Workshop: Digitale Slawistik
Ilia Afanasev, Elias Moncef Bounatirou, Maximilian Grübsch, Anna Jouravel

Mit dem Eintritt in die Big-Data-Ära des 21. Jahrhunderts haben sich die Forschungsmöglichkeiten und -methoden innerhalb der Geistes- und Sozialwissenschaften signifikant verändert. Das Training von Large Language Models (LLM) und die Entwicklung von Transformern wie BERT (Devlin et al. 2019) oder der GPT-Familie (Brown et. al 2020) beeinflussen alle linguistischen Bereiche, insbesondere die Verarbeitung natürlicher Sprache (NLP), und die Slawistische Linguistik ist hier keine Ausnahme (vgl. Nogolová et al. 2023). Das Ziel dieses Workshops ist es, die Auswirkungen von LLM auf die Fragestellungen und Arbeitsweisen innerhalb der slawistischen Forschung auszuloten.

Regina Guzaerova (Justus-Liebig-Universität Gießen)
Corpus-Based Analysis of the Concepts of Political Correctness and New Ethics in the Russian-Speaking Media Space

This study explores the concepts of political correctness and new ethics in the Russian-speaking media space through a comprehensive corpus-based analysis. Using advanced natural language processing (NLP) techniques alongside traditional corpus linguistic methods, the study examines how these concepts are represented and have evolved in the Russian media in recent years.
The research uses a diverse and representative corpus compiled from various sources, including Russian newspapers, online news platforms, blogs, and social media, spanning the years 2010 to 2024. This extensive time frame allows for a detailed exploration of the temporal dynamics and shifts in discourse related to political correctness and new ethics. Sentiment analysis assesses public attitudes and emotional tones, revealing how media coverage has evolved.
Advanced NLP techniques such as Named Entity Recognition (NER) and Topic Modeling identify key entities and underlying topics within the corpus. Discourse analysis critically examines media framing of political correctness and new ethics, highlighting variations by political orientation and media type.
Results provide insights into term frequency, distribution, and context, offering a nuanced understanding of public discourse. Trends illustrate the evolution of these concepts and correlate with major socio-political events.
This study contributes to the growing body of research on global manifestations of political correctness and evolving social norms. By focusing on the Russian-speaking context, we shed light on how these concepts are localized, contested, and reimagined within a specific cultural and linguistic sphere. Our findings have implications for understanding cross-cultural communication, media discourse analysis, and the global circulation of ideas related to social justice and cultural change.

Maksim Aparovich (KNOT Knowledge Research Group, Brno University of Technology), Volha Harytskaya, Vladislav Poritski, Oksana Volchek (independent scholar, Lithuania), Pavel Smrž (KNOT Knowledge Research Group, Brno University of Technology)
Towards a GLUE-type benchmark for Belarusian

Recent progress in language modelling gave rise to various kinds of natural language understanding benchmarks. Many of them are similar to GLUE [Wang et al. 2019a] and its descendant SuperGLUE [Wang et al. 2019b]; in particular, such benchmarks are available for Russian [Shavrina et al. 2020] and Polish [Rybak et al. 2020] but they are not yet available for some of the smaller, relatively low-resource Slavic languages, which hinders further development of multilingual capabilities in LLMs.
This presentation introduces a GLUE-type benchmark for Belarusian, an East Slavic language. The benchmark includes five novel datasets focused on the following tasks:
1. Sentence-level sentiment analysis. Sentences with positive and negative polarity (none neutral) have been manually selected from thematically varied online sources reflecting the real-world diversity of modern written Belarusian.
2. Named entity recognition. The dataset, derived from be_hse corpus in Universal Dependencies [Nivre et al. 2020; Shishkina & Lyashevskaya 2021], has been annotated in accordance with Universal NER guidelines [Mayhew et al. 2024].

