

CLIN36

The 36th Meeting of Computational Linguistics
in the Netherlands

Book of Abstracts

Friday 11 September 2026

Vrije Universiteit Brussel • hosted by the Artificial Intelligence Laboratory

CLIN36

Brussels, Friday 11 September 2026

The 36th Meeting of Computational Linguistics in the Netherlands took place on Friday 11 September 2026 at the Vrije Universiteit Brussel, hosted by the VUB Artificial Intelligence Laboratory.

This volume collects the abstracts of all talks and posters presented that day: two keynote lectures, 51 talks, and 66 posters. The abstracts appear as submitted by their authors, in the order in which they were presented.

CLIN36 was organised by Paul Van Eecke, Maxime Toquebiau, Jérôme Botoko Ekila, Jamie Wright and Lara Verheyen, and was made possible by the support of the Taalunie, De Taalsector, Textgain, the Instituut voor de Nederlandse Taal, SAY IT Labs and the HERMES programme.

The full programme remains available at clin36.ai.vub.ac.be.

Contents

Keynote lectures	4
Oral session 1	6
Semantics, Meaning Representation & Reasoning	6
Historical, Corpus & Stylometric Analysis	10
Multilingual & Dutch Language Resources and Tools	15
AI Policy, Governance & Societal Impact	21
Oral session 2	27
Reasoning, Structure & Grounding	27
Discourse, Communication & Language Evolution	32
Understanding & Improving Language Models	38
AI-Assisted Decision Support & Agentic Problem-Solving	42
Poster session 1	49
Clinical & medical NLP	49
Dutch language resources, benchmarks & tools	55
Sign language processing	61
Speech, assessment & interaction	63
Metaphor & figurative language	67
Multilingual & cross-lingual NLP	70
Poster session 2	78
Reasoning, puzzles & distillation	78
Model behaviour, evaluation & efficiency	83
Bias, society & political discourse	88
Emergent communication & cognition	94
Stylometry, literature & creativity	97
Corpora & annotation methods	102

Keynote lectures

Q building: Aula Q.C

09:15 – 10:15 · Q BUILDING: AULA Q.C

Emergent Linguistic Structure in Agent Communication

Tessa Verhoef

Leiden University

Why is linguistic structure shaped the way it is? To answer this, we need to disentangle the many forces, cognitive, communicative, and social, that shape languages as they are learned and used. Agent-based simulations offer a powerful way to isolate pieces of this puzzle and can model the spontaneous development of a communication system through repeated interactions between individuals. While this set-up is certainly not new, it is increasingly implemented with deep learning-based neural agents trained via reinforcement learning and this comes with a challenge: these neural agents still often behave differently from human learners and their languages fail to exhibit human-like properties, including a tendency toward destabilizing language drift. I will present recent work drawing on developmental trajectories in human language acquisition, showing that age-based plasticity, where younger agents learn quickly while older agents maintain more stable representations, can substantially reduce drift and help populations sustain a shared language even as new learners are introduced. Directly inspired by human artificial language learning studies, I will also present a set-up in which agents first learn an artificial language and then use it to communicate, with the aim of studying the emergence of a well-known language phenomenon: the word-order/case-marking trade-off. Finally, I will discuss recent results that reveal what a novel artificial language looks like when it has emerged to adapt to preferences of both humans and LLMs in a hybrid language game experiment.

About the speaker: Tessa Verhoef is an assistant professor at the intersection of language, cognition, cultural evolution and computation at the Leiden Institute of Advanced Computer Science of Universiteit Leiden, and director of its Emergent Communication group.

15:30 – 16:30 · Q BUILDING: AULA Q.C

Using Toy Models and Interpretability to Understand How Large Reasoning Models Reason

Willem Zuidema

University of Amsterdam

Large language models have evolved into large reasoning models, and have started to yield revolutionary results in mathematics and computer science. Meanwhile, Interpretability — the field trying to understand model internals — is struggling to keep up. Interpretability remains crucial, however, for ensuring safety, governability and crediting of sources, and new approaches to understanding reasoning models are therefore needed. I will show how experiments with toy transformers, surrogate models and a hypothesis-driven methodology shed light on how the Transformer architecture may implement reasoning strategies. Finally, using the planning puzzle

‘Tower of Hanoi’ as a case study, I will show that such a methodology can also be successful in frontier large reasoning models.

About the speaker: Willem Zuidema is an associate professor in natural language processing, explainable AI and cognitive modelling at the Institute for Logic, Language and Computation of the Universiteit van Amsterdam, and director of its Cognition, Language and Computation lab.

Oral session 1

11:15–12:30 · D building

Semantics, Meaning Representation & Reasoning

D.2.10 · Chair: Marie-Catherine de Marneffe

11:15–11:27

Between logic and contextualization: A new understanding of variation in Natural Language Inference

Erika Lombart, Patrick Watrin, Marie-Catherine de Marneffe

UCLouvain

It has been shown that annotators vary in NLI annotations, a task in which one identifies whether, given a premise (“It is, as you see, highly magnified”), a hypothesis (“You can see that it’s amplified”) is true, false, or undetermined. Several reasons for the variation have been put forth: different interpretations of the labels [1] or the items (i.e., [2], Kalouli et al. [3] who show that some items favor a logical interpretation). We focus on the construal of the task, arguing that annotators vary in how they approach inference. We conduct a qualitative analysis of annotators’ explanations from a pilot study in which annotators were asked to label NLI items and explain their label, following the ecologically valid annotation framework of the LiveNLI dataset [4]. Annotators approach the task in two ways: (i) a contextualizing approach in which different possible contexts are alluded to, grounded in pragmatic theories ; (ii) one interpretation is latched onto, based on pinpointing a precise meaning and clear semantic relations between terms, with no broader context taken into account. Compare two explanations for the item above: (a) What “it” is needs to be specified. It could be something concrete. In that case highly magnified/amplified are similar in meaning. Or it could relate to immigration issues/policies. (b) highly magnified and amplified are synonyms Linguistic markers associated with the contextualizing approach are identified and used to perform a quantitative analysis of the LiveNLI explanations, prompting Qwen3.7 to automatically classify these. We examine the distribution of NLI labels across approaches, testing the hypothesis that the contextualizing approach produces more indeterminate labels, whereas the non-contextualizing one leads to more decisive labels.

Finally, we explore whether LLMs can reproduce human annotators’ biases or generate more homogeneous inferences. Several LLMs are prompted with the same guidelines as humans, and their explanations are manually analyzed to determine the approach employed. This work highlights the importance of distinguishing between different construals of the NLI task, which are only identifiable when annotators provide explanations for their labels. More generally, it argues that ecologically valid annotations are essential for building robust and interpretable benchmarks. [1] A. Nigohjkar, A. Laverghetta Jr., and J. Licato. 2023. No Strong Feelings One Way or Another: Re-operationalizing Neutrality in Natural Language Inference. *Proceedings of LAW*, 199–210. [2] N. Jiang & M-C. de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *TACL*, 10:1357–1374. [3] A. Kalouli, H. Hu, A. F. Webb, L. S. Moss and V. de Paiva. 2023. Curing the SICK and Other NLI Maladies. *Computational Linguistics*, 49(1):199–243. [4] N. Jiang, C. Tan & M-C. de Marneffe. 2023. Ecologically Valid Explanations for Label Variation in NLI. *Findings of EMNLP 2023*, 10622–10633.

KEYWORDS Human Label Variation, natural language inference, annotation strategies

11:27–11:39

Exploring Subword Compositionality for Dutch Compounds

Tim Van de Cruys

KU Leuven

Large language models read text as sequences of subwords, so much of the word-level meaning they use is assembled from fragments. Dutch makes this assembly particularly demanding. Its productive, concatenative compounding generates arbitrarily deep words ('arbeidsongeschiktheidsverzekering') that subword tokenizers split with no regard for morpheme boundaries. The model must then recover a compositional structure that its own segmentation obscures. In this talk, we examine how well it manages, and what that might reveal about the relationship between tokenization and morphology. We evaluate a model's representation of a compound against the meaning its morphology predicts, using the right-hand head rule as a guide: a 'zonnepaneel' is a kind of 'paneel', not a kind of 'zon', so an adequate representation should sit near its head and shift within that neighbourhood along the modifier's semantics. Opaque compounds, where the head misleads (a 'vingerhoed' is not a kind of 'hoed'), are the hard case in which this structure should break down. Our stimuli combine attested compounds, including the opaque cases above, with controlled nonce compounds that no model can have memorized. We compare composition over the model's own subword tokens with composition over morphological constituents, treating the gap between them as a measure of how much tokenization departs from morphology.

KEYWORDS LLM, morphology, compounds, semantics, compositionality

11:39–11:51

CxGr-AMR: Incorporating argument structure constructions into Abstract Meaning Representation

Claire Bonial, Claire Benet Post, Harish Tayyar Madabushi, Paul Van Eecke & Katrien Beuls

DEVCOM U.S. Army Research Laboratory, University of Colorado, University of Bath, Vrije Universiteit Brussel, Université de Namur

Since standard Abstract Meaning Representation (AMR) annotation guidelines always tie argument structure relations to lexical rolesets, cases in which core semantic roles are expressed through clause-level structures rather than through specific verbs have been systematically misrepresented. To overcome this limitation, we have recently released CxGr-AMR, an extension to AMR that explicitly captures the semantics of various types of phrasal constructions, including argument structure constructions. Together with a collection of constructional rolesets and annotation guidelines, we also released a dataset containing 355 instances of four English argument structure constructions (resultative, caused motion, intransitive motion and ditransitive) annotated with both standard AMR and CxGr-AMR.

In this talk, we will first examine how the selected argument structure constructions were handled under current Standard-AMR guidelines and show that these analyses are often inadequate when constructionally contributed roles clash with those assigned by the verb. We will then provide a theoretical grounding for the novel CxGr-AMR rolesets that lay out the relationship between the

syntactic signatures of constructional slots and particular semantic roles associated with them. Finally, we will provide further details on our annotation-expert-in-the-loop pipeline for the semi-automatic annotation of sentences and the release of the CxGr-AMR corpus.

The paper that accompanies this talk has been presented at the 7th International Workshop on Designing Meaning Representations earlier this year:

Claire Bonial, Claire Benet Post, Paul Van Eecke, Katrien Beuls, and Harish Tayyar Madabushi. 2026. CxGr-AMR: Extending abstract meaning representation beyond lexically anchored relations with constructional rolesets. In Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 1–19.

KEYWORDS construction grammar, semantics, meaning representation, Abstract Meaning Representation

11:51-12:03

Revisiting the CommitmentBank in the Age of Reasoning Language Models

Sebastian Loftus, Marie-Catherine de Marneffe

UCLouvain

The CommitmentBank (CB) [1] is designed to evaluate pragmatic reasoning by measuring the degree to which a speaker commits to the truth of an embedded clause - a task requiring a nuanced understanding of conversational context and linguistic structure (e.g., Premise: Jane worries that it is snowing. Hypothesis: It is snowing. Label: Neutral). Early benchmarks by Jiang and de Marneffe (2019) [2] demonstrated that encoder-based models like BERT achieve strong performance (up to 85.3 F1) on CB by reframing speaker commitment as a Natural Language Inference task. However, feature probing revealed that BERT does not simply rely on surface artifacts [4] (e.g., a negation indicates the label contradiction), but on other, less superficial cues. The NLP landscape has since shifted towards generative models trained with post-training paradigms that encourage stable reasoning. These new types of reasoning paradigms raise the question of whether current LMs display a stronger understanding of complex pragmatics. Here, we investigate whether the mechanisms by which today's language models solve the CB task have evolved since the beginning of the transformer era. We analyze LLM's reasoning traces for occurrences of surface artifacts determined by Jiang and de Marneffe (2019) [2]. To automate the annotation process, we adopt an LLM-as-a-judge approach, where a secondary LLM checks for the occurrence of such artifacts. To ensure a valid evaluation, we calculate the IAA between a data-split annotated by a human annotator and the judge LLM. Preliminary tests using Qwen3-8B [5] indicate that it achieves competitive performance (F1 72.2), but underperforms compared to fine-tuned encoder models. Ultimately, this work sheds light on whether current post-training paradigms inherently improve or inadvertently compromise deep pragmatic language understanding.

[1] Jiang, N., and de Marneffe, M. C. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6086–6091. [2] de Marneffe, M. C., Simons, M., & Tonhauser, J. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung, Vol. 23, No. 2, pp. 107-124. [3] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., ... & Guo, D. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint

arXiv:2402.03300. [4] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. 2018. Annotation artifacts in natural language inference data. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 107-112. [5] Yang, An, et al. (2025). Qwen3 technical report. arXiv preprint arXiv:2505.09388.

KEYWORDS natural language understanding, reasoning, speaker intent, LLM evaluation

12:03–12:15

Do commonsense benchmarks measure commonsense? Correlating benchmark performance with downstream tasks

Ine Gevers, Walter Daelemans

Universiteit Antwerpen

Evaluating Large Language Models (LLMs) is a primary bottleneck in NLP, driven by validity problems, evaluation inconsistencies, model discriminability challenges, and dataset contamination. One of the core complaints against NLP benchmarks in the modern LLM-era is that they lack in construct validity; scores achieved on a task do not reliably reflect the acquisition of the underlying skill that was intended to be measured, which risks misinforming and misdirecting scientific research. This is especially relevant in testing common sense knowledge, an inherently broad task which suffers from a lack of clear definition, resulting in diverging and broad benchmarks. This study aims to make the link between common sense benchmarking and downstream performance explicit through an extensive correlation study. Specifically, to establish the practical usability of widely adopted common sense benchmarks, we introduce two research questions. First, are model performances on original (and, often argued to be flawed) commonsense benchmarks and their reworked / improved (e.g., paraphrased or filtered) editions correlated? In other words, what is the practical effect of making superficial amendments to original benchmarks, which is a popular method to patch up known deficits of benchmarks? Here, we focus on WinoGrande, HellaSwag, Social IQA, and Physical IQA and their reworked editions. The second research question intends to estimate how useful such standardized commonsense benchmarks are to assess LLMs' performance on downstream tasks. Specifically, we compute the correlations between commonsense benchmarks and a curated suite of downstream pragmatic tasks that rely on commonsense knowledge for success. As a control variable, we include three benchmarks that do not evaluate commonsense knowledge (i.e., general language skills or factual knowledge). Each benchmark is evaluated using averaged conditional log-probability scoring over answer options. This study contributes to the ongoing debate on benchmark validity by quantifying whether success on widely used commonsense benchmarks reflects transferable commonsense capabilities or merely task-specific performance.

KEYWORDS benchmarking, construct validity, commonsense evaluation

12:15–12:27

Revealing Uncertainty Through Questions in Defeasible Reasoning

Marzieh Abdolmaleki, Veronique Hoste, Els Lefever

Ghent University

Questions are fundamental cognitive tools for reducing uncertainty and guiding reasoning. When confronted with incomplete information, humans often ask questions to explore alternative explanations and identify missing contextual variables (GRAESSER et al., 1996; Chouinard et al., 2007). Inspired by this process, we investigate whether questions can function not merely as prompts, but as mechanisms for exposing uncertainty in defeasible reasoning. In defeasible reasoning, a premise may support multiple plausible hypotheses (Abdolmaleki et al., 2026) depending on unresolved contextual variables referring to context-dependent meaning components that cannot be inferred from the available context alone. We examine whether a single targeted question can reveal this latent ambiguity and simultaneously elicit competing defeasible hypotheses. We introduce a novel dataset for studying questions to highlight the role of inquiry as an intermediate reasoning mechanism for modeling uncertainty in generative reasoning systems.

KEYWORDS Defeasible Reasoning, Generative Models, Natural Language Processing

Historical, Corpus & Stylometric Analysis

D.2.16 · Chair: Luna De Bruyne

11:15–11:27

Just Letters: Stylometric Analysis of the Just Judges Ransom Letters

Loic De Langhe, Orphée De Clercq, Veronique Hoste

Universiteit Gent

The 1934 theft of the Just Judges panel from Jan and Hubert van Eyck's Adoration of the Mystic Lamb (1432) and the subsequent extortion correspondence constitute one of the most enduring unsolved cases in Belgian criminal history. Despite considerable scholarly attention to the theft itself, the authorship of the ransom letters has never been subjected to systematic computational analysis. This study addresses that gap by applying a multi-layered stylometric and natural language processing framework to the surviving correspondence.

