2025 LCT Annual Meeting Poster Abstracts

With the support of the Erasmus+ Programme of the European Union

Group I

1. Name: Elisa Forcada Rodríguez

Title: Colombian Waitresses y Jueces canadienses: Gender and Country Biases in Occupation Recommendations from LLMs

One of the goals of fairness research in NLP is to measure and mitigate stereotypical biases that are propagated by NLP systems. However, such work tends to focus on single axes of bias (most often gender) and the English language. Addressing these limitations, we contribute the first study of multilingual intersecting country and gender biases, with a focus on occupation recommendations generated by large language models.We construct a benchmark of prompts in English, Spanish and German, where we systematically vary country and gender, using 25 countries and four pronoun sets. Then, we evaluate a suite of 5 Llama-based models on this benchmark, finding that LLMs encode significant gender and country biases. Notably, we find that even when models show parity for gender or country individually, intersectional occupational biases based on both country and gender persist. We also show that the prompting language significantly affects bias, and instruction-tuned models consistently demonstrate the lowest and most stable levels of bias. Our

findings highlight the need for fairness researchers to use intersectional and multilingual lenses in their work.

2. Name: Ariun-Erdene Tumurchuluun

Title: Causal Investigation of Tense Encoding in Multilingual Transformers

Transformer models achieve state-of-the-art performance on many NLP tasks, yet their internal representations often entangle linguistic features in "superposition." In particular, how these models encode grammatical tense, as distinct from broader semantic notions of time, remains under-explored. In this study, we probe the multilingual Llama-3.1 8B model to (i) localize layers and attention heads that robustly capture syntactic tense features, (ii) apply causal tracing to validate their functional role, and (iii) disentangle these features via a Sparse Autoencoder (SAE). We first train lightweight classifiers on layer activations across typologically diverse languages to identify candidate components. We then train an SAE to extract sparse latent units corresponding to grammatical tense markers and perform targeted ablation: removing past-tense units should shift model continuations toward non-past constructions. Finally, we assess whether these disentangled features differ from potential semantic "time" representations. By combining probing, causal intervention, and sparse disentanglement, our work advances mechanistic interpretability and sheds light on how multilingual transformers represent temporal morphology.

3. Name: Álvaro Arozarena Gómez

Title: A computational phylogeny of South American languages

Traditionally, research in historical linguistics has been done mostly through the comparative method. This approach, while very robust, requires painstaking amounts of labor and a very high level of expertise that, for most of the world's languages, few people possess. In recent decades, computational methods have become a powerful tool to assist researchers in this area. Phylogenetic inference was first developed for research in evolutionary biology, and produces as output sets of trees that have a high probability of explaining patterns in the input data, which typically consists of

vocabulary sets coded for cognacy. This method can be leveraged to check for whether or not sets of languages are related at all, and how closely related they are or how they cluster together if that is the case. Our goal is to do this for a representative sample of the indigenous languages of South America using automatic cognate detection with the LexStat method.

4. Name: Roman Kovalev

Title: Live Dictionary for AI Chatbots

Due to their nature, LLMs often provide nonsensical or inaccurate responses, which is an especially acute problem in high-risk fields like medicine or finance. The current approaches to enhancing accuracy are time-consuming and computationally expensive, making them difficult to use in practical applications, such as heavy user-faced chat agents. We introduce the token-level RAG approach - Live Dictionary, which provides re-usable extra information to LLMs with the aim of improving their accuracy. By measuring the uncertainty of the model at the input and detecting the most confusing words, we update the prompt with their definitions. We implement sense disambiguation and experiment with prompt formats to see how they influence the accuracy. The change in model performance is measured on a QA dataset across multiple low- and high-resource languages.

5. Name: Blanca Alonso Gonçalves

Title: Exploring Perplexity as a Linguistic Marker of Psychosis in Speech Transcripts

The perplexity of a sequence given a context is the exponentiation of the negative log-likelihood of that sequence under a language model, and it measures how unexpected the sequence is for the language model given the context. This means that perplexity scores depend on the size of the context, the choice of the model, and the text for which it is calculated.

This study explores the feasibility of using perplexity to discriminate between the speech of patients suffering from psychotic disorders and healthy subjects. Previous studies suggest that perplexity may be a promising marker of linguistic differences in psychosis, but its generalizability remains unclear. Perplexity scores tend to decrease with increasing context size, but this relationship may be less pronounced in the case of psychotic speech.