3. Linguistic acceptability. Due to the limited availability of linguistic publications providing Belarusian utterances labeled as (un)acceptable, we chose a different route than in the original CoLA [Warstadt et al. 2019]. First, unacceptable Belarusian sentences were harvested opportunistically from various sources, including normative sources such as Belarusian language textbooks, machine translation outputs, and language model hallucinations. Then, a group of native speakers constructed acceptable counterparts for most instances.
4. Word in context. Like the original WiC [Pilehvar & Camacho-Collados 2019], this dataset is derived from a lexicographic resource – an explanatory dictionary of Belarusian that provides usage examples for each word meaning separately. Pairs of contexts attached to the same meaning are considered positive, and pairs of contexts attached to homonymic words are negative. 5. Winograd schema challenge. This is an expert-crafted translation of the original WSC-285 [Levesque et al. 2012] into Belarusian, mostly straightforward but with a few variations related to typological differences between English and Belarusian. An early-stage preliminary evaluation doesn't show any significant performance gap between commercial offerings, such as GPT-4o, and state-of-the-art free LLMs, such as Llama3-70B, on the new benchmark. The datasets included in the benchmark are not equally challenging for the current generation of computational tools: while sentence-level sentiment analysis and word-in-context tasks appear to be mostly solved for Belarusian, the models' acceptability judgments and named entity recognition skills are still lagging behind the human baseline, and the models struggle significantly with the Winograd schema challenge. Our plan is to perform a more detailed quantitative evaluation based on templates from LM Evaluation Harness [Gao et al. 2023].

Nataliia Cheilytko, Ruprecht von Waldenfels (Friedrich-Schiller-Universität Jena)
Lectometry Analysis for Standard Ukrainian Varieties in XX Century

The paper examines Ukrainian in the 20th century as a multistandard language, following Auer's framework (Auer, 2021). In the early 20th century, Ukrainian was characterized by two distinct standard varieties, a result of its division between Poland and the Russian Empire. However, in the latter half of the century, these varieties began to converge following World War II. Despite this convergence, regional linguistic features persist, particularly in western and central Ukraine.
The research project presented here aims to investigate the mechanisms and trends underlying this convergence process using a large, representative linguistic corpus. Specifically, we calculated aggregated uniformity (Geeraerts et al., 2023) to measure diachronic distances among language varieties. These calculations focused on onomasiological profiles of approximately 100 concepts, which reflect the distribution of synonymous lexemes across different lects. In the case of polysemous lexemes, only occurrences where a lexeme was used in the relevant sense representing the target concept were considered. Word sense disambiguation for these occurrences was conducted using the GPT-4o-mini model (Cheilytko & Waldenfels, 2024).
Regionally annotated data for the study were sourced from the GRAC corpus (Shvedova et al., 2024). To visualize the trajectory of convergence over time, pairwise distances among lects were analyzed using splits graphs.

Liudmila Radchankava (Universität Mainz)
Erforschung komplexer Adpositionen im Russischen

Die Untersuchung komplexer Adpositionen im Russischen hat in den letzten Jahren aufgrund ihrer komplexen semantischen und funktionalen Eigenschaften zunehmend an Bedeutung gewonnen. Diese Forschung nutzt einen digitalen linguistischen Ansatz, um die Entwicklungen und strukturellen Dynamiken komplexer Adpositionen im Russischen zu analysieren. Ein zentrales Ziel ist die Erforschung der diachronen Entwicklung und der semantischen Veränderungen sekundärer Adpositionen. Die Methodik umfasst die Nutzung umfangreicher Korpusdaten aus dem Russischen Nationalkorpus sowie die Anwendung fortgeschrittener NLP-Techniken (BERT). Dabei werden semantische Cluster gebildet, um die unterschiedlichen Verwendungsweisen und Kontextualisierungen komplexer Adpositionen zu erfassen. Frequenzanalysen ermöglichen es, Veränderungen in der Häufigkeit und Verwendung über verschiedene Zeitperioden hinweg zu dokumentieren. Besondere Aufmerksamkeit wird auf die semantische Entwicklung dieser Adpositionen gelegt, um ihre Rolle innerhalb des russischen Sprachsystems besser zu verstehen. Durch die Kombination von digitalen Analysetechniken und linguistischen Methoden wurden Muster der Adpositionsformation sowie ihre syntaktische Funktionalität untersucht.

Edyta Jurkiewicz-Rohrbacher (Universität Hamburg)

Accessing the language knowledge in the case of pre-trained generative: structures with Russian dative pronouns

The paper deals with the problem, how language competence of pre-trained generative models can be accessed for linguistic research by adjusting the established linguistic methods. I postulate that translation should be considered a reliable task enabling indirect access to linguistic competence of pre-trained generative models, similarly to psycholinguistic tests and translational questionnaires used in theoretical and typological linguistic studies. Here, I test how reliable is a translation task for testing the coreference resolution and syntactic constituency parsing in Slavic languages well-known for their morphological ritchness and word order flexibility.

