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.

Beatrice Bindi (University "G. D'Annunzio" of Chieti-Pescara)

On linguistic annotation of ancient Russian texts: performance evaluation of Stanza and UDPipe on selected works of Maximus the Greek (c. 1470-1555/56)

This report is part of the broad and widely problematic issue of linguistic annotation of ancient Russian texts, and proposes to evaluate the performance of Stanza and UDPipe on selected works of Maximus the Greek (c. 1470-1555/56) The work presented is part of my doctoral project, which aims to carry out a lexical-semantic analysis of the writings of Maximus the Greek, starting with the creation of a corpus of the author's texts available in modern critical editions in digital format (Sinicyna 2008; Sinicyna 2014). It is well known that a corpus, to be defined as such, must present itself as «a large collection of linguistic data, stored in electronic format, standardised, structured, annotated and philologically reliable» (Zacharov, Bogdanova 2013: 5). In the case of historical corpora of the Russian language, however, the annotation requirement is difficult to fulfil. Over the past decade, numerous specialists have devoted their efforts to the problem of linguistic annotation of ancient texts, leading to the adoption of individual solutions which, in the most successful cases (e.g. «Manuskript». Slavjanskoe pis'mennoe nasledie [http://mns.udsu.ru], Nacional'nyj korpus russkogo jazyka: Istoričeskie korpusa [https://ruscorpora.ru/]), have required large teams of specialists and a hybrid mode of work capable of combining the statistical approach, the approach based on electronic dictionaries and a good deal of manual work (cf. e.g. Baranov 2019; Gavrilova, Šalganova, Lyashevskaya 2016). What is still missing, however, are automatic taggers that can be widely used in the annotation of historical sources. In this context of great experimentation and lack of unambiguous solutions, we wanted to question the potential of some existing and freely available statistical taggers: our evaluations were carried out comparatively on Stanza and UDPipe. The annotation results were subjected to a quantitative and qualitative analysis with the aim of identifying the most common types of errors. The resulting insights allow for an assessment of potential future applications of the investigated taggers for the purpose of further automation of linguistic annotation procedures of Church Slavonic texts. At the same time, it provides a basis for possible considerations on what kind of preprocessing/retraining of the models could be introduced in order to avoid the identified problems.

Fabio Maion (Universität Innsbruck)

Zum Aufbau eines historischen Korpus des Balkanslawischen

Dass die Verwendung digitaler Methoden zu Fortschritten bei der Erforschung historischer Sprachvarietäten beitragen kann, würden wohl die wenigsten bestreiten. Der eng damit verbundene Aufbau diachroner Korpora ist in der Slawistik jedoch hauptsächlich auf das Altkirchenslawische als älteste Sprachstufe und auf Texte aus dem ostslawischen Raum fokussiert (Baranov, V. A., Varfolomeev, A. G. 2012; Eckhoff et al. 2018; Eckhoff/Berdičevskis 2015). Für südslawische Sprachvarietäten gibt es bislang nur wenige solche Ressourcen. An der Universität Sofia wurde eine Sammlung südslawischer Texte vom 11. bis ins 18. Jahrhundert zusammengestellt (Totomanova 2021). Diese ist jedoch auf die reine Widergabe der Texte beschränkt und ermöglicht nur eine Suche nach Wortformen. Zusätzlich aufbereitet und mit linguistischen Zusatzinformationen versehen wurden südslawische Texte von Šimko (2021), wobei der Fokus auf frühneubulgarischen Damaskini-Texten lag. Im Gegensatz zu Texten aus dem ostslawischen Raum wurden südslawische Texte aus dem Spätmittelalter allerdings noch nicht in einem annotierten und somit nach linguistischen Kriterien durchsuchbaren Korpus aufgearbeitet. Dies ist aus mehreren Gründen bedauerlich: Erstens gab es auf dem