Our methodology integrates both surface-level and deep structural approaches to authorial style. At the lexical level, we employ type-token ratios, word frequency profiling, and Latent Dirichlet Allocation (LDA)-based topic modelling to identify thematic clusters and idiosyncratic vocabulary patterns. At the syntactic level, we analyse sentence length distributions, dependency parse structures, and part-of-speech n-gram profiles as proxies for deeper, less consciously controlled stylistic signatures. Authorship attribution is subsequently performed through outlier detection techniques, including Burrows' Delta, Principal Component Analysis, and unsupervised clustering algorithms, allowing us to isolate letters that deviate significantly from the stylometric centre of the corpus.

The results of this multi-method analysis provide compelling evidence against single-author hypotheses: at least one letter exhibits a stylometric profile sufficiently divergent from the remainder of the corpus to suggest a distinct hand. This divergence persists across both lexical and syntactic

feature sets, rendering explanations based on register variation or noise unlikely. These findings carry significant implications for historical and criminological reconstructions of the case, and demonstrate the broader potential of computational authorship analysis as a tool in forensic art history.

KEYWORDS Stylometry, Forensic Linguistics, Syntax

11:27–11:39

Dutch without loanwords? A large-scale study of Dutch puristic word lists (1550–1850)

Nelle Simonet & Sara Budts

Fonds Wetenschappelijk Onderzoek - Vlaanderen; Vrije Universiteit Brussel

Over the past decades, historical (socio)linguistics has increasingly embraced empirical, quantitative, and corpus-based approaches, leading to substantial advances in our understanding of historical language variation and change. Within historical standardisation studies, however, research remains largely qualitative, particularly when metalinguistic sources constitute the primary object of study. This is especially true for lexicographical standardisation efforts, which have received very limited systematic quantitative attention. Within this field, research on lexical elaboration has predominantly relied on qualitative analyses of individual authors, works, or semantic domains, resulting in a fragmented understanding of the process. This paper demonstrates how computational and corpus-based methods can provide new insights into historical elaboration processes through the analysis of twenty Dutch puristic word lists published between 1550 and 1850. These lists propose Dutch alternatives to loanwords and were intended to strengthen the capacity of Dutch to function across a wide range of social and cultural domains without relying on other prestige languages. After digitising the word lists using OCR technologies, we normalised and lemmatised them for large-scale quantitative analysis, resulting in a corpus of 63,899 lemmas. This corpus allows us to systematically investigate patterns of lexical innovation, reuse, circulation, and similarity across three centuries of puristic lexicography. Our approach consists of two steps. First, we construct PCA plots based on the vocabulary contained in each word list. This allows us to map out how the word lists relate to one another and assess patterns of lexical reuse across the tradition. Second, we convert the vocabulary of the word lists into word embeddings and condensed them into a three-dimensional space. This enables us to explore how lexical representations of puristic tendencies travelled through the vocabulary of Early and Late Modern Dutch, one semantic subspace at a time. The results reveal that formal elaboration was characterised not only by lexical innovation but also by extensive reuse and recirculation of existing material. Rather than representing a series of isolated lexical interventions, the tradition emerges as a highly interconnected process in which works repeatedly draw on and reshape existing material. More broadly, the paper illustrates how computational approaches can uncover long-term patterns in historical metalinguistic traditions that remain difficult to identify through close reading alone.

KEYWORDS corpus linguistics, historical lexicography, distributional semantics, language change

11:39–11:51

Rags2Riches – Towards a Microdata Collection of Nineteenth-Century Dutch Wealth Distribution

Erik Tjong Kim Sang, Angel Daza, Abel Soares Siqueira, Auke Rijpma, Bas Machielsens, Amaury de Vicq, Ruben Peeters, Bram van Besouw, Michalis Moatsos

Netherlands eScience Center, Utrecht University, University of Groningen, Maastricht University, University of Antwerp

The Memories van Successie (Memories of Succession) is a collection of nineteenth-century and early twentieth-century Dutch inheritance tax records. They are a valuable source for studying historical wealth distribution in The Netherlands; however, the vast amount of documents in the collection makes exhaustive analysis infeasible, hence recent studies have been restricted to a sample of the data (Peeters, 2024).

Rags2Riches, a project of Utrecht University in collaboration with the Netherlands eScience Center, aims to extract digital microdata from the full set of digitized pages belonging to the Memories van Successie. A first challenge is that the digital scans are available at various provincial archives in The Netherlands, which calls for an intermediate data model that harmonizes the sources. A second challenge is the variety of layouts in which the documents are presented; the majority of the documents contain hand-written texts, but some sections may also contain printed text. Although good quality text recognition systems for Dutch exist, the variety of layouts, frequent use of lists, tables and margin notes present a big challenge to them.

We propose using vision language models (VLMs) to aid in the extraction of both biographical and wealth information from the tax records. Biographical information consists of deceased names, death dates and death places. Wealth information includes total taxable wealth, assets and liabilities.

In the first phase of the project, we focus on the data from the province of Noord-Brabant, as provided by the Brabants Historisch Informatie Centrum (bhic.nl). For reasons of consistency and completeness, we have selected the tax records from the years 1878 to 1927 (198,306 persons). In this data source, the biographical data of the deceased are available in a digital index.

Our first experiments using VLMs include Claude Sonnet 4.6 and Gemma 4. We are impressed by the recognition quality of Dutch hand-written text of a proprietary system like Claude. However, the associated costs for processing a single scan show already that a large-scale usage is not feasible. For this reason, we will focus on using locally-run models, and possibly build a hybrid pipeline which includes hand-written text recognition systems such as KNAW HUC's Loghi software for our tasks.

We plan to use the finished data collection to study at scale the development of the wealth distribution in The Netherlands in the nineteenth century and beyond, similar to earlier work by Piketty (2014) for France and other countries. The final data collection will also be shared with the research community.

References

(Peeters, 2024) Ruben L. Peeters & Amaury de Vicq. End of Life Wealth Portfolios in the Netherlands in 1921: The Memories Database. *Research Data Journal for the Humanities*, vol. 9, pages 1–12, 2024. ISSN 2452-3666.

(Piketty, 2014) Thomas Piketty. *Capital in the Twenty-First Century*. Harvard University Press, 2014.

KEYWORDS hand-written text recognition, vision language models, microdata

11:51–12:03

Evaluating the impact of source diversity for RAG in historical research

Ruhi Mahadeshwar, Andreas van Cranenburgh, Tommaso Caselli & Malvina Nissim

University of Groningen

Historical research increasingly benefits from large language models (LLMs) (González-Gallardo et al., 2024). However, LLMs are prone to factual inaccuracy, unreliability (Jaskulski et al., 2025), and biased interpretations of data (Stranisci et al., 2023). Retrieval-augmented generation (RAG) approaches (Lewis et al., 2020) have emerged as solutions, but may inadvertently perpetuate biased perspectives (Bal, 2009) embedded in historical collections.

We investigate how source diversity in RAG impacts perspective variation in historical question answering. We compile a multilingual corpus (English, French, Dutch) of historical documents from 1805–1819, spanning multiple countries and focus on Napoleon Bonaparte. We evaluate three Qwen3 models across ten questions using a multi-layered framework combining traditional metrics (BERTScore, ROUGE-L), frame semantics analysis (Minnema et al., 2022; Chanin, 2023), and syntactic profiling (Brunato et al., 2020).

Our results highlight that, while traditional similarity metrics suggest high semantic consistency, frame-semantic analysis exposes substantial perspective shifts. Baseline answers present “flattened” cross-lingual perspectives, whereas RAG introduces diversity. Critically, this diversity manifests differently across languages, demonstrating language-specific patterns. Our findings highlight limitations of traditional evaluation metrics for perspective-sensitive tasks and demonstrate that RAG constitutes as active perspective transformation rather than neutral augmentation

References:

Mieke Bal. 2009. *Narratology: Introduction to the theory of narrative*. UoT Press.

Dominique Brunato et al. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of the Twelfth LREC*, pages 7145–7151, Marseille, France. European Language Resources Association.

David Chanin. 2023. Open-source frame semantic parsing.

Carlos-Emiliano González-Gallardo et al. 2024. Yes but.. Can ChatGPT Identify Entities in Historical Documents? In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries, JCDL '23*, page 184–189. IEEE Press.

Piotr Jaskulski et al. 2025. Reliability of large language models as a tool for knowledge extraction from biographical dictionaries: the case of the Polish Biographical Dictionary. *Digital Scholarship in the Humanities*, 40(2):538–548.

Patrick Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*, volume 33, pages 9459–9474. Curran Associates, Inc.

Gosse Minnema et al. 2022. Responsibility framing under the magnifying lens of NLP: The case of gender-based violence and traffic danger. *CLIN Journal*, 12:207–233.

Marco Antonio Stranisci et al. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. ACL.

KEYWORDS Frame semantics, Linguistic profiling, Retrieval Augmented Generation, Large Language Models, Bias

12:03–12:15

DetectUA: AI Generated Text Detection in Dutch

Jens Lemmens, Jens Van Nooten, Walter Daelemans

University of Antwerp

This study compares the robustness of various methods for the detection of AI generated text in Dutch: a Support Vector Machine (classical machine learning), ModernBERT (fine-tuned encoder with large context window), and Binoculars (unsupervised method exploiting cross-perplexity). Specifically, it is investigated how these models perform during inference on out-of-distribution data. The latter includes genres, text generation models, and prompting strategies that were not seen during training.

Three datasets have been used for the presented work. First, the shared task dataset of CLIN33, which consists of 1,600 texts distributed across 4 genres. Secondly, we expand a subset of the existing CLiPS Stylometry Investigation corpus (CSI), which consists of reviews, with AI generated reviews. Finally, a collection of 15,000 news articles from De Standaard was collected, of which half was used as a basis to generate AI written news articles from.

The initial results show that supervised methods achieve F1-scores up to 97.5% on in-domain data for binary classification, and 92.4% for the prediction of the specific model used to generate the text. The results also show that introducing new generator models in the test set that have not been seen during training causes virtually no performance drops. Nevertheless, we observe that CLAUDE and llama remain more challenging to detect than the other utilized models (Qwen, Gemma, GPT-5, -nano, -mini, and -oss).

A significant accuracy reduction, however, can be observed in the supervised models when testing on new genres or prompting strategies, which mostly affects recall. Binoculars, however, provides a more stable result across datasets, but yields lower scores in general. The final model, a Support Vector Machine trained on all the data that was collected, is available for use in the DetectUA web interface where users can upload their data and obtain (paragraph-level) predictions. We invite interested parties to contact us if they would like to contribute to the project with additional data or have a specific use case they want to collaborate on.

KEYWORDS Large Language Models, AI generated text detection, Stylometry

12:15–12:27

Signs of Generative AI in Dutch Newspaper Articles

Sem Huisman & Rik van Noord

University of Groningen

Since the public release of ChatGPT in November 2022, generative AI has become part of news writing. By 2023, 85% of surveyed newsrooms reported experimenting with generative AI (Beckett and Yaseen, 2023), and journalists now use tools like ChatGPT across the journalistic workflow, from news gathering and drafting to distribution and moderation (Cools and Diakopoulos, 2026). While media companies' AI guidelines usually stress transparency, human oversight, and bias prevention (Becker et

al., 2025), they rarely prohibit AI use altogether. We ask the question: to what extent do Dutch newspaper articles published after the release of ChatGPT show measurable linguistic traces of generative AI?

To answer this question, we analyse 66,295 Dutch articles from six regional and national newspapers, published between December 2019 and December 2025. Our approach is twofold. First, we compare texts from before and after the release of ChatGPT, measuring changes in lexical diversity, lexical sophistication, LLM-associated words, and syntactic features. Previous studies have characterised the style of LLM-generated text in English, showing differences from human writing across lexical, syntactic, and structural dimensions. One well-known English example is the overuse of words such as “delve”, but LLM-associated words have been identified in 34 languages, including Dutch (Juzek, 2026), allowing us to analyse whether such overuse is also visible in Dutch news articles.

Second, we use several black-box AI-detection systems, including Pangram, on the full article collection. Although these detectors are imperfect, they can provide a useful signal at scale. Across our analyses, we find clear signs of AI influence on Dutch news texts. AI tools appear to have been adopted gradually from 2023 onward, followed by a sharp increase after June 2024, with no sign of levelling off. Finally, we examine whether these changes differ across topics, between regional and national newspapers, and whether they are driven by a small number of individual journalists.

Kim Björn Becker, Felix M. Simon, and Christopher Crum. 2025. Policies in parallel? a comparative study of journalistic AI policies in 52 global news organisations. *Digital Journalism*, 13(9):1578–1598.

Charlie Beckett and Mira Yaseen. 2023. Generating change: a global survey of what news organisations are doing with AI. Technical report, London, UK.

Hannes Cools and Nicholas Diakopoulos. 2026. Uses of generative ai in the newsroom: Mapping journalists’ perceptions of perils and possibilities. *Journalism Practice*, 20(3):878–896.

Thomas Stephan Juzek. 2026. AI-associated lexical shifts across 34 languages: Cross-lingual convergence and diachronic uptake in news writing. arXiv preprint arXiv:2605.25358

KEYWORDS generative AI, corpus linguistics, LLM-associated lexical shift, computational stylometry, longitudinal analysis

Multilingual & Dutch Language Resources and Tools

D.2.20 · Chair: Miryam de Lhoneux

11:15–11:27

What makes a tokeniser fair?

Sing Sing Ngai

KU Leuven

A fundamental question in multilingual NLP is: ‘What makes a tokeniser fair, and how can we measure fairness?’ This question is crucial because tokenisation sits at the very beginning of any language model pipeline, yet its downstream consequences are profound. The core argument is that tokenisation is not a neutral preprocessing but a consequential design choice that systematically disadvantages speakers of non-dominant languages before a language model even sees a single gradient. We propose a four-layer definition of fairness. Parity (Petrov et al. 2023) requires that

parallel texts of equal informational content yield comparable token counts—the ‘tokenisation premium’, ideally stands at 1.0. Representational adequacy concerns whether tokens can be more linguistically meaningful than fragmented bytes, measured by subword fertility (Rust et al. 2021), average rank and characters per token (Limisiewicz et al. 2023). Cross-lingual equity of transfer addresses vocabulary overlap, a double-edged property that aids sentence-level tasks (NLI, retrieval) but harms word-level tasks (POS, dependency parsing), thus requiring task-sensitive calibration (Limisiewicz et al. 2023), and it also depends on high or low resource languages (Patil et al. 2022, Itkonen et al. 2026). Morphological coherence—the deepest layer—demands that units correspond to comparable meaningful constituents, e.g. morphemes; and MYTE encoding (Limisiewicz et al. 2024) replaces UTF-8 byte assignment with morpheme-based mapping. The proposed methodology rests on parallel, non-English-centric corpora (FLORES-200, Bible corpus, EU/UN repositories, OPUS database, CC100), multi-dimensional and task-sensitive evaluation, and typological stratification by script family, resource level and morphology. Seven metrics span all layers: tokenisation premium, parity Gini coefficient, subword fertility, characters per token, average rank, JSD of token distributions and MYTE compression rate. A five-step protocol moves from corpus/tokeniser preparation through intrinsic evaluation, extrinsic downstream testing on BERT-scale models, Spearman correlation analysis and an unseen-language generalisation test. Potential strengths of this methodology include comprehensiveness across all fairness layers, grounding in standardised parallel corpora, revealing typological patterns, task diversity, and interpretability with clear remediation paths. Potential weaknesses include bias toward formal written registers, uneven language coverage and survivorship bias, MYTE’s segmentation limitations, the need for numerous data points and JSD variance for low-resource languages. Our tokeniser evaluation will include these factors. Future directions span dialect/register benchmarks, context-sensitive metrics, unseen-script generalisation, economic and accessibility impact assessment, morphological ground-truth validation and scaling-law, and ultimately, participatory evaluation involving native-speaker communities.

