperative)' – *kayta* (CVS-CV) 'he / she returns', *kaytaçak* (CVS-CV-CVC) 'he / she will return';

*kart* (CVSC) 'an old man' – *kartı* (CVS-VC) 'his / her old man', *kartım* (CVS-CVS) 'my old man'.

As a result, syllables with consonant SC cluster are relatively rare in our data. So word inflection in many cases simplifies syllabication.

Evidently, new text data, especially texts with numerous loanwords, will provide new patterns of syllable structures with more complex onsets and codas.

<sup>6</sup> NEG – Negative, PRES – Present, 1SG, 1<sup>st</sup> person, Singular.

## 5.2. Structure of the initial and final syllables of words

Analysis of the structure of syllables in word-initial and word-final positions interests us because it allows determining how differences in stems and affixes correlate to differences between syllable structures. In Tatar polysyllabic words, initial syllables are stems or parts of stems and final syllables are usually affixes or fragments of affixal chains. Thus, we proceed from the hypothesis that the distribution of syllable structures at the beginning and at the end of word forms should have statistically significant differences; this could be so due to different phonological arrangement of stems and chains of affixes.

The samples included words consisting of two or more syllables taken from 10 Tatar texts; the data was processed separately for each text. For example, in the text by Eniki we detected 1782 words having more than one syllable. So the samples of initial and final syllables comprised 1782 syllables, and the syllables in the middle of words were not considered. 14 types of syllables from total 18 types found in the text by Eniki (taking into account monosyllables and syllables in the middle of the word) are represented in the initial and final positions. Figure 3 represents a number of syllables of different types at the beginning and at the end of word forms in this text. The data indicate that the final syllables tend to fall into a fewer number of types: mainly types CV (474 times), SV (391 times), SVS (296 times), CVS (313 times), SVC (148 times), and CVC (148 times) are represented. Besides, final syllables rarely begin with a vowel (V pattern is not represented in the sample at all, VC pattern is encountered in 3 cases only and VS type is found in 6 cases). Another important feature is that final syllables tend to have a sonorant onset.

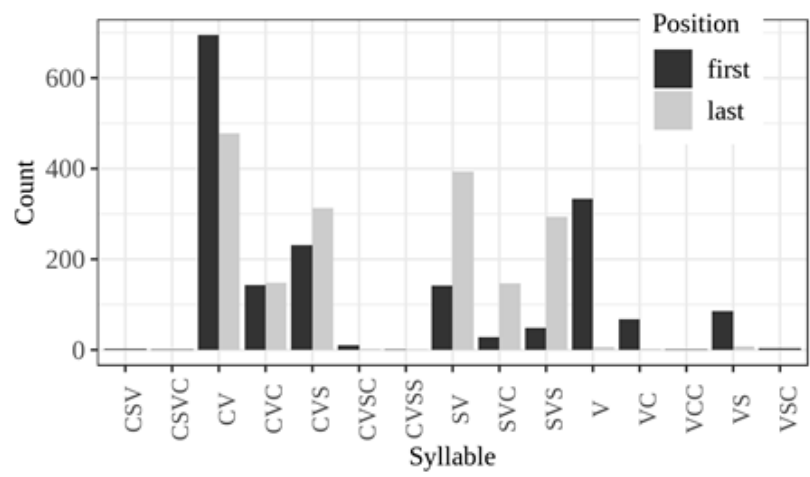


Figure 3. Initial and final syllables defined in the text by Eniki

Initial syllables are characterized by greater diversity, while the CV type dominates, occurring 683 times (26% of the entire sample).

The data presented in Figure 4 also support our assumption that the syllables at the beginning and at the end of words differ in quantitative and qualitative features.

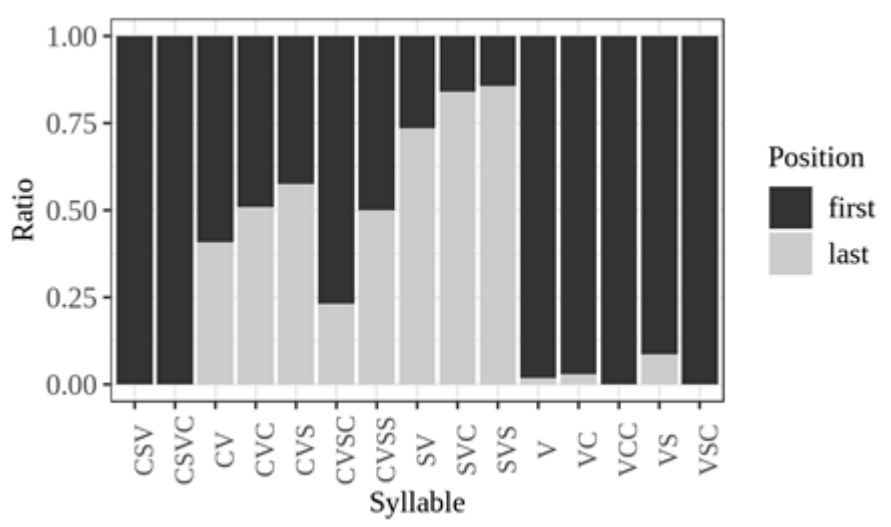


Figure 4. The ratio of initial and final syllables for each type in the text by Eniki

Of frequently occurring syllables, only for the CVC pattern a distribution close to 50% is observed: it occurs 143 times at the beginning of words and 148 times at the end. Although the CVSS type has a 50% distribution, it is characterized by a very low frequency (it occurs once at the beginning and once at the end of the word).

### 5.3. Initial and final syllables: $\chi^2$ test results

With the obtained data on building initial and final syllables of Tatar words, we can compare the arrangement of onsets and codas in both positions. Now we can ask ourselves whether the distribution of syllables with onsets and codas in initial and final syllables is random. An answer to this question can be provided by applying the  $\chi^2$  test, which would allow us to evaluate the statistical significance of differences between nominative variables in a contingency table. In particular,  $\chi^2$  criterion of Pearson is a nonparametric method that allows for assessing significance of differences between the observed number of qualitative characteristics of the sample falling into each category and the theoretical amount that can be expected in the studied groups if the null hypothesis is true. The null hypothesis of the  $\chi^2$  test is that there is no relationship between columns and rows in the contingency table: the event “an observation in row  $i$ ” is independent of the event “that same observation is in column  $j$ ” for all  $i$  and  $j$  (Conover, 1999: 205). So as far as our data is concerned, the null hypothesis may be formulated as “The proportions of onsets in initial and final syllables are independent”. We used the Yates's corrected version of Pearson's  $\chi^2$  statistics (Yates, 1934):

$$(2) \ x_{Yates}^2 = \sum_{i=1}^N \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where:  $O_i$  is observed frequency,  $E_i$  is expected frequency, asserted by the null hypothesis,  $N$  is the number of distinct events. The generic formula for computing the expected frequency in row  $i$  and column  $j$  is given below:

$$(3) E_{ij} = \frac{S_i S_j}{N}$$

where  $s_i$  is the marginal frequency of row  $i$ ,  $s_j$  is the marginal frequency of column  $j$  and  $N$  is the total number of observations.

We performed the  $\chi^2$  test twice, separately for the onsets and the codas. Table 3 presents the results for the onsets.

Table 3.  
 $\chi^2$  test for onsets results

<b>A. Eniki, Äytemägän wasıyät</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	1299	1530.5	1762	1530.5	3061
<b>No</b>	483	251.5	20	251.5	503
<b>total</b>	1782		1782		3564
$\chi^2 = 494.07, df = 1, p\text{-value} < 0.0001$					

<b>G. Tukay, Şüräle</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	512	591	670	591	1182
<b>No</b>	164	85	6	85	170
<b>total</b>	676		676		1352
$\chi^2 = 165.85, df = 1, p\text{-value} < 0.0001$					

<b>G. Tukay, Şüräle</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	512	591	670	591	1182
<b>No</b>	164	85	6	85	170
<b>total</b>	676		676		1352
$\chi^2 = 165.85, df = 1, p\text{-value} < 0.0001$					

*Initial and Final Syllables in Tatar: from Phonotactics to Morphology*

<b>A. Alish, Sertotmas ürdäk</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	523	625	727	625	1250
<b>No</b>	209	107	5	107	214
<b>total</b>	732		732		1464
$\chi^2 = 165.85, df = 1, p\text{-value} < 0.0001$					

<b>G. Tukay, Kücä belän sarık ükiyäte</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	391	438	485	438	876
<b>No</b>	95	48	1	48	96
<b>total</b>	486		486		972
$\chi^2 = 99.967, df = 1, p\text{-value} < 0.0001$					

<b>G. Tukay, Su anası</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	223	263.5	304	263.5	527
<b>No</b>	82	41.5	1	41.5	83
<b>total</b>	305		305		610
$\chi^2 = 89.253, df = 1, p\text{-value} < 0.0001$					

<b>F. Amirkhan, Häyät</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	359	408	457	408	816
<b>No</b>	100	51	2	51	102
<b>total</b>	459		459		918
$\chi^2 = 103.78, df = 1, p\text{-value} < 0.0001$					



<b>G. Ibragimov, Kızıl çäçäklär</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	290	334.5	379	334.5	669
<b>No</b>	92	47.5	3	47.5	95
<b>total</b>	382		382		764
$\chi^2 = 93.091, df = 1, p\text{-value} < 0.0001$					

<b>G. Gilman, Oçraşu</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	630	713	796	713	1426
<b>No</b>	188	105	22	105	210
<b>total</b>	818		818		1636
$\chi^2 = 148.73, df = 1, p\text{-value} < 0.0001$					

<b>Suleyman, Dürt mizgel</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	225	254	283	254	508
<b>No</b>	67	38	9	38	76
<b>total</b>	292		292		584
$\chi^2 = 49.146, df = 1, p\text{-value} = 0.0001$					

<b>Zulfat, Söyembikäneñ huşlaşu dogası</b>					
<b>Have onsets</b>	<b>Initial syllables</b>		<b>Final syllables</b>		<b>Row total</b>
	<b>Observed values</b>	<b>Expected values</b>	<b>Observed values</b>	<b>Expected values</b>	
<b>Yes</b>	107	124	141	124	248
<b>No</b>	34	17	0	17	34
<b>total</b>	141		141		281
$\chi^2 = 36.421, df = 1, p\text{-value} = 0.0001$					

An appropriate graphic way for visualizing data from two or more qualitative variables is a mosaic chart. Figure 5 demonstrates that syllables with onsets are strongly overrepresented in the final position and underrepresented in the initial position, whereas syllables with no onsets are strongly underrepresented in the final position. The colour of shading corresponds to the sign of standardized residuals, and the intensity of shading shows relative importance of the differences. The data are represented for the text by Eniki.

