Computational Linguistics in the Netherlands Journal 15 (2026) 1-3 Submitted 06/2026; Published 06/2026

Preface

Marijke Beersmans* MARIJKE.BEERSMANS@KULEUVEN.BE
Miryam de Lhoneux* MIRYAM.DELHONEUX@KULEUVEN.BE
Jens Lemmens** JENS.LEMMENS@UANTWERPEN.BE
Wessel Poelman* WESSEL.POELMAN@KULEUVEN.BE
Kushal Tatariya* KUSHALJAYESH.TATARIYA@KULEUVEN.BE
Maria Mihaela Trusca* MARIAMIHAELA.TRUSCA@KULEUVEN.BE
Tim Van de Cruys* TIM.VANDECRUYS@KULEUVEN.BE
Vincent Vandeghinste* VINCENT@CCL.KULEUVEN.BE
Bram Vanroy* BRAM.VANROY@KULEUVEN.BE

** KU Leuven*
Oude Markt 13, 3000 Leuven (Belgium)

*** University of Antwerp (CLiPS)*
Prinsstraat 13, 2000 Antwerp (Belgium)

The 15^{th} volume of Computational Linguistics in the Netherlands (CLIN) contains finished and reviewed work that was presented at the 35^{th} CLIN conference. CLIN is organized annually, and took place at the city campus of KU Leuven in 2025. In total, we received 126 abstracts, all of which were accepted; due to withdrawals, 115 were ultimately presented at the conference. 42 abstracts were presented as talks distributed over various parallel sessions, whereas the remaining 73 abstracts were presented as poster presentations throughout two afternoon sessions. The conference welcomed over 200 participants, reflecting the continued vitality of the computational linguistics community in the Low Countries.

It was our pleasure to welcome Prof. Dr. Marie-Catherine de Marneffe of UCLouvain as our keynote speaker. A FNRS research associate and one of the principal developers of the Universal Dependencies framework, de Marneffe gave a talk entitled *Consensus is a myth: Human label variation in Natural Language Inference*. Starting from the observation that humans often diverge in their interpretations of NLP tasks, she focused on the case of Natural Language Inference, in which one identifies whether a hypothesis sentence is true, false, or undetermined given a premise. Using examples such as whether *"My friend often travels with a heavy suitcase"* entails that *"My friend often travels with a light suitcase"*, she examined the various sources of human label variation in NLI and investigated whether they can be captured by current LLMs, arguing that, in the presence of such variation, labels without explanations are not sufficiently meaningful. The talk resonated strongly with several themes that recur throughout the present volume, from the evaluation of large language models to the role of context, perspective, and interpretation in language understanding.

After the conference, we received 15 submissions to the current volume of CLIN journal, of which 11 were accepted for publication after reviewing. Together, the accepted papers offer a representative cross-section of current research in the Low Countries, ranging from the probing and evaluation of large language models to applications in education and healthcare, from speech and multimodal processing to historical and diachronic NLP, and including a contribution on the broader research infrastructure that sustains the field.

A first group of papers investigates what language models learn and how their behaviour can be analysed, controlled, or evaluated. Klein, Manna and Vanmassenhove train a Dutch BERT model from scratch on the SoNaR corpus and use linear probes on intermediate checkpoints to track when and how gender information emerges in contextual embeddings. They find that male information is encoded diffusely across many dimensions while female information is concentrated in fewer, that stereotypical profession-gender pairings are classified far more accurately than anti-stereotypical

ones, and that generic (historically male) profession forms default to male interpretations even in the presence of explicit female cues. Van den Bent, Tepei and Bloem study whether Uniform Information Density-based regularizers, previously shown to be useful during pre-training, also help during the cheaper fine-tuning stage. Fine-tuning a Dutch GPT-2 on OpenSubtitles and Europarl with three UID-inspired regularizers, they observe small but consistent perplexity improvements, most visible on smaller data subsets, while noting trade-offs between information density and lexical diversity. Mohammadi and Bagheri, finally, probe 20 LLMs across 9 model families to assess the degree to which they mirror cross-cultural variation in moral attitudes as reported by the World Values Survey and the Pew Global Attitudes Survey. Earlier and smaller models often correlate weakly or negatively with human judgments, while instruction-tuned and larger models reach substantially higher positive correlations; alignment is consistently stronger for W.E.I.R.D. nations than for Sub-Saharan African, MENA, and South Asian regions.

A second group of papers explores how large language models can support concrete applications in education and healthcare. Moerman, Degraeuwe and Tezcan investigate retrieval strategies for LLM-based grading of short-answer grammar exercises, comparing rule-based and semantic-similarity retrieval (the latter adapted from translation memory fuzzy matching) across nine grammar topics in English, Spanish and Dutch. Semantic retrieval reaches 89.4% accuracy and up to 93.0% recall, with an accuracy-recall trade-off that maps neatly onto student-facing versus teacher-facing use cases. Schampheleer, Macken and De Wilde examine how dictionaries, machine translation (DeepL), and generative AI (ChatGPT) shape L2 English speaking performance in a multi-method experiment with university EFL students. They report that machine translation users produced the most fluent presentations, AI users the most lexically diverse and dense ones, and that lecturer evaluations clearly favoured both MT- and AI-prepared presentations over dictionary-prepared ones. In a healthcare setting, Ionescu, Han, Heijdra Suasnabar, Stiggelbout and Verberne apply embedding-based topic modelling (BERTopic and Top2Vec, with clinical embedding models and GPT-4o for labelling) to interview transcripts from pancreatic cancer patients. BERTopic combined with BioClinicalBERT yields the most coherent and interpretable topics, with care coordination and patient decision-making emerging as the most dominant recurring themes across the interviews.

