Abstracts Glottometrics 50, 2021

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

Tayebeh Mosavi Miangah, Relja Vulanović

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.*

English Loanwords in Mongolian Usage

Minna Bao, Saheya Brintag, Dabhurbayar Huang

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.*

A Quantitative Study on English Polyfunctional Words

Lu Wang, Yahui Guo, Chengcheng Ren

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 partd 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-Altmann function, as used in the previous investigation. The results show very good fit for all groups exzept conjunction group.

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

Initial and Final Syllables in Tatar: from Phonotactics to Morphology

Alfiya Galieva, Zhanna Vavilova

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 *χ2* test is applied.

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

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, Xenia Piotrowska, Vladimir Fomin

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.*

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.