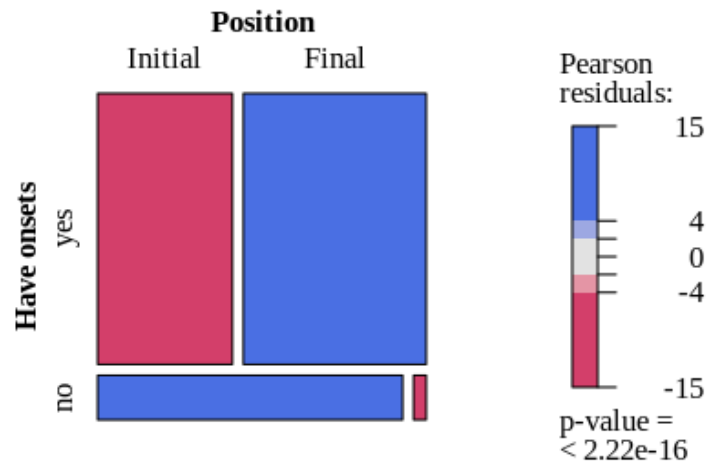


Figure 5. Syllables with and without onsets depending on syllable position

Next, we carried out the  $\chi^2$  test for open and closed syllables and found out that the differences between the initial and final syllables are not statistically significant (p-value > 0.05). The result for *Şüräle* by Tukay is presented in Table 4.

Table 4.

$\chi^2$  test results for syllables with coda

Syllable type	Initial syllables		Final syllables		Row total
	Observed values	Expected values	Observed values	Expected values	
Open	326	313	300	313	626
Closed	349	362	375	362	724
Column total	675		675		1350
$\chi^2 = 1.8617$ , $df = 1$ , p-value = 0.1724					

It should be noted that analyzing the first and the last syllables gives but a rough idea of the morphological structure of the Tatar word (the last syllable may be a part of a polysyllabic stem); nevertheless, the main trends can be traced. Thus we examined 10 texts and in all of them found statistically significant differences between patterns of the initial and final syllables.

## **6. Conclusion**

As the aim of the research was to evaluate syllable structures of Tatar word forms in actual use, we analyzed classical and present-day texts of Tatar literature, poetic and prose works or fragments from them, disregarding dictionary data for a large number of loanwords with syllable structures that are atypical for Tatar and for the lack of affixed word forms.

The research design relied upon distinguishing between sonorant and obstruent consonants. It has been found that in Tatar, simple syllable structures prevail (CV, SV, CVS, SVS, CVC, SVC). In many cases, available consonant clusters are broken into joining inflection affixes and fall into two adjacent syllables.

The main task of the study was to propose a way to assess possible correlations between syllable structures and morphology. We analyzed the structure of initial and final syllables of polysyllabic words and compared a number of syllables with and without onsets. The  $\chi^2$  test showed that the observed values were statistically significantly different from the expected values.

This study is preliminary in many respects. It is aimed at developing a methodology for studying the structure of the syllable in Tatar in order to create a comprehensive syllable model in the future as well as to disclose possible correlations between phonotactics and morphology. In Tatar, with its rich agglutinative morphology, such correlations should exist and could be quantified. We suppose that further research will be carried out taking into account the sonority scale; analysis of initial and final syllables distinguishing between the types of sonorants (nasals, liquids, and semivowels /w/, /j/), and obstruent consonants (fricatives, stops), it seems, should provide more detailed information about the structure of stems and affixal chains in Tatar.

**Acknowledgments:** The work is carried out according to the Russian Government Program of Competitive Growth of Kazan Federal University.

## References

- Altmann, E. G., Gerlach, M.** (2016). Statistical laws in linguistics. In: *Creativity and Universality in Language*. Springer, 7-26.
- Antić, G., Kelih, E., Grzybek, P.** (2007). Zero-syllable words in determining word length. Contributions to the science of text and language. In: *Word Length Studies and Related Issues*. Springer, 117 – 156.
- Aşliyan, R., Günel, K.** (2005). Design and implementation for extracting Turkish syllables and analyzing Turkish syllables. In: *International Symposium on Innovations in Intelligent Systems and Applications*. INISTA, 170-173.
- Conover, W. J.** (1999). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.
- Davis, S.** (1988). *Topics in syllable geometry*. New York: Garland.
- Davis, S.** (2006). Syllable constituents. In: *The Encyclopedia of Language and Linguistics* (2nd ed.). Vol. 12. Oxford & New York: Pergamon Press, 326-328.
- Fenk, A., Fenk-Oczlon, G., Fenk, L.** (2006) Syllable complexity as a function of word complexity. In *The VIII International Conference Cognitive Modeling in Linguistics*. Vol. 1, 324 - 333.
- Féry, C. & Vijver van de, R. (eds.)** (2003). *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press.
- Galieva, A. M.** (2018). Synonymy in modern Tatar reflected by the Tatar-Russian Socio-Political Thesaurus. In: Čibej, J. et al. (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, 585-994.
- Galieva, A. M.** (2020). Struktura sloga v tatarskom yazyke: ot dannykh k modeli [Syllable structure in Tatar: from data to modeling]. *International Journal of Open Information Technologies* 8 (1), 9-16.
- Grzybek, P.** (2007) History and methodology of word length studies. In: *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Springer, 15-90.
- Guzev, V. G., Burykin, A. A.** (2007) Obshchie stroevye osobennosti agglutinativnykh yazykov [General structural peculiarities of agglutinative languages]. In: *Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovaniy [Papers of Institute of Linguistic Studies, Russian Academy of Sciences]*. Vol. 3-1. Saint-Petersburg: Nestor-istoriya, 109-117.
- Haugen, E.** (1956). The syllable in linguistic description. In: M. Halle, H. Lunt, & H. McLean (eds.) *For Roman Jakobson*. The Hague: Mouton, 213-221.
- Hulst van der, H., Ritter, N. A.** (1999). *The Syllable: Views and Facts*. Berlin: Mouton de Gruyter.
- Khisamova, F. M. (ed.)** (2015). *Tatar grammatikası [Tatar Grammar]*. Vol. 1. Kazan: Institute of Language, Literature and Art.
- Knyazev, S. V.** (2006). *Struktura foneticheskogo slova v russkom yazyke: sinkhroniya i diakhroniya [Structure of the Phonetic Word in Russian: Synchrony and Diachrony]*. Moscow: Max Press.
- Kodzasov, S. V., Muravyova, I. A.** (1980). Slog i ritmika slova v alyutorskom yazyke [Syllable and word rhythmicity in Alutor]. In: *Publikatsii otdeleniya strukturnoi i prikladnoi lingvistiki MGU. Filologicheskii fakultet [Papers of Department of Structural and*

*Applied Linguistics of Moscow State University*] No. 9. Moscow: Lomonosov Moscow State University Press, 103-127.

- Mandelbrot, B. B.** (1965). Information Theory and Psycholinguistics. In: B. B. Wolman & E. Nagel (eds.) *Scientific Psychology*. New York: Basic Books, 550-562.
- Moroz, G. A.** (2019). Slogovaya struktura adygeyskogo yazyka: ot dannykh k obobshcheniyam [Adyghe syllable structure: from empirical data to generalizations]. *Voprosy yazykoznaniya [Issues of Linguistics]* 2, 82-95.
- Prince, A., Smolensky, P.** (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder. Available from <http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF>
- R Core Team** (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. Available from <https://www.R-project.org/>.
- Radojičić, M., Lazić, B., Kaplar, S., Stanković, R., Obradović, I., Mačutek, J., Leššová, L.** (2019). Frequency and length of syllables in Serbian. *Glottometrics* 45, 114-123.
- Russo, D.** (2015; ed.). *The Notion of Syllable across History, Theories and Analysis*. Cambridge: Cambridge Scholars Publishing.
- Vserossijskaya perepis naseleniya [All-Russian Census]** (2010). Available from [https://rosstat.gov.ru/free\\_doc/new\\_site/perepis2010/croc/perepis\\_itogi1612.htm](https://rosstat.gov.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm)
- Wickham, H.** (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. Available from <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H.** (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. Available from <https://CRAN.R-project.org/package=stringr>
- Yates, F.** (1934). Contingency tables involving small numbers and the  $\chi^2$  test. *Supplement to the Journal of the Royal Statistical Society*, 1(2), 217-235.
- Zakiev, M. Z. (ed.)** (1993). *Tatarskaya grammatika [Tatar Grammar]*. Vol. 1. Kazan: Tatar Publishing House.
- Zörnig, P., Stachowski, K., Ráková, A., Qu, Y., Místecký, M., Ma, K., Lupea, M., Kelih, E., Gröller, V., Gnatchuk, H., Galieva, A., Andreev, S., Altmann, G.** (2019). *Quantitative Insights into Syllabic Structures. Studies in Quantitative Linguistics* 30. Lüdenscheid: RAM-Verlag.

#### Appendix I

##### Basic information on the texts processed

No	Author	Title	Genre	Words	Syllables
1	Eniki, Amirkhan	Äytemägän wasıyät / Unspoken Testament, Chapter 1	novel, prose	2,169	4,962
2	Tukay, Gabdulla	Şüräle / Forest Spirit	fairy tale (verse)	925	1,917
3	Tukay, Gabdulla	Su anası / Aquatic Woman	fairy tale (verse)	419	854
4	Tukay, Gabdulla	Käcä belän sarık äkiyäte/ The tale of the goat and the ram	fairy tale (verse)	579	1,211
5	Amirkhan, Fatikh	Häyät Hayat, Chapter 1	novel, prose	548	1,310
6	Ibrahimov, Galimjan	Kızıl çaçäklär /	novel, prose	444	1,085

*Initial and Final Syllables in Tatar: from Phonotactics to Morphology*

		The Red Flowers, Chapter 1			
7	Alish, Abdulla	Sertotmas ürdäk / The Talkative Duck	fairy tale for children, prose	917	2,093
8	Gilman, Galimdzhän	Oçraşu / Встреча	story, prose	1,014	2,351
9	Suleyman	Dürt mizgel / Four moments	poem	355	815
10	Zulfat	Söyembikäneñ huşlaşu dogası / The farewell prayer of Suyumbike	poem	163	360

# Automatic Identification of Authors' Stylistics and Gender on the Basis of the Corpus of Russian Fiction Using Extended Set-theoretic Model with Collocation Extraction

Alexandr Osochkin<sup>1</sup>

Xenia Piotrowska<sup>2</sup>

Vladimir Fomin<sup>3</sup>

## Abstract

We present a novel quantitative approach for classification of authors' stylistics and gender differences based on extraction of word collocation. The proposed algorithm attenuates previously described issues of text processing using the vector models. We demonstrate the approach by analyzing a corpus of Russian prose. We discuss different approaches for classification and identification of the author's style implemented by currently-available software solutions and libraries of morphological analysis, methods of parameterization, indexing of texts, artificial intelligence algorithms and knowledge extraction. Our results demonstrate the efficiency and relative advantage of regression decision tree methods in identifying informative frequency indexes in a way that lends itself to their logical interpretation. We develop a toolkit for conducting comparative experiments to assess the effectiveness of classification of natural language text data, using vector, set-theoretic and the author's set-theoretic with collocation extraction models of text representation. Comparing the ability of different methods to identify the style and gender differences of authors of fiction works, we find that the proposed approach incorporating collocation information alleviates some of the previously identified deficiencies and yields overall improvements in the classification accuracy.