As a study case I use Russian authentic bipredicative stimuli with adjacent personal pronouns in Dative, as shown in (1):

(1) Istočnik takže upominaet nekotorye interesnye spekuljacii otnositel'no planov Intel i NVIDIA, no im2 nam1 by chotelos'1 posvjatit'2 otdel'nyj material. [Ex1]

'The source also mentions some interesting speculations regarding the plans of Intel and NVIDIA, but we would like to dedicate a separate article to them.' The constituent im nam by chotelos' posvjatit' otdel'nyj material is ambiguous without context. In the authentic example, the linearly first dative pronoun im 'them' is governed by the infinitival complement posvjatit 'dedicate', while the linearly second and adjacent pronoun nam 'us' is governed by the complement taking matrix, predicate by chotelos' 'COND wish.3SG.REFL'. Notice that this sentence does not contain an explicit subject in the nominative, which makes the structure more complicated for parsing than the structures containing a canonical nominative, agentive subject.

Basing on the analysis of web corpus data Jurkiewicz-Rohrbacher (2023) assumes such structures should still be challenging for pre-trained generative language models. In the present paper, I test the performance of

Table 1
Accuracy in the translation task.

Translation Systems				
	DL	**Google**	**Yandex**	
accuracy	0.85	0.74	0.66	
Chatbots				
	Omni	Turbo	Gemini	Perplexity
accuracy	0.95	0.89	0.89	0.89

Table 2
Distribution of the studied factors.

incorrectly classified			correctly classified		
hierarchy	**no**	**yes**	hierarchy	**no**	**yes**
word order			word order		
D1D2	24 (0.49)	43(0.17)	D1D2	25	216
D2D1	14 (0.15)	2(0.02)	D2D1	77	117

Table 3
Results of the performed regression model

Random effects:				
Groups	Name	Variance Std.Dev.		
SentID (Intercept)	3.5459	1.8831		
Translator (Intercept)	0.9743	0.9871		
Fixed effects:				
	Estimate	Std. Error	z value Pr(>|z|)	
(Intercept)	0.2053	0.9090	0.226	0.82133
D2D1	2.3771	1.0624	2.237	0.02525*
D1 locuphoric	2.5861	0.9511	2.719	0.00655*
D2D1:D1 locuphoric	0.9296	1.5092	0.616	0.53791

seven models commercially available as machine translation systems (DeepL, Yandex Translate, Google Translate) or agents capable of translating (ChatGPT4omni, ChatGPT3.5turbo, Google Gemini, Perplexity.AI). 74 stimuli representing different word orders are presented together with context of total length between 200 and 800 characters to the seven systems and agents. Each context contains enough information to disambiguate and correctly resolve the coreference to the adjacent dative pronouns in Russian. The results of translation tasks (508 observations) evaluated with a logistic regression mixed- effect model show that the task is not straightforward and leads to considerable errors in particular in structures which violate the prominence hierarchy principle (Haspelmath 2021) and or standard word orders. In general, translating agents perform better than translation systems with the accuracy 0.89-0.95. Translation systems reach the accuracy 0.66-0.85 in the task. The study suggests that both correspondence between the word order of the governors and pronouns, and hierarchical prominence should play a role in solving the task, since the available context does not suffice to solve correctly the task in certain syntactic environments. The phenomena studied in theoretical linguistics and typology seem, therefore, relevant and retrievable from the linguistic behavior of large language models, also by means of the established linguistic methods. Although language processing of pre-trained models is still comparable to a black box, I argue that methods used for studying the linguistic behavior of human species can be adjusted to studying the linguistic behavior of machines. Otherwise, the human natural language users can be considered a black box too, as linguistic knowledge is never accessible directly.