The study addresses four key questions: how perplexity scores differ between psychotic and control subjects, how these differences vary across language models, how context size impacts the scores of each group, and how perplexity relates to clinical psychotic symptom measurements.

By systematically analyzing speech transcripts using different models and context sizes, this research aims to contribute to more robust methods of psychosis detection and deepen our understanding of linguistic patterns in psychotic speech.

6.Name: Syed Muhammad Ibrahim Bukhari

Title: Synthetic Data Generation for Domain Specific Post-Training

This research investigates the generation of synthetic datasets to enhance the performance of a foundation LLM for specific domains and tasks. With the increasing capabilities of LLMs, synthetic data generation has become a widely adopted practice, particularly in scenarios where larger models are used to distill knowledge into smaller ones. This approach allows for the creation of data that is tailored to a particular domain or task.

A central challenge in this process lies in selecting optimal methods for both data synthesis and evaluation. Given the limited interpretability of LLMs, the quality and behavior of the generated datasets can only be assessed indirectly, through careful prompt engineering and task design.

In this study, we begin by narrowing the target domain and identifying relevant tasks, which is essential for selecting appropriate evaluation benchmarks. Our work focuses on generating datasets for RAG, reasoning-based tasks, and general NLP tasks within

the chosen domain. Findings indicate that RAG datasets coupled with reasoning for the answers generated significantly outperform those without such reasoning. Moreover, relying solely on an LLM for data generation does not guarantee high-quality outputs. Instead, using a domain-specific seed dataset consistently leads to better performance.

7.Name: Anna Smirnova

Title: Evaluating Temporal Reasoning in Multimodal Large Language Models through Action Order Comparison in Paired Videos

Despite recent progress in multimodal large language models (MLLMs), their ability to directly compare videos—particularly with respect to temporal understanding—remains unstudied. This work presents a novel approach to explore MLLMs' capacity for video-to-video comparison through the lens of temporal reasoning, using the order of actions as the central evaluation criterion. Leveraging a dataset of paired videos annotated across five semantic dimensions, we focus specifically on the temporal axis to test whether models can judge the similarity of action sequences. We begin by evaluating MLLMs' performance using straightforward similarity prompts, establishing a baseline for temporal comprehension. Building on this, we introduce a reasoning-augmented pipeline to investigate whether structured, step-by-step prompting improves the models' ability to capture temporal relationships. To our knowledge, this is the first systematic study to assess video comparison in MLLMs with an emphasis on temporal ordering. Our findings might shed light on current models’ limitations in temporal alignment and suggest that integrating reasoning may enhance their ability to understand the flow and structure of events over time.

8.Name: Ezgi Basar

Title: TR-BLiMP: A Turkish Benchmark of Linguistic Minimal Pairs

We introduce TR-BLiMP, the first Turkish benchmark of linguistic minimal pairs, designed to evaluate the linguistic abilities of monolingual and multilingual language models. Covering 16 linguistic phenomena with 1000 minimal pairs each, TR-BLiMP fills an important gap in linguistic evaluation resources for Turkish. Various LLMs are evaluated on their performance across the 16 linguistic phenomena, as well as on their alignment to human acceptability judgments. Through the development of additional experimental paradigms, the benchmark also addresses the challenges posed by word order flexibility and subordination.

9. Name: Maria Francis

Title: Graph-Augmented Field-of-Research Classification Using Citation Relations

This thesis addresses the task of Field-of-Research Classification (FoRC), a multi-label classification problem where academic papers are assigned research topics from a predefined taxonomy. We focus on classifying papers in computational linguistics (CL) into their respective subfields using the FoRC4CL dataset, a collection of 1,500 ACL Anthology papers, annotated with Taxonomy4CL, a hierarchical taxonomy containing approximately 180 CL (sub-)topics. Previous shared tasks on this dataset have reported a top macro-F1 score of 0.66. With the aim of improving performance, we explore incorporating information from the citation network surrounding each input paper by constructing a citation graph where nodes represent papers and edges represent citation-based relations, including direct citation, co-citation, and bibliographic coupling. We generate graph embeddings from this network and combine them with textual embeddings from the paper's title and abstract to form enriched input representations. Additionally, we investigate which citation relations, individually or in combination, contribute most effectively to improving FoRC performance.

