

Sektion 13

The quantitative turn: NLP and AI Methods in Romance Linguistics

Iris Ferrazzo (Rheinische Friedrich-Wilhelms-Universität Bonn) & Olga Kellert (Arizona State University)

Abstracts

María del Carmen Balbuena Torezano (Universidad de Córdoba) & Alba Montes Sánchez (Universidad de Córdoba)

Desterminologización y alfabetización en salud: la efectividad de Llama3 para producir textos para el paciente de cardiología / De-terminologisation and health literacy: the effectiveness of Llama3 in producing texts for the cardiology patient

This paper will address the effectiveness and reliability of a GPT model, such as Llama3, for the simplification of documents that the cardiology patient may receive, and thus facilitate their understanding. Using examples and original documents, we will try to determine through Llama3 (1) the use of specialised meaning units (USE); (2) the identification of comprehension problems for the patient, as a non-specialist receiver of the medical text; (3) the treatment that Llama3 makes to generate a simplified text, understandable by a non-specialist reader in cardiology; and (4) the possibilities that Llama3 offers to achieve a better health literacy in the field of cardiology.

For this purpose, we will use clinical cases, hospital discharge documents, or scientific articles related to different heart diseases, whose lexical density is particularly high, to analyse the de-terminologisation processes carried out after processing the language through the pre-trained Llama3 model. Because of the results obtained, we will try to draw conclusions regarding, firstly, the effectiveness of the model for the production of texts understandable by patients and non-cardiology specialists; secondly, how the use of Llama3 for the de-terminologisation and simplification of texts in the field of cardiology would improve doctor-patient communication.

Keywords: Llama3; de-terminologisation; medical language; machine translation; deep learning

References:

- Batthey, R.; Gupta, S. (2024). Training Llama: A Storage Perspective. [online]. URL: <https://atscaleconference.com/videos/training-llama-a-storage-perspective/>.
- Benchechrout et. al. (2023). WorldSense: a synthetic benchmark for grounded reasoning in large language models. CoRR, abs/2311.15930. DOI: 10.48550.
- Bizzoni et al. (2020). How human is machine translationese? Comparing human and machine translations of text and speech. En: Federico, M. et al. (eds.), *Proceedings of the 17th International Conference on Spoken Language Translation*. ACL, pp. 280-290. DOI: 10.18653/v1/2020.
- Campos, O. (2013). Procedimientos de desterminologización: traducción y redacción de guías para pacientes. *Panace@ XIV* (37), 48-52.
- Cobos, I. (2019). Traducir para el paciente: acercamiento y adaptación como modalidad de traducción. *Quaderns de Filologia. Estudis Lingüístics XXIV*, 211-228. <https://doi.org/10.7203/QF.24.16307>.
- Colantonio, C. (2023). Los géneros discursivos y los textos en el español para la salud. *AGON* 36, 5-39.
- García-Izquierdo, I., Muñoz-Miquel, A. (2015). Los folletos de información oncológica en contextos hospitalarios: la perspectiva de pacientes y profesionales sanitarios. *Panace@. Revista de Medicina, Lenguaje y Traducción* 16 (42), 225-231.
- Montalt-Resurrecció, V., García-Izquierdo, I. y Muñoz-Miquel, A. (2025). *Patient-Centred Translation and Communication*. Routledge.

Weber, L.; Bruni, E.; Hupkes, D. (2023). Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. En: Jiang, J.; Reitter, D; Deng, S. (eds.), *Proceedings of the 27th Conference on Computer Natural Language Learning (CoNLL)*, pp. 294-313. DOI: <https://doi.org/10.18653/v1/2023.conll-1.20>.

Zhou, J. et. al. (2023). Instruction-following evaluation for large language models. En: *arXiv: 2311/07911*. [online]. DOI: <https://doi.org/10.48550/arXiv.2311.07911>.

Sonja Böker (Universität Trier)

Song Lyrics in Linguistics with a Focus on Adaptations

Literal translations aim for a maximum of semantic proximity. However, when song lyrics are translated for singing, the rhythm with its fixed number of syllables places constraints on translations (Low 2017). This leads to creative deviations from the source texts in semantics, morpho-syntax and in the use of stylistic devices. The resulting target texts are therefore not called translations but song adaptations (Franzon 2021).

Previous research has examined features of translated fiction (Hansen 2003) and non-fiction (Baker 2004). They demonstrate similar “translation universals” when compared to their source texts, such as lengthening (Malmkjær 2011). The length of song adaptations, by contrast, is musically predetermined. Since features of song adaptations are unexplored, this paper will assess to which extent other translation universals like less lexical density and demetaphorization apply to song texts.

Previous research on song adaptations has been evaluative or prescriptive and mainly focussed on opera (Apter & Herman 2016). Conversely, this paper uses a descriptive approach and a larger corpus by choosing a genre with many digitally available song adaptations. The global practice of singing in church has led to hundreds of interlingual adaptations. Conveniently, the website shir.fr provides digital access to over 1000 French song lyrics including adaptations from English. Among these, a rather recent pop genre called “Praise & Worship” arose in French around 1975. Theological and hymnological scholars on Praise & Worship texts have analyzed small corpora of up to 77 English song lyrics (Woods & Walrath 2007) and nearly 250 German songs (Scheuermann 2023) but none in French.

Therefore, this paper will compare the Praise & Worship lyrics of 725 original French compositions with 333 French adaptations of English, and the corresponding 333 English source texts from songselect.ccli.com. Preliminary studies with NLP tools show significant differences between these three subcorpora. Frequency lists based on part-of-speech tags indicate for example that French adaptations overuse interjections compared to original French and English compositions. Moreover, counts of stylistic devices reveal more metaphors and repetitions in English lyrics than their French adaptations, whereas the latter exhibit more rhymes.

In summary, the quantitative linguistic approach of this paper offers a descriptive perspective on the song genre of Praise & Worship by comparing French compositions with French adaptations and their English source texts.

