Workshop: Natural language and AI

Nicholas Catasso (Bergische Universität Wuppertal), Thomas Scharinger (Friedrich-Schiller-Universität Jena)

New perspectives for linguistic studies In recent years, Artificial Intelligence (AI) has made significant strides across various disciplines. The integration of AI applications, particularly Large Language Models (LLM), into linguistic studies has opened new horizons for the analysis of natural language. From morphosyntax to semantics and variation linguistics, these technologies provide linguists with the opportunity to explore complex linguistic phenomena. This development has led to the automation of linguistic tasks such as text generation, translation, and corpus annotation to an extent that was previously unimaginable. However, the application of AI in linguistic research also reveals challenges. One significant issue lies in the need to provide adequate training data for AI models, covering a wide range of linguistic phenomena and structures. Often, these data are incomplete, uneven, or even erroneous, which can compromise the reliability of AI systems. Another obstacle is the fact that AI models may inherit implicit biases from the existing data on which they are trained. This can result in distortions in the results and compromise the neutrality and objectivity of linguistic analyses. Furthermore, language variation seems to pose a challenge. AI models must be able to recognize and process this diversity appropriately. This is often difficult as the models may be constrained by certain linguistic patterns or norms. The planned workshop will be a forum to discuss how and to what extent AI applications (such as ChatGPT, DialoGPT, Meena, BlenderBot, etc.) may be relevant for linguistic studies. By merging theoretical approaches in linguistics with modern AI methods, the potentials and challenges of these technologies for linguistic research will be explored. The workshop will focus on – but will not be limited to – the following research questions:

• How can AI applications be used to investigate and compare grammaticality in different languages? Which applications come closest to native speaker intuition?
• To what extent can Large Language Models (LLM) be used for automatic corpus annotation and analysis to identify and understand linguistic variation?
• What role do semantic models and neural networks play in translating between languages with different syntactic structures?
• How can Natural Language Processing (NLP) techniques be used to examine the meaning and usage of language variation in different social contexts?

Samuel Innes
(Universität Heidelberg)

Quantifying the aspectualisation-contextualisation spectrum through multilingual language models

Verbal aspect is an important part of language systems. In general, "aspects are different ways of viewing the *internal temporal constituency of a situation*" (Comrie 1976, p.3, italics my own). While some languages (such as Slavic languages) have an overt aspectual encoding in the verb, others (like Vietnamese) rely solely on the context or inherent semantics of the verb phrase to deduce the aspectual class of an event. Other languages still, such as French, seem to fall between these two extreme examples, for example with more frequent use of periphrastic means. From a theoretical standpoint, it is therefore clear that there is a spectrum of grammaticalisation of aspect, from languages that have more explicit grammaticalised mechanisms to those that rely more on context. The empirical validation of a proposed ordering is, however, difficult with traditional methods in empirical linguistics. I show how the advent of newer language technologies, more specifically multilingual language models (MLMs), can be used to compare linguistic tendencies in written (and thus attested) data on a language level, removing the need for the less reliable method of native speaker questionnaires. More broadly, I aim to provide a proof of concept of how language models can be used for answering similar questions in linguistics.

There has been a lot of previous theoretical work on the topic of verbal aspect; indeed, it is famously one of the most studied areas in theoretical linguistics. However, due to the complex and often covert nature of the subject, more empirical studies are harder to come by. Östen Dahl's monograph *Tense and Aspect Systems* (1985) remedies this situation, providing a large-scale comparative study of TMA systems across languages based on questionnaires filled out by native speakers. This makes it possible to compare the usage and prevalence of certain TMA mechanisms in a variety of different languages with a unified terminology. However, since the construction of these questionnaires runs the inevitable risk of injecting researcher bias, it is hard to draw reliable conclusions from the true behaviour of aspect in practice. There has also been a fruitful line of work on aspect in psycholinguistics, which is complementary to the experiments I present here. However, here too, the results from the methodologies used, on the one hand, are still of varying validity, and on the other hand provide insights into responses to artificially crafted stimuli, rather than attested usage. Neither approach is sufficient to provide the desired quantification of verbal aspect grammaticalisation introduced above.

