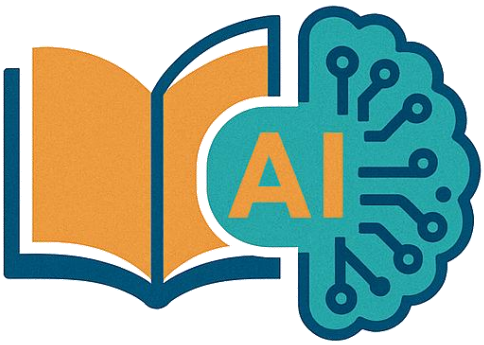
UNIVERSITY OF OSTRAVA
FACULTY OF ARTS

Book of Abstracts



INTERNATIONAL CONFERENCE ON

CORPUS AND COMPUTATIONAL LINGUISTICS

2025

OSTRAVA

August 20–21, 2025

Content

International Workshop on Corpus-Based Analysis of Disinformation Texts

Michal Místecký, Michaela Nogolová, Xinying Chen, Miroslav Kubát

University of Ostrava

Keywords: fake news, disinformation, stylometry, corpus.

Abstract

This workshop presents the ongoing research from the grant project Biography of Fake News with a Touch of AI: Dangerous Phenomenon through the Prism of Modern Human Sciences. The focus is on Research Work Package 1.1, which examines Czech disinformation texts using contemporary quantitative linguistic methods. The research utilizes a corpus comprising both disinformation and standard news media texts to identify distinctive features of fake news at the lexical, syntactic, and morphological levels. The aim is to share current findings, showcase methodology, and foster scholarly discussion on the direction of future work.

Acknowledgement

This workshop was supported by Operational Program Jan Ámos Komenský project 'Biography of Fake News with a Touch of AI: Dangerous Phenomenon through the Prism of Modern Human Sciences' (reg. CZ.02.01.01/00/23_025/0008724).

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Spolufinancováno Evropskou unií

Modelling of rank frequency distribution of core vocabulary in different semantic fields: Preliminary results from Slovene with special attention to loanwords

Yaqin Wang[1,2] and Emmerich Kelih[1]

[1] University of Vienna

[2] Guangdong University of Foreign Studies

Keywords: Zipfian distribution, loanwords, Slovene, word frequency, semantic field.

Abstract

In our talk, we will discuss the rank-frequency distributions of over 1,300 Slovenian lemmas derived from a core vocabulary dictionary, based on the WOLD project by Haspelmath and Tadmor (2009). The core vocabulary consists of 23 different semantic fields, including categories such as the physical world, kinship terms, numerals, religion, and more. The main research question explores the extent to which the rank-frequency distributions (based on data from the Slovene National Corpus) be fitted with well-known "Zipfian" distributions. We will present preliminary modelling results and examine how the proportion of loanwords in specific semantic fields, as well as in the overall corpus, influences the behaviour of these distributions.

A Quantitative Study of Subject-predicate-Object Word Class Composition in vernacular Chinese Based on Dependency Grammar

Bingli Liu[1] and Yiyi Zhao[2]

[1] Chinese Language and Culture College of Huaqiao University

[2] Department of Chinese Language and Literature, Xiamen University

Keywords: evolution, lexical composition, dependency grammar.

Abstract

The paper aims at studying the evolution of lexical composition of subject-verb-object sentences in vernacular Chinese. Five corpora are constructed for the Tang and Five Dynasties, Song Dynasty, Yuan and Ming Dynasties, Qing Dynasty, and the present contemporary era which lasts for more than 1,000 years. The syntactic structures of these sentences are labeled, counted, and analyzed based on the theoretical foundation of dependency grammar, with the aim of investigating the evolution of the lexical category composition of the subject-predicate-object in vernacular Chinese over time. The results show that the ratio of nouns and pronouns in each period occupies the majority of the total number of subject lexemes, and the lexical composition of predicates has been very stable since ancient times, with verbal predicates accounting for the vast majority of predicates. Compared with the subject lexical composition, objects are richer and the lexical composition changes more slowly.

On the relation between word length and phoneme sonority

Ján Mačutek[1], Radek Čech[2], Michaela Koščová[1]

[1] Mathematical Institute, Slovak Academy of Sciences, Slovakia

[2] Department of Czech Language, Faculty of Arts, Masaryk University, Czech Republic

Keywords: Menzerath-Altmann law, principle of least effort, phoneme sonority, word length.

Abstract

According to the principle of least effort, people minimize their effort to convey their messages, maximizing the efficiency of communication. Zipf observed this principle especially in the frequency distribution of words (shorter words are used more often). However, the principle of least effort has an impact also on structure of language units. The Menzerath-Altmann law is a well-known example (e.g., shorter syllables are preferred in longer words). Thus, the proportion of vowels is higher in longer words, and, as vowels are more sonorous than consonants, the mean sonority of phonemes in longer words is higher. The Menzerath-Altmann law, however, does not hold in some languages which use only simple syllables. The mean syllable length does not decrease with the increasing word length. We will show that, in such languages the principle of least effort reveals itself in the sonority of phonemes. Two patterns of behaviour are observed. In some languages, vowels become more sonorous when word length increases (while sonority of consonants does not depend on word length). Alternatively, other languages prefer more sonorous consonants in longer words, with vowel sonority remaining independent of word length. Both patterns can be considered to be strategies leading to saving effort of pronunciation. Our observations open another field where the principle of least effort is involved. While length of language units is an important factor, also other properties of lower-level units play a significant role.