KEYWORDS tokenisation, parity, tokeniser fairness

11:27–11:39

QQ: A Toolkit for Language Identifiers and Metadata

Wessel Poelman, Yiyi Chen & Miryam de Lhoneux

KU Leuven, Aalborg University

The growing number of languages considered in multilingual NLP, including new datasets and tasks, poses challenges regarding properly and accurately keeping track of which languages are used and how. For example, datasets often use different language identifiers; some use BCP-47 (e.g. en_Latn), others use ISO 639-1 (en), and more linguistically oriented datasets use Glottocodes (stan1293). Mapping between identifiers is manageable for a few dozen languages, but becomes difficult when dealing with thousands. We introduce QwanQwa, a Python toolkit for unified language metadata management. QQ integrates seven metadata sources into a single interface, provides convenient normalization and mapping between language identifiers, and affords a graph-based structure that enables traversal across families, regions, writing systems, and other linguistic attributes. QQ serves both as (1) a tool for multilingual NLP research to make working with many languages easier, and (2) as an intuitive way for exploring languages, such as finding related ones through shared scripts,

regions or other metadata. The demo website is available here: <https://wesselpoelman.nl/qq/> and the source code here: <https://github.com/WPoelman/qwanqwa>

KEYWORDS multilingual, language resources, tooling

11:39-11:51

Grammatical Error Detection and Reporting in SASTA

Jan Odijk

Utrecht University

SASTA (<https://sasta.hum.uu.nl/>) is a web application that automates the analysis of transcripts of spontaneous language sessions in Dutch. It supports established methods such as TARSP (for children 1-4, [Schlichting 2017]), STAP (for children 4-8 [van Ierland et al. 2008, Verbeek et al. 2007]) and ASTA (for patients with aphasia [Boxum et al. 2013]). Until recently, the focus of SASTA was on the grammatical analysis of such transcripts. However, some methods also require detection and reporting of grammatical errors. Some work on grammatical errors was already done, because the grammatical analysis required this [Odijk 2021, Odijk et al. 2026]. Here we describe how we detect grammatical errors and report on them in annotation forms and profile charts. Grammatical errors dealt with include subject-verb agreement errors, absent determiners, wrong determiners, wrong adjectival agreement, absent or wrong auxiliaries, omitted word groups, omitted or wrong 'er', overgeneralized inflection and other morphological errors, and several others. We will compare the output of SASTA with reference data, and indicate what kind of errors are still very difficult to deal with properly.

References [Boxum et al. 2013] Elsbeth Boxum, Fennetta van der Scheer, and Mariëlle Zwaga. 2013. Analyse voor spontane Taal bij Afasie. Standaard in samenwerking met de VKL. VKL, October. <https://klinischelinguistiek.nl/uploads/201307asta4eversie.pdf>. [van Ierland et al. 2008] Margreet van Ierland, Jeannette Verbeek, and Leen van den Dungen. 2008. Spontane Taal Analyse Protocol. Handleiding van het STAP-instrument. UvA, Amsterdam. Odijk, J. (2021). Towards Semi-Automatic Analysis of Spontaneous Language for Dutch. In Selected papers from the CLARIN Annual Conference 2020 (Vol. 180, pp. 165-175). (Linköping Electronic Conference Proceedings). Linköping University Press. <https://doi.org/10.3384/ecp18018>

[Odijk et al. 2026] Jan Odijk, Jelte van Boheemen, Xander Vertegaal, Tessel Boerma and Marijn Schraagen. 2026. 'SASTA Self Assessment: An efficient human-in-the-loop strategy for developmental and pathological language analysis'. In Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, p. 103-112. <http://lrec-conf.org/proceedings/lrec2026/workshops/clinicalnlp/2026.clinicalnlp-1.0.pdf> [Schlichting 2017] Liesbeth Schlichting. 2017. TARSP: Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar met aanvullende structuren tot 6 jaar. Pearson, Amsterdam, 8th edition. [Verbeek et al. 2007] Jeannette Verbeek, Leen van den Dungen, and Anne Baker. 2007. Spontane Taal Analyse Protocol. Verantwoording van het STAP-instrument, ontwikkeld door Margreet van Ierland. UvA.

KEYWORDS grammar error detection, child language, syntax, morphology, aphasia

11:51-12:03

Automated essay scoring for L2 dutch

Kshitij Malatpure, Joni Kruijsbergen, & Orphée De Clercq

KU Leuven, Gent Universiteit

The field of Automated Essay Scoring (AES) has grown significantly over the past two decades (Li&Ng, 2024), with greater progress being made in languages beyond English in recent years (eg. Seßler et al., 2024). For Dutch as a second language(L2 Dutch), existing work has focused on binary (pass/fail) problems (Chen 2024). This study builds on this foundation by introducing an ordinal-aware, cross-prompt, multi-trait AES framework for L2 Dutch. We utilize a dataset of 2136 B2-level essays drawn from the Certificaat Nederlands als Vreemde Taal (CNaVT) corpus. Given the constraints inherent in working with a low-resource language such as L2 Dutch and the limitations and inconsistencies present in the available corpus, this study focuses exclusively on non-content-based traits. We investigate approaches to assess four key linguistic traits: Vocabulary, Grammar, Cohesion, and Language Technique, and an aggregate score for these traits. We evaluate the performance of source-available Large Language Models (LLMs), and we compare them to a range of feature-based approaches with features extracted using T-scan (PanderMaat, 2014) and LanguageTool (Naber, 2003). Here, we experiment with multi-class classifiers, regression models with thresholding, and finally, an ordinal-aware FabOF (Buczak, 2024) classifier using a variety of feature selection setups. Models are evaluated using QWK and macro-F1 scores to address the significant class imbalance in the dataset as well as the ordinal nature of the scoring systems. The results strongly favour feature-based approaches, with a 46% relative improvement in the aggregate macro-F1 score for the best feature-based model when compared to the best-performing LLM. We further apply SHAP analysis to identify the direction and magnitude of each feature's contribution per trait. This provides a foundation for interpretable, trait-level feedback systems that can give learners actionable insight into their specific writing weaknesses.

References Buczak, P. (2024). fabof: A novel tree ensemble method for ordinal prediction. OSF Preprints h8t4p, Center for Open Science. Chen, S. (2024). Automated pass/fail classification for dutch as a second language using large language models. Master's thesis, KU Leuven, Faculty of Engineering Science, Leuven, Belgium. Master of Artificial Intelligence, option Speech and Language Technologies. Shengjie Li and Vincent Ng. 2024. Automated Essay Scoring: A Reflection on the State of the Art. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics. Naber, D. (2003). A rule-based style and grammar checker. Diplomarbeit, Universität Bielefeld, Technische Fakultät, Bielefeld, Germany. Seßler, K., Fürstenberg, M., Bühler, B., and Kasneci, E. (2024). Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring.

KEYWORDS Automated Essay Scoring, educational nlp, L2 Dutch

12:03–12:15

Extending Dutch benchmarks in EuroEval

Simone van Bruggen, Collin Aldaibis, Jean Paul Dingemanse, Eveline Kalff, Anne Fleur van Luenen, Thomas van Osch, Martijn Spitters, Leandra Swiers, Hannah Tops & Edwin Rijgersberg
SURF, NFI, TNO

Developing large language models requires diverse, high-quality benchmarks to evaluate their performance. Although the number of Dutch benchmarks is growing, they do not always address cultural knowledge specific to the Dutch language and their quality varies due to a lack of manual review. This leads to a misalignment between benchmark scores and real-life usefulness, and is a challenge when developing Dutch or multilingual LLMs. Benchmarks are also typically hosted independently with their own codebases, hindering reproducible and comparable evaluation. In response to similar problems across many European languages, the open-source EuroEval framework [1] seeks to create a standardized, reproducible suite of benchmarks for all European languages.

As part of GPT-NL, an initiative by TNO, NFI and SURF to develop a sovereign, transparent, and ethically driven Dutch LLM, we extend EuroEval by introducing simplification and bias as new task types and contributing Dutch benchmarks, including a new dataset on Dutch proverbs and existing datasets for simplification, bias, and reasoning. We release all code and datasets under permissive open-source licenses.

We manually created and reviewed the proverb dataset, pairing short scenarios with two proverbs: one applicable to the situation and one that is not. To evaluate simplification, we use a subset of the Duidelijke Taal dataset [2], drawing on its crowd-sourced annotations to select the most distinct and accurate pairs of complex and simple sentences. Bias is assessed through MBBQ [3], a curated multilingual version of the Bias Benchmark for Question-answering that includes Dutch, from which we also adopt the bias metric for multiple-choice stereotype questions. Finally, for common-sense reasoning we integrate the COPA-NL dataset from the DUMB benchmark [4]. We evaluate a set of multilingual and monolingual Dutch LLMs on these benchmarks.

We present our approach and considerations for adopting these datasets, our findings, and opportunities for new Dutch benchmark datasets in the future, as well as the challenges involved in accessing and publishing these datasets.

[1] Nielsen, D. (2023) ScandEval: A Benchmark for Scandinavian Natural Language Processing. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library. [2] Vandeghinste, V., Vanroy, B., & van Doeselaar, J. (2025). Human evaluation of automated text simplification through crowdsourcing. In CLARIN Annual Conference Proceedings (pp. 143-147). CLARIN. [3] Neplenbroek, V., Bisazza, A., & Fernández, R. (2024). MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. arXiv preprint arXiv:2406.07243. [4] de Vries, W., Wieling, M., & Nissim, M. (2023, December). Dumb: A benchmark for smart evaluation of Dutch models. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 7221-7241).

KEYWORDS large language models, benchmark, Dutch

12:15–12:27

Synthetic Monolingual Data for Back-Translation: Which Properties Predict Translation Quality

Thomas Moerman, Els Lefever, Arda Tezcan

LT3, Ghent University, Belgium

Fuzzy match (FM) augmentation improves neural machine translation by retrieving similar translations from existing parallel corpora to guide the translation of new source sentences, and it has proven particularly effective in domain-specific scenarios. Recent work has shown that FM augmentation can be combined with back-translation when monolingual target-language data is available, and that, when such data is scarce, the missing monolingual resources can be synthesised with large language models (LLMs): the generated target text is back-translated to produce synthetic source-target pairs for FM augmentation. This removes a hard data-availability constraint, but it introduces a new and largely unexamined question. Because synthetic monolingual data can be generated with arbitrary prompts, seeds, and sampling settings, a generator can produce text that varies widely in vocabulary, structure, domain proximity, and naturalness. Which of these properties actually make the generated data useful for back-translation augmentation?

It is commonly assumed that more diverse or more in-domain target-language data yields better augmentation, but for synthetic data used in back-translation this assumption has not been tested directly. Understanding which measurable properties of a generated corpus predict downstream translation quality would let practitioners design generation recipes deliberately, rather than by trial and error, and would clarify how far synthetic data can substitute for real in-domain text.

This work investigates that question by generating synthetic monolingual data with an instruction-tuned LLM across a range of prompting and sampling regimes, and characterising each synthetic source along several axes. These include lexical, syntactic, and semantic diversity, proximity to the target domain, coverage of the evaluation distribution, and general-language fluency. We relate these measurements to downstream translation quality, evaluated with both BLEU and COMET, through the full FM augmentation and back-translation pipeline with a fine-tuned translator. To assess how well these relationships generalise, we evaluate across multiple language pairs and several specialised domains of varying resource availability.

The result is a characterisation of what does and does not predict the back-translation value of synthetic data, distinguishing the properties that genuinely drive downstream gains from those that are commonly assumed to but do not, with practical implications for generating synthetic resources for low-resource and domain-specific machine translation.

KEYWORDS machine translation, synthetic data, data augmentation, back-translation

AI Policy, Governance & Societal Impact

D.2.18 · Chair: Frieda Steurs

11:15–11:27

Presenting the Action Plan AI for the Dutch Language

Vincent Vandeghinste, Bram Vanroy & Suzan Verberne

Instituut voor de Nederlandse Taal, KU Leuven, CLARIN-ERIC, Leiden University

In this talk we present the Action Plan Artificial Intelligence for the Dutch Language, which is a blueprint for a sovereign and sustainable AI ecosystem for Dutch. This plan was written on the request of the Nederlandse Taalunie, the intergovernmental organisation for Dutch language policy representing the Netherlands, Flanders, and Suriname, by a task force on AI for Dutch. The task force consisted of Tanguy Coenen, Saskia Lensink, Antal van den Bosch, Catia Cucchiarini, Walter Daelemans, Martijn Kleppe, Roeland Ordelman, Vincent Vandeghinste, Hugo Van hamme, Bram Vanroy and Suzan Verberne. The action plan describes an integrated approach to enable cooperation between the Dutch-speaking regions and to set common priorities for work on AI for Dutch.

The action plan sets a shared agenda and maps the needs and opportunities within five coherent pillars: 1. Data: the systematic disclosure and legal safeguarding of representative text, speech and multimodal data; 2. Infrastructure: setting up and maintaining secure and scalable facilities in which data can be accessed in a responsible manner and models can be developed and used reproducibly; 3. AI development: testing, managing and continuously improving models with attention to quality, bias and public values from the Dutch-speaking area; 4. Valorization: translating technologies into concrete applications in public and private sectors that strengthen economic innovation and social participation; 5. Knowledge sharing: sharing knowledge about AI for Dutch through an expertise center, where policymakers, companies and the general public can be consulted.

For the CLIN audience, pillar one and three are of particular importance, as they focus on the research and development agenda for AI for Dutch. The action plan describes a continuous and testable development cycle for Dutch-language AI models, grounded in robust benchmarking, representativity, and public values. It emphasizes the need for authentic Dutch benchmarks that evaluate linguistic competence, cultural knowledge, and conversational language use across the Dutch-speaking regions. The plan advocates building on existing multilingual and Dutch models through continued pre-training, post-training and fine-tuning, with explicit attention to regional, social and ethnic language variation. In addition, the plan highlights the importance of transparency, bias mitigation, ethical governance and digital sovereignty, positioning AI development for Dutch as a shared public infrastructure aligned with the linguistic and cultural diversity of the Dutch-speaking community.

KEYWORDS infrastructure, data, policy, benchmarking, GenAI

11:27-11:39

Six Dimensions of Linguistic Complexity in Corporate Disclosure

Guy Mathys , Geoffrey Aerts , Kris Boudt

SMIT-imec, Department of Applied Economics, FARI – AI for the Common Good Institute, Solvay Business School, Vrije Universiteit Brussel

Scalar readability measures are widely used in accounting research to proxy for disclosure quality. However, they often produce inconsistent results across studies. A single readability score cannot capture the different ways in which readers make sense of a text. This paper defines linguistic complexity as a construct with six theoretically grounded dimensions: lexical sophistication, syntactic organisation, semantic dispersion, discourse cohesion, sequential unpredictability, and epistemic complexity. The framework is tested using management and auditor reports from 115,840 Belgian annual accounts in Dutch and French, covering 2019 to 2023. A stratified sample of 55,460 document-year observations is analysed separately by language. Six dimensions are confirmed by parallel analysis and the Kaiser criterion. These dimensions vary systematically across document types, institutional periods, and economic sectors. The dimensions show different associations with economic outcomes. The associations with the implicit cost of debt are language-specific: syntactic organisation is the dominant predictor in Dutch, whilst epistemic complexity and lexical sophistication drive the association in French. Scalar readability measures align primarily with syntactic organisation and suppress this structure by construction. This suggests that prior findings capture one mechanism while concealing others. These results indicate that linguistic complexity is inherently multidimensional. Aligning measurement with the underlying mechanism is a necessary condition for construct validity.

KEYWORDS linguistic complexity; readability; corporate disclosure; textual analysis; construct validity; natural language processing

11:39-11:51

Which price are professionals willing to pay for confidentiality? Processing confidential data with SLM's

Michael Bauwens, Heike Pauli & François Remy

UCLL, Parallia

“Where can I download Ollama to use GenAI on my HR data?” This question which we’ve received seems simple, but it captures a broader challenge many organisations are currently facing: generative AI offers clear opportunities for business processes, but its adoption becomes far more complex when sensitive organisational data are involved. In HR contexts, this tension is especially visible: professionals want to work more efficiently with personnel, contract, payroll, sickness and leave data, while also respecting confidentiality, internal governance and GDPR requirements. Small Language Models (SLMs) offer a viable alternative to cloud-based Large Language Models (Vaes, 2024), but their feasibility remains a question for non-technical teams.

The research focus in this project lies on professionals without deep technical expertise who want to use AI on internal data (mainly in Dutch) while retaining control over governance in order to achieve more sovereignty in data and AI (Accenture, 2025). Rather than only asking how local models perform, we ask a broader and more practice-oriented question: what price are professionals willing

to pay for confidentiality? That price may take many forms, including lower output quality, slower performance, additional technical complexity, limited functionality, or extra investment in hardware, training of personnel, maintenance and data governance.

As a first pilot case study, we've provided our own HR department with a demo solution where confidential exports from operational systems are processed in internal file environments with the help of SLMs, using RAG and text-to-SQL operations. The solution as such is not new, so our research focus mainly lies on investigating adoption barriers and organisational conditions: what kinds of workflows are realistic, what level of technical knowledge is required, and which governance measures are needed to deploy such systems responsibly in practice.

The second case study is situated in healthcare, more specifically in real-time medical environments, where local AI infrastructure may offer advantages in terms of latency, reliability and data control. This contribution provides insight into the practical opportunities and limitations of privacy-conscious SLM workflows in applied settings. This way, it frames confidentiality as an important yet complex design choice when applying GenAI responsibly in organisations.