**Keywords:** *Natural language processing, frequency and morphological analysis, text-mining, gender linguistics, collocation extraction, set-theoretic model, vector text analysis.*


## 1. Introduction

Recent studies on the use of deep machine learning in the field of natural language processing (NLP) and text-mining (Kang, et al., 2020; Moschitt, 2004) have shown that statistical methods can be more effective (Grekhov, 2012) when used in combination with linguistic (morphological and parsing) analysis (Khalezova, et al., 2020). This concept has given rise to a separate direction in linguistics, which studies language based on statistical regularities, including the use of linguistic and semantic analysis, expanding the statistical approach to text analysis through the use of latent semantic connections between text elements (Maheshan, et al., 2018; McCann et al., 2017; Yang, et al., 2019). Increasing availability of computational resources and advanced algorithm implementations has now enabled individual researchers to process large volumes of data, and employ sophisticated computational methods for their analysis. One of the most promising avenues for improving NLP technologies is through incorporation of parsing-based quantitative methods (for example, the relationship of compositional construction and word formation, the length of compounds and the length of their components).

---

<sup>1</sup> Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation [osa585848@bk.ru](mailto:osa585848@bk.ru),  <http://orcid.org/0000-0001-9449-5603>.

<sup>2</sup> Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation, [kpr62@mail.ru](mailto:kpr62@mail.ru),  <http://orcid.org/57207357482>.

<sup>3</sup> Herzen State Pedagogical University of Russia, Moika Emb., 48, Saint-Petersburg, 191186, Russian Federation, [v\\_v\\_fomin@mail.ru](mailto:v_v_fomin@mail.ru),  <http://orcid.org/0000-0001-7040-5386>.

The modern quantitative approach to text analysis arose from the development of many different models of text representation that were focused on solving highly specialized problems (Devlin, et al., 2018; Belinkov & Bisk, 2018; Belinkov & Glass, 2019; Iyyer et al., 2018). Modern NLP data analysis and processing packages rely on complex linguistic algorithms for text analysis (Piotrowska, 2014). In 2019 Google introduced Bidirectional Encoder Representations from Transformers (BERT), which has shown to be highly efficient in solving a wide range of tasks (Macro, et al., 2020), and formed the basis of NLP digital services. Recent studies, however, have demonstrated that in some settings BERT can pose a number of notable disadvantages.

Microsoft Azure Machine Learning, based on the BERT model, is part of the Cortana Intelligence Suite that enables predictive analytics and interaction with data using natural language and speech. One of the promising BERT applications is the improvement search systems based on the classification and indexing of texts on sites and repositories.

The paper by T. Macro (Macro, et al., 2020) received an award for identifying critical flaws in modern text processing models at the "Association for Computational Linguistics" (ACL) in 2020. This critical survey examined performance of advanced applications of BERT in "Google AI", "Microsoft Azure Text Analysis", "Amazon Comprehend", "Facebook RoBERTa AI", etc. The survey noted shortcomings in the ability to capture the grammatical and lexical cohesion structure of the text, its integrity, and incorporation of term collocations in texts. These limitations of the modern text representation models suggest that further research is needed to improve text representation models, procedures for generating and extracting significant digital indicators, as well as in the development of artificial intelligence algorithms.

The current study aims to evaluate the effectiveness of technology in identifying and classifying the author's style and gender differences in literary works using a quantitative approach based on collocation algorithms and regression decision trees.

## **2. Data analysis models**

Most importantly, a quantitative approach to text analysis requires a formalized representation of textual data. There are several notable paradigms of text representation that rely on various mathematical models, including the vector model, probability word distribution, and the set-theoretic model (Wang & Zhu, 2019; Martin & Jurafsky, 2019).

A specific mathematical model for text representation enables extraction of quantitative characteristics from text data (Kashcheyeva, 2013). Specific quantitative representations include, frequency-based models (Beel et al., 2017), frequency-morphological (Osochkin, et al., 2018), vector (Salton G., Allan J. & Buckley, 1994), topic vector (Devlin et al., 2018; Belinkov & Bisk, 2018; Belinkov & Glass, 2019; Iyyer, et al., 2018), and set-theoretic models (Allahyari, et al., 2017; Harish, 2012; Marcus, 1967). Despite the large number of approaches for text conversion, all models can be divided into two types: vector and set-theoretic.

A number of advanced computational models are utilized in the computer text processing industry. Here we will consider and compare the vector, and the set-theoretic models, together with our extension of the set-theoretic model incorporating term collocation.

### **2.1. Vector model of text representation**

The vector model became popular at the beginning of the 20th century, and nowadays it remains relatively unchanged despite the appearance of alternative models (Wang & Zhu, 2019; Popescu, et al., 2010). A vector text model represents each word or sentence in a text as a vector that captures the underlying meaning (Popescu et al., 2010). The vector model is often referred to as a topic vector model, because the basis of text class division is rooted in a semantics of the words, which in aggregate represent the subject field. Vector text representation can use



different text elements for analysis: words, sentences, paragraphs, particles of speech, etc., though sentences are the most commonly used text elements.

Topic vector text representation supposes that text contains chapters that have a common subject and include paragraphs. Paragraphs, in turn, contain sentences.

$$G = \{G_1, G_2, \dots, G_n\}; Vg = \{Vg_1, Vg_2, \dots, Vg_n\}, (1)$$

where  $G$  represent multiple topic chapters in the text,  $G_i$  represents an  $i$ -th text chapter,  $i = 1 \dots n$ ,  $Vg$  – multiple vectors of topic chapters,  $Vg_i$  topic vector of  $i$ -th chapter.

Therefore:

$$A_i = \{A_{i1}, A_{i2}, \dots, A_{i3}\}; Va_i = \{Va_{i1}, Va_{i2}, \dots, Va_{i3}\}, (2)$$

where  $A_i$  are multiple paragraphs of  $i$ -chapter,  $A_{ij}$  are paragraphs of  $i$ -th chapter,  $j = 1 \dots m$ ;  $Va_i$  are multiple vectors of paragraph topics.  $Va_{ij}$  are topic vector of the  $j$ -paragraphs of the  $i$ -th chapter. The mathematic model of sentence interpretation is represented as:

$$P_{ij} = \{P_{ij1}, P_{ij2}, \dots, P_{ijk}\}; Vp_{ij} = \{Vp_{ij1}, Vp_{ij2}, \dots, Vp_{ijk}\}, (3)$$

where  $P_{ij}$  are multiple sentences of  $i$ -th chapter of  $j$ -th paragraph,  $P_{ijh}$  are  $h$ -th sentence of  $i$ -th chapter of  $j$ -th paragraph,  $h = 1 \dots k$ ;  $Vp_{ij}$  are multiple vectors of sentences topics of  $i$ -th chapter of  $j$ -th paragraph;  $Vp_{ijh}$  are the topic vector of  $h$ -th sentence of  $i$ -th chapter of  $j$ -th paragraph.

The vector models of text representation were initially able to overcome key disadvantages of frequency and theoretical models of data representation, including the homonym problem and the consideration of the semantics of sentences. Further development of the vector model of data representation, however, could not solve a number of outstanding challenges, including time-consuming vector calculation needed for analysis of large texts. Therefore, the vector model is mainly applied to processing of small texts.

Aside from the computational requirements, the significant disadvantage of vector models lies in the lack of consideration for the language specificity of the word order, position of the subject and predicate, characteristics of parts of speech, forms and other text features.

We choose the "Word2Vec" library as the main tool for studying the vector model of text representation, because it:

- supports more than 40 languages, including Russian,
- makes use of an embedded model of replacing associative words and homonyms (Bag of Words),
- does not require supervised training data.

## 2.2. Set theoretic model of text representation

Set-theoretic models assume that a text is composed of distinct terms (words, n-grams, sentences), possessing common characteristics and unique traits. The main concept of such models is the reflection of different text characteristics in relative indicators, to which mathematical methods for identification of common and unique characteristics of each sample analyzed text are applied. Set-theoretic models commonly rely on analysis of frequency and measures of metric proximity (e.g. Dice, Ochiai, Jaccard, Simpson etc.) (Marcus, 1967; Zakharov & Khochlova, 2010; Kolesnikova, 2016; Belyaeva, et al., 2019).

In our experiments on classification, the Jacquard similarity index was used as a metric for evaluating the similarity of words in texts (Jadhao, et al., 2016). This index is the easiest to calculate and is widely employed in linguistic analysis. Its values are equivalent in particular cases to other similarity measures: Sokal-Sneath and Serensen distance measures.

The values of the Jacquard coefficient vary from 0 to 1. The Jacquard coefficient measures the similarity between the sets of words used by two texts, and is defined as the size of the intersection (i.e words used in both texts) divided by the size of the union of the word

sets (i.e. total number of unique words used in both texts). To compare the proximity of two texts A and B, the Jaccard similarity index can be calculated using the formula:

$$K(A, B) = \frac{|A \cup B|}{|A \cap B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \quad (5)$$

In addition to words, such similarity indexes can be calculated for n-grams, sentences, etc. Detailed information on how indicators are calculated on the basis of Jaccard similarity index can be found in Moulton & Jiang (2018). We used the freely-licensed Python library "Jaccard-index", which is able to calculate the similarity index between texts to implement calculations of the Jaccard similarity index. The library is both fast and is well-supported by the text and image analysis communities. Words were used as the main unit of analysis.

### **2.3. Set-theoretic model with collocation extraction**

The limitations identified in the work of T. Macro (Macro, et al., 2020) suggested that to improve the classification accuracy of author's style and gender identification (Mikros, 2013; Vincze, 2015), it is important to identify not only the topic aspects of the text, but also the morphological features of the words as well as their collocations.