References:

- Apter, Ronnie & Mark Herman. 2016. *Translating for Singing: The theory, art and craft of translating lyrics*. London: Bloomsbury.
- Baker, Mona. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2). 167–193. <https://doi.org/10.1075/ijcl.9.2.02bak>.
- Franzon, Johan. 2021. The liberal mores of pop song translation: Slicing the source text relation six ways. In Johan Franzon, Annjo K. Greenall, Sigmund Kvam & Anastasia Parianou (eds.), *Song Translation. Lyrics in Contexts*, 83–122. Berlin: Frank & Timme.
- Hansen, Silvia. 2003. *The nature of translated text: an interdisciplinary methodology for the investigation of the specific properties of translations*. Saarbrücken: Dt. Forschungszentrum für Künstl. Intelligenz.
- Low, Peter. 2017. *Translating Song. Lyrics and texts*. London: Routledge.
- Malmkjær, Kirsten. 2011. Translation Universals. In *The Oxford Handbook of Translation Studies*, 83–93. Oxford, New York: Oxford University Press.
- Scheuermann, Andreas. 2023. *Praise and Worship: zur Bedeutung populärer Lobpreismusik für den Gottesdienst: eine praktisch-theologische Untersuchung*. Giessen: Brunnen.

Woods, Robert & Brian Walrath (eds.). 2007. *The Message in the Music: Studying Contemporary Praise & Worship*. Nashville: Abingdon Press.

Johnatan E. Bonilla (Humboldt-Universität zu Berlin)

Mapping Register Variation in Canarian Spanish Using NLP and Emerging Language Technologies

This presentation aims to (1) examine how morphosyntactic variation—specifically, the pluralization of the Spanish existential *haber* (*hubieron problemas* vs. standard *hubo problemas*)—interacts with register in Canarian Spanish; (2) demonstrate the utility of large-scale, multi-register corpora combined with NLP methodologies to analyze linguistic variation; and (3) highlight the value of researcher-free, technology-driven data elicitation tools for understanding sociolinguistic variation in a transitional Romance variety.

Canarian Spanish occupies a unique space between European and Latin American varieties (Almeida, 2014; Samper Padilla, 2008), creating a complex environment of *diaglossia* where intermediate and standard forms coexist, influenced by urban diffusion centers promoting linguistic innovation (Peña Rueda, 2024). The pluralization of existential *haber*, once stigmatized, reveals that while standard usage dominates formal contexts, non-standard variants appear in informal or spontaneous settings, indicating subtle register shifts.

Four corpora were compiled to analyze these patterns: (1) PARCAN (~41K tokens) from parliamentary proceedings; (2) JABLE (~251K articles), a digital press archive; (3) COPACAYO (~16.9M tokens) from Canarian parliamentary YouTube sessions mixing formality and spontaneity; and (4) COMECAYO (~228M tokens) from spoken media on YouTube. To overcome limitations of traditional methods (e.g., interviewer bias, Labov 1972; Guerrero & Ramada, 2019), *HablaCanariaBot*—a Telegram chatbot—elicits morphosyntactic data *in situ*, capturing authentic registers. Data analysis employs NLP methods, including morphological tagging, syntactic parsing, and transformer-based models (Vaswani et al., 2017; Devlin et al., 2019), facilitating a systematic, quantitative assessment of language variation (Kortmann, 2021).

In sum, by integrating multiple corpora, employing researcher-free elicitation, and applying advanced NLP, this study illuminates ongoing changes in Canarian Spanish and offers a scalable framework for exploring linguistic variation, register, and evolving norms in Romance languages.

References:

- Almeida, M. (2014). El concepto de 'hablas de tránsito' y el español canario. *Revista de Filología Románica*, 31(1), 37–47.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
- Guerrero, A. C., & Ramada, G. U. (2019). Diseño y construcción de un corpus oral multidialectal. *El corpus amaresco. Normas: revista de estudios lingüísticos hispánicos*, 9(1), 17–36.
- Kortmann, B. (2021). Reflecting on the quantitative turn in linguistics. *Linguistics*, 59(5), 1207–1226.
- Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, 1(1), 97–120.
- Peña Rueda, C. (2024). Desorientación normativa y variación gramatical en el español de Canarias. *ENERGEIA. Online Journal for Linguistics, Language Philosophy and History of Linguistics*, 57–90.
- Samper Padilla, J. A. (2008). Sociolinguistic aspects of Spanish in the Canary Islands. *International Journal of the Sociology of Language*, 193–194, 161–176.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* pp. 6000–6010).

Iris Ferrazzo (Rheinische Friedrich-Wilhelms-Universität Bonn)

Are Large Language Models the future crowd workers of Linguistics?

Data elicitation from human participants is one of the core data collection strategies used in empirical linguistic research. The amount of participants in such studies may vary considerably, ranging from a handful to crowdsourcing dimensions. Even if they provide resourceful extensive data, both of these settings come alongside many disadvantages, such as low control of participants' attention during task completion, precarious working conditions in crowdsourcing environments (Van Zoonen, Sivunen, and Treem 2024), and time-consuming experimental designs. Furthermore, with the widespread of computational models, especially Large Language Models (LLMs), that can perform human-like tasks and even outperform the accuracy of crowd workers (Kocoń et al. 2023; Gilardi, Alizadeh, and Kubli 2023), illicit evidence of their usage in crowdsourcing studies is currently under observation (Veselovsky, Ribeiro, and West 2023).

For these reasons, this research aims to answer the question of whether Large Language Models (LLMs) may overcome those obstacles if included directly in empirical linguistic pipelines. Two reproduction case studies are conducted to gain clarity into this matter: Cruz (2023) and Lombard, Huyghe, and Gygax (2021). The choice of the studies aims to overcome the English-centric training and usage of LLMs by studying their performance on Romance languages. The two forced elicitation tasks, originally designed for human participants, are replicated in the proposed framework with the help of OpenAI's GPT-4o-mini model. Its performance with our zero-shot prompting baseline shows the effectiveness and high versatility of LLMs, which tend to outperform human informants in linguistic tasks. Even if GPT-4o-mini outperforms human informants in all experimental conditions tested, the findings of the second replication further highlight the need to explore additional prompting techniques, such as Chain-of-Thought (CoT) prompting (Wei et al. 2022), which, in a second follow-up experiment, demonstrates higher alignment to human performance on both critical and filler items.