References: Accenture. (2025). Sovereign AI: From managing risk to accelerating growth. Vaes, M. (2024). Aan de slag met lokale, privacyvriendelijke Large Language Models. Kenniscentrum Data & Maatschappij.

Disclosure of use of generative AI: During the preparation of this abstract, the authors used Copilot for Microsoft 365 for the purpose of rephrasing and rewriting of the text. The authors reviewed and edited the content as needed and take full responsibility for the content of the abstract.

KEYWORDS Small Language Models; confidential data; local AI; sovereign AI; applied research

11:51-12:03

Unlearning Hazardous Knowledge in Attention-Linear-Recurrent Hybrids

Alexandre Le Mercier, Pranaydeep Singh

Ghent University

Reliable removal of hazardous knowledge is crucial for the safe deployment of large language models, which can otherwise operationalize dual-use expertise relevant to chemical, biological, and nuclear weapons design and illicit drug synthesis. Inference-time guardrails remain brittle to jailbreaks because the underlying knowledge persists in the model weights; hence machine unlearning, which excises such knowledge at the parameter level, has emerged as a more principled remedy. However, existing localization and unlearning techniques have been developed and validated almost exclusively on dense Transformers, even though frontier systems increasingly adopt hybrid architectures that interleave full self-attention with sub-quadratic linear-recurrent layers (e.g., Nemotron-3, Qwen, Jamba, Falcon-H, etc.). How hazardous knowledge is stored, localized, and removed in such hybrids is, to the best of our knowledge, entirely unstudied.

In this paper, we present the first systematic study of hazardous-knowledge localization and unlearning in hybrid linear-attention models. We use the five-size Qwen family (0.8B-27B), whose blocks combine Gated DeltaNet linear-attention layers with periodic full-attention layers, as a controlled testbed spanning more than an order of magnitude in scale under a single fixed architecture. Our contributions are as follows. (1) We localize hazardous knowledge across layer types

and depth using sparse autoencoder (SAE) features, causal tracing, activation patching, layer-type knockout, and probing classifiers, and we show whether such knowledge concentrates in the full-attention layers or is distributed into the recurrent state of the Gated DeltaNet layers despite their improved associative recall. (2) We adapt and benchmark selective feature gating against gradient-based forgetting and targeted weight interventions, evaluating removal efficacy on hazardous-knowledge benchmarks alongside the preservation of general utility. In sum, our study establishes how architectural composition governs the localization and removability of hazardous knowledge, laying a foundation for safety interventions across the broad and growing class of attention-linear-recurrent language models.

References:

[1] A. Sen Sharma, D. Atkinson, and D. Bau. "Locating and Editing Factual Associations in Mamba." COLM, 2024. arXiv:2404.03646. (The closest precedent — localization on a state-space model — defining the gap your hybrid study extends.) [2] S. Yang, J. Kautz, and A. Hatamizadeh. "Gated Delta Networks: Improving Mamba2 with Delta Rule." ICLR, 2025. arXiv:2412.06464. (The Gated DeltaNet linear-attention mechanism in the Qwen blocks; the "improved associative recall.")

KEYWORDS Machine unlearning, Mechanistic interpretability, Hybrid language models, AI safety, Linear-attention models

12:03–12:15

Cross-lingual consistency as a groundedness guardrail for multilingual RAG

Eduardo Brito

Brito Chacón SRL/BV

Multilingual retrieval-augmented generation (RAG) systems must ensure factual correctness across all supported languages, yet verifying answers at inference time is expensive: a naive approach requires one LLM judge call per language, multiplying cost linearly with the number of languages served. However, many real-world deployments (e.g., for EU institutions publishing in 24 official languages, federal agencies in multilingual countries like Belgium or Switzerland, or commercial services offering dual-language interfaces) require parallel multilingual outputs by policy, creating an opportunity for low-cost quality control.

We propose measuring cross-lingual LaBSE consistency (Feng et al., 2022) as a groundedness guardrail at two stages of the multilingual RAG pipeline. The input is a set of equivalent queries in the supported languages. At the retrieval stage (pre-generation), the top-1 retrieved passage per language is embedded with LaBSE. Low cross-lingual passage consistency signals divergent retrieval and triggers abstention before any LLM call. At the generation stage (post-generation), the same metric is applied to generated answers, adding only a single LaBSE encoding pass. The signal rests on differential retrieval failure: when retrieval succeeds in some languages but fails in others, grounded and ungrounded answers diverge in the shared cross-lingual embedding space. When retrieval fails uniformly across all languages, answers converge on the same wrong content (the guardrail's principal blind spot).

We evaluate on MuPLer (Steinberger et al., 2014; Enevoldsen et al., 2025), a parallel multilingual legal retrieval benchmark built on EUR-Lex, using a four-language subset (English, French, Dutch, and Spanish) chosen so that all generations and judge verdicts may be manually verified. We compare

three retrieval configurations: in-language (C1), English-pivot (C2), and mixed (C3), with Aya Expanse 8B and GPT-4o-mini as generators (2,400 answers each). Correctness judgments were validated on a 100-example human-annotated sample.

The pre-generation guardrail achieves AUROC 0.69–0.72 with no LLM call. The post-generation guardrail amplifies the retrieval-stage signal, achieving AUROC 0.74–0.87. Both stages substantially outperform retrieval confidence baselines (top-1 cosine score and score margin, AUROC 0.63–0.75), indicating that the signal captures information beyond standard retrieval-confidence measures. Spearman ρ between the retrieval- and answer-level signals is 0.36–0.54 (all $p < 10^{-4}$), and the $C1 \geq C3 > C2$ ordering holds at both stages, consistent with the mechanism originating at the retrieval step. At 10% abstention (flagging the 10% lowest-consistency queries for review rather than answering automatically), the guardrail identifies wrong answers at 90% precision, making it a practical quality-control tool for the multilingual legal and public-sector services.

KEYWORDS multilingual retrieval-augmented generation, cross-lingual consistency, groundedness guardrail, legal NLP, parallel text evaluation

12:15–12:27

Applicability of the ValuesML human value detection model on student progress monitoring

Franceina van Zalk, Maya Sappelli & Renate Wesselink

HAN/ WUR, HAN (Hogeschool Arnhem Nijmegen) en WUR (Wageningen University and Research)

Values are widely recognized as key drivers of human behaviour, societal transformation, and sustainability transitions but methods—such as surveys, self-assessment or manual coding—are limited in scale, reproducibility and are sensitive to bias. Therefore we use ValuesML (Legkas, et al. 2024), developed for the CLEF 2024 shared task on value detection. The ValuesML model was, however, trained on multilingual texts from political sources, reflecting mostly political stances. In addition, the authors found a large class imbalance of the data. The applicability of the model outside this context has yet to be established. In our work we aim to analyze values expressed by students in student material to assess their progress in learning.

We applied the model on longitudinal collected dataset of 323.797 Dutch texts (400–600 words) coming from 96 students of the master Circular Economy, written as part of a personal development programme. A subset of 200 texts was manually annotated by 2 annotators in order to establish reliability of model classifications. This subset was selected by sampling texts based on ValuesML classification scores to ensure 10 examples for each of the 19 values (selection criteria confidence > 0.8 or > 0.55). In addition, 10 samples with no values were selected to establish whether the model correctly identified these as well. Classification results showed that with a threshold of > 0.8 , 3.6% of the texts are classified as value laden, attained or constrained. In our data the value Universalism: Nature attained was most prominently present (22.2%), followed by Self-direction: action attained (13.6%) and Self-direction: thought attained (9.8%). No or minimal classification (both attained or constrained) of Tradition, Humility, Security: personal, Face. These results align with theoretical expectations of values and relations between values occurring in student texts.

An interesting aspect was that humility was highly prevalent in the full dataset but with lower confidence scores (around 0.5) suggesting frequent but semantically weak occurrences. This could be

a side effect of the original ValuesML dataset as humility was least represented there (151 examples; Legkas et al., 2024). This discrepancy underscores a critical challenge in computational value detection: some values manifest explicitly in text, whereas others remain implicit, necessitating context-aware interpretation of model outputs.

Literature Legkas, S., Christodoulou, C., Zidianakis, M., Koutrintzes, D., Petasis, G., Dagoglou, M.: Hierocles of alexandria at touché: multi-task & multi-head custom architecture with transformer-based models for human value detection. In: Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org (2024)

KEYWORDS Value Detection, ValueML, Sustainability Education

Oral session 2

14:00–15:25 · D building

Reasoning, Structure & Grounding

D.2.10 · Chair: Anaïs Tack

14:00–14:12

From human annotators to LLM personas: Interpretation of English scalar implicatures across different L1 backgrounds

Namrah Zaman, Sebastian Schuster & Marie-Catherine de Marneffe

UCLouvain, University of Vienna, FNRS

English NLI benchmarks, such as SNLI or MultiNLI, are crowdsourced, but annotators' first language (L1), English proficiency, and cognitive profile are usually not treated as main variables. It remains unclear whether speakers with different linguistic backgrounds interpret a given English NLI item the same way. Unlike XNLI, which translates items into multiple languages, this study keeps all items in English and varies annotators' language background. We focus on scalar implicatures, a pragmatic phenomenon open to different interpretations: if a premise says that a stone looks “good” and the hypothesis says it looks “great”, some annotators may infer “good but not great” and reject the hypothesis, while others may treat “good” and “great” as compatible or leave the stronger interpretation open.

We re-annotate 200 English SIGA items (Nizamani et al. 2024), targeting 5 pairs of scalar adjectives (good/great, small/tiny, uncomfortable/painful, uncommon/rare, possible/practical), gathering 25 annotations per item, 5 for 5 different L1: English, French, German, Urdu and Chinese. These languages display different semantic and syntactic features, which may influence the interpretation of scalar implicatures in English. In Urdu for instance, “great” is the same adjective as “good” quantified by a degree adverb (~ “very good”). Does the L1 structure impact the implicature in English? To separate lack of evidence from genuine interpretive ambiguity (Nighojkar et al. 2023), annotators are asked to choose between 4 labels (instead of the standard three-way NLI scheme): true, false, cannot be determined, and can be true or false depending on interpretation. Annotators also provide an explanation of their label. English proficiency is determined via IELTS/TOEFL scores or a short English proficiency test. Cognitive profile is assessed with the Verbal CRT and Need for Closure Scale, which capture annotators' tendency to override intuitive interpretation and tolerate ambiguity. We thus build a resource allowing us to examine whether scalar implicatures vary across language background (L1 and English proficiency) and cognitive profiles.

Second, we test whether frontier LLMs, when persona prompted for L1, English proficiency and cognitive profile, reproduce human labels and generate similar explanations. To evaluate explanation similarity, we use both quantitative measures (e.g. Jaccard) and a qualitative analysis. We show that LLMs still lack the ability to capture the full range of human language interpretations, especially when the L1 is low-resource.

R. Nizamani, S. Schuster and V. Demberg. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. Proceedings of LREC 2024. A. Nigohjkar, A. Laverghetta Jr. and J. Licato. 2023. No Strong Feelings One Way or Another: Re-operationalizing Neutrality in Natural Language Inference. Proceedings of LAW.

KEYWORDS Natural Language Inference; scalar implicatures; L1 background; cognitive profiling; persona-based LLM prompting

14:12–14:24

Step-by-step annotation of natural language reasoning

Lasha Abzianidze, Xander Vertegaal, Julian Gonggrijp & Ben Bonfil

Utrecht University

We present an online annotation environment for annotating natural language inference problems with step-by-step reasoning. The step-by-step reasoning is represented as a tree structure, an offshoot of a reasoning framework based on natural logic and semantic tableaux. Besides the inference steps, the lexical or world knowledge required for reasoning is also part of the reasoning tree.

The annotation environment facilitates the collection of comprehensive and structured explanations for natural language inference problems. The pipeline behind the annotation environment includes syntactic parsing in the style of Combinatory Categorical Grammar, followed by reasoning with a natural-logic-based tableau theorem prover. The initial, fully automatically generated reasoning can be modified through an interactive annotation environment.

The collected step-by-step reasoning trees can be exploited to design explainable natural language inference tasks with varying depths of structured explanations.

KEYWORDS annotation environment, natural language inference, structured explanation

14:24–14:36

Modelling Huai'an Mandarin tone changes and degrees of neutralization

Aleksei Nazarov

Utrecht University

In incomplete neutralization, a contrast collapsed in the surface phonology resurfaces in the phonetics. However, the degree of phonetic differentiation can vary with a competing linguistic task (Zeng et al. 2025) or pragmatic context (Nelson & Heinz 2025). We demonstrate numerically that scalable reference to lexical diacritics can account for a complex case of this, while also allowing for variable incompleteness. In Huai'an Mandarin (Du & Durvasula 2022), two T(one) 1 syllables in a row are changed to T3 T1, and two T3 syllables in a row are changed to T2 T3. A sequence like T3 T1 T1 is changed into T2 T3 T1, so T1 and T3 are truly neutralized in surface phonology, since derived T3 triggers avoidance of two T3 syllables in a row. However, the phonetic realization of derived and underlying T3 syllables is markedly different: a case of incomplete neutralization that escapes existing grammatical accounts (Van Oostendorp 2008, Braver 2019). Du & Durvasula suggest a processing account (cf. Zeng et al. 2025 for standard Mandarin) and Nelson & Heinz (2025) suggest a separate part of grammar that relates the lexicon and phonetics directly. I show the feasibility of a model (Maximum Entropy, Goldwater & Johnson 2003) that uses only the assumptions of a standard

grammar framework. The complexity of this case arises from special (phonetic/)phonological constraints that have access to diacritics associated with specific vowels/syllables in the lexicon (indices a la Round's 2017). By allowing these special constraints to be variably weighted, degrees of incomplete neutralization can be derived, as in Nelson & Heinz's (2025) account. Variable weighting is achieved through scaling factors (Coetzee & Kawahara 2013) and phonetic/phonological constraints are based on Braver (2019). References Braver, A. 2019. Incomplete neutralization as paradigm uniformity with weighted phonetic constraints. *Phonology* 36(1), 1–36. Coetzee, A., & Kawahara, S. 2013. Frequency biases in phonological variation. *NLLT* 31, 47–89. Du, N., & Durvasula, K. 2022. Phonetically incomplete neutralisation can be phonologically complete: evidence from Huai'an Mandarin. *Phonology* 39(4), 559–595. Goldwater, S., & Johnson, M. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, 111–120. Nelson, S., & Heinz, J. 2025. The blueprint model of production. *Phonology* 42, e12. Van Oostendorp, M. 2008. Incomplete devoicing in formal phonology. *Lingua* 118(9), 127–142. Round, E. 2017. Phonological exceptionality is localized to phonological elements: The argument from learnability and Yidiny word-final deletion. In *On looking into words (and beyond): Structures, relations, analyses*, 59–97. Zeng, Y., Chang, W., & Zhang, J. 2025. Cascading activation in spoken word production drives incomplete neutralization: An internet-based study of Mandarin 3rd tone sandhi. *Journal of Phonetics* 115, 101428.

KEYWORDS computational phonology, quantitative modeling, maximum entropy, variation

14:36–14:48

BERTje neemt een badje, but for how long? Studying event duration of Dutch light verb constructions

Lin de Huybrecht & Geraint A. Wiggins

Vrije Universiteit Brussel

According to psycholinguistic research, the perceived duration of an event is influenced by the syntactic construction used to describe it. Wittenberg & Levy (2017) found that punctive events in count syntax (e.g., give a kiss) and durative events in mass syntax (give advice), both in Light Verb Construction (LVC), are perceived as taking less time than when described by their corresponding Full Verb Construction (FVC; kiss and advising, resp.). Liu & Chersoni (2023) found similar results using computational methods. They semantically projected contextualised BERT embeddings onto a Duration scale (Devlin et al., 2019; Grand et al., 2022). Huybrecht & Wiggins (2026) further developed their work by instead using explicit word embeddings from a fully transparent co-occurrence count-based vector space. We expand this research by studying event duration of Dutch LVC-FVC pairs using embeddings from BERTje (de Vries et al., 2019). Dutch LVCs are distinguished from English LVCs by the use of diminutives for certain expressions (i.e., -je suffix), which can already emphasise the duration of the event being described. Some expressions have a non-diminutive alternative (een bad(je) nemen – take a (small) bath), while others do not (een praatje/*praat maken – to have a little chat). We semantically project 153 LVC-FVC pairs onto our duration scale. Using a right-tailed paired T-test ($\alpha=0.05$, $p<2.2\times10^{-16}$, $t(152)=9.74$, CI [0.51,+inf]), we find events in LVC are modelled as significantly shorter in duration than events in FVC. This preliminary work on Dutch data provides a starting point for gaining more insight into modelling event construal in Dutch.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model (arXiv:1912.09582).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proc. of the 2019 Conf. of the NAACL-HLT, Vol. 1, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nat. Hum. Behav., 6(7), 975–987. <https://doi.org/10.1038/s41562-022-01316-8>

Huybrecht, L., & Wiggins, G. A. (2026). How Long Does a Quick Kiss Take? Studying Event Duration of Light Verb Constructions Using Explicit Word Embeddings. Proc. of the 15th LREC. 9618–9634. <https://doi.org/10.63317/5gsno5o8o3ve>

Liu, C., & Chersoni, E. (2023). On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT. Proc. of the 6th BlackboxNLP Workshop (pp. 367–378). ACL. <https://aclanthology.org/2023.blackboxnlp-1.28>

Wittenberg, E., & Levy, R. (2017). If you want a quick kiss, make it count: How choice of syntactic construction affects event construal. J. Mem. Lang., 94, 254–271. <https://doi.org/10.1016/j.jml.2016.12.001>

KEYWORDS light verb constructions, event construal, language modelling

14:48–15:00

What makes linguistic representations good models of high-level visual perception in the human brain?