Martin Meindl, Elena Renje (Universität Freiburg)

Vorstellung eines Workflows zur Verarbeitung und Analyse vormoderner slavischer Handschriften und Drucke

Mit dem Aufkommen öffentlich zugänglicher Technologien zur automatischen Transkription von handschriftlichen Texten und auch ihrer Verbreitung in den verschiedenen historischen Disziplinen (u.a. Camps et al. 2019, Franzini et al. 2018, Polomac 2024, Rabus 2019) erschließen sich auch der Paläoslavistik neue Analysewege. Der Beitrag beschreibt einen Workflow, der von ausgewählten digitalisierten Texten (in Bildformat) bis zu potenziellen Analyseergebnissen führen soll. Zunächst werden verschiedene Aspekte der automatischen Texterkennung mithilfe der Plattform Transkribus (https://readcoop.eu/de/transkribus) besprochen, die vom Modelltraining bis zur eigentlichen Texterkennung (Handwritten Text Recognition, HTR) reichen. Dabei nimmt die Auswahl von Trainingsdaten eine zentrale Rolle ein. Außerdem wird auf die Menge der Trainingsdaten so- wie das Training von spezifischen und generischen Modellen mitsamt ihren Vor- und Nachtei- len eingegangen. Als nächster Schritt des Arbeitsprozesses wird das Postprocessing der generierten Daten beleuchtet. Hierbei geht es um die Evaluation der Transkriptionsqualität anhand der Berechnung einer Character Error Rate (CER) oder Word Error Rate (WER) sowie die damit einhergehende Analyse der verschiedenen Fehlerarten, um in einem weiteren Schritt gezielte Korrekturen, die sowohl manuell als auch automatisiert erfolgen können, in den Transkriptionen vorzunehmen. Die überarbeitete Transkription wird dann mithilfe unterschiedlicher Verfahren für weitere Analysen vorbereitet. Hier sind zum einen die Tokenisierung mithilfe des Tools UDPipe zu nennen (https://github.com/bnosac/udpipe). Zum anderen wird das generierte HTR-Material mithilfe des Stanzataggers getaggt (https://github.com/yvesscherrer/stanza-tagger). Außerdem werden Einblicke in die Grenzen und Möglichkeiten der Analyse unkorrigierter HTR-Daten gegeben und es wird anhand verschiedener Methoden demonstriert, dass die Analyse derartiger Daten – mit Einschränkungen – befriedigende Ergebnisse liefern kann. Zu den diskutierten Verfahren zählen hier vor allem stilometrische Ansätze (Textclustering, Nearest Shrunken Centroids Algorithmus) und deskriptive statistische Verfahren, die ebenfalls kurz vorgestellt werden. Abschließend soll die Bedeutung der Qualitätskontrolle der Analyseergebnisse hervorgehoben werden. Ein zentraler Bestandteil des vorgestellten Workflows ist eine manuelle Überprüfungsschleife, um sicherzustellen, dass die Ergebnisqualität nicht eingeschränkt ist. Auch wenn ein quantitativer Ansatz zu robusteren Ergebnissen führt, ist der qualitative Aspekt nicht zu vernachlässigen. Daher streben wir bei der bisherigen Arbeit mit vormodernen Handschriften einen Mixed-Methods-Ansatz an (vgl. Rabus & Petrov 2023).

Vladimir Polomac (University of Kragujevac)

Universal Dependencies for Serbian medieval charters and letters: towards defining a tagset for morpho-syntactic annotation

The paper represents a part of the research devoted to the theoretical and methodological foundations of the creation of an electronic corpus of Serbian medieval charters and letters. The basic concept of the corpus, the principles of text selection, the principles of arranging texts in electronic form and the definition of metadata about the texts were previously discussed in Polomac 2021. The current paper focuses on the most important challenges of morphosyntactic annotation of the corpus. Given that the texts from the corpus mix Old Serbian and Serbian Church Slavonic languages, as well as that the corpus covers the period from the 12th to the 15th centuries, the choice of morphosyntactic categories to be annotated represents the main challenge. Having in mind that the electronic corpus of Serbian medieval charters and letters is conceived as the most important part of the future referenct historical corpus of the Serbian language, which should serve as a fundamental resource for the development of a corpus-based historical grammar and the Serbian language, the author's intention is for the annotation to include not only parts of speech (PoS tagging) but also as many morphosyntactic categories as possible. The most important result of the paper is the definition of a tagset for the morphosyntactic annotation of the electronic corpus of Serbian medieval charters and letters, obtained by adapting the Old Church Slavonic, Old East Slavic (Eckhoff and Berdičevskis 2015), Old Czech (Zeman et al. 2023) and modern Slavic languages (Zeman 2015) tagsets developed within the Universal Dependencies project.