References:

- Cruz, Abel. 2023. Linguistic factors modulating gender assignment in spanish–english bilingual speech. *Bilingualism: Language and Cognition* 26, no. 3 (May): 580–591. ISSN: 1366-7289, 1469-1841, accessed December 9, 2024. <https://doi.org/10.1017/S1366728922000839>.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. Publisher: arXivVersion Number: 2, accessed January 30, 2025. <https://doi.org/10.48550/ARXIV.2303.15056>.
- Lombard, Alizée, Richard Huyghe, and Pascal Gygax. 2021. Neological intuition in french: a study of formal novelty and lexical regularity as predictors. *Lingua* 254 (April): 103055. ISSN: 00243841, accessed December 9, 2024. <https://doi.org/10.1016/j.lingua.2021.103055>.
- Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, et al. 2023. ChatGPT: jack of all trades, master of none. *Information Fusion* 99 (November): 101861. ISSN: 15662535, accessed January 29, 2025. <https://doi.org/10.1016/j.inffus.2023.101861>.
- Van Zoonen, Ward, Anu E. Sivunen, and Jeffrey W. Treem. 2024. Algorithmic management of crowdworkers: implications for workers' identity, belonging, and meaningfulness of work. *Computers in Human Behavior* 152 (March): 108089. ISSN: 07475632, accessed January 28, 2025. <https://doi.org/10.1016/j.chb.2023.108089>.
- Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: crowd workers widely use large language models for text production tasks. Version Number: 1. Accessed December 9, 2024. <https://doi.org/10.48550/ARXIV.2306.07899>.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. Version Number: 6. Accessed January 29, 2025. <https://doi.org/10.48550/ARXIV.2201.11903>.

Annette Gerstenberg, Christian Löser, Marta Lupica Spagnolo & Friederike Schulz (Universität Potsdam)

Mutual benefit in transcription workflows: Improving manual transcription with AI, and fine-tuning a model for spoken French with manual transcriptions

Transcribing spoken language is widely recognized as both time-consuming and frustrating, as each cycle of listening-transcribing carries the risk of unearthing alternative readings. In contrast, transcriptions generated with Large Language Models (LLMs) offer apparently perfect transcriptions – sacrificing each and every feature unique to spoken language. Additionally, when it comes to less represented languages and varieties, which is the case with language data from older speakers, LLMs have been found to perform poorly (Hekkel, Schulz, and Lupica Spagnolo 2024).

In our presentation, we propose a way to draw upon the mutual benefit of both approaches, based on LangAge corpora, a longitudinal collection (2005–2023) consisting of oral biographical interviews with French speakers, mostly over the age of 70 living in and around the city of Orléans and accessible upon registration.

Its manually created transcriptions (corpus LangAgeMt) consist of 948 000 tokens, the result of a long, complex manual process of reiterated stages of transcription, correction, segmentation, and anonymization (El Sherbiny Ismail et al. 2022). These transcriptions conserve features of spoken language such as false starts, repetitions, hesitation markers, and so on (Gerstenberg, Hekkel, and Kairet 2018).

With the support of Hasso Plattner Institute (University of Potsdam), using open-source neural net Whisper with the models Large and Large V2 (Radford et al. 2022),¹ a complete transcription of the entire corpus recordings (approx. 150 audio files) was undertaken in 2024 (corpus LangAgeWh).

We begin the presentation with a general description of LangAgeWh with regard to spoken language features. We then present a script-based workflow aimed to map LangAgeMt and LangAgeWh to the fullest extent possible. A collation script automatically compares the resulting versions of the transcripts. Based on these results, we discuss the benefits of confronting manual vs. ai-generated transcripts, and how the new readings improve the existing transcriptions.

In the final section, we explore ways to train the transcription algorithm, based on the “gold standard” of the improved transcripts, to conserve the core features of spoken language. In such a way, human intelligence can be supported by LLMs, in a linguistically valid and efficient way.

References:

- El Sherbiny Ismail, Eman, Annette Gerstenberg, Marta Lupica Spagnolo, Friederike Schulz, and Anne Vandenbroucke. 2022. “L’âge avancé en perspective longitudinale et ses outils: LangAge, un corpus au pluriel.” *SHS Web Conf. (SHS Web of Conferences) - 8e Congrès Mondial de Linguistique Française* 138:10003 [1–14]. <https://doi.org/10.1051/shsconf/202213810003>.
- Gerstenberg, Annette, Valerie Hekkel, and Julie Kairet, eds. 2018. *Corpus LangAge: Transcription Guide*. University of Potsdam: Department of Romance Studies. <https://doi.org/10.5281/zenodo.6444538>.
- Hekkel, Valerie, Friederike Schulz, and Marta Lupica Spagnolo. 2024. “‘Nous fêterons’ or ‘On va fêter’? Mimicking Age-Sensitive Variation with ChatGPT.” *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1 (1): 1–36. <https://doi.org/10.62408/ailing.v1i1.11>.

¹ With a Razor Blade 15 with RTX 3080. The project on which this report is based was funded by the Federal Ministry of Education and Research under the funding code KI-Servicezentrum Berlin-Brandenburg” 01IS22092. Responsibility for the content of this publication remains with the author.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision [eess.AS]." *arXiv*, <https://doi.org/https://doi.org/10.48550/arXiv.2212.04356>.

Yoshifumi Kawasaki (川崎義史 / University of Tokio)

Diachronic Studies of Romance Languages in the Era of Deep Learning

Recent advancements in Deep Learning and Large Language Models have renewed interest in linguistic investigations, including diachronic language change. The relevant techniques involve representing linguistic units (character, word, sentence, etc.) as fixed-length dense vectors (embeddings), thereby allowing for measuring quantitatively semantic and functional similarity between the units. This property paves the way for novel computational approaches that would otherwise be impossible. This talk will briefly overview embeddings-based methods and present my research addressing diachronic semantic and morphological changes from Latin to Romance languages. I will also explore potential applications of the related techniques to philological studies and linguistic investigation in general.

Anna Ladilova (Justus-Liebig-Universität Gießen) & Katharina Müller (Justus-Liebig-Universität Gießen) & Erdal Ayan (Leibniz-Institut für Informationsinfrastruktur Karlsruhe)

Language change in times of the COVID-19 pandemic: A corpus analysis of the emergence of neologisms

The COVID-19 pandemic left in its wake a veritable linguistic laboratory reflecting the profound social, political, and economic effects of language change: On the one hand, this new lexicon is the clearest linguistic output of the pandemic and on the other hand a testimony to the exceptional nature of the pandemic, which is also underlined by the fact that many of the new terms have already been included in major dictionaries. The dynamic emergence of neologisms has already been explored in Romance languages, but mostly from a qualitative point of view (Müller/Ladilova/Kiegel-Keicher/Born 2024; Pietrini 2021). The existing quantitative studies focus mainly on (popular) discourse about COVID-19 in social media or have a lexicographic aim (Klosa-Kückelhaus/Kernermann 2022). However, there have been few studies that take a cross-cultural perspective by comparing language change during the pandemic in several languages so far (Ladilova/Müller/Gomes/Born 2024; Cartier et al. 2022) and moreover take a quantitative corpus linguistic approach with machine learning methods.