Anna Bavaresco, Ina Klarić, Raquel Fernández, Marie-Francine Moens

University of Amsterdam, KU Leuven

Withdrawn by the authors.

15:00–15:12

Co-creation and Feature Optimisation of Sign Language Data

Mirella De Sisto, Dimitar Shterionov, Lisa Lepp, Phillip Brown, Ifigenia Mavridou, Yves Duppen & Anastasja Rosanoff

Tilburg University

There is a growing emphasis on inclusion in natural language processing as well as in machine translation in recent years. Not only do we want language technology to support as many language communities as possible, it is also important to take all stakeholder perspectives into account. This is especially important for sign language research, where we are not just processing sentences or words in some written form, but rather the recordings of a human individual signing. Along with privacy and ethical considerations, this raises a question of the extent of human involvement in machine translation and natural language processing projects for sign languages. This paper presents the final findings of the CoCoS project and discusses key post-project considerations for the involvement of

community members in the collection of sign language data. We summarise our key findings and share practical solutions for data collection, and evaluation of sign language processing pipelines.

KEYWORDS Dimensionality reduction, sign language NLP, sign language data, sign language recognition, co-creation

15:12–15:24

Conception, Development, and Evaluation of a Visualisation and Exploration Tool for Skeleton-augmented Sign Language Data

Margaux Leleu, Adélaïde Couplet & Benoît Frénay

Université de Namur

Research in linguistics applied to sign languages and in machine learning and deep learning has in common the need for access to a large amount of data. Therefore, the two fields often have the opportunity to collaborate on this issue, allowing the creation of new data adapted to machine learning and deep learning from video data used in sign languages linguistics. In particular, at the University of Namur, a team specialised in machine learning and deep learning had the opportunity to work with a dataset made by a linguistic team of conversations held in LSFB, the LSFB corpus [1]. This collaboration resulted in two new datasets containing data adapted for machine learning and deep learning purposes [2]. This data takes the form of skeletons, extracted from videos of signers, thanks to a tool called MediaPipe [3]. These skeletons, although created for computer science purposes, are also of interest for research in linguistics. However, these new data are too specialised and difficult to handle for an audience who is not an expert in the field. The purpose of this work is to propose a solution to ease access to skeleton data. The proposed solution takes the form of a complementary tool to ELAN [4], a software that allows the annotation of videos and is currently the most used in sign languages linguistics. The tool proposed in this work, developed in collaboration with linguist experts in sign languages, allows for easy access to skeletons. It also offers data manipulation with visualisation, which allows users to become familiar with these new techniques. The tool proposed within this work was assessed through user studies with the participation of experts in the field.

[1] Laurence Meurant. Corpus LSFB: Un corpus informatisé en libre accès de vidéos et d'annotations de la langue des signes de Belgique francophone (LSFB), 2015. [2] Jerome Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. LSFB-CONT and LSFB- ISOL: Two New Datasets for Vision-Based Sign Language Recognition. Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN 2021), 2021. [3] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. [4] The Language Archive. ELAN (Version 7.1). Max Planck Institute for Psycholinguistics, Nijmegen, 2026. Computer software.

KEYWORDS sign language, LSFB, visualisation, data accessibility, user experience

Discourse, Communication & Language Evolution

D.2.16 · Chair: Katrien Beuls

14:00–14:12

From Argument Mining to Debate Analytics: Modeling Dynamic Interaction in Competitive Debates

Ali Al-Zawqari

Vrije Universiteit Brussel

Competitive debate combines sequences of claims and evidence with interactive exchanges in which speakers build cases, answer opponents, and repeatedly return to issues that shape the round. This makes debate difficult for standard argument-mining pipelines that treat argumentative spans mainly as isolated classification outputs.

I present a graph-based view of competitive debate as a step from argument mining to debate analytics. In this framing, argumentative spans are represented as nodes, enriched with speaker role, team, and position in the debate, while directed edges capture relations such as rebuttal and support under simple time and cross-team constraints. This representation makes it possible to ask interaction-level questions that are central to debate practice: who answered which argument, which issues persisted across speeches, and which claims became structurally important in the round.

The approach is grounded in competitive debate following the 3×3 Australia–Asia format, but the proposed framing is more general. It supports three forms of analysis: tracing rebuttal links between later responses and earlier opponent arguments, modeling how topics gain or lose momentum across speeches, and ranking arguments by their role in the evolving debate graph.

The broader goal is to move from “snapshot” argument mining toward dynamic debate understanding. Such models can support computational-linguistic analysis of argumentative interaction while also providing practical value for debate education, coaching feedback, and adjudication support.

KEYWORDS argument mining, debate analytics, graph-based NLP, rebuttal linking, educational feedback

14:12–14:24

Mining Implicit Causality in Climate Discourse

Liesbeth Allein

Universiteit Gent

Causality is key in climate change discourse. At the explicit level, causation is a profound framing device in statements on climatic processes and mitigation strategies. They typically report multiple cause-effect relations that intertwined in implicit and complex causal graph structures. At the implicit level, climate change discussions heavily rely on people's ability to form mental causal pathways to make sense of and ultimately (dis)agree with the reported causal relation between two events.

I will present collaborative work on implicit causal graph construction, causal chain discovery, and LLM causal reasoning in climate discourse [1, 2, 3]. Next to in-depth discussion of new causal discovery datasets, benchmarking experiments, graph construction methodologies and database-based evaluation frameworks, special attention will be given to the wider implications for argumentation mining and polarization studies.

[1] Allein, Liesbeth, Nataly Pineda-Castañeda, Andrea Rocci and Marie-Francine Moens (2026). “Assessing LLM Reasoning Through Implicit Causal Chain Discovery in Climate Discourse”. Proceedings of the 15th Language Resources and Evaluation Conference. [2] Allein, Liesbeth, Nataly Pineda-Castañeda, Andrea Rocci and Marie-Francine Moens (2026). “ClimateCause: Complex and Implicit Causal Structures in Climate Reports”. Findings of the Association for Computational Linguistics: ACL 2026. [3] Allein, Liesbeth and Marie-Francine Moens. “Implicit Causal Graph Construction in Text via Chain Discovery”. (Under review).

KEYWORDS Causal discovery from text, causal reasoning, climate change, argumentation mining

14:24–14:36

xLiMe-MELT: A Multi-layer dataset for Computational Affective Science using SFL-guided Sentiment Analysis

Lorella Viola

Vrije Universiteit Amsterdam

Affective meaning is often subtle, yet most classification frameworks collapse this complexity in to coarse polarity labels. As a result, heterogeneous affective phenomena such as caution, humour, mild sadness, or evaluative distancing are frequently conflated into the “neutral” class, for example in sentiment analysis. From the perspective of computational affective science, this ambiguity can reflect genuine interpretive plurality rather than annotation noise. We introduce xLiMe-MELT, a multi-layer dataset designed to capture affective nuance through linguistically grounded, explainable annotations. The dataset comprises 8,601 Italian language posts from X (formerly Twitter) annotated across multiple perspectives: human gold labels, a transformer-based sentiment model, a large language model using standard prompting, and the same model guided by Systemic Functional Linguistics (SFL) and Appraisal theory. In addition to polarity labels, the SFL-guided layer assigns fine-grained affective subtypes and provides explicit textual rationales anchored in interpersonal meaning cues. Our analysis shows that affective nuance is concentrated in instances traditionally labelled as neutral, and that linguistically grounded explanations help surface consistent evaluative signals even when categorical agreement is low. While overall classification performance remains comparable to standard approaches, the proposed framework substantially increases affective interpretability and coverage of nuanced meanings. We argue that theory-informed, explainable annotation frameworks such as SFL offer a valuable complement to accuracy-driven sentiment modelling, supporting more transparent and cognitively plausible analyses of affect in language. With over 34,000 annotations and explanations, xLiMe-MELT provides a rich resource for research on affective nuance, the interpretive gains of linguistic theory, and explainability in computational affective science. xLiMe-MELT is released open source.

KEYWORDS Systemic Functional Linguistics, Appraisal Theory, Computational Affective Science, Italian, Social Media

14:36–14:48

Emotional dynamics in political communication: Large-scale emotion annotation of politicians' social media posts and citizen responses

Luna De Bruyne, August De Mulder, Willem Buyens, Zeljko Poljak, Jonas Lefevere, Evelien Willems, Peter Van Aelst

Universiteit Antwerpen, Aarhus University

Recent advances in natural language processing have created new opportunities for studying political communication at scale. Yet, most computational research on emotions in politics has focused on sentiment or small annotated datasets, limiting our understanding of how specific emotions evolve over time and how citizens respond to them. This study presents a computational framework for analyzing emotional dynamics in political communication through large-scale emotion annotation of politicians' social media messages and their audiences' reactions.

We constructed a corpus of 239,409 Facebook posts published by 347 Flemish politicians between 2010 and 2024, together with engagement data and over 1.5 million citizen comments. To move beyond traditional sentiment analysis, we developed an annotation scheme covering nine discrete emotions: anger, fear, disgust, sadness, compassion, joy, hope, pride, and gratitude. In addition, citizen comments were annotated for stance (support versus criticism), enabling the study of emotional interactions between political elites and audiences while controlling for political agreement.

The annotation framework was developed through iterative manual coding and subsequently scaled using large language models (LLMs). We evaluated thirteen commercial and open-source LLMs against a manually annotated subset and selected models based on intercoder agreement with human annotators. The resulting annotation pipeline achieved human-level reliability and was used to automatically label the full corpus (GPT-5-mini for emotion annotation and GPT-5.4 for stance). This produced one of the largest emotion-annotated datasets of political communication currently available.

We demonstrate the usefulness of this resource through two applications. First, we analyze the longitudinal evolution of emotional expression in politicians' Facebook communication between 2010 and 2024. The results reveal a substantial decline in emotionally neutral communication and a marked increase in emotional expression over time. This increase is mainly driven by positive emotions, although also negative emotions (particularly anger) have increased among politicians from opposition and radical parties. Second, we investigate emotional dynamics between politicians and citizens by linking emotions in posts, comments, and reaction buttons. We examine emotional alignment and disagreement between politicians and citizens while distinguishing supportive from critical audience responses. The analyses show that: a) emotional posts generate significantly more emotional engagement; b) emotional posts lead to aligned emotion in the comments (but mostly on valence level) among supporters; and c) emotions are generally not aligned when the commenter has a critical stance towards the politician's post, except when the original post contains anger.

KEYWORDS emotion detection, computational social science, political communication, social media, LLM annotation

14:48–15:00

Struggling LLMs and creative AI: a corpus-linguistic study of agency attribution to “AI” in Digital Humanities research papers

Tess Dejaeghere, Els Lefever & Julie Birkholz

Universiteit Gent

Terminology surrounding artificial intelligence has undergone radical linguistic abstraction in recent years, shifting from technical to conceptual umbrella terms. This linguistic erosion risks obscuring the statistical realities of these models, allowing for anthropomorphist “folk theories” in the public and scientific debate (Bearman et al., 2023; Guest et al., 2026; Inie et al., 2026). We present a reproducible corpus-linguistic pipeline designed to map whether and how scientific communities linguistically frame, anthropomorphise and attribute agency to AI technologies. As a proof of concept, we apply this methodology to a dataset of research papers from the Digital Humanities (DH), a field balancing computational adoption and tool-criticism. Our pipeline is two-fold: first, a macro-analysis utilizes the OpenAlex API (Priem et al., 2022) to compile a diachronic corpus of scientific abstracts in DH (N=15,427, 2015-2025) to map broader temporal frequencies and normalization of AI lexicon. Second, a micro-analysis zooms in on a subset of available full texts (N=539 documents; 17,705 AI mentions). Using dependency parsing (en_core_web_trf) (Ines Montani et al., 2023), we extract sentences containing AI terminology and analyse three key linguistic features: grammatical agency (mapping syntactic roles such as active/passive subjects to measure modal responsibility), epistemic modality (tracking bare versus hedged assertions), and verbal valency (categorizing material, cognitive, and instrumental processes). Preliminary observations on the DH dataset indicate a rapid lexical normalization of umbrella terms over specific model architectures, accompanied by complex shifts in cognitive verb associations. Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, 86(2), 369-385. <https://doi.org/10.1007/s10734-022-00937-2> Guest, O., Suarez, M., Müller, B., van Meerkerk, E., Oude Groote Beverborg, A., de Haan, R., Reyes Elizondo, A., Blokpoel, M., Scharfenberg, N., Kleinherenbrink, A., Camerino, I., Woensdregt, M., Monett, D., Brown, J., Avraamidou, L., Alenda-Demoutiez, J., Hermans, E., & van Rooij, I. (2026). Against the Uncritical Adoption of “AI” Technologies in Academia. *Digital Culture & Education*, 16(2), 85-118. <https://doi.org/10.5281/zenodo.20082828> Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, & Henning Peters. (2023). explosion/spaCy: V3.7.2: Fixes for APIs and requirements (Versie v3.7.2) [Software]. Zenodo. <https://doi.org/10.5281/ZENODO.1212303> Inie, N., Zukerman, P., & Bender, E. M. (2026). De-anthropomorphizing “AI”: From wishful mnemonics to accurate nomenclature. *First Monday*. <https://doi.org/10.5210/fm.v31i2.14366> Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (arXiv:2205.01833). arXiv. <https://doi.org/10.48550/arXiv.2205.01833>

KEYWORDS artificial intelligence, corpus-driven discourse analysis, digital humanities

15:00–15:12

Cultural evolution drives the emergence of conventions in neural emergent communication

Maxime Toquebiau, Jérôme Botoko Ekila, Paul Van Eecke & Katrien Beuls

Vrije Universiteit Brussel, Université de Namur

Computational models of language emergence have been used since the 1990s to simulate theories on the origins of human languages in controlled environments (Steels, 1995; Cangelosi and Parisi, 2002). This approach is rooted in the view of languages as complex adaptive systems that evolve to fit the communicative needs of populations of individuals (Smith et al., 2003).

Recently, neural models of language evolution have been proposed, using populations of agents modelled as neural networks trained with deep learning techniques. However, most of these methods use unrealistic assumptions with regard to the cultural constraints that shape emergent languages. They propose models where agents can either speak or listen but not both (Chaabouni et al., 2022; Rita et al., 2022), where there are only two agents in the population (Cao et al., 2018; Choi et al., 2018), or where agents can share internal representations (Tucker et al., 2022; Gualdoni et al., 2024).

Therefore, we present a new neural model of language emergence that highlights the dynamics of cultural evolution in the emergence of linguistic conventions. This model consists of a population of bi-directional agents, i.e. they can both speak and listen. Agents take part in a referential game. They are trained in a fully distributed fashion with an imitation objective inspired by cultural learning in humans.

We demonstrate that this model is able to solve a language game in various domains, including image datasets, and with populations with up to 1000 agents. We show dynamics of cultural evolution at play in our model: more conventional languages are easier to learn, varying interlocutors and updating after each interaction leads to better languages, and learning to both speak and listen enables a co-adaptation process characteristic of cultural evolution.

Overall, this work provides a novel approach to neural emergent communication, more rooted in evolutionary linguistics, that opens up new directions for future work.