In this context, we developed the FaM software, which was described in details in (Osochkin, et al. 2018). It uses an algorithm for text representation as a frequency-morphological set of indicators, considering collocations, and can improve the accuracy of classification in NLP applications.

The mathematical model of text representation with collocation extraction models texts as interconnected sequences of terms. It is based on the hypothesis that taking into account sustainable links in phrases and the relationship between text elements will create a more accurate model of text representation.

In order to take into account collocation, FaM calculates a number of custom indicators based on the usage frequency word sequences (n-grams) in the text. It is expected that some of the words in the sentence will not have a semantic connection with other members, such as: prepositions, parenthesis, etc. To accommodate such cases, FaM incorporates an algorithm based on morphological libraries, which identifies and omits the functional words and words that were not in a semantic connection with the sentence members when calculating n-gram sequences. A normalized text is represented as an array of objects, where each word is described as an object with properties: part of speech and morphological characteristics. For each sequence of objects and for each combination of their morphological characteristics, a separate frequency indicator is then calculated. This indicator is defined as a number of times a given object sequence occurrence occurs in a normalized text, divided by the total number of objects.

Thus, the set of n-gram indicators is determined by the natural language in which the text is written and by the length of the sequence (i.e. by  $n$ ). The total set of indicators of bigrams ( $n=2$ ) extracted for the Russian language can reach more than 200.

A key factor in improving the efficiency of classification is the conversion of text into a set of numerical indicators using frequency-morphological analysis. In FaM, morphological analysis is performed by a hybrid algorithm that uses two morphological analysis modules – Natural language processing (AOT)<sup>4</sup> and "Solarix Engine"<sup>5</sup>.

The morphological module AOT is based on the multi-level representation of data in a natural language, and was first used in the French-Russian automated translation system (FRAM). The module contains a Russian morphological dictionary: about 161,000 words with various forms. It also incorporates syntactic and semantic analysis of the text.

---

4 Official website of the library "AOT" URL: <http://www.aot.ru>.

5 Website of the "Solarix Engine" library URL: [http://www.solarix.ru/for\\_developers/api/grammar-engine-api.shtml](http://www.solarix.ru/for_developers/api/grammar-engine-api.shtml).

Solarix Engine is a morphological analysis module that includes a dictionary of 1,800,000 words and 218,000 thesaurus articles, containing information about possible subordination and associative relationships between words, suitable for machine learning. The main advantage of this module is the support of different languages: English, French, German, etc.

A custom algorithm embedded in FaM enables the use these two libraries simultaneously, allowing one to obtain aggregated information about the analyzed word, its semantic relationship with other words in the sentence, and carry out morphological, syntactic, and frequency analysis. To identify semantic connections, the algorithm carries out syntactic analysis which identifies parts of speech and functional words in a sentence, and builds a syntactic tree. In the subsequent stages, the algorithm searches for words that are syntactically related to the subject or predicate in the sentence, and checks for semantic connections. The semantic relationship is evaluated by synthesizing a new sentence without the analyzed word, building a new syntactic tree, and analyzing the resulting node changes. If no context changes were observed for the tree nodes associated with the deleted word, the word was omitted from the normalized text.

#### 2.4. Normalization and relevance of indicators

Almost all text analysis packages perform pre-processing to normalize the data. Text pre-processing enables more accurate and reliable extraction of the features present in the text. In this regard, the lemmatization procedure is a key pre-processing step that can significantly reduce the size of the vector space, by removing inflectional endings and collapsing words into their basal forms. This word variant reduction is also beneficial for estimation of the vector indexes, as it reduces the dimensionality of the vector space. The NLTK4Russian library was used to carry out text lemmatization <sup>6</sup>.

Normalization of the data is carried out using the TF-IDF technology, the Scikit-learn library <sup>7</sup>.

$TF_{ij}$  indexes are defined as the frequency of word's use in the analysed text, relative to the total number of words in the text:

$$TF_{ij} = \frac{f_{ij}}{f_{i1} + f_{i2} + \dots + f_{in}}, i = 1, m, (6)$$

where  $TF_{ij}$  is the index for the  $j$ -th word in the  $i$ -th text,  $f_{ij}$  is the frequency of use of the  $j$ -th word in the  $i$ -th text.

The TF-IDF method (Roul, et al. 2017) calculates the value of the  $j$ -th term  $IDF_{ij}$  in the  $i$ -th text as the product of the frequency of term usage in the  $TF_{ij}$  document and the normalized inverse frequency of term content in the documents.

$$IDF_{ij} = TF_{ij} * \log \left( \frac{|D|}{Df_i} \right), i = 1, m, j = 1, n, (7)$$

where  $D$  is the total number of documents in the collection.  $Df_i$  is the number of documents in which the term  $f_j$  occurs. If the term is not present in any of the documents  $Df_i$  is taken to be equal to 1.

This approach allows one to determine the importance of the term in the entire collection of the analyzed documents. Terms with high uniqueness, which are less common in other documents, and often occur in the analyzed document, have the highest  $IDF$  value.

---

<sup>6</sup> Website of NLTK4 Russian developer: Department of mathematical linguistics SPbSU. URL: <http://mathling.phil.spbu.ru/node/160>.

<sup>7</sup> Official website of the developer "SciKit-learn" URL: <https://scikit-learn.org/stable/index.html>.

## 2.5. Artificial intelligence algorithms

For tasks of parametric analysis, regression, classification, identification and knowledge extraction, NLP uses an extensive toolkit of artificial intelligence algorithms (neural networks, genetic algorithms, metric algorithms, reference vectors, decision trees, etc.). Earlier studies of classification methods (Osochkin et al., 2018; Fomin & Osochkin, 2016) have shown that regression decision tree algorithms were effective in identifying the style and gender of the author of literary works. The advantage was due to their ability to attain higher classification accuracy when using small texts (less than 80,000 objects), compared to neural networks and the support vector machines. A significant advantage of all decision tree methods is their ability to represent the results as a hierarchical set of logical rules "if-then", which allows for meaningful identification, interpretation, verification of the classification results, as well as quantitative estimation of significance of each indicator. A variety of algorithms for constructing decision trees exist (Random Forest, ID3, C4.5, C5.0, CRT, CHAID, etc.), allows application of full potential of statistical analysis in the framework of a quantitative approach to natural language text processing.

In this paper, several algorithms were used for building decision trees in the IBM SPSS data analysis package. When identifying the author's gender, the CRT algorithm was used, as it is most suitable for binary classification. When classifying texts by the author's style, the CHAID algorithm was used, as it is the most suitable for classification of a large number of clusters.

## 3. Research materials

Two sets of texts in Russian were collected and analyzed in this study. The first corpus (*Corpus 1*) of texts contains Russian and Soviet literary prose of the 19th-20th centuries. Novels and stories were divided into chapters, containing several texts and each author is represented by 30 texts, as detailed in Table 1. The column "Average quantity of text symbols" shows the average number of characters contained in each text (without spaces).

Table 1.  
Corpus of Russian fiction

№	Class	Number of texts	Average quantity of text symbols
1	V.I. Belov	30	67,024
2	A.P. Beliaev	30	54,029
3	M.A. Bulgakov	30	89,525
4	D.A. Granin	30	48,078
5	F.M. Dostoyevsky	30	92,031
6	I.A. Efremov	30	53,089
7	A.I. Kuprin	30	56,380
8	A.N. Ostrovsky	30	62,022
9	Strugatsky brothers	30	72,097
10	L.N. Tolstoy	30	110,705
11	A.A. Fadeev	30	55,092
12	A.P. Chekhov	30	56,092
13	M.A. Sholokhov	30	105,032

The second corpus (*Corpus 2*) was compiled to evaluate identification of gender based on the author's style. It consists of 120 works of fiction by Russian writers of the XXI century, for example, a series of novels and fantasy by O.M. Sergeeva, A. Sergeeva, Surzhevskaya Marina Eff IR, K.G. Nazimov, M. Kamensky etc. The literary works included in the corpus of texts were taken from various Internet sites dedicated to fiction and scientific literature. Table 2 shows the characteristics of the Corpus 2.

Table 2.  
Author's gender text corpus

№	Class	Number of texts	Average quantity of text symbols
1	M	60	487 792
2	F	60	589 126

## 4. Results and Discussion

### 4.1 Author's style classification experiment

The experiment was conducted to identify the author's style based on fiction prose in order to evaluate the accuracy of the proposed data extraction method. The objective was to classify the corpus of texts presented in Table 1 according to authors' style identification.

The classification is based on the data extracted using the text data processing models described in the previous sections. The "exhaustive CHAID" algorithm using the Gini coefficient was chosen as the main algorithm for building the decision tree. This algorithm was chosen due to its ability to handle large number of classes (more than 10 clusters at the same time), high accuracy (Osochkin et al., 2020; Piotrowska, 2012), and a lower complexity of the resulting decision tree due to its use of non-binary tree-splitting algorithm.

The ratio of the training and test samples was taken to be 50%. The maximum tree depth was limited to 10, to avoid internal nodes with low number of texts. The Pierson Chi-square test was used to check the hypothesis of finding common characteristics. Since all indicators are relative, the node split significance criteria was set to 0.005.

The results of the experiments<sup>8</sup> on the author's style identification based on the text Corpus 1 using different mathematical models are shown in Figure 1.

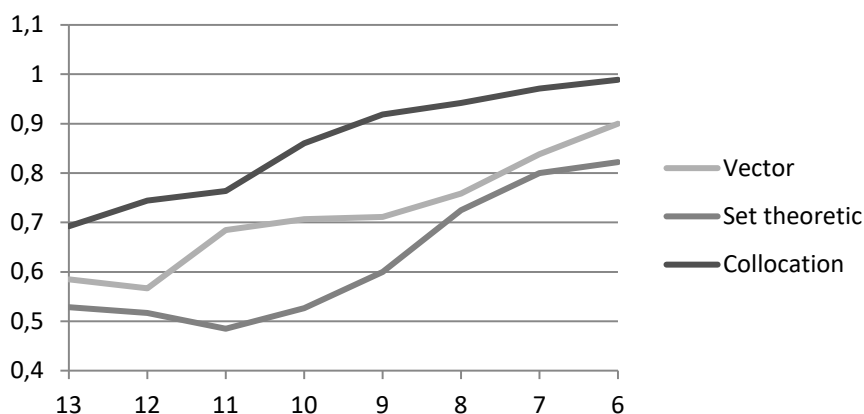


Figure 1. Chart of the author's style identification accuracy.  
(x-axis shows the number of clusters and y-axis shows the average classification accuracy)

Figure 1 shows the dependence of the classification accuracy on the number of clusters used. It is seen that the trend of classification accuracy increases when related authors (authors who had similar stylistic features) are removed from the classification, as the use of similar stylistic forms led to misclassification of texts among such authors.