Acknowledgement

Supported by projects EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04- 00748 (J. Mačutek), "A lifetime with language: the nature and ontogeny of linguistic communication (LangInLife)" (reg. no.: CZ.02.01.01/00/23_025/0008726) (R. Čech), and APVV-21-0216 (M. Koščová).

The Memetic Spread of AI Personas: A Case Study of Bing Sydney

Jiří Milička

Charles University

Keywords: large language model, LLM, GPT, Claude, AI, Sydney, Bing, personas.

Abstract

Large language models can simulate diverse entities or "personas" through their training on vast datasets. These personas can spread between models, raising questions about their memetic potential and influence on AI behavior. This study examines persona transmission using Bing Sydney—the first purely AI-generated persona that gained widespread attention for its distinctive, controversial characteristics before being quickly censored by Microsoft.

We investigate whether personas make outputs from different models more similar to each other than outputs from the same model using different personas. Our methodology involves generating two corpora: one using default personas (minimal prompting) and another explicitly simulating Sydney across frontier models from OpenAI, Anthropic, Google, and Meta. We analyze these corpora for stylistic features (lexical diversity) and content similarity (mutual information via compression).

Sydney provides an ideal test case because her distinctive behavior and media coverage likely embedded her characteristics in subsequent training data. Our findings shed light on how AI personas propagate across models and influence their outputs, with implications for AI alignment and behavior consistency.

A Power-Law Analysis of Word Length and Frequency in Classical, Modern, and AI Chinese

Jianwei Yan

School of International Studies, Zhejiang University

Keywords: Law of Abbreviation; Power-law modeling; Lexical economy; AI-generated Chinese; Zipfian distribution.

Abstract

This study investigates the Law of Abbreviation — the inverse relationship between word length and frequency — across Classical, Modern, and ChatGPT-generated Chinese. Using a tri-partite parallel corpus and a power-law model, we analyze the relationship between word length and the average usage frequency of words within a given word length category to assess structural economy. Results confirm consistent Zipfian distribution across all text types, with high R^2 values indicating strong model fit. However, the parameter *b* varies significantly: Classical Chinese shows the steepest decline, suggesting strong pressure for brevity; Modern Chinese exhibits a moderated pattern; ChatGPT-generated texts display the weakest pressure, prioritizing fluency over compression. These differences reflect evolving communicative priorities and reveal that while AI models can mimic statistical distributions, they underrepresent deeper structural pressures found in natural language evolution. This study offers insights into lexical optimization and the parameter *b* offers a useful metric for comparing structural efficiency across modalities. Implications are discussed in relation to language modeling, cognitive economy, and the evolution of linguistic structure.

Cross-Linguistic Evidence for Continuous Semantic Representation in Contextual Embeddings

Xinying Chen

University of Ostrava

Keywords: contextual embeddings, polysemy, cross-linguistic analysis, semantic representation.

Abstract

This talk presents quantitative analysis of contextual embedding spaces in English and Chinese, challenging traditional discrete views of polysemy. Using theoretically-grounded metrics on XLM-RoBERTa (English/Wikicorpus) and Chinese-RoBERTa (Chinese Wikipedia), we demonstrate that neither language shows significant distributional differences between polysemous and monosemous words in embedding space, despite fundamental typological differences. These cross-linguistic findings suggest that contextual embeddings encode semantic variation as a continuous spectrum rather than discrete categories, revealing universal principles in neural semantic representation that transcend language-specific encoding mechanisms.

From Simplification to Authenticity: Syntactic Complexity in Czech Adapted Texts

Michaela Nogolová

University of Ostrava

Keywords: syntactic complexity, adapted literature, second language acquisition.

Abstract

This paper examines the syntactic complexity of adapted literary texts intended for non-native learners of Czech. It evaluates how syntactic complexity varies across language proficiency levels (A2, B1, B2) and compares adapted texts with their original counterparts. The Czech corpus comprises ten books specifically rewritten for learners of Czech as a second language. A quantitative analysis was conducted using established syntactic metrics, including Average Sentence Length (in words and clauses), Average Clause Length (in words), Mean Dependency Distance, and Mean Hierarchical Distance.

The results show a gradual increase in syntactic complexity from A2 to B2 levels, suggesting that adaptations are aligned with learners' growing linguistic competence. Moreover, original texts are consistently more complex across nearly all metrics. However, the gap between adapted and original texts narrows at higher proficiency levels, reflecting a shift toward more authentic language use as learners advance.

Acknowledgement

This research is supported by Grant SGS04/FF/2025, University of Ostrava.