Title: LLMs for Low-Resourced Languages: Hands-On Approaches to Cultural and Genre Diversity in NLP

Abstract

This course provides linguists with a practical and critical introduction to the use of large language models (LLMs) for under-resourced and typologically diverse languages. It emphasizes how these models behave with linguistically, and culturally rich data, and guides participants to use NLP tools for their own languages.

The course also covers the anatomy and behaviour of multilingual LLMs, hands-on prompting for linguistic tasks, corpus creation and augmentation, evaluation of bias and failure modes, and the ethical implications of AI applied to low-resourced languages.

Through case studies, group activities, and guided labs, participants will:

- Learn to tokenize and probe LLMs for multilingual and genre sensitivity
- Craft genre-aware and culturally framed prompts
- Build and augment corpora tailored to low-resourced scenarios
- Evaluate LLM outputs across dialects, registers, and NLP tasks
- Evaluate LLM performances using both standard and alternative metrics

Level: Introductory to Intermediate

The course is designed for linguists with little or no background in machine learning. All coding is minimal or done through interactive notebooks.

Session 1 – Introduction to LLMs

Topics Covered:
This session introduces participants to the core concepts behind large language models and why they have transformed the field of NLP. It explores the anatomy of an LLM, tokenizers, embeddings, decoding, and unpack the difference between multilingual "coverage" and true capability. Recent LLMs are used as examples to explain how under-represented languages are handled.

Hands-on Activities:
Participants will tokenize sentences written in Romanian, Arabic, Thai, and Armenian, and [their own language] using multilingual tokenizers to observe how the model splits unfamiliar texts. They will visualize sentence embeddings and explore their dimensionality. Also in this session the trainees will reflect to what the model likely “knows” about their language.

Session 2 – Prompt Engineering for Linguistic and Cultural Tasks

Topics Covered:
This session focuses on how to effectively prompt LLMs for linguistically relevant tasks such as glossing, translation, and named entity recognition. A taxonomy of prompt types (zero-shot, few-shot, chain-of-thought) is presented, along with a comparison of their behavior. The

lecture emphasizes the difference between surface-level compliance and deeper linguistic understanding, and encourages participants to think about how genre and context shape expected model behaviour.

Hands-on Activities:
Participants design their own prompts to perform tasks such as glossing agglutinative phrases (e.g., in Turkish or Quechua), translating ceremonial or legal expressions, or rewriting a sentence from formal to colloquial Romanian. They experiment with few-shot examples and structured templates to test model output. We will also discuss whether the model's success stems from linguistic competence or pattern imitation, and explore what constitutes a genre-aware prompt.

Session 3 – Corpus Design and Augmentation

Topics Covered:
This session introduces the principles of genre- and register-aware corpus creation for participants working in low-resource contexts. It covers the importance of metadata (e.g., speaker info, dialect, genre) and shows how mismatches between genre and model expectations can lead to failure. We discuss practical augmentation techniques such as back-translation, paraphrasing, and using dialectal alternates to enrich small corpora without large-scale annotation.

Hands-on Activities:
Participants begin by tagging a small sample corpus with genre and register metadata using a shared annotation format. They then create parallel sentence versions by rewriting utterances across formality levels or dialects (e.g., "You must go" in informal, polite, ceremonial forms). Using LLMs, they generate paraphrased or back-translated versions of their original examples. The aim is to build a prompt-ready or training-ready corpus that better reflects the variation found in their community's speech and text practices.

Session 4 – LLMs in a Diverse World: Failures, Biases, and Cultural Gaps

Topics Covered:
This session investigates how LLMs behave when exposed to linguistic, cultural, or social contexts they were not designed for. It distinguishes between structural bias (model architecture or data imbalance) and social bias (e.g., gender stereotypes, colonial erasure). Real-world failure cases are analysed, such as the mistranslation of proverbs or hallucination of folk narratives. Participants are encouraged to consider how their own language may be affected by similar issues.

Hands-on Activities:
Participants run test prompts across models using culturally specific or genre-rich input from their language: folktales, dialectal phrases, legal language, gendered professions. They document when and how the models fail whether by refusing output, introducing bias, or hallucinating content. These hands-on investigations help surface what parts of their linguistic tradition are "invisible" to the models. A final discussion unpacks these cases and generates community-specific insights into what's missing or misrepresented in mainstream LLMs.

Session 5 – Ethics and Evaluation

Topics Covered:

This session introduces participants to structured evaluation of LLM outputs using both automated metrics and alternative evaluation metrics (e.g., cultural fit, user validation), highlighting the gap between surface-level performance and real-world applicability in under-represented languages and genres. It covers data misuse, and the risks of over-relying on monolithic models.

Hands-on Activities:

In this hands-on session, participants evaluate LLM outputs using both standard and alternative metrics to understand the limits of traditional evaluation frameworks in culturally diverse NLP contexts. In the first part, they compute automated metrics such as BLEU, chrF++, and perplexity across genres (e.g., news, folktale, legal) and languages, analyzing how these scores reflect, or fail to reflect actual quality. In the second part, they shift to a qualitative evaluation, manually assessing outputs for cultural fit, user acceptability using an annotation template. Participants compare metric-based scores with human judgment and explore how prompt revisions could improve performance in linguistically and culturally grounded applications.

Necessary Infrastructure

Individual laptops are **required** for hands-on sessions.

Reliable internet access.

Prerequisites for the course

Trainees will use Google Colab, Google AI Studio, and various LLM-based chat interfaces such as Gemini or ChatGPT (free tier or similar). A personal Google account is required to access the Google ecosystem.
No prior programming or machine learning expertise is strictly required, but a willingness to engage with simple code snippets (Python) during demonstrations and hands-on labs is expected. Basic familiarity with Python would be beneficial for hands-on labs but not mandatory for participation.

Bibliography

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in Neural Information Processing Systems* 35 (2022): 24824-24837.

Lin, Z. (2024). *Prompt Engineering for Applied Linguistics: Elements, Examples, Techniques, and Strategies*. English Language Teaching

Vatsal, S., Dubey, H. and Singh, A., 2025. Multilingual Prompt Engineering in Large Language Models: A Survey Across NLP Tasks. *arXiv preprint arXiv:2505.11665*.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N.K., 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), pp.1097-1179.

Chiang, C.H. and Lee, H.Y., 2023. Can large language models be an alternative to human evaluations?. *arXiv preprint arXiv:2305.01937*.

Instructors, their affiliation, contact details and experience

Instructor: Maria Carp

Affiliation: Research Institute for Artificial Intelligence „Mihai Drăgănescu”, Romanian Academy, Bucharest, Romania

Contact: maria@racai.ro

Experience: Senior Researcher in NLP

https://scholar.google.com/citations?user=LjcZnEgAAAAJ&hl=en