<sup>8</sup> Classification was performed using the IBM SPSS software package.

Table 3.  
Author's style identification accuracy

Number of clusters	Vector model of text representation	Set-theoretic model	Set-theoretic model with collocation extraction
13	58.46%	52.82%	69.23%
12	56.67%	51.67%	79.65%
11	68.48%	48.48%	78.18%
10	70.67%	52.67%	87.72%
9	71.11%	60.00%	95.74%
8	75.83%	72.50%	95.00%
7	83.81%	80.00%	97.14%
6	90.00%	82.22%	98.88%

The data of Table 3 shows the identification accuracy of the author's style based on the parameters of the mathematical models and the number of utilized clusters. The results show that when classifying texts using 13 classes, the largest number of errors in the classification was encountered in the works of V.I. Belov and A.A. Fadeev. The total share of correctly identified works of A.A. Fadeev was 38.89%, when using the collocation method. An erroneous attribution of V.I. Belov is observed, since when classifying indicators extracted from a text using vector and set-theoretical text representation, the greatest percentage of false identifications of the author's style is seen in the works of V.I. Belov.

To increase the accuracy of classification, Fadeev's texts were removed from the corpus, since they were often identified as the works of V.I. Belov. The results of the removal of these texts slightly increased the overall accuracy of classification with the collocation method, but in other text representation methods the accuracy decreased. Furthermore, V.I. Belov has remained as the most mis-identified author.

Also, removing of Belov's texts from the text corpus had a positive effect on the general accuracy of the set-theoretic model with collocation extraction and vector representation methods. As the general accuracy of the set-theoretic method decreased, the largest proportion of errors in the classification with 12 clusters was made in the identification of V.I. Belov and A.R. Belyaev. Therefore, their works were removed from the subsequent classification based on the 11 and 10 clusters.

In the subsequent experiments, the authors' works classified with the least accuracy were removed from the corpus. The general accuracy reached 95.74% by using the set-theoretic text method with collocation extraction, when classifying by 9 clusters. That is 22.5% more accurate than in the set-theoretic text representation model and 19.17% more accurate than in the vector model.

In order to improve further classification accuracy, such authors as A.I. Kuprin, M.A. Bulgakov, Strugatsky Brothers, A.N. Ostrovsky, were removed. Each author's removal increased the accuracy of the general classification.

Table 4 shows detailed classification results using a mathematical model of text representation based on a set-theoretic text representation with collocation extraction.

Table 4.  
Classification of Russian prose by 13 authors

	Belov	Beliaev	Bulgakov	Granin	Dostoyevsky	Efremov	Kuprin	Ostrovsky	Strugatsky	Tolstoy	Fadeev	Chekhov	Sholokhov	Accuracy (%)
<b>Belov</b>	15	4	2	0	0	1	0	0	0	0	1	0	0	<b>65.22</b>
<b>Beliaev</b>	2	7	0	0	0	0	0	0	0	0	0	0	0	<b>77.78</b>
<b>Bulgakov</b>	1	0	19	0	2	0	0	0	0	0	3	0	1	<b>73.08</b>
<b>Granin</b>	3	1	0	18	0	0	0	0	0	0	0	0	0	<b>81.82</b>
<b>Dostoyevsky</b>	1	1	0	1	9	0	1	0	0	0	0	0	2	<b>60.00</b>
<b>Efremov</b>	5	0	0	0	0	6	0	0	0	0	0	0	0	<b>54.55</b>
<b>Kuprin</b>	0	3	0	0	0	0	13	0	0	0	2	0	0	<b>72.22</b>
<b>Ostrovsky</b>	1	0	0	0	0	0	1	8	0	0	1	0	0	<b>72.73</b>
<b>Strugatsky</b>	0	0	0	0	0	4	0	0	10	0	0	0	0	<b>71.43</b>
<b>Tolstoy</b>	0	0	0	0	0	0	0	0	0	10	0	1	2	<b>76.92</b>
<b>Fadeev</b>	4	5	0	0	0	2	0	0	0	0	7	0	0	<b>38.89</b>
<b>Chekhov</b>	0	0	0	0	0	0	2	0	0	0	0	7	0	<b>77.78</b>
<b>Sholokhov</b>	0	0	0	0	0	0	0	0	0	0	0	0	6	<b>100.00</b>
<b>Share of author's material in the test sample (%)</b>	16.41	10.77	10.77	9.74	5.64	6.67	8.72	4.10	5.13	5.13	7.18	4.10	5.64	<b>69.23</b>

Our results (Table 3-4) show that the data classification using the set-theoretic model, with collocation increased the accuracy of author identification by 24.63%, and the average increase was 15.81%, which indicates the effectiveness of the method. The vector text representation model showed on average 25.15% lower accuracy compared to the set-theoretic text representation with collocation extraction.

Table 5 shows the indicators and their values that were used by the exhaustive CHIAD decision tree construction method.

Table 5.  
The most important nine indicators of the author's identification

№	Indicator	Weight(%)
1	<i>Noun in accusative form + verb 1-st person</i>	6.26
2	<i>Noun in accusative form + verb 2-nd person</i>	5.58
3	<i>The use of Latin symbols</i>	5.53
4	<i>Adjective + Adjective</i>	4.99
5	<i>Adverb + Adverb</i>	4.87
6	<i>Numerals per sentence</i>	4.87
7	<i>Personal pronouns per sentence</i>	4.51
8	<i>Adjective + unanimated noun</i>	4.21
9	<i>Punctuation marks per sentence</i>	4.09

The main attributes used to identify the author were related not only to the frequency of individual parts of speech, but to their sequences. For example,

- the frequency of using a pair of nouns in the accusative form with the verb in the 1-st and the 2-nd personal;
- the use of Latin symbols allows one to distinguish most of the authors by their time periods at the first levels of decision trees;
- the authors of the Soviet period do not use Latin characters, which makes it possible to uniquely identify the works of Strugatsky, Belyaev, etc.

#### 4.2 Author's gender identification experiment

It was shown in (Macro, et al., 2020) that different digital services based on the Bert language model<sup>9</sup> made mistakes when identifying the author's work by gender.

We conducted an experiment to identify the author's gender. Specifically, we classified the Corpus 2 according to the author's gender. A binary algorithm for building a CRT decision tree was chosen, due to its efficiency in carrying out binary classifications. The results are presented in the Table 6.

Table 6.  
Classification by author's gender

Vector model	Books quantity	Gender	F	M	Accuracy
	37	F	25	11	69.44%
	23	M	12	12	50.00%
	Total 60	Share (%)	61.67	38.33	61.67%
Set-theoretic model	Books quantity	Gender	F	M	Accuracy
	36	F	19	12	61.29%
	24	M	17	12	41.38%
	Total 60	Share (%)	60.00	40.00	51.67%
Set-theoretic model with collocation extraction	Books quantity	Gender	F	M	Accuracy
	31	F	29	2	93.50%
	29	M	2	27	93.10%
	Total 60	Share (%)	51.70	48.30	93.30%

As can be seen from the classification results, the best accuracy was shown by the collocation method, with the total accuracy of 93.3%. The main characteristics that were used to identify the text were: the frequency of particles usages, the frequency of constructions such as “*a noun + verb in the 2-nd person*”, the average length of word, adverbs and punctuation marks per sentence, etc.

The main features for identifying the author's gender were the number of adjectives used, the frequency of adverbial verbs, and the number of n-grams: “*a noun in accusative + verb in the 1-st person, a noun in accusative + verb in the 2-nd person*”.

At the first level of binary classification, it was possible to divide the samples almost in half, thanks to the feature of particles from the total number of words. It was found that the female writers used particles in their works much more often than the male writers. In cases where particles did not accurately identify the cluster, it was found that the authors could be identified using the average word length in the text, with the male writers using on average longer words. A distinctive feature of the female authors was a more frequent use of the n-gram

<sup>9</sup> It is mentioned modifications of the Bert language model such as Google BERT and Facebook RoBERTa AI.



“*nouns in the instrumental case + verbs in the 1-st person*”. At the last level of the classification tree, punctuation marks were used; the male writers employed fewer figures of speech, direct speech, and other constructs where punctuation marks are used.

Table 7.  
Indicators value

№	Indicator	Weight(%)
1	<i>Percentage of the total number of particles</i>	27.67
2	<i>Nouns in the instrumental case + verb in the 1 - st person</i>	20.63
3	<i>Average words length</i>	20.27
4	<i>Percentage of participles</i>	11.79
5	<i>Punctuation marks per sentence</i>	11.32
6	<i>Vowel letters per word</i>	8.28

As it can be seen from the Table 7, the extraction of indicators related to the use of parts of speech and their features from the text has significantly increased the accuracy of classification when identifying the author's style in different literary works. The second most important indicator identified by the regression trees was the template “*nouns in the creative case + a verb in the 1-st person*”, as this sequence was more often used by the female authors.

When identifying the author's gender, the template that took into account the morphological form of the bigram: “*nouns in the instrumental case + verb in the 1-st person*”, which had an indicator value of 20.633 and was often used in the algorithm of regression decision trees to determine the author's gender.

The conclusions, which are based the author’s gender identification evaluations, confirm that the text representation model based on the set-theoretic model with collocation is more effective compared to other models.

## 6. Conclusion

The use of a quantitative approach with collocation extraction allows one to increase the accuracy of the authorial style identification. The experimental results of the style and gender identification accuracy confirmed the effectiveness of the proposed modification of the set-theoretic model of text processing of Russian prose. Algorithms for frequency-morphological extraction of numerical indicators and the formation of text indexes that reflect the frequency of individual parts of speech and n-gram parts of speech use, can be successfully used to identify the style. Our experiments have confirmed an increase in the total classification accuracy using collocation compared to the vector model of text representation.

Using the set-theoretic model with collocation extraction allows one to eliminate some of the disadvantages of the BERT data representation model that were pointed out earlier, and in conjunction with the methods of regression decision trees, the potential of text mining can be expanded. We also plan to conduct further experiments to analyze the accuracy of identifying the emotional colors of messages.

## Acknowledgements

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (project No. FSZN-2020-0027).