Balkan im 13. und 14. Jahrhundert eine reiche Übersetzungsliteratur und zweitens ist gerade das Balkanslawische, das seit seiner ersten Bezeugung im Zuge der Slawenmission tiefgreifende strukturelle Veränderungen erfahren hat, aus einer historischen Perspektive für Linguisten besonders interessant.
Ich möchte daher in meinem Vortrag auf aktuelle Projekte der Innsbrucker Slawistik eingehen, die sich zum Ziel gesetzt haben, diese Lücke zu füllen, und zum Aufbau eines historischen Korpus des Balkanslawischen beizutragen. Dabei werde ich zum einen vorstellen, mit welchen Methoden sich Texte aus der Zeit des Zweiten Bulgarischen Reiches linguistisch aufarbeiten lassen und welche Ergebnisse dabei zu erwarten sind. Zum anderen möchte ich zeigen, wie gemeinsam mit bulgarischen Kollegen am Aufbau einer Infrastruktur für ein historisches Korpus gearbeitet wird (Dimitrova et al. akzeptiert). Abschließend werde ich an einem Fallbeispiel, den Lang- und Kurzformen von Adjektiven, illustrieren, welche Möglichkeiten sich durch digitale annotierte Korpora ergeben und welche neuen Erkenntnisse sich aus entsprechenden Abfragen gewinnen lassen.

Vladimir Neumann (Staatsbibliothek zu Berlin)

Effektiver Einsatz von NLP- und Suchmaschinentechnologien für die Analyse diachroner slawistischer Texte am Beispiel des kirchenslawischen Textkorpusmaterials (als Volltext) und der historischen slawistischen Wörterbücher (als Dataframe)

Im Zentrum dieser Untersuchung steht die Fragestellung, wie sich lexikalische Vergleiche effizient bewerkstelligen lassen und welche Methode dabei besser geeignet ist: die bewährten Methoden aus dem Bereich des Natural Language Processing (NLP) oder moderne Ansätze wie die semantische Suche mittels Embeddings. Diese Fragestellung wird anhand von diachronen Texten des Altkirchenslawischen, die nach Unicode konvertiert wurden, untersucht. Der lexikalische Vergleich erfolgt durch den Abgleich mit Textmaterial aus verschiedenen Korpora, konkret aus den konvertierten Materialien des SofiaTrondheimCorpus (STC), des Corpus Cyrillo-Methodianum Helsingiense (CCMH), TITUS und Syntacticus, sowie mit Datenbanken digitalisierter Wörterbücher wie Prager Wörterbuch (SJS, Gorazd), LexiconPGL (Miklosich), Sadnik-Digital, Vostokov-Sreznevskij und SlavaComp-Metaglossar, die als Dataframe vorliegen. Der Vortrag behandelt folgende Methoden und Technologien: Lexikalischer Textvergleich mittels NLP-Algorithmen zur Identifikation und zum Vergleich des Vokabulars, Implementierung von Indexierungs- und Retrieval-Mechanismen aus der Suchmaschinentechnologie zur effizienten Textanalyse, Einsatz von Machine Learning-Techniken zur Klassifikation und Analyse von Texten, inklusive der Verwendung von Transformer-Modellen wie BERT und GPT, Anwendung von Embeddings zur semantischen Suche und zur besseren Erfassung der Semantik und Kontextualisierung von Texten sowie Entwicklung und Implementierung von Visualisierungswerkzeugen in Python-Notebooks zur Darstellung und Analyse der Ergebnisse. Obwohl die semantische Suche und Embeddings moderne und vielversprechende Ansätze darstellen, zeigt dieser Vortrag, dass traditionelle NLP-Methoden, wie die Extraktion von Lemmata und morphologischen Annotationen, sowie die Nutzung von Bi- und Trigrammen in Kombination mit Suchmaschinentechnologien, nach wie vor äußerst effektive und wertvolle Ergebnisse liefern können. Diese traditionellen Technologien werden oft zugunsten der Transformer-basierten Methoden vernachlässigt, doch es gibt viele interessante und ertragreiche Aspekte, die weiterhin aus diesen Ansätzen herausgeholt werden können. Der Vortrag stellt mehrere Anwendungsfälle vor, darunter die intratextuelle Analyse und der Vergleich historischer slawistischer Texte zur Identifikation diachroner Sprachwandel und Entwicklungen sowie die intertextuelle Untersuchung zeitgenössischer slawistischer Texte zur Erkennung diatopischer und diastratischer Unterschiede. Zudem werden Techniken zur Ausrichtung von lexikalischen und syntaktischen Einheiten zwischen unterschiedlichen Textkorpora vorgestellt (Alignment), um eine präzise Auffindung und Lokalisierung der Lexik im diachronen Zusammenhang zu ermöglichen.
Die vorgestellten Methoden bieten innovative Werkzeuge zur Untersuchung und Visualisierung slawistischer Texte. Sie ermöglichen eine detaillierte Analyse sowohl innerhalb einzelner Texte als auch im Vergleich zwischen verschiedenen Texten über längere Zeiträume hinweg. Diese Ansätze fördern die Forschung in der slawistischen Linguistik und bieten neue Möglichkeiten für die Textanalyse.