MLMs provide a useful and as of yet unused tool for empirical typology. They have been trained on data from multiple languages, originally with the hope of transferring knowledge from high-resource to low-resource languages. This gives them the convenient feature of having an aligned latent semantic space across languages, allowing for cross-linguistic comparison. While not perfect I find that these models offer valuable insights and a useful resource for further linguistic research on verbal aspect, both of individual languages and in a typological setting. More generally, I show how modern language models can be utilised to shed a more quantitative light on previously purely theoretical questions in linguistics.

References

Comrie, Bernard: Aspect. An Introduction to the Study of Verbal Aspect and Related Problems, Cambridge 1976.
Dahl, Östen: Tense and Aspect Systems. Oxford 1985.

Gergely Pethő[1,2] Esra' Abdelzaher[2] Ragini Menon[4], Gustavo Campedelli[5] & Sabine Tittel[4]
([1]University of Regensburg, [2]University of Debrecen, [3]Hungarian Research Institute for Linguistics, [4]Heidelberger Akademie der Wissenschaften, [5]Ruprecht-Karls-Universität Heidelberg)

Evaluating generative language models on historical texts: A comparative study of Arabic, Latin and two low-resource medieval languages

Our proposed talk investigates the machine translation (MT) performance of generative language models (LLMs) in historical languages, focusing on two high-resource and highly standardised languages: Arabic and Latin. This research builds on and extends our earlier work, which evaluated the capabilities of LLMs in translating Old French and Old Occitan, both of which are non-standardised and low-resource language variants, into English.
Our previous study revealed a significant performance gap between large commercial models like GPT-4o and free, smaller models like Llama 3 and Gemma 2 when translating low-resource historical languages. The commercial models demonstrated a reasonably good understanding of Old French and Old Occitan, presumably thanks to their similarity to modern Romance languages that they have been trained on, whereas the free models performed considerably worse. This paper aims to systematically compare the MT quality of free and commercial models when applied to high-resource languages, and evaluate whether the performance gap persists under these conditions.
We specifically examine Albucasis' *On Surgery and Instruments,* comparing the original classical Arabic text, as presented in the Spink & Lewis (1973) critical edition, with its Latin translation by Gerard of Cremona, as well as the Old French and Old Occitan vernacular translations that are believed to be based on the Gerard translation. The Arabic original is considered one of the most influential medical texts of the Middle Ages, and its Latin translation by Gerard was pivotal in disseminating Arabic medical knowledge in medieval Europe. For this study, we used the aforementioned edition of the Arabic text, the 1532 Strasbourg printed edition of Gerard's Latin translation, and the critical editions of the Old French (Trotter 2005) and Old Occitan (Elsheikh 1992) version as the basis for MT into English.
We apply a variety of free LLMs ranging in size from 7 billion to 70 billion parameters (e.g. Phi-3, Mistral, Mixtral, Gemma 2, Llama 3.1), as well as three commercial models (GPT-4o, Gemini 1.5, Claude Sonnet 3.5) to translate these different language versions of the same historical text into English, comparing the MT to the reference translation by Spink & Lewis, and evaluate the quality of these MTs using standard word- and n-gram-based translation quality metrics such as BLEU and ROUGE, as well as neural metrics such as BERTScore. By comparing MT quality of the models on Arabic and Latin to their performance on Old French and Old Occitan, we aim to assess whether the challenges faced by free models are unique to non-standardized, low-resource languages or if they also apply to more resource-rich languages. We regard MT quality not as our primary interest, but rather as a relatively easy to measure proxy for the capability of different models to “understand” a language.
Therefore we hope that our results will help guide future research in developing more effective tools for historical language processing and will be of interest to scholars in the fields of computational linguistics, digital humanities, and medieval studies.

References

Elsheikh, M.S. (ed.): *Abu’l Qasim Halaf Ibn Abbas az-Zahrawı, La Chirurgia. Versione occitanica della prima met`a del Trecento.* Firenze (1992)
Spink, M.S., Lewis, G. (eds.): *Albucasis on Surgery and Instruments: A Definitive Edition of the Arabic Text with English Translation and Commentary.* Wellcome Institute of the History of Medicine, London (1973).
Trotter, D. (ed.): *Albucasis, Traitier de Cyrurgie: Edition de la traduction en ancien français de la Chirurgie d’Abu’l Qasim Halaf Ibn Abbas al-Zahrawi du manuscrit BNF, français 1318.* Niemeyer (2005).