## References

- Allahyari, M. et al.** (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 13–17, 2017, Halifax, Nova Scotia, Halifax, Canada, CoRR abs/1707.02919/*.
- Beel, J. et al.** (2017). TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections. *Conference Preliminary Results Papers*, 1–8.
- Belinkov, Y. & Bisk, Y.** (2018). Synthetic and natural noise both break neural machine translation. *International Conference on Learning Representations*. URL: <https://arxiv.org/abs/1711.02173>
- Belinkov Y. & Glass, J.** (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*. 7, 49–72.
- Belyaeva, L.N. et al.** (2019). *Setevyye lingvisticheskiye tekhnologii. Kollektivnaya monografiya* (Network linguistic technologies. Collective monograph), 111. Saint-Petersburg: Herzen State University of Russia.
- Devlin, J., et al.** (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google AI Language*. URL: <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>
- Fomin, V.V. & Osochkin, A.A.** (2016). Text classification in the creative category with application of the frequency-morphological analysis algorithms and regression trees. *Current issues and prospects for the development of mathematical and natural Sciences. Collection of scientific papers on the results of the international scientific and practical conference. May 11, 2016 Omsk*, 64–66.
- Grekhov, A.V.** (2012). Kvantitativnyy metod: poisk latentnoy informatsii (Quantitative method: searching for latent information). *Vestnik Nizhegorodskogo universiteta im. Lobachevskogo*. 1 (3), 94–100.
- Harish, B.** (2012). Text Document Classification: An Approach Based on Indexing. *International Journal of Data Mining & Knowledge Management Process*, 1, 43–66. DOI: 10.5121/ijdkp.2012.2104
- Iyyer M. et al.** (2018). Adversarial example generation with syntactically controlled paraphrase networks. *Proceedings of NAACL-HLT*, 1875–1885. <https://www.aclweb.org/anthology/N18-1170>
- Jadhao, A. et al.** (2016). Text Categorization using Jaccard Coefficient for Text Messages. *International Journal of Science and Research (IJSR)*, 5, 2046–2050.
- Kang, Y. et al.** (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7 (12), 1–34. DOI: 10.1080/23270012.2020.1756939
- Kashcheyeva, A.V.** (2013). Kvantitativnyye i kachestvennyye metody issledovaniya v prikladnoy lingvistike (Quantitative and qualitative methods of research in applied linguistics). *Sotsial'no-ekonomicheskiye yavleniya i protsessy*, 3 (49), 18.
- Khalezova, N., et al.** (2020). Cross-sectional Study of Clinical and Psycholinguistic Characteristics of Mental Disorders in HIV Infection. *R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019)*. Proceedings of the III International Conference on Language Engineering and Applied Linguistics. CEUR-WS, 2552, 161–178 URL: <http://ceur-ws.org/Vol-2552/Paper14.pdf>
- Kolesnikova, O.** (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Comp. y Sist.* [online]. 2016, 20 (3), 327–344. URL: <https://doi.org/10.13053/cys-20-3-2456>

- Macro, T., et al.** (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList», *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 49024912 URL: <https://www.aclweb.org/anthology/2020.acl-main.442>
- Maheshan, M., et al.** (2018). Indexing-Based Classification: An Approach Toward Classifying Text Documents Information Systems. *Design and Intelligent Applications*, 1, 894902. DOI: 10.1007/978-981-10-7512-488
- Marcus, S.** (1967). *Algebraic Linguistics; Analytical Models*. Academic Press: New York.
- Martin, D. & Jurafsky, D.** (2019). *Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice Hall. URL: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
- McCann, B., et al.** (2017). Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems* 30 (NIPS 2017), 6294–6305.
- Mikros, G.** (2013). Systematic stylometric differences in men and women authors: a corpus-based study. Köhler, R.; Altmann, G. (eds.): *Issues in quantitative linguistics*. Lüdenscheid: RAM, 206–223.
- Moschitt, A.** (2004). Complex Linguistic Features for Text Classification: a comprehensive study. *Lecture Notes in Computer Science. 26 European Conference on IR Research, Sunderland, UK*, 181–196.
- Moulton, R. & Jiang, Y.** (2018). Maximally Consistent Sampling and the Jaccard Index of Probability Distributions. *International Conference on Data Mining, Workshop on High Dimensional Data Mining 2018*, 347–356. URL: <https://arxiv.org/abs/1809.04052>
- Osochkin, A.A., et al.** (2018). Eksperimenty text-minig po klassifikacii tekstov v ramkah zadach personalizacii obrazovatel'noj sredy (Text-minig experiments on the classification of texts in the framework of the problems of personalization of the educational environment). *Informatizaciya obrazovaniya i nauki*, 2 (38), 38–50.
- Osochkin, A.A., et al.** (2020). Comparative Research of Index Frequency - Morphological Methods of Automatic Text Summarisation. *NESinMIS-2020. Proceedings of the XV International Conference "New Educational Strategies in Modern Information Space", Saint-Petersburg, Russia, March 25, 2020. Vol. 2401*, 73–86. URL: [http://ceur-ws.org/Vol-2630/paper\\_8.pdf](http://ceur-ws.org/Vol-2630/paper_8.pdf)
- Piotrowska, X.R.** (2012). Kvantitativnyy psikholingvisticheskiy analiz khudozhestvennogo tvorchestva (Quantitative psycholinguistic analysis of artistic creativity). *Nauchnoye mneniye (Scientific opinion)*, 6-7, 16–20.
- Piotrowska, X.R.** (2014). Tekst mayning: perspektivy razvitiya (A Survey of Text mining). *Izvestiya RGPU im. A.I. Gertsena*, 168, 128–134.
- Popescu, I.-I. et al.** (2010). Vectors and codes of text. Lüdenscheid: RAM.
- Roul R. et al.** (2017). Modified TF-IDF Term Weighting Strategies for Text Categorization. *Proceedings of 14th IEEE India Council International Conference (INDICON)*, 16.
- Salton, G., Allan, J. & Buckley, C.** (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2), 97–108.
- Yongchang, W. et al.** (2019). Research on improved text classification method based on combined weighted model. *National Natural Science Foundation of China*, 7(11), 783–796.
- Vincze, V.** (2015). The relationship of dependency relations and parts of speech in Hungarian. *Journal of Quantitative Linguistics*, 22(1), 168–177.
- Wang, Y. & Zhu, L.** (2020). Research on improved text classification method based on combined weighted model. *Concurrency and Computation: Practice and Experience*, 32 (6), 783–796.

- Yang, Zh., et al.** (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of Advances in Neural Information Processing Systems 32*  
URL: <https://arxiv.org/abs/1906.08237>
- Zakharov, V.P., Khokhlova, M.V.** (2010). Analiz effektivnosti statisticheskikh metodov kollokatsiy v tekstakh na russkom yazyke (A Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts). *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii. Proceedings of the Annual International Dialogue Conference. Bekasovo. 26–30 May 2010, 9 (16)*. Moscow: Russian State University of Humanities, 137–14

## **Glottometrics, 31–50: Bibliography**

Using the jubilee of *Glottometrics*, we are glad to present a continuation of a complete bibliography of all publications of the issues 31–50.

The contributions are ordered in five sections: (1) general articles, (2) history, (3) reviews, (4) bibliographies, and (5) miscellanea. Within each of these sections, the contributions are ordered according to authors' names and year of publication. For a copy of the bibliography as RIS or BibTeX file please contact [emmerich.kelih@univie.ac.at](mailto:emmerich.kelih@univie.ac.at).

### **(1) General articles**

- Altmann, Gabriel** (2016). Types of Hierarchies in Language. *Glottometrics*, 34; 44–55.
- Altmann, Gabriel** (2018). Some Properties of Adjectives in Texts. *Glottometrics*, 41; 67–79.
- Altmann, Gabriel** (2018). The Nature and Hierarchy of Belza-Chains. *Glottometrics*, 42; 75–85.
- Andreev, Sergej; Fan, Fengxiang; Altmann, Gabriel** (2018). Adnominal Aggregation. *Glottometrics*, 40; 63–76.
- Andreev, Sergej; Lupea, Mihaiela; Altmann, Gabriel** (2017). Belza Chains of Adnominals. *Glottometrics*, 39; 72–87.
- Andreev, Sergej** (2018). Distribution of Syllables in Russian Sonnets. *Glottometrics*, 41; 13–23.
- Andreev, Sergej** (2016). Verbal vs. Adjectival Styles in Long Poems by A.S. Pushkin. *Glottometrics*, 33; 25–31.
- Andreev, Sergej** (2017). Verbal vs. Adjectival Styles in Belkin Tales by A. S. Pushkin. *Glottometrics*, 36; 17–21.
- Andreev, Sergej** (2018). A Study of Russian Adnominals. *Glottometrics*, 42; 56–74.
- Andreev, Sergej** (2018). Adnominal Valency Motifs in Sonnets. *Glottometrics*, 42; 46–55.
- Andreev, Sergej** (2019). Types of Syllable Distribution in Russian Long Poems. *Glottometrics*, 46; 29–40.
- Andreev, Sergej** (2020). Nominal vs. Adjectival Adnominals in Russian Fiction: Relationship and Distribution. *Glottometrics*, 49; 1–12.
- Andreev, Sergej** (2020). Syllabic Identity of Verse Lines in Russian Long Poems: Skinner's Hypothesis. *Glottometrics*, 48; 1–8.
- Andreev, Sergej; Celano, Giuseppe; Yang, Jiang; Altmann, Gabriel** (2018). Some Properties of Polysemy. *Glottometrics*, 43; 77–90.
- Andreev, Sergej; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2016). On Russian Adnominals. *Glottometrics*, 35; 64–84.
- Andreev, Sergej; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2017). Skinner's Hypothesis Applied to Russian Adnominals. *Glottometrics*, 36; 32–69.
- Andreev, Sergej; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2017). Some Properties of Adnominals in Russian Texts. *Glottometrics*, 38; 77–106.
- Andreev, Vadim** (2020). Fitting the Distribution of the Syllabic Types in Different Positions of Verse. *Glottometrics*, 49; 87–97.
- Best, Karl-Heinz** (2015). Malay Borrowings in English. *Glottometrics*, 31; 50–53.
- Best, Karl-Heinz; Místecký, Michal; Zörnig, Peter; Altmann, Gabriel** (2019). Quantifying the Quantitative Meter: On Rhythmic Types in the Dactylic Hexameter. *Glottometrics*, 46; 83–98.
- Bortolato, Claudia** (2016). Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature. *Glottometrics*, 34; 28–43.