Therefore, in our project we will develop a Media Corpora platform to perform a corpus-based discourse analysis of language change in times of the COVID-19 pandemic in four Romance languages (French, Italian, Portuguese, Spanish). In addition to classical corpus linguistic methods, we will also apply two NLP and machine learning methods as semantic predictors: dynamic topic modelling (DTM) (Egger/Yu 2022; Grootendorst 2023) and sentiment analysis (SA) (Taboada, 2016). By combining the sentiment polarization of SA with the probabilistic topic models of DTM, we will show which topics come to the fore in different languages and countries during the pandemic, what kind of reactions they receive from society and which new linguistic patterns these topics and reactions are associated with. This will allow us to contribute to the field comparing how different languages are exposed to evolution in the given context with sentimental interventions.

References:

- Cartier, Emmanuel et al. 2022. Linguistic repercussions of COVID-19: A corpus study on four languages. *Open Linguistics* 8(1), 751–766. <https://doi.org/10.1515/opli-2022-0222>.
- Egger, Roman; Yu, Joanne. 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7: 886498. <https://doi.org/10.3389/fsoc.2022.886498>.
- Grootendorst, Maarten P. 2023. Dynamic Topic Modeling—BERTopic. *Dynamic Topic Modeling*. https://maartengr.github.io/BERTopic/getting_started/topicovertime/topicovertime.html
- Klosa-Kückelhaus, Annette; Kernermann, Ilan (eds.). 2022. *Lexicography of coronavirus-related neologisms*. Berlin/Boston: De Gruyter. <https://doi.org/10.1515/9783110798081>.
- Ladilova, Anna; Müller, Katharina; Gomes, Simone; Born, Joachim. 2024. Linguistic change in times of the COVID-19 pandemic: a corpus linguistic comparison of language contact phenomena in Romance languages. *Zeitschrift für romanische Philologie* 140(1), 1-29. DOI: <https://doi.org/10.1515/zrp-2024-0001>.
- Müller, Katharina/Ladilova, Anna/Kiegel-Keicher, Yvonne/Born, Joachim. 2024. Sprache, Gesellschaft und Covid in romanischsprachigen Ländern. *Quo Vadis, Romania?* 63 (Themenheft: Sprache, Gesellschaft und COVID-19 in romanischsprachigen Ländern), 5-21. <https://quovadisromania.univie.ac.at/wp-content/uploads/2024/07/2QVR-63-Einleitung-5-21.pdf>.
- Pietrini, Daniela. 2021. *La lingua infetta: L'italiano della pandemia*. Roma: Treccani.
- Taboada, Maite. 2016. Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, 2(1), 325–347. <https://doi.org/10.1146/annurev-linguistics-011415-040518>.

Elen Le Foll (Universität zu Köln)

Reproducible semantic annotation using open LLMs

Large, multifactorial corpus studies in Romance languages typically rely on automatic NLP tools to label parts-of-speech and syntactic dependencies. In contrast, semantic annotation remains the task of either human annotators or rule-based algorithms (e.g., *pymusas*; Piao et al., 2016; UCREL, 2022). The former is time-consuming, expensive and requires trained annotators to achieve acceptable inter-rater agreement, while the latter approach is inherently limited by the finite size of the lexicons on which these tools rely.

Recent studies suggest that Large Language Models (LLMs) may be suited to such semantic annotation tasks (see, e.g., Fonteyn et al., 2024; Gilardi et al., 2023; Koeva, 2024; Kuzman et al., 2023; Nasution & Onan, 2024). However, to date, most research workflows involving instruction-tuned text generators have involved proprietary LLMs – most commonly, ChatGPT–, which means that their outputs cannot be reproduced (see, e.g., Kapoor & Narayanan, 2023; Liesenfeld et al., 2023).

In a preliminary study, a reproducible workflow for semantic annotation was created for annotating of French tweets for topics (Hollmann & Le Foll, 2024). Although the results were promising, several challenges were identified. In the present study, we investigate whether a reproducible workflow using a local, open-source LLM can be used to automate the annotation of semantic features typically used in the study of morphosyntactic alternations (e.g., animacy, definiteness, semantic gender) in different varieties and registers of French and Spanish. These varieties range from Old Spanish to French social media texts, none of which have not been orthographically standardised. The results are evaluated against the manual annotation of trained linguists.

References:

- Fonteyn, L., Manjavacas, E., Haket, N., Dorst, A. G., & Kruijt, E. (2024). Could this be next for corpus linguistics? Methods of semi-automatic data annotation with contextualized word embeddings. *Linguistics Vanguard*. <https://doi.org/10.1515/lingvan-2022-0142>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Hollmann, J., & Le Foll, E. (2024). *An LLM-based Approach to Categorising French Tweets for Linguistic Analyses* [Poster]. Sustainable archiving of social media data - Twitter and beyond, German National Library (Frankfurt am Main).
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9). <https://doi.org/10.1016/j.patter.2023.100804>
- Koeva, S. (2024). Large Language Models in Linguistic Research: The Pilot and the Copilot. *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria*, 319–329. <https://doi.org/10.47810/CLIB.24.35>
- Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification* (No. arXiv:2303.03953). arXiv. <https://doi.org/10.48550/arXiv.2303.03953>
- Liesenfeld, A., Lopez, A., & Dingemans, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6. <https://doi.org/10.1145/3571884.3604316>
- Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, 12, 71876–71900. IEEE Access. <https://doi.org/10.1109/ACCESS.2024.3402809>
- Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, M. A., Shafi, J., Teh, P. L., & Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, 2614–2619.

UCREL. (2022). *pymusas: PYthon Multilingual Ucrel Semantic Analysis System* (Version 0.3.0) [Python].
<https://ucrel.github.io/pymusas/>

Alessandro Lenci (Università di Pisa)

Towards Bio-inspired Language Models: Building A BabyLM for Italian

The cognitive implausibility of Large Language Models (LMs) is a recurring topic in the Natural Language Processing (NLP) literature. From a purely quantitative perspective, LMs are trained on corpora that contain several orders of magnitude more words than the linguistic input to which a human being is exposed. From a qualitative perspective, the training data of Large LMs are primarily composed of web-derived written texts. This paper presents BAMBI (BAby language Models Boostrapped for Italian), a series of Baby Language Models (BabyLMs) trained on data that mimic the linguistic input received by a five-years-old Italian-speaking child. The BAMBI models are tested on a benchmark specifically designed to evaluate LMs, which takes into account the amount of training input the models received.

