Invited Talk: The Cross-Lingual Transfer Myth: Why Modern LLMs Still Fail Without Comparable Corpora and Representations

Els Lefever
LT3, Ghent University
Els.Lefever@UGent.be

Abstract

Comparable corpora have long served as a foundation for multilingual NLP, supporting transfer across languages in tasks such as classification, retrieval, translation, and argument mining. Yet in the era of multilingual transformers and generative models, a central question is no longer simply whether texts are comparable, but what kinds of internal representations and downstream behaviors that comparability actually enables. In this keynote, I argue that cross-lingual transfer is best understood as a continuum oscillating between shared semantic structures and language-specific realizations. Drawing on two complementary studies, I demonstrate how this tension manifests both in the data models learn from and in the representations they develop.

The first case study investigates multilingual stance and argument mining using the new Russian LoveHate corpus alongside English debate data. The results indicate that translated or multilingual resources are useful but insufficient proxies for language-specific corpora: local topics, culturally situated argumentation patterns, and stance expression still shape model performance and generalization. The second case study presents a neuron-level analysis of multilingual emotion detection, showing that multilingual encoders such as XLM-R develop both polyglot neurons, which respond consistently across languages, and monolingual neurons, which remain tied to particular linguistic systems. This reveals that even successful cross-lingual emotion transfer depends on only partial internal alignment.

Together, these findings suggest that multilingual NLP needs corpora that preserve culturally specific meaning while supporting robust transfer, as well as interpretability frameworks that can diagnose where multilingual systems genuinely share representations and where they merely approximate them. Comparable corpora are not just training material; they are essential to understand how cross-lingual generalization succeeds, where it breaks down, and how truly multilingual NLP can move beyond English-centric assumptions and conclusions.

Keywords: Comparable Corpora; Representation; Cross-lingual Transfer; Stance and Argument Mining; Emotion Detection; Culturally Specific Meaning

Bio

Els Lefever is associate professor at the LT3 language and translation technology team (Ghent University). She holds a PhD in computer science on ParaSense: Parallel Corpora for Word Sense Disambiguation (2012). Els has a strong expertise in machine learning of natural language and multilingual natural language processing, with a special interest for computational semantics, language modeling of lower-resourced languages and multilingual terminology extraction. She currently supervises research on complex reasoning in large language models, argumentation mining in social media, the automatic detection of irony, stereotypes and bias in online text, multimodal emotion detection and generation, and NLP approaches for low(er)-resourced languages, such as cuneiform, Byzantine Greek, or historical travelogues.

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, page 1
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026