- Buk, Solomija; Rovenchak, Andrij** (2019). Simple Definition of Distances between Texts from Rank–Frequency Distributions. A Case of Ukrainian Long Prose Works by Ivan Franko. *Glottometrics*, 46; 1–11.
- Cai, Huiying; Qu, Yunhua; Feng, Zhiwei** (2019). A Corpus-Based Study of the Semantic Prosody of Chinese Light Verb Pattern Across Registers: Taking *jinxing* and *shoudao* as Examples. *Glottometrics*, 46; 61–82.
- Chen, Xinying; Gómez-Rodríguez Carlos; Ferrer-i-Cancho, Ramon** (2018). A Dependency Look at the Reality of Constituency. *Glottometrics*, 40; 104–106.
- Coloma, Germán** (2016). An Optimization Model of Global Language Complexity. *Glottometrics*, 35; 49–63.
- Dai, Zheyuan; Liu, Haitao** (2019). Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages over 50 Years (1967–2018). *Glottometrics*, 45; 63–88.
- Fang, Yu; Liu, Haitao** (2015). Comparison of Vocabulary Richness in Two Translated Honglougong. *Glottometrics*, 31; 54–75.
- Ferrer-i-Cancho, Ramon** (2016). The Meaning-Frequency Law in Zipfian Optimization Models of Communication. *Glottometrics*, 35; 28–37.
- Ferrer-i-Cancho, Ramon** (2017). Random Crossings in Dependency Trees. *Glottometrics*, 37; 1–12.
- Ferrer-i-Cancho, Ramon** (2017). The Optimality of Attaching Unlinked Labels to Unlinked Meanings. *Glottometrics*, 36; 1–16.
- Ferrer-i-Cancho, Ramon** (2017). The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach. *Glottometrics*, 39; 38–71.
- Ferrer-i-Cancho, Ramon; Gómez-Rodríguez, Carlos** (2016). Liberating Language Research from Dogmas of the 20th Century. *Glottometrics*, 33; 33–34.
- Gao, Zhao** (2019). A Quantitative Lexical Study on Commercial English. *Glottometrics*, 47; 16–35.
- Gnatchuk, Anna** (2016). A Quantitative Analysis of English Compounds in Scientific Texts. *Glottometrics*, 33; 1–7.
- Gnatchuk, Hanna** (2015). A Quantitative Investigation of English Compounds in Prose Texts. *Glottometrics*, 32; 1–8.
- Gnatchuk, Hanna** (2015). Anglicisms in the Austrian Newspaper *Kleine Zeitung*. *Glottometrics*, 31; 38–49.
- Gnatchuk, Hanna** (2015). Sound Symbolism: Myths and reality. *Glottometrics*, 31; 1–30.
- Gnatchuk, Hanna** (2016). The Relationship between English Synonyms and Compounds. *Glottometrics*, 34; 9–13.
- Gnatchuk, Hanna** (2019). Measuring Lexical Richness of the USA Presidents' Inauguration Speeches. *Glottometrics*, 44; 87–93.
- Gnatchuk, Hanna** (2019). The Classification of English Styles on the Basis of Lexical Parameters: A Case of Clustering Analysis. *Glottometrics*, 45; 57–62.
- Gnatchuk, Anastasia; Gnatchuk, Hanna** (2018). Linking Elements of German Compounds in the Texts of Technical Science. *Glottometrics*, 40; 46–50.
- Gnatchuk, Anastasia; Gnatchuk, Hanna** (2020). Identification of English Styles on the Basis of Parts of Speech: A Case of Principal Component Analysis and Factor Analysis. *Glottometrics*, 48; 52–66.
- Holm, Hans J.** (2017). Steppe Homeland of Indo-Europeans Favored by a Bayesian Approach with Revised Data and Processing. *Glottometrics*, 37; 54–81.
- Hou, Renkui; Huang, Chu-Ren; Zhou, Mi; Jiang, Menghan** (2019). Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis. *Glottometrics*, 45; 24–57.

- Huang, Wei** (2015). Quantitative Studies in Chinese Language. *Glottometrics*, 31; 76–84.
- Hůla, Jan; Kubát, Miroslav; Čech, Radek; Chen, Xinying; Číž, David; Pelegrinová, Kateřina; Milička, Jiří** (2019). Context Specificity of Lemma. Diachronic Analysis. *Glottometrics*, 45; 7–23.
- Ishutin, Denys; Gnatchuk, Hanna** (2017). Ukrainian Compounds in the Texts of Computer Science. *Glottometrics*, 39; 88–92.
- Kelih, Emmerich** (2019). Segmental and Suprasegmental Vowel Frequencies in Slovene: Statistical Modeling. *Glottometrics*, 45; 1–6.
- Kelih, Emmerich; Altmann, Gabriel** (2015). A Continuous Model for Polysemy. *Glottometrics*, 31; 31–37.
- Kelih, Emmerich; Andreev, Sergey; Altmann, Gabriel** (2018). Polysemy of some Parts of Speech. *Glottometrics*, 42; 39–45.
- Kelih, Emmerich; Köhler, Reinhard; Altmann, Gabriel** (2020). Obituary. Peter Grzybek (1957 – 2019). *Glottometrics*, 48; 1–2.
- King, Adam** (2018). The Lexicon and the Noisy Channel: Words are shaped to avoid confusion. *Glottometrics*, 43; 58–67.
- Köhler, Reinhard; Naumann, Sven** (2016). Syntactic Text Characterisation Using Linguistics S-Motifs. *Glottometrics*, 34; 1–8.
- Kolenčíková, Natália; Altmann, Gabriel** (2020). Analysis of Prepositions in Marína (Slovak Romantic Poem). *Glottometrics*, 48; 88–107.
- Kolenčíková, Natália; Místecký, Michal; Altmann, Gabriel** (2020). Polysemy of Rhyme Words: A Case Study of Two Slovak Poems. *Glottometrics*, 49; 13–31.
- Kubát, Miroslav; Čech, Radek** (2016). Quantitative Analysis of US Presidential Inaugural Addresses. *Glottometrics*, 34; 14–27.
- Li, Jingjie** (2019). Inter-textual Vocabulary Growth Patterns for Marine Engineering English. *Glottometrics*, 47; 36–51.
- Liu, Haitao; Xu, Chunshan; Liang, Junxing** (2016). Dependency Length Minimization: Puzzles and Promises. *Glottometrics*, 33; 35–38.
- Liu, Ziqi; Liu, Haitao** (2020). Quantitative Analysis of Academic Writing as to Informality and Vocabulary Features. *Glottometrics*, 49; 32–51.
- Lu, Fang; Liu, Haitao** (2015). Probability distribution of interlingual lexical divergences in Chinese and English: dao and said in Honglougong. *Glottometrics*, 32; 63–87.
- Ma, Hong; Liu, Haitao** (2019). Probability Distribution of Causal Linguistic Features. *Glottometrics*, 44; 77–86.
- Mehler, Alexander; Gleim, Rüdiger; Uslua, Tolga; Stegbauer, Christian** (2018). On the Self-similarity of Wikipedia Talks: A Combined Discourse-analytical and Quantitative Approach. *Glottometrics*, 40; 1–45.
- Melka, Tomas** (2018). Stylistic study of Omnilingual by H. Beam Piper. *Glottometrics*, 43; 31–57.
- Miangah Mosavi, Tayebbeh; Rezai Javad, Mohammad** (2016). Persian Text Ranking Using Lexical Richness Indicators. *Glottometrics*, 35; 6–15.
- Michels, Christopher** (2015). The Relationship between Word Length and Compounding Activity in English. *Glottometrics*, 32; 88–98.
- Místecký, Michal** (2018). Belza Chains in Machar's Letní sonety. *Glottometrics*, 41; 46–56.
- Místecký, Michal** (2018). Counting Stylometric Properties of Sonnets: A Case Study of Machar's Letní sonety. *Glottometrics*, 41; 1–12.
- Místecký, Michal; Altmann, Gabriel** (2019). Tense and Person in English: Modelling Attempts. *Glottometrics*, 46; 98–104.

- Místecký, Michal; Altmann, Gabriel; Andreev, Sergey** (2018). Piotrowski Law in Sequences of Activity and Attributiveness: A Four-Language Survey. *Glottometrics*, 42; 21–38.
- Místecký, Michal; Radková, Lucie** (2020). School and Gender in Numbers: A Stylometric Insight into the Lexis of Teenagers' Description Essays. *Glottometrics*, 49; 52–65.
- Místecký, Michal; Yang, Jiang; Altmann, Gabriel** (2018). Belza-Chain Analysis: Weighting Elements. *Glottometrics*, 43; 68–76.
- Mohanty, Panchanan; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2019). Script Complexity in Indian Languages. *Glottometrics*, 44; 94–99.
- Pavel, Kosek; Čech, Radek; Navrátilová, Olga; Mačutek, Ján** (2018). On the Development of Old Czech (En)clitics. *Glottometrics*, 40; 51–62.
- Pelegrinová, Kateřina; Altmann, Gabriel** (2017). The Study of Adverbials in Czech. *Glottometrics*, 37; 34–53.
- Pelegrinová, Kateřina; Altmann, Gabriel** (2020). Concept Realization in Texts. *Glottometrics*, 48; 9–16.
- Petrack, Fabienne; Gnatchuk, Hanna** (2017). Lexicographic Problems of Collocations in Quantitative Linguistics. *Glottometrics*, 38; 21–29.
- Poiret, Rafaël; Liu, Haitao** (2017). Mastering the Measurement of Text's Frequency Structure: an Investigation on Lambda's Reliability. *Glottometrics*, 37; 82-100.
- Popescu, Ioan-Iovitz; Mosavi Miangah, Tayebah; Gnatchuk, Hanna; Čech, Radek; Bodoc, Alice; Altmann, Gabriel** (2017). On Rank-Frequency Distributions in Poetry. *Glottometrics*, 38; 30–54.
- Rácová, Anna; Zörnig, Peter; Altmann, Gabriel** (2019). Syllable Structure in Romani: A Statistical Investigation. *Glottometrics*, 46; 41–60.
- Radojičić, Marija; Lazić, Biljana; Kaplar, Sebastijan; Ranka, Stanković; Obradović, Ivan; Mačutek, Ján; Leššová, Livia** (2019). Frequency and Length of Syllables in Serbian. *Glottometrics*, 45; 114–123.
- Reina, Francesc; Castellón, Irene; Padró, Lluís** (2019). Towards the Prepositional Meaning via Machine Learning: A Case Study of Spanish Grammar. *Glottometrics*, 44; 1–15.
- Rinkeit-Vit, Lyubov; Gnatchuk, Hanna** (2016). Euphemisms in Political Speeches by USA Presidents. *Glottometrics*, 35; 16–21.
- Roelcke, Thorsten; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2017). Aspects of Text Concentration. *Glottometrics*, 36; 70–86.
- Rostin, Tim** (2016). List of Journals Containing Contributions to Quantitative linguistics. *Glottometrics*, 33; 73–100.
- Rottmann, Otto A.** (2018). On Word Length in German and Polish. *Glottometrics*, 42; 13–20.
- Rovenchak, Andrij; Rovenchak, Olha** (2018). Quantifying Comprehensibility of Christmas and Easter Addresses from the Ukrainian Greek Catholic Church Hierarchs. *Glottometrics*, 41; 57–66.
- Rovenchak, Andrij; Vydrin, Valentin** (2020). Syllable Frequencies in Manding: Examples from Periodicals in Bamana and Maninka. *Glottometrics*, 48; 17–36.
- Sanada, Haruko; Altmann, Gabriel** (2018). Word Length and Polysemy in Japanese. *Glottometrics*, 41; 40–45.
- Savoy, Jacques** (2017). Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. *Glottometrics*, 38; 55–76.
- Shmidt, Ekaterina; Gnatchuk, Hanna** (2016). German Compounds in the Texts of Technical Science. *Glottometrics*, 35; 1–5.