Liviu P. Dinu (Universitatea din București)

RoBoCoP: a Comprehensive Romance Database for Computational Historical Linguistics

The identification of cognates is a fundamental process in historical linguistics, on which any further research is based (Campbell 1972). Even though there are several cognate databases for Romance languages, they are incomplete, noisy, built with automated translation methods, or are of uncertain availability.

To overcome as much as possible these weaknesses, we have decided to build from scratch a fully available database of Romance cognates, for the main five Romance languages (It, Es, Fr, Pt, Ro), starting with the available machine-readable reference dictionaries², which contain etymological information. The process was semi-automated, guided and verified by human experts, to ensure the quality and coverage of the data.

For each of the five Romance languages, the database contains lists of words, with their etymologies. The database comprises a total of 125,598 words across all languages and 90,853 cognate pairs. Our framework includes: a linguistically informed and computationally usable definition of cognate words, a methodology for extracting cognate pairs automatically in a robust way, a comprehensive dataset of word etymologies for Romance languages based on etymological information given by dictionaries, and a comprehensive database of cognate pairs, as well as benchmark results for automatic cognate detection, based on a series of machine learning experiments (using a variety of features and models: graphical and phonetical features, including prior feature engineering to obtain word alignment information, or alignment-agnostic, and several types of model architectures) for automatically detecting cognates. The best results are obtained using the ensemble models with alignment features across all experimental settings, while the transformer based model generally comes second. We show that the combination of both graphic and phonetic features in the ensemble model surpasses the ensembles that were limited to only one kind of features.

We extract pairs of cognates between any two Romance languages by parsing electronic dictionaries of Romanian, Italian, Spanish, Portuguese and French (Table 1).

	Ro	It	Es	Pt	Fr
Ro	-	4,999 6,683	7,588 9,056	5,855 8,211	7,360 8,120
It	3,139	-	7,863 8,627	12,198 13,343	7,105 7,361
Es	209	770	-	9533 10,731	10,220 10,543
Pt	103	620	1,201	-	7,783 8,179
Fr	33,311	2,896	1,690	2,450	-

Table 1: Number of cognate pairs (above the main diagonal) and borrowing pairs (below the main diagonal) for each Romance language pairs. For cognate pairs we report total number (first number in each cell) and pairs of Latin etymology only (the second line).

Based on this resource, we propose a strong benchmark for the automatic detection of cognates, by applying machine learning and deep learning based methods on any two pairs of Romance languages. We find that automatic identification of cognates is possible with accuracy averaging around 94% for the more difficult task formulations.

² ldizionario.internazionale.it, rae.es/drae, www.infopedia.pt/lingua-portuguesa, www.cnrtl.fr, dexonline.ro.

Our main contributions are:

1. We introduce a comprehensive database of Romance cognate and loanword pairs (pairs of cognates/ loanwords between any two Romance languages).
2. We propose a strong benchmark for the automatic detection of cognates, by applying a set of machine learning models (using various feature sets and architectures) on any two pairs of Romance languages.

Keywords: cognates, borrowings, Romance languages

References:

Campbell, L 1998. *Historical Linguistics. An Introduction*. MIT Press.

Ciobanu, A & L Dinu: Automatic Identification and Production of Related Words for Historical Linguistics. *Comput. Linguistics* 45(4): 667-704 (2019)

Dinu, L, A Uban, A Ciobanu, A Dinu, B Iordache, S Georgescu, L Zoicas. 2023. RoBoCoP: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *Proc EMNLP 2023*, Singapore, 2023, 7610-7629.

Nemika Tyagi (Arizona State University) & Olga Kellert (Arizona State University)

Structural Analysis of Code-Switching using UD-Parsers and ChatGPT

This study examines code-switching (CS) behaviors among bilingual Spanish-English users in Miami, Florida, addressing the opportunities and challenges of applying Natural Language Processing (NLP) and Artificial Intelligence (AI) to structural analysis of bilingual communication. CS, defined as the alternating use of two or more languages within a single conversation, as in (1) and (2) (Poplack 1980; Myers-Scotton 2002; Lipski 2014), represents a dynamic, spontaneous phenomenon that requires complex computational modeling to analyze effectively.

Using the Bangor Miami dataset (<https://biling.talkbank.org/access/Bangor/Miami.html>), this study analyses the structural relation of code-switch points of English and Spanish bilingual communication. The linguistic patterns are analyzed to distinguish between true CS and other phenomena such as lexical borrowing (e.g., el *doctoral* in (3)) or discourse constructions (e.g., *so* in (4)). These analyses demonstrate the challenges posed by unstructured data, including the need for advanced parsing tools capable of identifying subtle differences in bilingual language use. The present study discusses various challenges in using UD- parsers (Kellert et al. 2023) and ChatGPT on authentic communication.

- (1) FLA: creo que la hija de Laura **is gonna get baptised on Sunday**.
'I think that Laura's daughter **is gonna get baptised on Sunday**.'
- (2) pero en todas las escuelas si tú quieres ir de **school counsellor you have to have a masters in psychology en en behavioural education or one of these things like that**
'but in all the schools if you want to become a **school counsellor you have to have a master's in psychology, in behavioural education or one of these things like that**.'
(Deuchar et al. 2014, <http://bangortalk.bangor.ac.uk/sastre11.mp3>, English in bold)
- (3) DIE: y sacar el **doctoral**. 'And get the doctoral'
- (4) DIE: **so** es todo tecnología. '**So**, it's all about technology'
(Deuchar et al. 2014, <http://bangortalk.bangor.ac.uk/sastre11.mp3>)

References:

- Kellert, O., M. Zaman, N. H. Matlis, C. Gomez-Rodriguez (2023). Experimenting with UD Adaptation of an Unsupervised Rule-based Approach for Sentiment Analysis of Mexican Tourist Texts. CEUR Workshop Proceedings, Vol. 3496, Rest-Mex paper 15, Alvarez-Carmona et al. 2023 (Eds.).
<http://www.grupolys.org/biblioteca/KelZamMatGom2023a.pdf>
- Lipski, J. (2005). Code-switching or borrowing? No sé so no puedo decir, you know. In L. Sayahi & M. Westmoreland (Eds.), *Selected Proceedings of the 2nd Workshop on Spanish Sociolinguistics* (pp. 1–15). Somerville, MA: Cascadilla.
- Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford, New York: Oxford University Press.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: Toward a typology of code-switching. *Linguistics*, 18(7-8), 581-618.
- Poplack, S. (1985). Contrasting patterns of codeswitching in two communities. In M. Heller (Ed.), *Codeswitching: Anthropological and Sociolinguistic Perspectives* (pp. 215-243). Berlin: Mouton De Gruyter.