References: Cangelosi and Parisi, 2002. *Simulating the Evolution of Language*. Springer London. Cao et al., 2018. Emergent communication through negotiation. In *International Conference on Learning Representations (ICLR)*. Chaabouni et al., 2022. Emergent communication at scale. In *ICLR*. Choi et al., 2018. Multi-agent compositional communication learning from raw visual input. In *ICLR*. Gualdoni et al., 2024. Bridging semantics and pragmatics in information-theoretic emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*. Rita et al., 2022. On the role of population heterogeneity in emergent communication. In *ICLR*. Smith et al., 2003. Complex systems in language evolution: The cultural emergence of compositional structure. In *Advances in Complex Systems*. Steels, 1995. A self-organizing spatial vocabulary. In *Artificial Life*. Tucker et al., 2022. Trading off utility, informativeness, and complexity in emergent communication. In *NeurIPS*.

KEYWORDS Language evolution, Emergent communication

15:12–15:24

Cultural transmission, individual learning, and the emergence of compositionality: an agent-based model with Bayesian learners

Francijn Keur, Phong Le, Raquel G. Alhama

Universiteit van Amsterdam, University of St. Andrews

Compositionality, the principle that the meaning of a complex expression can be determined by the meaning of its constituents and the way in which they are combined, is an important property of human languages. How compositionality emerged, and how specific factors such as generational transmission, peer-to-peer communication, population size and social network structure affect this emergence, has been investigated using many different computational models and laboratory experiments (e.g., Kirby & Hurford, 2002; Kirby et al., 2008, 2015; Raviv et al., 2019). A limitation of this wide range of studies is that the results are difficult to compare due to the differences in methodology and underlying assumptions, which are not always made explicit.

To bridge this gap, we use a single framework, a Bayesian Iterated Learning Model, to systematically examine the impact of multiple factors on the emergence of compositionality, including cultural transmission as well as the assumed individual learning mechanisms. Specifically, we investigated the effects of transmission type (horizontal and vertical transmission), the rate of agent replacement, population size, social network structure —alongside the magnitude of assumed biases and the strength with which observed evidence shapes behaviour. An advantage of this framework is that the assumed biases of the agents are explicitly encoded in the form of the prior, making the underlying assumptions transparent. The two core assumptions are that the agents have a prior preference for more compressible languages (a simplicity bias), and the agents are pragmatic, meaning that they strive to be understood.

The results revealed that several previously reported effects replicate within this framework. Compositionality emerges due to a trade-off between a pressure for expressivity and compressibility. Furthermore, when agents learned from multiple teachers a relatively weak simplicity bias was sufficient for compositionality to arise. However, previously found effects of population size effects were not replicated. We argue that this might be attributed to the limited input variability in the current model, suggesting that a sufficiently rich signalling space may be necessary for these effects to emerge.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105 (31), 10681–10686.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). Springer.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proc. R. Soc. B*, 286 (1907), 20191262.

KEYWORDS Bayesian iterated learning model, compositionality, cultural evolution

Understanding & Improving Language Models

D.2.20 · Chair: Tim Van de Cruys

14:00–14:12

Interpretable n-gram tensor models

Stef Accou

KU Leuven

Large-scale transformer models have become the dominant paradigm for NLP applications, achieving high performance across a wide range of tasks. However, these models remain fundamentally difficult to interpret and steer due to their black-box nature. Tensor-based language models offer an interpretable alternative, but earlier work has been limited by severe scalability problems, resulting in restricted vocabularies or requiring important simplifying assumptions that undermine practical applicability. We revisit this tradition of linguistically informed language modelling in the light of current compute and algorithmic advances. Our approach exploits the fact that the vast majority of n-way co-occurrences in language are never attested to construct sparse n-gram tensors, and decomposes them using a new GPU-accelerated multiplicative update implementation for non-negative Tucker Decomposition. The decomposition results in factor matrices that can function as interpretable embedding lookup tables, alongside a core tensor that explicitly models the interactions between latent dimensions. As the decomposition is non-negative and the structure is interpretable by design, the resulting representations are easily inspectable and allow for direct, targeted steering. We present a range of evaluations of this interpretable alternative to black-box neural architectures, covering predictive performance and controlled generation alongside a breakdown of the scalability characteristics of the decomposition approach. Finally, we apply the method to creative language applications, demonstrating its performance in metaphor generation and interpretation.

KEYWORDS Tensor decomposition, Interpretable language models, Non-negative Tucker decomposition, n-gram language models

14:12–14:24

Olifant: An Eco-friendly Memory-based Language Model and Its Utility as a Speculative-Decoding Draft Model

Antal van den Bosch, Ainhoa Risco Patón, Maarten van Gompel, Peter Berck

Utrecht University, KNAW Humanities Cluster, Lund University

Large language models deliver strong next-token prediction at a high computational and ecological cost: they depend on GPU clusters, remain largely opaque, and consume substantial energy per token. We present Olifant, a memory-based alternative to neural language modeling, and argue that beyond being a viable stand-alone model, a relevant practical utility is as a draft model for speculative decoding.

Olifant implements language modeling as fast approximate k-nearest-neighbor classification using TiMBL's IGTREE algorithm: instances mapping a fixed-width token context to the next token are stored in an information-gain-ordered decision tree, allowing millisecond CPU lookup. It needs no GPU and no backpropagation, its internal workings are fully transparent, and next-token prediction accuracy scales roughly log-linearly with training data. Against GPT-2 and GPT-Neo, Olifant reaches competitive

accuracy at a fraction of the estimated emissions and with far lower latency (Van den Bosch et al., 2025).

Building on this, we use Olifant as the draft model in speculative decoding, where a cheap draft proposes k tokens that an expensive verifier (Mistral-7B) accepts or rejects in a single forward pass (Leviathan et al., 2023). Unlike neural draft models, Olifant runs as a CPU process; unlike n -gram models, IGTREE generalizes across similar-but-not-identical contexts. Across general, medical, and legal text, domain-matched Olifant variants reach up to 38.1% token acceptance, and speedups of 2.8x without KV caching and 1.3x with. Because the CPU draft adds virtually no extra power, these gains convert directly into energy savings: Olifant cuts energy per 100 tokens by 55% versus plain generation and is 1.8x more energy-efficient than a TinyLlama-1.1B draft model on the same task.

Together these results position Olifant as a scalable, fast, and eco-friendly language model of which the memory-based design is not only viable on its own but also turns the draft-model bottleneck of speculative decoding into an opportunity for lossless, low-energy LLM acceleration.

References

Van den Bosch, A., Risco Patón, A., Buijse, T., Berck, P., & van Gompel, M. (2025). Memory-based Language Models. arXiv:2510.22317.

Leviathan, Y., Kalman, M., & Matias, Y. (2023). Fast Inference from Transformers via Speculative Decoding. Proceedings of the 40th International Conference on Machine Learning, PMLR 202:19274-19286, 2023.

KEYWORDS memory-based learning; language modeling; speculative decoding; energy efficiency; k-nearest neighbor

14:24–14:36

SEED: Self-Explanation-Enhanced Distillation for Common Sense Reasoning in SLMs

Elke Vandermeerschen, Tinne De Laet, Tim Van De Cruys

KU Leuven

Improving the reasoning capabilities of small language models (SLMs) through learning from natural language explanations has emerged as a promising direction for enhancing reasoning performance and interpretability. Existing approaches have shown that explanations can function as effective supervision signals, intermediate reasoning structures, and carriers of reasoning behaviour during distillation. However, most existing methods rely either on human annotated explanation datasets or on explanations generated by large external language models. Both approaches are costly, difficult to scale, and often fail to capture reasoning quality directly. This motivates the need for frameworks in which SLMs learn directly from their own generated explanations. However, self-generated explanations introduce a fundamental challenge: smaller language models are often poorly calibrated, making confidence, consistency, or self-agreement unreliable proxies for reasoning quality. As a result, iterative self-training approaches risk reinforcing incorrect or unstable reasoning patterns over time. We introduce SEED (Self-Explanation-Enhanced Distillation for Commonsense Reasoning in SLMs), a self-improvement framework in which an opensource SLM iteratively learns from its own generated natural language explanations without relying on external teacher models. At each iteration, the model generates answers together with reasoning traces, after which candidate explanations are evaluated through a multi-signal filtering framework before being reused as supervision for

subsequent fine-tuning. Rather than relying solely on confidence or correctness filtering, SEED estimates reasoning quality across complementary dimensions. Specifically, the framework combines epistemic signals (logit-based evidence strength and sequence-level entropy), semantic signals (natural language inference consistency between explanation and prediction), and robustness signals based on Contrastive Explanation Invariance (CEI), which measures whether explanations remain valid under meaning-preserving perturbations. Within SEED, explanations serve multiple functional roles simultaneously: they act as intermediate reasoning representations guiding prediction, as supervision targets during iterative self-distillation, and as structured signals for reasoning-guided training set construction. By treating explanations as central computational objects rather than auxiliary outputs, SEED enables explanation-centered self-improvement while reducing the accumulation of self-reinforcing erroneous rationales. This unified use of explanations enables the model to bootstrap its reasoning capabilities from its own generated signals, without external supervision.

KEYWORDS SLM Self-Improvement, Multi-Agent Frameworks, Commonsense Reasoning

14:36–14:48

Probing cross-lingual differences in how LLMs represent natural language inference

Nicolas Ramos Fernandez, Michiel van der Meer & Gijs Wijnholds

Universiteit Leiden

It is unclear whether modern LLMs reflect language-independent understanding. We probe the internal representations of Olmo 3 7B Instruct, trained exclusively on English, and Tiny Aya Global, explicitly multilingual, on a parallel dataset for Natural Language Inference (NLI) in English, Spanish, Japanese, and Dutch. We (1) measure which layers encode NLI concepts most strongly, (2) how similar representations are across languages, and (3) whether multilingual pretraining changes how NLI is encoded. While prompting performance is language-dependent, probes recover entailment with high selectivity compared to a control task in the middle and later layers for all four languages in both models. Crosslingual probe transfer and direct comparison of probe weights show that representations are most alignable in the middle layers. Orthographically similar languages (English, Spanish, and Dutch) group together, while Japanese is a consistent outlier even in Tiny Aya Global, where Japanese pretraining data is equally common as Dutch and Spanish. We thus show that multilingual pretraining is not a straightforward road to representational alignment for NLI.

KEYWORDS natural language inference, probing classifiers, interpretability, cross-lingual transfer, large language models

14:48–15:00

Recovering Information from 'Impossible' Languages with LLMs

Amirhossein Mohammadi, Laurence Frank, Albert Gatt, Robert Bagheri

Utrecht University

While there has been research into whether LLMs can learn linguistically impossible languages, it remains unclear whether they can systematically recover linguistic information from degraded input, and whether recovery difficulty reflects the degree of locality violation. We investigate whether LLMs can translate impossible languages back into possible forms and whether recovery difficulty varies by perturbation type. By fine-tuning GPT-2 models pre-trained on impossible languages on three

perturbation types that violate information locality to different degrees, we find that models can reconstruct grammatically well-formed output, with performance systematically modulated by perturbation nature. We further examine how training sample size and sentence length affect recovery: models trained on larger datasets show improved performance across all perturbations, though at different rates, while longer sentences provide richer contextual information that benefits perturbations with preserved structure but cannot overcome the fundamental difficulty when locality is completely disrupted. Our findings demonstrate that information locality violations create processing difficulty for LLMs, providing empirical support for information locality as a fundamental principle of efficient sequential processing rather than a categorical learnability boundary.

KEYWORDS Large Language Models, Information Locality, Impossible Language

15:00–15:12

Traces of social competence in large language models

Tom Kouwenhoven, Michiel van der Meer & Max van Duijn

Leiden Universiteit

The False Belief Test (FBT) has been the main method for assessing Theory of Mind (ToM) and related socio-cognitive competencies in a wide variety of human and non-human populations, including, recently, Large Language Models (LLMs). Throughout, the test subjects' ability to appreciate that a character may hold a False Belief has been taken as an indication of their broader social-cognitive competence [1]. Though for LLMs, the debate centres on whether distributional patterns in their linguistic training data amount to "social-world models" that generalise beyond highly specific contexts, such as the FBT, or whether they only learn to solve such tests using superficial, context-specific heuristics. Contributions to this debate often test only a small number of LLMs, for which few details are available, complicating control for leakage of test data into the training data and the systematic mapping of the effects of differences in model size, architecture, pre-, mid-, and post-training [2]. Leveraging the increasing availability of open-source and open-weight model families, we address these issues by testing 17 open-weight models on a balanced set of 192 FBT variants [3].

Our analyses amount to the main finding that FBT performance of LLMs in our sample shows nuanced associations with scale and post-training, but is not robust to False vs. True and explicit vs. implicit variations. We find that scaling model size benefits performance, but not strictly. A cross-over effect between these variations reveals that explicating propositional attitudes (e.g., 'X thinks') fundamentally alters response patterns. Instruction tuning partially mitigates this effect, but further reasoning-oriented fine-tuning amplifies it.

In a case study analysing the social reasoning ability of OLMo 2 (for which we rule out data leakage), we show that this cross-over effect emerges during pre-training, suggesting that models acquire stereotypical response patterns tied to mental-state vocabulary that can outweigh other scenario semantics. Motivated by the strong influence of explicating propositional attitudes, we apply vector steering [4] to test whether the instruction-tuned OLMo 2 encodes a representational direction that can causally influence predictions. In doing so, we isolate a 'think' vector as the causal driver of observed FBT behaviour.

In conclusion, it appears that the interaction between learned scenario patterns and information contained in the verb 'think' interferes with genuine social reasoning, sometimes in productive and

sometimes in detrimental ways.

[1] Apperly. 2010. Mindreaders: the Cognitive Basis of "Theory of Mind". [2] Hu et al. 2025. Re-evaluating theory of mind evaluation in large language models. [3] Trott et al. 2023. Do large language models know what humans know? [4] Subramani et al. 2022. Extracting latent steering vectors from pretrained language models.

KEYWORDS theory of mind, socio-cognitive reasoning, large language models

15:12–15:24

Do multilingual Large Language Models think in Greek? English mediation in semantic similarity judgments

Markella Englezou & Jelke Bloem

Universiteit van Amsterdam

Multilingual Large Language Models (LLMs) have demonstrated strong semantic understanding across many languages, yet concerns remain regarding their performance in low-resource languages and the extent to which they rely on English-centered representations. This thesis investigates whether multilingual LLMs compute semantic similarity in Greek using language-internal representations or rely on English-mediated semantics. Prior research has explored multilingual embeddings, cross-lingual transfer, and semantic similarity evaluation, but limited work has directly examined English semantic interference in low-resource language similarity judgments. Using a quantitative cross-lingual design, this study evaluates multiple multilingual LLMs on the Greek SimLex-999 dataset and its English equivalent through correlation, regression, and divergence-based analyses. The results show that all models strongly correlate with Greek human semantic similarity judgments, while also exhibiting significant cross-lingual alignment with English, as English similarity scores partially predict Greek model outputs. However, divergence testing reveals that top-performing models more often align with Greek rather than English judgments when the two languages differ, indicating that English influence is present but not dominant. Overall, the findings support a hybrid representational structure in which multilingual LLMs combine shared cross-lingual semantic spaces with language-specific knowledge.

KEYWORDS multilingual LLMs, semantic similarity, Greek, English interference, cross-lingual semantic alignment

AI-Assisted Decision Support & Agentic Problem-Solving

D.2.18 · Chair: Veronique Hoste

14:00–14:12

Natural language processing and language models for Dutch clinical text: a systematic review

Artuur M. Leeuwenberg, Ruurd J.A. Kuipers

University Medical Center Utrecht, Utrecht University

Increasingly natural language processing (NLP) tools and applications - including those using large language models (LLMs) - are developed and used in electronic health records (EHRs). As generalization to specific languages and EHR settings or departments is not guaranteed [1], having a

comprehensive overview about tool availability and accuracy across EHR settings is increasingly important for their effective application and reuse of existing tools [2]. No such overview was available for Dutch, focused on language technology for EHRs that covers the past decade [3].

This study's objective is to identify and describe existing NLP tools, including those using LLMs, that have been developed or evaluated in real-world Dutch EHRs.

A literature search was conducted in Scopus and Pubmed up to November 11, 2025. Information about the NLP task, architecture, patient group, text type, healthcare setting, code and model availability were extracted.

A total of 44 studies were included, describing 794 models and 792 evaluations. Most studies focused on information extraction (73%), followed by de-identification (14%), generative applications (11%), and language modeling tasks (9%). Rule-based methods were most frequently used at the study level (50%), while transformer-based approaches accounted for the majority of models and evaluations (55%). Prompting LLMs was used in 16% of studies and accounted for 32% of models and evaluations. Code was shared in 43% of studies, covering 91% of models, whereas only 4% of models were publicly available.