Nicola Brocca[1], Joseph Wang[1], Tamara Walder[1], Corrado Schininà[1], Lukas Zehetgruber[1], Nadine Mair[1], Elena Nuzzo[2] & Diego Cortes[2]
([1]Universität Innsbruck, [2]Università Roma Tre)

AI meets pragmatic annotation. LadderWeb: accuracy and applications

Producing pragmatically annotated corpora offers significant benefits for researching language variations and contrastive pragmatics (Weisser, 2018, p. 2). However, this process is often time-consuming and requires trained personnel. The LadderWeb project, funded by Clariah.at, aims to semi-automate this process. Users input text and receive annotated versions. The model's training process will be outlined through the following key stages:

1. **Designing an Annotation Scheme**: The annotation scheme is derived from taxonomies following speech-act theory, previously used in intercultural pragmatic studies (Brocca et al., 2023; Cortés Velásquez & Nuzzo, 2022; Nuzzo & Cortés Velásquez, 2020).
2. **Executing Manual Annotation**: Data for manual annotation come from (partially) annotated corpora DidDir (Cortés Velásquez & Nuzzo, 2024) and Ladder (Brocca, 2024), focusing on cancellations and requests in Italian L1, Italian L2, and German.
3. **Creating an AI-Based Routine**: Utilizing decision trees and statistical methods, the routine categorizes words based on annotation and features. The application integrates the Corpus Query Processor (CQP) of the IMS Open Corpus Workbench, offering a user-friendly interface for accessing annotated texts. Meta-information enables the creation of subcorpora based on speech act type, social context, medium, languages, participant data, etc.
4. **Training the Routine**: The routine is trained with the aforementioned manually annotated data, consisting of approximately 800 cancellations and 200 requests.
5. **Manually Correcting the Results**: The model is able to roughly tag new cancellations and requests autonomously, but the results still need fine-tuning by hand.

This paper assesses the accuracy of the annotation before and after the manual correction of the results, aiming to determine how reliable automatic annotation of pragmatic categories can be. We define the accuracy of the annotation as: Accuracy=Number of Correct Predictions/Total Number of Predictions. The presentation will also showcase some examples how LadderWeb can be used to conduct research in corpus-based (contrastive) pragmatics. While LadderWeb cannot yet claim definitive results due to its relatively small sample size, it provides a foundation for future projects with larger datasets. The online application, supports scholars in producing annotated material and examining speech acts across various languages and sociolinguistic variables. LadderWeb is programmed to be interoperable and support different types of data and annotation procedures.

References

Brocca, N. (2024). *Ladder: a corpus for pragmatic competences in Italian L1/L2* [Data set]. https://hdl.handle.net/21.11115/0000-0011-BF1F-5

Brocca, Nicola, Elena Nuzzo, Diego Cortés Velásquez, & Maria Rudigier (2023). Linguistic politeness across Austria and Italy: Backing out of an invitation with an instant message, *Journal of Pragmatics,* 209, pp. 56-70, https://doi.org/10.1016/j.pragma.2023.02.018.

Cortés Velásquez, Diego, & Nuzzo, Elena (2022). Declining an invitation: The pragmatics of Italian and Colombian Spanish. In S. Gesuato, G. Salvato, & E. Castello, (Eds.), *Pragmatic Aspects of L2 Communication. From Awareness through Description to Assessment*, pp. 143–163. Cambridge Scholars.

Cortés Velásquez, D., & Nuzzo, E. (2024). DisDir: Cancellations and Other Refusal Strategies [Data set]. https://hdl.handle.net/21.11115/0000-0011-BF16-E

Nuzzo, Elena, & Cortés Velásquez, Diego (2020). Canceling Last Minute in Italian and Colombian Spanish: A Cross-Cultural Account of Pragmalinguistic Strategies. *Corpus Pragmatics, 4*, pp. 1–26. https://doi.org/10.1007/s41701-020-00084-y

Weißer, Martin (2018). *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech acts and beyond*: John Benjamins.