10. Name: Salwan Aziz

Title: Emotion Detection in Social Media: A Hybrid NLP Approach with Transformers and Rule-Based Heuristics

Emotion detection from text, particularly in social media, has become a critical task for understanding human behavior, mental health, and public sentiment. Traditional emotion detection methods, based on handcrafted fea- tures or deep learning, struggle with the com- plexities of informal language, sarcasm, slang, and mixed emotions prevalent in social media content. This paper introduces a hybrid ap- proach that combines transformer-based mod- els with rule-based heuristics to enhance emo- tion detection accuracy, computational effi- ciency, and real-time scalability. This paper proposes a lightweight transformer architec- ture, leveraging dynamic lexicon updates to handle the evolving nature of social media lan- guage. The methodology addresses key chal- lenges such as sarcasm, slang, and overlap- ping emotional states, while ensuring efficient real-time processing in resource-constrained environments. Additionally, the model is de- signed to be adaptable to emerging trends in social media language, with transfer learning mechanisms ensuring its long-term relevance. The proposed approach demonstrates improved accuracy, broader applicability across diverse datasets, and enhanced scalability compared to existing methods.

11. Name: Elizaveta Ershova

Title: Error Type Sensitivity to the Volume of Manual Annotation in GEC

This thesis aims to identify the threshold at which adding more manually annotated training data yields diminishing returns in model performance. While GEC systems generally benefit from more training data, the high cost of manual annotation makes it essential to determine if there is a point beyond which additional human-annotated examples provide minimal gains. The experiments will involve training the baseline model on synthetic data only, then training models with increasing amounts of human-annotated data. Model performance will be evaluated, with particular focus on error-type specific analysis using ERRANT. This approach allows for systematic observation

of how model performance on different error types changes as the volume of manually annotated data increases. This research is important because it addresses a critical practical question in the development of GEC systems: how to optimally allocate expensive human annotation resources. By understanding which error types benefit most from human-annotated examples and at what point additional annotation efforts produce minimal gains, researchers and developers can make more informed decisions about data collection strategies. The findings will contribute to more efficient development of GEC systems by helping to balance the use of costly human annotation with synthetic data generation.

Group II

1. Name: Jose Maldonado Rodríguez

Title: Detecting language leakage in machine translation into low-resource languages: a case study on Galician

Multilingual neural models have the ability of extracting abstract linguistic information which can then be applied in inference time on languages that are unseen or underrepresented in the training data. This inherent ability to apply language transfer can be beneficial for low-resource languages in terms of the model's ability to produce an understandable output, but can also result in unnatural linguistic constructions. Evaluation metrics such as COMET or TER fail to account for these confabulations, and therefore have a tendency to provide inflated scores for subpar translations.

Machine Translation into Galician, a regional language spoken in the north of Spain, suffers from this issue. Featuring syntax similar to that of Portuguese and spelling rules closer to those of Spanish, translations into this language often include syntactic and orthographic calques which affect the correctness of the end result.

This investigation studies the extent to which these cases are detected by existing measures, with the aim of developing new strategies to take this dimension into

account. Work is still ongoing, so the methodology and results are still to be determined.

2. Name: Md Abdur Razzaq Riyadh

Title: Multi-modal Speech Language Model for Speech Recognition in Basque

Multi-modal speech language models (SpeechLM) are a recent advancement in natural language processing. These SpeechLMs are instruction tuned and optimized toward general tasks. There usefulness toward automatic speech recognition, particularly, in relatively low-resource scenarios have not been explored. In this work, we adapt a SpeechLM for speech recognition in Basque and study the impact of langue-adapted large language model in SpeechLM for ASR. Using supervised learning, we fine-tune LLaMA-Omni, a SpeechLM model, for automatic speech recognition. We conduct comprehensive hyperparameter tuning to improve performance and evaluate our best models on both in-distribution and out-of-distribution datasets. Our results show that SpeechLM are efficient in ASR and language-adapted LLM gives significant performance boost in out-of-distribution settings.

3. Name: Nam Luu

Title: End-to-end Automatic Speech Recognition and Speech Translation: Integration of Speech Foundational Models and LLMs

Speech Translation (ST) is a machine translation task that involves converting speech signals from one language to the corresponding text in another language; this task has two different approaches, namely the traditional cascade and the more recent end-to-end. This paper explores a combined end-to-end architecture of pre-trained speech encoders and Large Language Models (LLMs) for performing both Automatic Speech Recognition (ASR) and ST simultaneously. Experiments with the English-to-German language pair show that our best model not only can achieve better translation results than SeamlessM4T, a large foundational end-to-end, multi-modal translation model, but can also match the performance of a cascaded system with Whisper and NLLB, with up to a score gain of 8% in COMET-22 metric.