Giulia Mantovani (Universität Augsburg) & Franz Meier (Universität Augsburg)

Analyzing Italian gerunds in 18th century multilingual text production – How far can we go with AI?

The history of gerunds in 18th century Italian lacks systematic, corpus-based studies. Nevertheless, scholars have repeatedly pointed to the influence of French on Italian gerund prepositional constructions (like *in leggendo*) in this period (cf. e.g. Setti 1953; Folena 1983; Dardi 1992). While there are many contact phenomena of French on 18th century Italian (cf. e.g. Dardi 1992; Morgana 1994; Meier 2024), its impact on the gerund has largely remained unexplored. In fact, Mantovani (in prep.) provides preliminary findings based on a reduced corpus of French-Italian translations which give some evidence for a possible influence of French on the syntax and semantics of the Italian gerund; however, further investigation is needed to confirm these results. This paper explores the possibilities of applying AI as a tool for linguistic analysis in historic corpora. Our data set consists of a corpus of 60 French to Italian scientific translations (300 000 words) published between 1770 and 1795 and a corpus of 30 non translated Italian texts (110 000 words) published in the same period. Our aim is twofold: to further assess the diachronic influence of French on the Italian gerund in scientific translation and to evaluate the performance of AI models – especially chatbots such as ChatGPT – as analytical tools for diachronic studies in the Romance languages.

Our semi-automatic analysis is based on zero-shot/few-shot learning methods, that is, digital training methods with which AI models learn to make accurate predictions by using a very small amount of data (Suissa *et al.* 2022). While our syntactic analysis focuses on the word order in the gerund clause and its position in the matrix sentence, the semantic analysis deals with the meaning of the gerund clause. We expect the AI analysis of this latter feature to be especially challenging, not at least due to the many nuanced possibilities of classification (cf. Frenguelli 2001) which follows traditional adverbial categories.

References:

- Dardi, Andrea (1992), *Dalla provincia all'Europa. L'influsso del francese sull'italiano tra il 1650 e il 1715*, Firenze, Le lettere.
- Folena, Gianfranco (1983), *L'italiano in Europa. Esperienze linguistiche del Settecento*, Torino, Einaudi.
- Frenguelli, Gianluca (2001), "Tra narrazione e argomentazione: il gerundio nella prosa d'arte dei primi secoli", in: Rocchetti, Alvaro/Giacomo-Marcellesi, Mathée (ed.), *Il verbo italiano: studi diacronici, sincronici, contrastivi, didattici. Atti del 35. congresso internazionale di Studi (Parigi, 20-22 settembre 2001)*, Roma, Bulzoni, 23-42.
- Mantovani, Giulia (in prep.): "Gerund forms in 18th century language contact: The case of Targioni's scientific translations from French into Italian", *Perspectives* (Special Issue).
- Meier, Franz (2024), "Les phrases clivées dans la langue scientifique italienne de la fin du 18^e siècle: un cas de contact de langues à travers la traduction", in: Benjamin, Peter (ed.), *Contact des langues et plurilinguismes dans la Romania/Contacto de lenguas y plurilingüismo en la Romania*, Berlin, Frank & Timme, 131-155.
- Morgana, Silvia (1994), "L'influsso francese", in: Serianni, Luca/Trifone, Pietro (ed.), *Storia della lingua italiana*, vol. 3 (*Le altre lingue*), Einaudi, Torino, 671-719.
- Setti, Maria Vittoria (1953), "Francesismi trecenteschi nella lingua di F. Algarotti", *Lingua Nostra*, XIV 8-13.
- Suissa, Omri, Avshalom Elmalech und Maayan Zhitomirsky-Geffet (2022), "Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science", *Journal of the Association for Information Science and Technology*, 73 (2), 268–287.

Jonathan Sakunkoo (Stanford University OHS) & Kyle Gorman (City University of New York)

Mind the Gap: Computational Validation of Crowd-Sourced Linguistic Knowledge on Morphological Gaps in a Romance Language and Latin

Wikipedia and its sister websites, prominent examples of user-generated content, rank consistently among the most popular websites worldwide, attracting over 4.5 billion unique monthly visitors and 345 edits per minute. Despite their extensive reach and usage, they are generally perceived as unreliable by domain experts. However, for scarce linguistic phenomena such as word defectivity and morphological gaps in less-studied Romance languages, Wikipedia and Wiktionary often serve as two of the few widely accessible and frequently utilized resources. Documenting such defectivity and morphological gaps typically requires significant expertise and manual effort, making crowd-sourced content a potentially valuable but underexplored resource. In this study, we conduct computational analyses of inflectional gaps by customizing UDTube [Yakubov et al., 2024], which is a scalable and state-of-the-art neural morphological analyzer trained with Universal Dependencies (a corpus of morphologically annotated text in different languages), to incorporate mBERT as an encoder and annotate large corpora of text in certain languages [Conneau et al., 2020], particularly Latin (640MB, 390 million words) and Italian (8.3GB, 5 billion words). The resulting massive annotated data are then used to measure the frequency of certain inflectional forms of interest and validate lists of defective verbs scraped and compiled from Wiktionary's Latin and Italian pages to verify which verbs are confirmed computationally to be inflectional gaps. The aim is to conduct quality assurance on the linguistic information provided in crowd-sourced Wiktionary, assessing its reliability for linguistic research and knowledge. Preliminary findings indicate that over 50% of inflectional gaps listed in Wiktionary align with our computational morphology results, thus suggesting a degree of reliability in Wiktionary's linguistic data, despite coming from unreferenced, user-generated sources. This work highlights the potential of leveraging crowd-sourced, user-generated content as a supplementary linguistic resource and contributes to the intersection of computational methods and Romance linguistics by providing a novel, scalable tool and methods for validating and expanding knowledge of defectivity and morphological gaps in Romance languages and their parent language, Latin. By bridging computational NLP techniques with linguistic analysis in Romance languages, the study contributes to the quantitative turn in Romance linguistics as it offers novel insights and computational methodologies for scalable quality assurance and validation and addresses gaps in linguistic knowledge.