To conclude, a diverse set of NLP models has been developed for Dutch clinical text, with an increasing use of transformer-based and LLM-based approaches. Model availability remains limited. This review provides a structured overview of available tools and evaluations to support their application and reuse in Dutch clinical settings. A list of all included studies and models can be found at: <https://tuur.github.io/files/dutchnlpoverview.html>.

References:

- [1] Klug, Katrin, et al. "From admission to discharge: a systematic review of clinical natural language processing along the patient journey." *BMC Medical Informatics and Decision Making* 24.1 (2024): 238.
- [2] Gong, Eun Jeong, et al. "Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks." *Journal of Medical Internet Research* 27 (2025): e84120.
- [3] Cornet, Ronald et al. "Inventory of tools for Dutch clinical language processing." *Studies in health technology and informatics* vol. 180 (2012): 245-9.

KEYWORDS NLP, Electronic health records, Systematic review

14:12-14:24

Measuring the practice of shared-decision making (OPTION12): An Investigation into Open-sourced Smaller LLMs (OS-sLLMs) for Better Privacy and Sustainability

T Wit, L Han, C Heipon, D Lindevelt, A Stiggelbout, S Verberne

Leiden University Medical Center; Leiden University

Shared decision-making (SDM) is important in clinical consultations in which patients discuss and decide on treatment options with clinicians. This SDM process is often coded with the OPTION12 instrument, an observer-based tool, consisting of 12 items, rated on a 5-point Likert scale. This coding is conducted by human coders. Coding is time-intensive and frequently accompanied by disagreement between coders. We explore the capability of open-source privacy-preserving smaller LLMs (OS-sLLMs) to perform the coding task, potentially automating the process, with humans in the loop.

Methods: 26 transcripts of Dutch melanoma patients consultations with clinicians were double-coded by two coders. Two human coders resolved their disagreements after independent coding. To evaluate OPTION12 coding using OS-sLLMs, the consultation data was divided into development and testing sets. We designed the complete investigation framework. It includes 1) a pilot study of the development set (11 interviews) for prompt-refinement and sLLM selection. 2) deployment of the fine-tuned prompts and best performing sLLM (judge-sLLM) augmented with few-shot examples from the development set on the testing set (15 interviews) and asking the judge-sLLM to resolve the disagreement on other OS-sLLMs' scores. First, for prompt refinement, we use chain-of-thoughts (CoTs), LLM-assisted prompting, human-in-the-loop with sample output (few-shots) feedback. For OS-sLLMs, we use both 1) general domain models Llama, Gemma, and Mistral7b, and 2) medical domain models Meditron and Medllama. Second, for judge-sLLM selection, on the system-level, we measure the overall correlation of each sLLM and human coding using Spearman and Pearson correlation scores. On the segment-level, we also look into the most agreed-upon and disagreed-upon items. We will perform both qualitative and quantitative analysis by discussing the evaluation scores and categorise the OS-sLLMs' behaviours with examples. Finally, we deploy the OS-sLLMs on the testing set and use the judge-sLLM to resolve disagreements.

Preliminary results: Three general domain OS-sLLMs perform better than the two medical domain ones, which both generate hallucinations and do not follow prompts precisely, indicating further developments is needed for medical OS-sLLMs. Mistral7b outperformed the other two Gemma3:12b and Llama3.1:8b by 4 consensus with human coding, vs 3 items. The overall correlation with human coding on these 12 items is (0.83, 0.80, 0.64) using Pearson correlation, and (0.81, 0.78, 0.61) using Spearman rank correlation from the three models (gemma3:12b, llama3.1:8b, mistral7b). For the items for which OS-sLLMs agree with human coders, OS-sLLMs can generate the same sentences as humans in some cases, but in other cases, they generate better quotes than humans.

We reported the first research findings using OS-sLLMs on scoring SDM with the OPION12 in Dutch melanoma patients' consultation transcripts.

KEYWORDS NLP for Healthcare, AI for Medical, Shared Decision Making, Human-AI Correlations

14:24-14:36

Controllable emotional expression in humanoid social robots: A prompt-based framework for Furhat

Sadegh Jafari, Els Lefever, Veronique Hoste

LT3, Ghent University

As social robots increasingly assume roles in educational, healthcare, and domestic settings, their ability to exhibit natural, contextually appropriate emotional behavior is essential to fostering user engagement and trust. The Furhat robot [5], with its back-projected 3D face and rich expressive capabilities, provides a promising platform for investigating human-like non-verbal communication [1]. While traditional approaches to emotional expression in social robots have largely relied on predefined facial configurations and rule-based behaviors [2], recent advances in Large Language Models (LLMs) have opened new opportunities to generate adaptive, context-aware interactions [3]. This abstract outlines a proposed framework for controllable emotional expression on the Furhat platform. Inspired by recent work on LLM-driven facial expression generation for Furhat [4], the

framework aims to leverage the reasoning capabilities of LLMs to translate high-level emotional prompts, such as “express sadness” or “convey happiness,” into synchronized verbal responses and facial action units. Through a prompt-based control mechanism, the system is intended to dynamically map semantic intent to expressive robot behaviors while maintaining coherence with the conversational context. Building on evidence that expressive capabilities play an important role in human–robot interaction, studies have shown that the physical Furhat robot can be effectively used to investigate the impact of facial expressions on children’s trust [6]. Motivated by such findings, the proposed framework combines LLM-based reasoning with embodied emotional expression to support more adaptive and empathetic social robots. The effectiveness of the framework will be evaluated through human-subject experiments.

References: [1] Samer Al Moubayed et al. 2012. Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. *ACM Trans. Interact. Intell. Syst.* 1, 2, Article 11 (January 2012), 25 pages. [2] Rawal N, Koert D, Turan C, Kersting K, Peters J, and Stock-Homburg R (2022) ExGenNet: Learning to Generate Robotic Facial Expression Using Facial Expression Recognition. *Front. Robot.* [3] M. A. Azad, "Developing an Emotionally Intelligent Human-Robot Interaction System with Furhat (EmotionBot): A Multimodal Deep Learning Approach," 2025 International Conference on Electrical, Communication and Computer Engineering (ICECCE), Istanbul, Turkiye, 2025, pp. [4] Abbo, Giulio Antonio, et al. "Expressive Furhat: Generating Real-Time Facial Expressions for Human-Robot Dialogue with LLMs." *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026. [5] <https://www.furhatrobotics.com/> [6] N. Calvo-Barajas, G. Perugia and G. Castellano, "The Effects of Robot’s Facial Expressions on Children’s First Impressions of Trustworthiness," 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 2020, pp.

KEYWORDS human-robot interaction, large language models, emotional expression, social robots, Furhat

14:36–14:48

Topics as proxies for sociodemographics: How conversational context affects LLM answers

Vera Neplenbroek, Gabriele Sarti, Arianna Bisazza & Raquel Fernández

Universiteit van Amsterdam, Northeastern University, Rijksuniversiteit Groningen

When large language models (LLMs) are used in high-stakes scenarios, such as legal, medical and financial advice, even a single conversation history is enough to drive differences in outcomes between users. Prior work has demonstrated that this results in outcome disparities between sociodemographic groups, with some groups receiving more advantageous outcomes than others. In this work, we demonstrate that LLMs actually struggle to infer user sociodemographics from a single conversation history and that although there are disparities between sociodemographic groups, they are minimal in magnitude. To investigate what the main driver of these disparities is, we compare user sociodemographics to a range of (psycho)linguistic features of conversations, including conversation topic, emotions, and readability. We find that conversation topics are most predictive of LLM-generated advice within a conversational context, which, to some extent, function as proxies for sociodemographic groups and often affect advice in unpredictable ways. This is cause for concern and highlights the need for future research to better understand and, if needed, mitigate the effect of conversational context on LLM outputs in high-stakes scenarios.

KEYWORDS sociodemographics, high-stakes advice, conversation, topic, conversational context

14:48–15:00

Automated composition of agentic networks for multi-metric user preference satisfaction

Lize Pirenne, Gaspard Lambrechts, Norman Marlier, Maxence de la Brassinne Bonardeaux, Jonathan Pisane, Gilles Louppe & Damien Ernst

ULiège, Keyes

The construction of agentic networks to satisfy user-specific preferences is becoming a key component for value creation beyond model serving. In this paper, we use the evaluation of an agentic network over a QA dataset to stir its architecture towards the maximal satisfaction for different user profiles. We model a user profile as a distribution of preference over a set of metrics and the network as a composition of stochastic processes. We add constraints on token budget and response time in the framework by adding constraints on the searchable domain. Our algorithm first analyzes the sensitivity of each metric as the network is modified and then suggests atomic edits to its composition. This process is repeated until a desired satisfaction or convergence is reached. In our experiments, we compare manual compositions for an industrial Retrieval Augmented Generation (RAG) solution with our stirred networks for multiple user profiles. Additionally, we compare the speed, cost, and performance of our algorithm with SOTA frameworks in deep search datasets.

KEYWORDS Language Model, Retrieval Augmented Generation, User Preference, Hallucination, Factuality

15:00–15:12

Beyond benchmarks: Optimizing LLMs and agents for dutch cryptic crosswords

Pauline van Nies, Keze Hu, Joanna Pasiarska, Rey Rashid, Akif Berber, Janneke van der Zwaan, Antske Fokkens & Paul Verhaar

Sopra Steria NL, Vrije Universiteit Amsterdam, Radboud Universiteit Nijmegen

Standard benchmarks such as EuroEval tell us how LLMs handle summarization or reasoning, but not how they cope with tasks demanding genuine linguistic creativity - clever associations and deliberate wordplay. Building on our CLIN 2025 talk, where state-of-the-art models solved only ~25% of cryptic clues in a zero-shot setting, we present a systematic study of how far different optimization techniques can close this gap, evaluated on both Dutch and English datasets.

We first compare reasoning across the latest model families (GPT, Claude, Gemini, Mistral, DeepSeek) and quantify the effect of moving beyond basic prompting. Detailed, expert-guided chain-of-thought prompts with examples yield a consistent ~10% improvement over simple prompts across models. This raises a deeper question: can the reasoning itself - not just the final answer - be optimized? We define a reasoning-quality metric combining decomposition quality, inference soundness, and structural coherence. The criteria are first defined and applied by a human expert on a small annotated set, then operationalized as an LLM-as-judge so the metric can be scored automatically at scale and validated together with final-answer accuracy.

We then ask whether reasoning ability can be transferred from a teacher to a weaker student model through knowledge distillation, finetuning the student on high quality chain-of-thought traces for

cryptic clues, and compare this to a few-shot approach using similar expert-style demonstrations.

The largest gains, however, come from changing *how* the model works rather than how it reasons in isolation. Human solvers rarely work unaided - they reach for dictionaries, anagram solvers, and synonym lists, and revisit clues as crossing letters fall into place. Giving a single agent access to such lookup tools adds a further ~10% over chain-of-thought, with the strongest configuration (Gemini-3-Flash) reaching nearly 80% of solving Dutch clues correctly. We scale this up by mirroring the human solvers with a multi-agent architecture (built with LangGraph) in which specialized agents collaborate and iteratively complete an entire Dutch cryptic crossword, using the constraints of already-solved answers to crack the harder clues. We discuss the design choices behind the framework and demonstrate it solving a full crossword in real time.

KEYWORDS cryptic crossword clues, agents, reasoning optimization, prompt engineering, evaluation

15:12-15:24

Multi-Agent System for Solving Dutch Cryptograms

Joanna Pasiarska, Pauline van Nies, Akif Berber

Sopra Steria

One of the innovations in the field of artificial intelligence is using human cognition as an inspiration for mimicking intelligent behaviours. This study applies a similar approach discovering whether reproducing human reasoning steps improves the performance of agentic AI solving complex reasoning tasks. For example, cryptic clues require a combination of abstract thinking, logical reasoning, wordplay understanding and finding loose associations. While this subject is gradually explored for English cryptograms, research lacks evidence for low-resource languages such as Dutch. For this project, a multi-agent system was developed, aiming to simulate different stages of solving Dutch word puzzles in the same manner a human would do. It consists of a solving agent equipped with tools, being a thinking brain, a validation agent, verifying whether a proposed solution is reasonable, and a judging agent, allowing for correcting mistakes. Implementation of the proposed solution included testing three frameworks, Google ADK, LangGraph and DSPy, while the evaluation was performed with Inspect. Additionally, through systematic experiments, the most optimal settings for prompts, parameters, schemas and architecture were discovered allowing for most cost and time efficient solutions. A successful multi-agent system requires a thoughtful split of tasks between agents and some degree of guiderails. The latter comes from a noticed trade-off between creativity and reliability. More creative models tend to reason better but cannot be trusted with executing exact instructions and use tools as what they were meant for. Therefore, adding a deterministic component was necessary to ensure keeping constraints of puzzle requirements in place. Another innovation contributing to the successful application was incorporating information from the puzzle grid, updated with each iteration of the solving process. The agent aware of already filled-in letters was able to come up with more accurate answers, just as humans do when they are prompted with an additional information. However, the important difference with how humans solve it, is that they can realise that the previous answer might have been incorrect. To emulate this behaviour, a judging agent was implemented, evaluating two conflicting answers. Interestingly, the choice of the best prompt was not only dependent on the number of details and instructions included. What stood out, was the fact that kind and gentle prompts allowing agent to think without a strict condition to always return the correct answer, yielded the best result. Finally, implementing a cost-tracking feature allowed for

identifying points for improvement. Overall, this project contributed to understanding how agentic AI can be directed to solve cryptic clues and what are the most efficient approaches to achieve it. The presented results evaluate the performance of multi-agent system in Dutch-speaking environment.

KEYWORDS cryptograms, multi-agent system, human-like reasoning, Chain-of-Thought, tool-augmented agents

Poster session 1

10:15 – 11:15 · Q building: Nelson Mandela Hall

Clinical & medical NLP

POSTER 1

MediScribes: combining knowledge graphs and LLMs in Dutch clinical summarization

Maaïke H. T. de Boer, Roos M. Bakker & Quirine T.S. Smit

TNO, Universiteit Leiden, VU

Medical documentation is currently burdening the medical treatment process. One way to ease the process is to automate the documentation process and summarization. Because of the domain, it is essential that the documentation and summaries are factually correct. In this project, we explore how we can use knowledge graphs and Large Language Models (LLMs) to enhance clinical summarization and ensure factual correctness. We test a collection of knowledge graph based summarization methods on the MultiClinSum2 challenge: a challenge to automatically create summaries from case reports.

We compare 2 methods: 1. Prompting an LLM to create a summary, with a prompt based on analysis of the medical documents, 2. Creating a knowledge graph based on the medical documents and based on the knowledge graph creating a summary. For the second method, we explore 4 variations such as going from text to a knowledge graph with the help of an ontology, and building a knowledge graph with GraphRAG and using a syntactic parser. We use zero-shot and few-shot prompts.

We validate performance on the MultiClinSum2 challenge, on the Dutch and English test set. Evaluation is done with automatic metrics for text generation such as Rouge and BERTScore, and LLM-as-a-Judge metrics.

The results show that using few-shot prompting with an LLM is more powerful than zero-shot with a knowledge graph, at least for the summaries evaluated in this task. Furthermore, for this test set, the inclusion of a knowledge graph did not lead to better automatic evaluation scores. However, these evaluation scores do not focus on factual correctness, therefore, we are currently investigating an additional evaluation focused on factuality.

Ongoing and future work involves additional experimentation with our methods on conversational summarization of medical consults, and a manual evaluation of results to focus on factual correctness.

KEYWORDS Summarization, Knowledge Graphs, GraphRAG, LLM, MultiClinSum

POSTER 2

BijsluiterBot: a Dutch medical leaflet knowledge RAG assistant

Daan L. Di Scala, Roos M. Bakker, Simon P.J. van de Fliert, Caspar J. Meijer & Maaïke H.T. de Boer

TNO, Universiteit Utrecht, Universiteit Leiden

AI chatbots are increasingly used for medical advice, offering quick and interactive access to healthcare information. Patient information leaflets are often difficult to read, leading to more people moving away from reading leaflets to asking questions to chatbots. However, in the medical domain, where incorrect information can have serious consequences, it is essential to ensure access to trustworthy answers grounded in reliable evidence, as chatbots have been shown to provide inaccurate or potentially unsafe recommendations in real-world scenarios. For this reason, we introduce BijsluiterBot, an AI assistant that answers medicine-related questions based on Dutch patient information leaflets.