4. Name: Hoa Quynh Nhung Nguyen

Title: Semantic graph in the era of Large Language Model

Semantic representations (SRs) such as Abstract Meaning Representation (AMR) have long supported natural language processing (NLP) by providing structured and interpretable encodings of sentence meaning. These representations have shown clear benefits for a range of NLP tasks, especially when paired with earlier models like BERT and T5. However, the rise of large language models (LLMs), such as LLaMA and GPT, has reshaped the field, raising new questions about the continued utility of SRs. This thesis explores how SRs—specifically AMR graphs—can be integrated with LLMs to improve performance on downstream NLP tasks. While prior research has demonstrated potential, it has mostly focused on single-sentence inputs, leaving multi-sentence applications like summarization underexplored. These tasks require the model to process and reason over multiple semantic graphs, posing additional challenges. This work investigates whether LLMs can effectively leverage multiple AMRs without losing coherence or depth. Another critical dimension is how AMR graphs are integrated. Most existing methods rely on linearization, which may obscure structural information essential for accurate understanding. To address this, the thesis will compare different strategies for graph integration: linearized AMRs versus graph-specific encoders, aiming to assess their relative effectiveness in preserving semantic structure and enhancing LLM performance in multi-sentence settings.

5. Name: Anna Taylor

Title: Modeling Emphasis Area Prediction for Text-to-Speech Synthesis

Emphasis is critical in speech comprehension, but its implementation varies across languages and contexts. This research aims to model emphasis in text-to-speech (TTS) synthesis, specifically for American English presentation-style speech. Using a multimodal dataset of TED Talks, I develop a pipeline of fine-tuned TTS models to predict emphasis areas from text and synthesize speech with appropriate emphasis patterns, evaluated objectively and by human judgments.

Although modeling expressive speech is challenging due to its one-to-many nature, where multiple speech variations can convey emphasis for the same text, recent work suggests that duration alone suffices for modeling, as other relevant prosodic dimensions (pitch/intensity) tend to covary in context. Correspondingly, my study focuses on duration-based prosodic control, training a transformer-based model to predict emphasis probabilities from text and using these to replicate patterns in synthesized speech with various duration modeling strategies. Evaluation includes speech/pause rate alignment, emphasis prediction accuracy, Jensen-Shannon Divergence (real vs. synthesized duration distributions), and human discrimination and preference tests (MUSHRA/MOS).

This research contributes to expressive TTS synthesis by demonstrating duration-driven emphasis modeling in a real-world, multimodal domain. Applications include accessibility, public speaking assistance, and automatic summarization. Future work could explore cross-linguistic generalization and multimodal co-speech gesture generation.

6. Name: Fairooz Azim

Title: Evaluating Multimodal Large Language Models on Visual Grounding

Multimodal large language models (MLLMs) are going through a rapid surge in popularity, demonstrating impressive capabilities across a variety of vision-language tasks, including visual question answering (VQA), visual navigation, image captioning, and beyond. These models hold significant promise for enhancing assistive technologies aimed at supporting visually impaired individuals. However, current MLLMs still face substantial limitations—most notably, their tendency to generate hallucinated answers that are not grounded in the actual visual input which can have a critical impact on a visually impaired person. In this research, we aim to systematically investigate and quantify the limitations of state-of-the-art MLLMs in the context of visual grounding, i.e., identifying the specific region of the image that supports a model's generated answer. We conduct a comparative study using the VizWiz-VQA-Grounding dataset, a benchmark created using images from blind users.

Specifically, we evaluate the performance of two leading MLLMs, Gemma and Qwen, on visual grounding, by integrating an existing SHAP-based explainability framework with a modern segmentation model SAM to probe which image regions the models attend to when answering questions. Our analysis not only evaluates the performance of these models but also explores how well their explanations align with human-annotated ground truth.

7. Name: Michael Atamakira Awanah

Title: A Grounded Memory System for Personal Assistants

This research presents a cognitive assistant designed to support individuals with dementia in locating misplaced objects in real time. The system integrates three key components to enable memory-based perception and reasoning. First, it uses Vision-Language Models (VLMs) and Large Language Models (LLMs) to process visual inputs and extract consistent information about people, actions, and objects. Second, this information is stored in a structured memory combining a knowledge graph with semantic embeddings, enabling relational understanding over time. Third, the system retrieves relevant memories using a hybrid of semantic search and graph-based querying. Together, these components allow for intuitive interaction—via voice, touch, or images—and provide personalized, context-aware assistance grounded in the user's real-world environment.