References:

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. [Universal Dependencies](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.
- Daniel Yakubov. 2024. [How Do We Learn What We Cannot Say?](#) Ph.d. dissertation, City University of New York (CUNY). Available at CUNY Academic Works.
- Daniel Yakubov, Kyle Gorman, and Github Contributor Jonathan Sakunkoo. 2024. [UDTube: A tool for universal dependency-based linguistic analysis](#).

Mathilde Regnault (Universität Stuttgart) & Achim Stein (Universität Stuttgart)

Investigating the Conditions of Language Change with Historical Language Models

Corpora offer precious insights into the successive stages of a language and enable linguists to develop focused theories and grammars. However, they face challenges due to the heterogeneity of historical data, particularly regarding balanced corpora and difficulties with annotation (Grobol and al. 2022a). For both Natural Language Processing (NLP) and linguistics, these issues make domain adaptation and accurate analysis complex.

LMs have proven useful for automatic annotation of diachronic corpora, for example the BERTrade model for Old French (Grobol and al. 2022b). Following the example of the BabyLM challenge (Huand al. 2024), we train small LMs for the study of language evolution, in order to represent singular states of a language. For example, we select only texts from a specific time period, optionally selecting certain genres or dialects.

We use this kind of small historical LMs as proxies of human competence in a precise historical context to perform linguistic tests (Ettinger 2019), such as the selection of an auxiliary in front of a past participle. For example, Burnett et al. (2015) observed variation in the auxiliary for intransitive verbs like *corir* (to run). A completion task with a model trained on 12th-century data (*il <mask> coruz*) yielded predictions such as:

- est (13.08%)
- ad (9.23%).

By permitting such tests, LMs offer new opportunities for linguistic descriptions. It is then possible to change the input sentence with various word orders, lexical items, etc. in order to interrogate the conditions of appearance of phenomena according to a given model.

Unlike general-purpose LMs, historical LMs are limited by their smaller context windows. This constraint enables them to represent individual states of a language more precisely. By focusing on localized data, these models facilitate the reconstruction and understanding of linguistic phenomena. This enables us to try and replicate the conditions under which language change occurs by using LMs to simulate linguistic competence. This approach allows us to investigate the mechanisms behind specific phenomena, such as auxiliary selection or dative alternation, and offers new insights into the drivers of linguistic change over time.

References:

- Burnett, H., Caudal, P., & Troberg, M. (2015). Les facteurs de choix de l'auxiliaire en ancien français: étude quantitative et comparative. In *Conférence Diachro-VII-Paris, 5-7 February 2015* (p. 12).
- Ettinger, A. (2019). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8, 34-48.
- Grobol, L., Prévost, S., & Crabbé, B. (2022a). Is Old French tougher to parse?. In *20th International Workshop on Treebanks and Linguistic Theories* (pp. 27-34). Association for Computational Linguistics.
- Grobol, L., Regnault, M., Suarez, P. O., Sagot, B., Romary, L., & Crabbé, B. (2022b). Bertrade: Using contextual embeddings to parse old french. In *13th Language Resources and Evaluation Conference*.
- Hu, M.Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A.S., & Wilcox, E.G. (2024). *Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora*. ArXiv, abs/2412.05149.

José-María Sánchez-Sáez (Universidad de Málaga), Álvaro Ramajo-Ballester (Universidad Carlos III de Madrid) & Gabriel Pišvejc (Universidad de Málaga)

Application of Artificial Intelligence and Machine Learning Models in Language Analysis: RRR Principles for Modular Automation

Recent advances in Artificial Intelligence (AI), particularly in generative AI and language models, have revolutionized how traditionally human tasks are approached. While these technologies have limitations, their ability to enhance productivity is undeniable when used strategically and appropriately.

Over the last decade, the use of Machine Learning models for processing large volumes of data has become widespread, enabling breakthroughs that were unimaginable just a few years ago. However, in the field of linguistics, studies are often constrained by the manual nature of data collection and initial processing mechanisms. This limitation hinders the availability of significant amounts of information, which is essential for this type of analysis.

Today, we have tools—some of them AI-based—that significantly simplify data acquisition and processing from real-world samples, effectively addressing the issue of data scarcity. However, this introduces the challenge of managing and processing information efficiently and adaptively. Overly rigid process designs can limit their reuse in similar projects, reducing their applicability.

An example is the work conducted by the VUM group at the University of Málaga, which processes diverse databases such as audio recordings, phonological records nearly a century old, and already-transliterated documents. All of these contain real samples of the non-standardized variety of spoken Spanish in the city of Málaga. In these contexts, processing mechanisms must be flexible and adaptable to different situations.

Based on the RRR principles (Reliability, Reproducibility, and Replicability), this presentation explores the integration of AI-based tools to optimize modularity and efficiency across various domains. A practical and conceptual approach to automated processing management will be presented, addressing case studies, applied methodologies, and the challenges associated with their implementation. Additionally, strategies to ensure their impact will be proposed.

One of the most notable advantages of using these tools is the increased scalability, a key aspect of this research. Automating the analysis of large datasets significantly reduces the bottleneck associated with manual processes, enabling access to broader and more diverse samples. This not only improves the representativeness of studies but also enhances their statistical rigor, allowing for more robust and generalizable findings. This capability to handle and analyze large-scale data represents a crucial advancement in domains where the quantity and quality of information are critical.

The presentation aims to inspire professionals and academics to adopt these tools strategically, promoting their productive and responsible use across different application fields. It also seeks to share ideas, suggestions, and perspectives that can contribute to the continuous improvement of the tools, fostering a collaborative approach.

Matthias Schöffel (Bayerische Akademie der Wissenschaften) & Marinus Wiedner (Universität Freiburg)

Exploring the changes in grammatical gender from Latin to Old Occitan through simulation

This communication draws on the work of Polinsky/Everbroeck (2003), who used a connectionist model to simulate the evolution of grammatical gender from Latin to Old French. Building on their approach, we aim at simulating the development of gender from Latin to Old Occitan, starting on the character level. To achieve this, we employ a long-short-term memory (LSTM) architecture with an attention mechanism, in contrast to the heuristic models proposed by Marr and Mortensen (2020).