A third group of papers addresses speech and multimodal communication. Lathouwers, Gao, Cucchiarini and Strik evaluate nine ASR models from the Whisper, Parakeet and Wav2Vec2 families on Dutch child speech, finding that a fine-tuned Whisper-medium achieves the best word error rates on the JASMIN and DART corpora. They additionally propose an utterance-level selection method in which ASR output is matched against the original read prompt to automatically identify utterances with reliable orthographic transcriptions, yielding precision of 98.3% or higher and covering 42.0% of JASMIN and 18.1% of DART, and thereby substantially reducing manual transcription effort. Van der Heiden and van Miltenburg turn to multimodal generation and accessibility, asking whether LLM-generated image descriptions in Dutch news contexts actually meet the needs of blind and visually impaired users. Through a 2×2 design varying length and readability, combined with LiNT readability scores, ratings from 47 participants, and a content analysis grounded in established image description guidelines, they find that ChatGPT 4o follows length and readability instructions and produces detailed contextualised descriptions, but also that descriptions frequently contain inaccuracies and add subjective atmospheric commentary, and that human readability ratings do not correlate with automatic LiNT scores.

A fourth group brings computational methods to bear on historical and diachronic language data. Bouma, Coussé and van Noord apply the Alpino parser to the diachronic Dutch C-CLAMP corpus, using surface-level meta-annotations (spelling variants, accusative case marking, multiword units, phantom tokens) and iterative error mining to raise parser coverage from 78% to 84%. The result is a parsed corpus of 173M tokens and a Verb Construction Database of 16M verb chains with high extraction precision and recall, illustrated through a case study tracing shifting word order preferences in three-verb clusters across two centuries of written Dutch. Lemay, Lefever and Bentein develop two complementary pipelines for large-scale unsupervised similarity detection in historical

Greek: MinHash with locality-sensitive hashing for surface-level similarity, and transformer-based SHLM sentence embeddings with FAISS-based nearest-neighbour clustering for semantic similarity. Applied to a corpus that spans roughly 400 BC to 1500 AD and includes papyri, PHI inscriptions, and Byzantine Inscriptional Epigrams, the surface-level method reconstructs 86.9% of the DBBE Verse Groups and surfaces repeated formulae and shared manuscript traditions, while the semantic method captures broader thematic affinities, demonstrating that the two are complementary rather than competing.

Finally, Vandeghinste, Maegaard and Lušicky offer a panoramic view of the CLARIN Knowledge Infrastructure, complementing CLARIN's well-known technical infrastructure of language resources and tools. Their paper situates the Knowledge Centres, Best Practice papers, Tour de CLARIN, CLARIN Cafés, Learning Hub, Mobility Grants and Annual Conference against the knowledge infrastructures of related research infrastructures such as DARIAH, CESSDA, SSHOC, ELIXIR and OpenAIRE, and illustrates the impact of the ecosystem through a K-Dutch use case involving the e-Lex lexicon.

The 11 papers in this volume have benefited greatly from the careful and constructive feedback of the editorial board of CLIN journal and a number of external reviewers. Reviewing is often invisible work, and we would like to sincerely thank all colleagues who took the time to engage thoughtfully with the submissions: Marijke Beersmans, Katrien Beuls, Jelke Bloem, Gosse Bouma, Tommaso Caselli, Luna De Bruyne, Orphée De Clercq, Peter Dirix, Iris Hendrickx, Alek Keersmaekers, Florian Kunneman, Els Lefever, Lieve Macken, Ilia Markov, Jelena Prokic, Manon Reusens, Marijn Schraagen, Ineke Schuurman, Anaïs Tack, Kushal Tatariya, Arda Tezcan, Erik Tjong Kim Sang, Maria Mihaela Trusca, Sander van den Bent, Paul Van Eecke, Hugo Van hamme, Emiel van Miltenburg, Rik van Noord, Menno van Zaanen, and Vincent Vandeghinste.

To conclude, we would like to thank everyone who has played a role in turning CLIN35 into a grand success, which first and foremost are all presenters, authors and over 200 participants. We also thank our sponsors for their generous support:

- Gold sponsors: Instituut voor de Nederlandse Taal, Taalunie
- Bronze sponsors: NOTaS, Textgain, Zeta Alpha
- Copper sponsors: CrossLang, De Taalsector

Additionally, we want to convey our appreciation towards the students of the AI master's program who kindly volunteered to assist during the conference: Henry Grafé, Rohin Vijayakumar, Stef Accou, Robin Kokot, Mehak Sharma, Isha Thombre and Ángela María Gómez Zuluaga. And finally, CLIN35 would not have been possible without the organizational committee, consisting of Marijke Beersmans, Miryam de Lhoneux, Wessel Poelman, Kushal Tatariya, Maria Mihaela Trusca, Tim Van de Cruys, Vincent Vandeghinste and Bram Vanroy. We wish you an interesting reading experience, and we are looking forward to seeing you again at the 36^{th} edition of CLIN, organized in Brussels on the 11^{th} of September 2026.