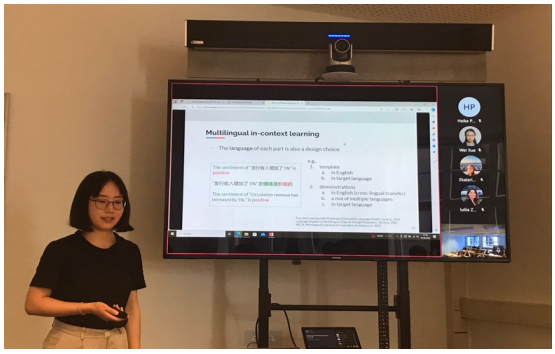
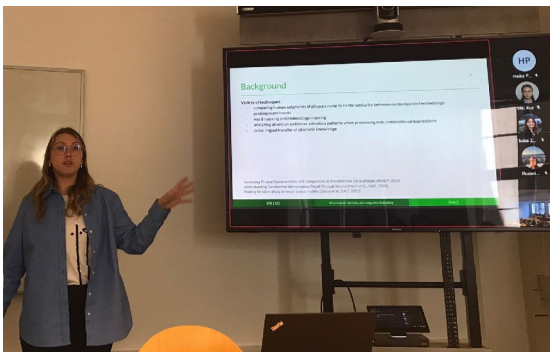
10 June 2024
Multilingual Modelling Workshop (MM-WS)

On 7 June 2024, B7 project had the pleasure of hosting the Multilingual Modelling Workshop. The event attracted researchers from five SFB projects, where the research goals are set in multilingual and/or cross-lingual contexts (see the programme). The workshop opened with an invited talk by **Dr Antonio Toral (University of Groningen)**, who shared his concerns about the machine-translation-exacerbated translationese as a factor in the target language change, especially for peripheral languages. Dr Toral presented the results of his recent study aimed at preserving the lexical richness of the source in automatic translation of literary texts. He called for a more differentiated approach to the trends in translator's behaviour in studies that explore ways to mitigate translationese.

The talks from the SFB projects were arranged in two sessions and allowed online participation.
The focus on the properties of translations was upheld in **Prof. Josef van Genabith**'s talk based on a RANLP-2023 publication extended to include a sneak preview of some follow-up studies from B6 project. This research tried to explain the amazing ability of BERT to pick up on the differences between translations and non-translations even at sentence-level by testing the hypothesis that BERT exploits the topical domain dissimilarities between the two text categories.
Also, in the morning session, Miaoran Zhang and Uliana talked about the properties and capabilities of various pre-trained multilingual and cross-lingual models. **Miaoran Zhang (B4)** presented a large-scale study of in-context learning, which was accepted to be presented at the ACL 2024. The results indicate that large language models seem to have outgrown in-context learning as a technology to adapt them to a new task.
Uliana Sentsova (B6) talked about the current results of her PhD research. They confirm that the multilingual XML-RoBERTa model performs well in the zero-shot cross-lingual transfer scenarios. In her work, the model was applied to the task of idiomaticity detection. The model was fine-tuned to detect English sentences with opaque idioms (as opposed to sentences without idiomatic components) and tested on the comparable sets of sentences in German and Italian.





The talks in the afternoon session used language models as a means to approach linguistic research questions from the perspective of Information Theory.
The project led by **Prof. Annemarie Verkerk (C7)** uses a multilingual corpus of out-of-English literary translations. Their study looks into the differences in the amount of context information associated with personal pronouns in the subject and object syntactic functions. The reported results defy straightforward universal interpretation across the languages involved. The talk, co-presented by Julius Steuer, provided interesting insights into calculating information estimates across a range of context sizes for individual tokens in targeted syntactic functions.
The **C4 talk, presented by Iulia Zaitova and Dr Irina Stenger**, explored the relationship between human subjects' performance in Slavic intercomprehension experiments and LLM-based metrics.

They find that Phonologically Weighted Levenshtein Distance (PWLD) predicts intelligibility scores a bit better than surprisal scores from ruGPT3Large or ruBertaLarge. The response time results are mixed across the tested Slavic languages.
Finally, a work-in-progress report by **Dr Maria Kunilovskaya (B7)** discussed the need for a cross-lingual approach to modelling translational data. In translation, the generation of the target is conditioned on the source, hence, the information-theoretical indices should take into account the source language input as the premise of the target. Inter alia, the report presented some reasoning for calculating the word-based memory and surprisal indices from an encoder-decoder machine translation model, using the weights of N context words from encoder and decoder attention layers to get the estimate of memory for sources and targets, respectively. It remains to be seen if this approach stands the empirical testing in the proposed evaluation setups.

The B7 team hopes that the event was a useful checkpoint that helped the participants to reflect on the progress in multilingual modelling achieved by the individual projects and to look around to see what else is going on in this area in the SFB. We are convinced that semi-formal topic-focused workshops like MM-WS are an excellent opportunity to showcase achievements, share finding and discuss issues for people working on related topics.

Maria Kunilovskaya
on behalf of B7 team