A gender reduction from three to two genders took place in the transition from Latin to Old Occitan, during which the neuter disappeared. The former neuter nouns (e.g. Latin third declension MARE) had to be reattributed to either masculine or feminine (cf. it. *il mare* vs. fr. *la mer* vs. both genders in Old Occitan).

For this simulation, we utilize nouns from the *Dictionnaire de l'occitan médiéval* (DOM), the most comprehensive lexicographical resource on Old Occitan. Additionally, we incorporate variants that were digitized using a customized OCR model (cf. Garcés Arias/Pai/Schöffel/Heumann/Aßenmacher 2023). As the foundation for model training, we begin with the linked etyma from the *Französisches Etymologisches Wörterbuch* (FEW).

In addition to the lexicographic (and thus normalized) data, we incorporate nouns extracted from original 13th and 14th-century manuscripts, which were semi-automatically transcribed using a Transkribus model for Old Occitan Handwriting (cf. Wiedner 2023). The texts were then annotated with a part-of-speech tagger, with manual corrections applied to the results. Additionally, we linked these nouns to their respective etyma, incorporating information on gender (and potential variation) and accusative forms taken from both FEW and *Thesaurus Linguae Latinae* (TLL). Our goal is to evaluate whether the simulation outcomes differ when using non-normalized, "authentic" data, which includes both lexical and graphical variants, as opposed to the standardized data from the DOM.

We will present and discuss the foundational concepts as well as preliminary findings.

References:

- DOM = *Dictionnaire de l'occitan médiéval*. <<http://www.dom-en-ligne.de/>>.
- FEW = Wartburg, Walther von, et al. (1922–2022): *Französisches Etymologisches Wörterbuch (FEW)*. Eine Darstellung des galloromanischen Sprachschatzes. 25 Bände, Bonn/ Heidelberg/ Leipzig/ Berlin/ Basel, Klopp/ Winter/ Teubner/ Zbinden. <<https://apps.atilf.fr/lecteurFEW/>>
- Garcés Arias, Esteban, Pai, Vallari, Schöffel, Matthias, Heumann, Christian, Aßenmacher, Matthias. 2023. Automatic Transcription of Handwritten Old Occitan Language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, 15416–15439.
- Marr, Clayton, Mortensen David (2020): „Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction“, *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, 28– 36.
- Polinsky, Maria, Everbroeck, Ezra van (2003): „Development of Gender Classifications: Modeling the Historical Change from Latin to French“, *Language* 79, 356–390.
- TLL = *Thesaurus Linguae Latinae (TLL) open acces*. München. <<https://publikationen.badw.de/de/thesaurus>>
- Wiedner, Marinus (2023). OldOccitanHandwriting, (Modell-Nr. 52822, CER=3,51 %), PyLaia-model for handwritten Occitan of the 13th and 14th century. <<https://readcoop.eu/de/modelle/old-occitan-handwriting/>>

Nolan Welch (Honors Carolina)

CS-Aware Sentiment Analysis: Enhancing Mixed-Code NLP for Spanish-English Bilingual Data

Within natural language processing (NLP), sentiment analysis—the classification of emotional states (e.g., positive, negative, neutral)—has emerged as a key task with applications in recommendation systems, customer feedback analysis, and brand perception studies. However, sentiment analysis models often underperform on underrepresented languages, as their effectiveness typically depends on the variety and quantity of training examples in the target language.

This study addresses these limitations by focusing on Spanish-English mixed-code text, commonly referred to as “Spanglish”. Despite its growing cultural and linguistic relevance, Spanglish is often overlooked in NLP research, presenting opportunities for generalizing findings to other Romance languages where code-switching remains underexplored. Prior research (Dewaele; Pavlenko; Resnik) highlights the interplay between emotional state and code-switching (CS) in bilingual speakers, yet current sentiment analysis models fail to capture this relationship. We propose “CS-aware” sentiment analysis models that integrate sociolinguistic priors to improve sentiment classification for mixed-code data. These insights can inform NLP methods for Romance languages in diverse bilingual contexts.

We evaluate traditional sentiment analysis models and large language models (LLMs) on Spanish-English mixed-code data, benchmarking their performance using the LinCE dataset (Aguilar et al.). To capture emotional states tied to CS, we analyze sociolinguistic patterns with established quantitative methods (Gambäck and Das; Lal et al.; Lipski; Poplack and Meechan; Qin et al.; Torres Cacoullos et al.; Vilares et al.). We then integrate these patterns into CS-aware models to assess their impact on sentiment classification accuracy.

Preliminary findings indicate that incorporating sociolinguistic priors significantly improves sentiment analysis performance on mixed-code data compared to CS-naive methods. This advancement is particularly promising for Romance languages, as their shared linguistic characteristics, such as flexible syntax and morphological agreement, present opportunities for cross-linguistic model adaptation. We argue that future NLP research must integrate linguistic knowledge and domain-specific insights to more effectively model natural language. This research highlights the potential of NLP tools to address the complexities of Romance languages and the value of integrating sociolinguistic priors for computational analysis.

Keywords: Spanglish, sentiment analysis, code-switching, NLP, bilingualism

References:

- Aguilar, G., Kar, S., & Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1803–1813. [Link](#)
- Dewaele, J.-M. (2010). *Emotions in multiple languages*. Palgrave Macmillan.
- Gambäck, B., & Das, A. (2016). Comparing the level of code-switching in corpora. *Proceedings of LREC'16*, 1850–1855. [Link](#)
- Lal, Y. K., et al. (2019). De-mixing sentiment from code-mixed text. *Proceedings of ACL SRW*, 371–377. [DOI](#)
- Lipski, J. M. (2008). *Language Mixing and Code Switching*. Georgetown University Press.
- Pavlenko, A. (2008). Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition*, 11(2), 147–164. [DOI](#)
- Poplack, S., & Meechan, M. (1998). Introduction: How Languages Fit Together in Codemixing. *International Journal of Bilingualism*, 2(2), 127–138. [DOI](#)
- Resnik, P. (2018). *Multilinguals' Verbalisation and Perception of Emotions*. Multilingual Matters. [DOI](#)
- Torres Cacoullos, R., et al. (2021). How to mix: Confronting “mixed” NP models and bilinguals' choices. *Linguistic Approaches to Bilingualism*, 12. [DOI](#)

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2016). EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. *Proceedings of LREC'16*, 4149–4153. [Link](#)