Maximilian Grübsch (Universität Wien)

Multidimensional Scaling and Russian Future Periphrases

The imperfective future tense construction *budu* + Infinitive, a loan from Polish (Moser 1998), entered the Russian language in the late Middle Russian period, a time of profound changes in the verbal paradigm. In

the 16th and 17th centuries, there still was a variety of infinitive periphrases referring to future time such as *stanu* and *učnu* with infinitive. While a syntactic distribution has been observed (Pen'kova 2019), many questions remain as to the history of their productivity and the semantic relations between these periphrases, as future tense itself is at the intersection of tense, aspect and mode and therefore not easy to evaluate by traditional methods (Hilpert 2008). This research strives to uncover some answers by creating corpus-based semantic maps via multidimensional scaling as demonstrated in Hilpert (2021). Multidimensional scaling is a computational approach that visualizes (in this case, semantic) proximity and therefore allows to compare the different periphrases to phasal and modal verbs in a systematic manner. The talk will be dedicated to the results of this method in the form of a tentative semantic description of the individual periphrases but also to the shortcomings of this method.

Ilia Afanasev (Universität Wien / Columbia School of Linguistics Society)

Multi-lect automatic detection of Swadesh list items from raw corpus data in East Slavic languages

The talk presents an analysis of the multi-lect automatic detection of Swadesh list items (Swadesh, 1955) from raw corpora, using the material of contemporary East Slavic standard lects (Ukrainian, Belarusian, and Russian). This task aids the early stage of historical linguistics study by helping the researcher compile word lists for further analysis, combining in a unique way semasiological and onomasiological approaches for the corpus data exploration. The main question that the talk explores is the notions of «basicness» and «swadeshness» (Dellert and Buch, 2019) in the vocabulary, and whether it is possible to detect them with the help of machine learning methods.

The conducted experiments utilise the ASJP list, one of the most well-known variations of the Swadesh list (Holman et al., 2008), enriched with additional words from the 110-item Swadesh list (Kassian et al., 2010), in particular, *woman, kill, eat, all, man, me, and you (indirect) (genitive stem).* This list is then split into two halves, α-list and β-list, with the underlying distributional semantic criterion, namely, the words may be either direct antonyms (*fire/water*), or very closely semantically related (*see/eye, ear/hear)*. The corpus is split into three parts, α-corpus (that contains only clauses with α-list items), β-corpus (that contains only clauses with β-list items), and neutral part that contains clauses without any Swadesh list items. The latter part is used only for training (to increase the size of the corpus for the models to get more information about the lects), the first two – for both training and prediction. It is also possible to augment the dataset with token 3-grams that consist of a Swadesh list item, surrounded by the previous (or beginning of the clause token) and the following (or the end of the clause token).

The task of Swadesh list items detection is formalised as a binary (Swadesh list item/not Swadesh list item) classification with two heavily skewed classes (the probability of random Swadesh list item encounter in the dataset is only 0.02). Thus, the utilised methods are different classification models, that are known to be robust enough to overcome this issue: Random Forest (Ho, 1995) is a baseline, and the tested against it main methods are hybridised Hidden Markov Model, h-HMM (Lyashevskaya and Afanasev, 2021), Conditional Random Fields (CRF) (Silfverberg et al., 2014), and mBERT (Devlin et al., 2019) with Named Entity Recognition (NER) head.

From a quantitative point of view, the models perform rather poorly (the best F1-score is 0.36 by mBERT trained on β-corpus for α-corpus – still significantly higher than the probability of random encounter), and data augmentation harms their performance heavily (for mBERT, it falls to 0), with the exception being h-HMM, which retains the same level of efficiency. The qualitative evaluation, however, shows the more complex picture, especially of the relationship between the tokens with similar semantics in a set of closely-related languages, while allowing for better evaluation of the term *basic vocabulary* in historical comparative research (Borin, 2012). Future research will expand both the dataset and the list of analysed tokens, in order to achieve better computational results.