8. Name: Željka Ćiraković

Title: Decoding Speech Imagery: Investigating Linguistic Features and ERP Correlates with Machine Learning and Pre-trained Language Models.

Speech imagery, the internal simulation of speech without articulation, plays an important role in language and cognition, yet its neural underpinnings remain relatively underexplored. This study aims to investigate how linguistic features—specifically word frequency, semantic similarity, and syntactic complexity—modulate

event-related potentials (ERPs) during imagined speech. EEG data will be recorded as participants engage in speech imagery tasks, and single-trial analyses will focus on established ERP components such as the N400 and P600. In parallel, machine learning classifiers, including support vector machines and basic deep learning models, will be applied to examine the feasibility of decoding imagined speech based on ERP patterns. Natural language processing (NLP) tools, such as pre-trained language models, will be explored to assess possible alignments between linguistic structure and neural signals. While this is an initial study with a limited scope, it seeks to contribute to a better understanding of the relationship between language processing and covert neural activity. The findings may offer preliminary insights into how linguistic information is represented during speech imagery and inform future work in neural decoding and brain-computer interface (BCI) applications.

9. Name: Helen Schmehl

Abstract: This poster presents a proposed thesis project exploring how machine learning might support the annotation of Gulf Coast research documents, such as oral histories, and social science dissertations. These documents were originally tagged by researchers at the Bureau of Ocean Energy Management (BOEM), but the work was cut short in 2025 due to funding losses. The goal is to develop an automated tagging system trained on existing human-coded data and evaluate its performance against manual annotation. Along the way, the project will explore research questions about annotation consistency, model generalizability, and the challenges of applying ML in low-resource, domain-specific settings. Sitting at the intersection of computational linguistics and the digital humanities, this work aims to make archival materials more accessible while contributing to methods for scalable qualitative analysis.

10. Name: G.M. Arafat Rahman

Title: New Interactions for Machine Translation

Post-editing refers to the process of refining the output generated by an automatic machine translation (MT) system by making minimal manual modifications to produce high-quality translations. The efficiency of post-editing can be significantly improved if the system not only allows for manual edits but also highlights specific areas in the translation that require correction. Despite this potential, existing post-editing systems do not currently offer integrated functionality that combines both edit suggestion and a dedicated editing interface. This research investigates the effectiveness of visually highlighting errorprone segments in machine translation (MT) output within a post-editing interface. We leverage both an automatic error detection model (XCOMET) and corrections made by professional translators to identify these segments. The study also explores the post-editing effort according to the number of editable areas displayed. Furthermore, this explores the use of areas displayed contributes to improved translation quality, particularly under time constraints.

11. Name: M. Belén Saavedra

Title: Neural Network Lesioning and the Loss of Meaning: Modeling Semantic Impairments in Speech Processing

This thesis explores semantic impairments in speech processing by modelling aphasia-like neural damage in an LSTM recurrent neural network. Using EARSHOT1—a model that maps acoustic inputs to semantic representations—we designed and applied three lesioning methods (each related to a theoretical basis for language impairments observed in aphasia) to study their effects on meaning retrieval in isolated word recognition. We tested connection severing at input, hidden, and output layers, unit ablation across the same layers, and activation alteration (attenuation and noise) applied to the hidden layer only. We based this approach on techniques used to model neurological impairments in RNNs2.

We trained 10 different EARSHOT instances (i.e., simulated subjects, each with different randomly-initialized weights) on 1000 high-frequency English words spoken by one talker. The inputs were 256-channel spectral slices, presented in 10-ms frames. The model's task was to activate the defined semantic vector (SkipGram) for the current target word at each frame. The model has 512 hidden nodes (long short-term memory nodes, or LSTMs3), and 300 semantic output nodes. As each word was presented, we identified the word with peak cosine similarity to the output at any time step, and operationalized this as the model's response. The undamaged model reached 95% accuracy. Each subject was evaluated across multiple lesion severity levels per method and layer.

Our goal was to compare the impact of different lesioning strategies and identify whether they produce distinct degradation profiles measuring accuracy progression. Preliminary results suggest method- and layer-specific effects, reinforcing the importance of lesion type and location in simulating linguistic deficits. Further work could provide insight into modeling language comprehension deficits akin to aphasia and contribute to the cognitive plausibility of lesioning techniques in neural networks. These findings may help us better understand how specific types of brain damage affect language, by showing how similar patterns can emerge in artificial systems.