- Stachowski, Kamil** (2020). Tools for Semi-Automatic Analysis of Sound Correspondences: The Soundcorrs Package for R. *Glottometrics*, 49; 66–86.
- Su, Hong** (2019). A Study on Inter-textual Vocabulary Growth Patterns for Maritime Convention English. *Glottometrics*, 47; 52–65.
- Vasilev, Alexei; Vasileva, Iiona** (2018). Text Length and Vocabulary Size: Case of the Ukrainian Writer Ivan Franko. *Glottometrics*, 43; 1–10.
- Wang, Lin; Čech, Radek** (2016). The Impact of Code-Switching on the Menzerath-Altmann Law. *Glottometrics*, 35; 22–37.

## (2) History

- Best, Karl-Heinz** (2017). Weitere Autoren für eine Geschichte der Quantitativen Linguistik. *Glottometrics*, 36; 87–119.
- Best, Karl-Heinz; Altmann, Gabriel** (2018). Word Length with G. Herdan. To the Memory of G. Herdan who Died 16. 11. 1968. *Glottometrics*, 42; 86–90.
- Hernández-Fernández, Antoni; Ferrer-i-Cancho, Ramon (2018). José María de Oleza Arredondo, S.J. (1887–1975). *Glottometrics*, 41; 80–86.
- Wang, Yaqin; Liu, Haitao** (2018). In Remembrance of Fengxiang Fan, 1950–2018. A Pioneer of Quantitative Linguistics in China. *Glottometrics*, 43; 91–96.
- Wang, Yawen; Liu, Haitao** (2019). The Effects of Source Languages on Syntactic Structures of Target Languages in the Simultaneous Interpretation: A Quantitative Investigation Based on Dependency Syntactic Treebanks. *Glottometrics*, 45; 89–113.
- Wei, Aiyun; Liu, Haitao** (2019). Typological Features of Zhuang from the Perspective of Word Frequency Distribution. *Glottometrics*, 44; 59–75.
- Wilson, Andrew** (2020). Lengths and L-motifs of Rhythmical Units in Formal British Speech. *Glottometrics*, 48; 37–51.
- Xu, Yingying** (2019). The Distribution of Word Families in Chinese College English Textbooks. *Glottometrics*, 47; 1–15.
- Xu, Yingying; Yu, Yang; Fan, Fengxiang** (2018). Quantitative Linguistics and R. *Glottometrics*, 42; 1–12.
- Yan, Jianwei; Liu, Siqui** (2017). The Distribution of Dependency Relations in Great Expectations and Jane Eyre. *Glottometrics*, 37; 13–33.
- Yan, Yaobin** (2019). A Corpus-Based Comparative Study of Lexis in Hong Kong and Native British Spoken English. *Glottometrics*, 47; 66–82.
- Yang, Yu; Jhang, Se-Eun** (2018). A Menzerath-Altmann Model for NP length and Complexity in Maritime English. *Glottometrics*, 40; 91–103.
- Yu, Biyan; Jiang, Yue** (2018). Probability Distribution of Syntactic Divergences of Determiner his-(Adjective)-Noun Structure in English-to-Chinese Translation. *Glottometrics*, 40; 77–90.
- Zenkov, Andrei V.; Místecký, Michal** (2019). The Romantic Clash: Influence of Karel Sabina over Mácha's Cikáni from the Perspective of the Numerals Usage Statistics. *Glottometrics*, 46; 12–28.
- Zhang, Cong; Liu, Haitao** (2015). A Quantitative Investigation of the Genre Development of Modern Chinese Novels. *Glottometrics*, 32; 9–20.
- Zhang, Fangfang** (2019). Computational Stylistic Characteristics of American English. *Glottometrics*, 47; 123–137.
- Zhang, Guoqiang; Liu, Haitao** (2019). A Quantitative Analysis of English Variants Based on Dependency Treebanks. *Glottometrics*, 44; 16–33.

- Zhang, Hongxin; Liu, Haitao** (2016). Quantitative Aspects of RST Rhetorical Relations across Individual Levels. *Glottometrics*, 33; 8–24.
- Zhang, Xiaojin; Liu, Haitao** (2020). Function Words in Male and Female Authors: A Diachronic Investigation of Modern Chinese Prose. *Glottometrics*, 48; 67–87.
- Zhang, Xiaowen; Qu, Yunhua; Feng, Zhiwei** (2019). The Diachronic Relationship Between the Contemporary American English Present Perfect and Past Simple across Registers. *Glottometrics*, 44; 34–58.
- Zhou, Pianpian** (2019). A Study on the Subjectival Position and the Syntactic Complexity in Spoken English. *Glottometrics*, 47; 83–103.
- Zhu, Jieqiang; Liu, Haitao** (2018). The Distribution of Synonymous Variants in Wenzhounese. *Glottometrics*, 41; 24–39.
- Zhu, Yujia** (2019). A Comparative Study on NP Length, Complexity and Pattern in Spoken and Written English. *Glottometrics*, 47; 104–122.
- Zörnig, Peter** (2017). On the Arc Length in Quantitative Linguistics: a Continuous Model. *Glottometrics*, 36; 22–31.
- Zörnig, Peter; Altmann, Gabriel** (2016). Activity in Italian Presidential Speeches. *Glottometrics*, 35; 38–48.
- Zörnig, Peter; Místecký, Michal** (2018). Quantifying the Importance of Stylometric Indicators: A Principal Component Approach to Czech Sonnets. *Glottometrics*, 43; 11–30.
- Zörnig, Peter; Popescu, Ioan-Iovitz; Altmann, Gabriel** (2015). Statistical Approach to Measure Stylistic Centrality. *Glottometrics*, 32; 21–54.

### **(3) Reviews**

- Altmann, Gabriel** (2017). Kelih, Emmerich (2016): Phonologische Diversität - Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax. Frankfurt am Main: Peter Lang Verlag. *Glottometrics*, 37; 101.
- Beaudouin, Valérie** (2016). Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis. *Glottometrics*, 33; 56–72.
- Chen, Heng** (2018). Review: Mikhail Kopotev, Olga Lyashevskaya, & Arto Mustajoki. (Eds.) (2017). Quantitative Approaches to the Russian Language. New York: Routledge. ISBN:978-1-138-09715-5, 220 pp. *Glottometrics*, 41; 91–95.
- Gnatchuk, Hanna** (2017). Liu, Haitao; Liang, Junying (eds.) (2017): Motifs in Language and Text. Berlin/Boston: de Gruyter Mouton. 271pp. (= Quantitative Linguistics, vol. 71). *Glottometrics*, 37; 102–105.
- Gnatchuk, Hanna** (2018). Haitao Liu, Junying Liang (eds.) (2017), Motifs in Language and Text. Berlin/ Boston: De Gruyter Mouton, pp. 271. (Quantitative Linguistics Vol. 71). *Glottometrics*, 41; 87–90.
- Ishutin, Denys** (2016). Hanna Gnatchuk: Sound Symbolism. A Phonosemantic Analysis of German and English consonants. Akademiker Verlag. 2015. 96 pp. *Glottometrics*, 33; 101–102.
- Místecký, Michal** (2017). Kubát, Miroslav: Kvantitativní analýza žánrů [A Quantitative Analysis of Genres]. Ostrava: Ostravská univerzita, 2016, 141 pp. *Glottometrics*, 39; 93–94.

### **(4) Bibliographies**

- Chen, Ruina** (2015). Bibliography of Quantitative Linguistics of Chinese Researches in International Academic Journals. *Glottometrics*, 31; 85–88.
- Grzybek, Peter; Kelih, Emmerich** (2015). *Glottometrics*, 1-30: Bibliography. *Glottometrics*, 31; 89–102.

**(5) Miscellanea**

**Ferrer-i-Cancho, Ramon** (2017). A Commentary on “The Now-or-Never Bottleneck: a Fundamental Constraint on Language”, by Christiansen and Chater (2016). *Glottometrics*, 38; 107–111.

**Futrell, Richard; Mahowald, Kyle; Gibson, Edward** (2016). Response to Liu, Xiu, and Liang (2015) and Ferrer-i-Cancho and Gómez-Rodríguez Dependency Length Minimization. *Glottometrics*, 33; 39–44.

**Léon, Jacqueline; Loiseau, Sylvain** (2016). Interview with Jean Petitot. *Glottometrics*, 34; 56–78.

**Liao, Shengyu; Lei, Lei** (2017). What We Talk about When We Talk about Corpus: A Bibliometric Analysis of Corpus-related Research in Linguistics (2000-2015). *Glottometrics*, 38; 1–20.

**Lin, Yanni; Liu, Haitao** (2017). Bibliometric Analysis of *Glottometrics*. *Glottometrics*, 39; 1–37.

**Wang, Yaqin** (2019). Preface. *Glottometrics*, 47; 1.

Compiled by *Emmerich Kelih*<sup>1</sup> & *Peter Grzybek*<sup>2</sup>

---

<sup>1</sup> University of Vienna, emmerich.kelih@univie.ac.at.  <http://orcid.org/0000-0002-8315-8916>

<sup>2</sup> University of Graz, Austria.

Other linguistic publications of RAM-Verlag:

### Studies in Quantitative Linguistics 1-30

The following volumes appeared up to now:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp.
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3*. 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language*. 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis*. 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1*. 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences*. 2015, IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015, III+146 pp.
22. P. Zörnig et al. *Positional occurrences in texts: Weighted Consensus Strings*. 2016. II+179 pp.
23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4*. 2016, 287 pp.

24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme*. 2017, VI+125 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6*. 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries*. 2018, 129 pp.
30. P. Zörnig, K. Stachowski, A. Rácová, Y. Qu, M. Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann, *Quantitative Insights into Syllabic Structures*. 2019, IV+134 pp.