To evaluate our AI assistant, we have developed the PIL-QA-NL benchmark of over 100 Dutch patient information leaflets (“bijsluiters”), alongside the PIL (Patient Information Leaflet) ontology, and a Question-Answer (QA) dataset with more than 5,000 pairs. The patient information leaflets are based on the most commonly used medicines in the Netherlands [1,2], and consist of both prescription drugs and over-the-counter medicines. To link and structure the data, we have created the PIL ontology. The ontology structures and links the data, and its design is guided by a set of 40 Competency Questions (CQs). Using the PIL ontology, we populate a knowledge graph, with subgraphs for each medical leaflet. This knowledge graph then serves as a basis for the QA dataset creation, which includes questions, corresponding answers, and their source document attributions.

We compare different Retrieval Augmented Generation (RAG) approaches for this QA task, ranging from well-known approaches such as BM25 and Vector RAG, to GraphRAG methods with Neo4J and RDF. Additionally, we introduce CompetentRAG, a retrieval approach which searches for matching human-made CQs and corresponding SPARQL queries to query the knowledge graphs. We employ a variety of state-of-the-art embedding and generation models for comparison, ranging in sizes, openness, and multilingual capabilities.

We aim to evaluate both retrieval and generation capabilities of BijsluiterBot by examining performance across different chunk sizes, measuring the impact of retrieval granularity and its effect on downstream answer generation. Overall, with this design and evaluation, BijsluiterBot offers a promising approach for grounded retrieval, delivering trustworthy medical question-answering that patients can depend on.

Citations: [1] Zorginstituut Nederland. Zorgcijfers Databank. <https://www.zorgcijfersdatabank.nl/> [2] Van Dijk, L., Brabers, A., & Vervloet, M. (2023). Informatie bij de aankoop en het gebruik van zelfzorggeneesmiddelen: Een peiling binnen het Nivel Consumentenpanel Gezondheidszorg. CBG-MEB. <https://www.cbg-meb.nl/documenten/2023/01/20/zelfzorgmedicijnen-onderzoek-nivel>

KEYWORDS Large Language Models, Retrieval Augmented Generation, GraphRAG, Question Answering

POSTER 3**Building a Dutch patient-centered medical lexicon from cardiology discharge summaries**

Dirk van Nimwegen, Robert Vander Stichele, Pascal Coorevits & Els Lefever

Department of Public Health and Primary Care, Ghent University, Belgium, LT3, Language and Translation Technology Team

Medical discharge letters are primarily written to support communication between healthcare professionals, but patients increasingly gain direct access to these documents through electronic health records. Although discharge letters contain important information about diagnoses, treatment, follow-up care, and outcomes, health information documents are often difficult for non-experts to understand (Okuhara et al., 2025). Previous work on patient-friendly medical communication has identified several sources of difficulty, including lexical, semantic, syntactic, and discourse-level complexity (Van Nimwegen et al., 2024). In this study, we focus on the lexical dimension as a first step towards improving the accessibility of Dutch medical discharge letters.

Lexical barriers include specialist terminology, abbreviations, loanwords, semi-technical expressions, and ordinary words that acquire a different meaning in a clinical context. Such items can hinder comprehension even when the overall sentence structure is relatively simple. At the same time, replacing medical terms with simpler alternatives is not always straightforward. Explanations must remain clinically correct and context-sensitive, while avoiding additional complexity. This creates a practical need for a controlled inventory of difficult terms and suitable patient-oriented explanations.

In this research, we present a workflow for identifying and analysing potentially difficult lexical items in Dutch medical documents. Candidate terms are extracted from a corpus of 4441 anonymized cardiology discharge summaries and linked, where possible, to Dutch medical terminology resources. These include Pinkhof's *Medisch Woordenboek* (General Medical Dictionary), the *Thesaurus Zorg en Welzijn* (definitions of medical related terms), a multilingual medical glossary (Dutch part), and a primary-care terminology list (*College ter Beoordeling van Geneesmiddelen*), as resources with validated definitions or explanations of the medical terms. The resulting inventory is enriched with contextual information and variants, including abbreviations, alternative spellings, and expressions that may be ambiguous outside a clinical context. We then explore the use of generative AI to produce short lay definitions and paraphrases for selected items. Clinical experts will assess a sample of explanations with respect to accuracy, clarity, and risk of misinterpretation.

The resulting lexicon is intended as a resource for patient-oriented lexical simplification and as a basis for training and supporting domain-specific language models. Beyond cardiology, the approach provides a reusable methodology for developing interpretable medical language resources in other clinical domains.

KEYWORDS medical terminology, patient empowerment, consumer health information, Dutch clinical NLP, lexical simplification

POSTER 4

Adapting clinical event annotation to Dutch primary care: an event annotation framework for post-acute infection syndromes

Sara Mazzucato, Artuur Leeuwenberg, Sander van Doorn, Joost van Rosmalen, Isabel Slurink

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Dutch general practitioner (GP) free-text notes capture important information on infections, post-acute infection syndromes (PAIS) and related symptoms for epidemiological studies such as the Post-Infectious Chronic Outcomes and Risk (PICOR) study; especially valuable as relevant structured diagnostic codes were introduced late and used inconsistently. Yet, no publicly available annotated corpus exists to develop or evaluate natural language processing tools for extraction in this domain, which we address by adapting the COVID-19 Annotated Clinical Text (CACT) framework [1] to Dutch GP notes, extending its scope to the PAIS spectrum and pragmatics of SOEP (Subjective, Objective, Evaluation, Plan) documentation. The proposed framework uses an event-based model where each event has a trigger span and typed argument relations capturing contextual modifiers. Three annotation layers are introduced: (i) a diagnostic expression typology (ii) an eleven-subtype evidence argument capturing how each diagnosis was established, including laboratory tests, self-tests, clinical examination context, and patient-reported information; (iii) explicit decision rules for the Dutch SOEP structure: Subject arguments extend beyond the patient to include family members and other contacts (e.g., 'neighbour had RSV'); Change arguments capture symptom evolution (improvement, resolution); clinician hedging is distinguished from patient-side hypotheticals (Actuality); and Dutch negation (geen/no) receives dedicated rules distinguishing plain absence, temporal resolution, and diagnostic uncertainty.

Using the routine health care database of the Julius General Practitioners' Network [2], we piloted the framework by annotating 200 notes (in duplicate, by two medical students). Inter-annotator agreement was assessed under token-level Cohen's κ , span-level F1 (Lybarger criterion [1]), and F1 conditional on annotation overlap [3]. Across six core entities, conditional F1 averaged 0.78 [95% CI: 0.76–0.80] and one annotator replicated 0.89 of the other's annotations. The Lybarger-conditional F1 gap (0.51 vs. 0.78) reveals that disagreement was driven almost entirely by coverage: when both annotators marked the same span, type agreement was high (0.78–0.96). Mention-level F1 is a conservative proxy for downstream performance; patient-level cohort assignment F1 (our study-relevant metric) is expected to be substantially higher. The released guidelines, INCEPTION schema, and synthetic SOEP examples provide a reusable template for cross-lingual clinical annotation adaptation. Future work will fine-tune MedRoBERTa.nl across ~2,000 notes from four Dutch GP networks.

[1] Lybarger et al. (2021) Extracting COVID-19 diagnoses and symptoms from clinical text. JAMIA Open
[2] Smeets et al. (2018) The Julius General Practitioners' Network (JGPN). BMC Health Serv Res [3] Hripcsak & Rothschild (2005) Agreement, the F-measure, and reliability in information retrieval. JAMIA

KEYWORDS clinical NLP, Dutch, post-acute infection syndrome, event extraction, manual annotation

POSTER 5

Multimodal AI-driven approaches to patient-centered quality of life assessments: A systematic review

Yee Man Ng, Kiana Shahrasbi, Bram van Dijk, Gerard van Oortmerssen & Marco Spruit

Universiteit Leiden, Leids Universitair Medisch Centrum

Quality of life (QoL) is generally defined as “an individuals’ perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns” (WHO definition). In other words, QoL is an inherently subjective and multidimensional construct that reflects an individual’s perceived physical health and functioning, mental wellbeing, and the individual’s social relationships. In clinical practice, clinicians use patient-reported outcome measures (PROMs), i.e., standardized questionnaires, to measure a patient’s experienced QoL. Considering the structured and inflexible nature of PROMs, Natural language processing (NLP) methods allow us to analyze patient-generated language data directly and gather insights on a patient’s QoL in a way that might not be captured by standardized PROMs. Patient language generally contains linguistic features that are indicative of a patient’s wellbeing: for instance, Drougkas et al. (2024) found that depression can be detected from a combination of multimodal language markers, e.g., slower speech rate, more speech pauses, and negative emotion words, among other. NLP thus shows great promise to analyze these complexities in patient text or speech to detect various areas of QoL. We present our current progress on a systematic literature review that provides an overview of works that perform automatic QoL assessment from multimodal data based on patient-authored text or speech in combination with other modalities. Our preliminary results suggest that current efforts in multimodal NLP for QoL in patient-generated markers tend to focus on technical improvements and validation, with most works focusing only on mental or neurological disorder (dementia) detection. Future directions for this area would be leveraging LLM-based conversational agents to support data collection, more validation of systems in a clinical setting, and moving beyond the interpretation only on mental health or neurological disorders as a binary presence or absence, but towards detecting level of quality of life as a multi-dimensional and dynamic construct. References G. Drougkas, E. M. Bakker, and M. Spruit, “Multimodal machine learning for language and speech markers identification in mental health,” *BMC Med Inform Decis Mak*, vol. 24, no. 1, p. 354, Nov. 2024, doi: 10.1186/s12911-024-02772-0

KEYWORDS quality of life, clinical nlp, multimodal nlp

POSTER 6

From rules to large language models: revisiting the literature mining of enriched SNP-disease associations

Kiana Shahrabi, Armel Lefebvre & Marco Spruit

Leiden University

Every day, thousands of biomedical papers are published. In 2025, over 1.79 million biomedical papers were indexed in OpenAlex (Priem et al., 2022). Although this growing literature is valuable, extracting specific information from it remains difficult and time-consuming. Natural Language Processing (NLP) helps researchers and medical practitioners identify relevant information more efficiently. For example, Tawfik and Spruit (2018) developed SNPcurator, a pipeline using rule-based regular

expressions to extract Single Nucleotide Polymorphism (SNP) (a single DNA variant that can affect diseases), along with p-value, and odds ratio from biomedical abstracts. They introduced a benchmark subset derived from the SNPPhenA corpus (Bokharaeian et al., 2017). A key question remains whether prompt-engineered LLMs outperform rule-based pipelines, and where their failures differ. In this work, we extend the SNPcurator pipeline as follows: First, we refine the dataset by re-annotating instances using a stricter criterion requiring a significance threshold of $p < 0.05$. Second, we implement a pipeline using LangExtract (Goel, 2026), an extraction framework that links LLM outputs directly to source text, using few-shot prompting and carefully designed guidelines to handle edge cases. We then evaluate three systems: SNPcurator pipeline, LangExtract with Gemini 3.5 Flash, and LangExtract with Gemma 3:4B, at three levels: full association extraction (SNP, p-value, and odds ratio), partial SNP and p-value extraction, and individual field-level extraction. Finally, we perform a detailed error analysis to characterize the limitations of each approach. Our results show that Gemini significantly outperforms other methods across all evaluation levels (full association micro F1: 0.86 vs. 0.52 rule-based vs. 0.40 local). The largest improvements are in recall, showing better coverage of relevant associations while maintaining high precision. The rule-based system frequently misses valid associations, and the local model shows signs of overprediction. Error analysis shows rule-based failures are mainly due to variability in the expression of statistical results, whereas LangExtract struggles with more structural issues such as sentence truncation at extraction boundaries. We also test whether ensembling could improve performance, but find that even the best combination (rule-based + Gemini) overlaps with 47.4% of correct extractions, and the rule-based approach introduces more false positives than true positives when merged. Most associations missed by Gemini are due to LangExtract's chunking strategy rather than a linguistic limitation, suggesting that improving chunking would close much of the remaining gap. Overall, these findings show that the remaining extraction challenges in LLMs are concentrated in fixable pipeline components rather than in language understanding itself, providing a clear path to improve structured biomedical information extraction.

KEYWORDS biomedical information extraction, large language models, named entity recognition

POSTER 7

Classifying utterances with CPS strategy labels in VR-based collaborative medical training

Xinran Li, Siem Buseyne, Ine Windey & Anaïs Tack

KU Leuven, imec

Collaboration is widely recognised as a key 21st-century skill, but supporting students effectively during collaborative training remains challenging. In this work, we explore one possible solution: using NLP methods to turn spoken teamwork interactions into interpretable signals by classifying utterances with Collaborative Problem Solving (CPS) strategy labels. By identifying how team members communicate while working together, such models can support future feedback on team communication during or after training.

Building on previous work on Dutch CPS classification in a more general teamwork task (Tack, Özturan & Buseyne, 2024), this study extends utterance-level CPS classification to VR-based collaborative medical training. The dataset consists of 22 VR conversation sessions, in which teams of two students collaboratively solve urgent patient cases. In total, the corpus contains 5,506 utterances,

with about 250 utterances per group on average. The interactions were produced in a West Flemish variety, first automatically transcribed using Whisper large-v3, and then manually corrected and diarised. Data preparation revealed challenges at both the transcription and annotation stages. For transcription, the two recording sources involved different trade-offs: headset audio captured individual speakers more clearly but was more prone to Whisper hallucinations during long silent intervals, while camera audio captured the full interaction more completely but with lower recording quality. For annotation, each utterance was labelled with a hierarchical CPS framework. Agreement was higher for broad CPS categories than for fine-grained labels, showing that specific CPS roles are harder to distinguish.

For modelling, we frame the task as a transfer learning problem from general teamwork to medical training simulation. We first examine cross-domain transfer by applying models developed in earlier CPS classification experiments to the new medical corpus. In those earlier experiments, fine-tuned Dutch transformer models (RobBERT-v2 and BERTje) performed better than feature-based n-gram baselines (TF-IDF + logistic regression), while zero-shot prompting with Gemma 3B was less stable. Testing these model types on the medical corpus allows us to assess how well they generalise to a more complex and domain-specific setting and to identify typical transfer errors. We then study domain adaptation by training and evaluating models directly on the medical corpus using cross-validation. In this stage, we further refine Dutch transformer-based classifiers and explore parameter-efficient fine-tuning methods such as LoRA. Few-shot LLM prompting is considered as a possible extension to zero-shot prompting.

References Tack, A., Özturan, T. & Buseyne, S. 2024. Classifying utterances with collaborative problem-solving strategies in student teamwork interactions. Presentation at CLIN34.

KEYWORDS spoken dialogue analysis, utterance classification, hierarchical classification, Dutch NLP, collaborative problem solving

Dutch language resources, benchmarks & tools

POSTER 8

BeterBERT-2026: training Dutch models in the age of the AI Act

Edwin Rijgersberg

AI Studio Delta

BeterBERT-2026 is a modern Dutch masked language model designed for AI Act compliance from the ground up.

Recent years have seen a resurgence of masked language models (MLMs) built with modern techniques such as Rotary Position Embeddings (RoPE) and Flash Attention. ModernBERT, EuroBERT and mmBERT have demonstrated that the encoder paradigm remains competitive with autoregressive alternatives for many downstream tasks. However, no equivalent model existed for Dutch. BeterBERT-2026 is a modern Dutch MLM that fills that gap.

BeterBERT-2026 is based on the EuroBERT architecture and is trained from scratch on a 175-billion-token Dutch-only dataset assembled from publicly available sources in compliance with European copyright law. It is available in base (150M) and large (520M) variants and has a context length of

8192 tokens. It outperforms multilingual alternatives on every Dutch benchmark in EuroEval. The weights are released under the Apache 2.0 license.

We present what it takes to train a compliant model in the age of the AI Act. We show how we do data gathering and filtering of copyrighted data according to the EU General Purpose AI Code of Practice, show bias testing and mitigation, and demonstrate resistance against re-identification under the GDPR through data extraction and membership inference attacks.

KEYWORDS masked language modeling, BERT, AI Act, Dutch AI

POSTER 9

Towards MTEB-NL (v2): Improving Dutch text embedding evaluation

Nikolay Banar, Ehsan Lotfi & Walter Daelemans

Universiteit Antwerpen

Text embeddings have become a central component of many language applications. Hence, reliable benchmarks are essential for understanding the progress of these models across languages, tasks, and domains. For Dutch, evaluation resources remain limited compared to those available for English and larger multilingual settings. The recently introduced Massive Text Embedding Benchmark for Dutch (MTEB-NL) [1] addressed this gap by providing a dedicated benchmark for evaluating Dutch text embeddings. While it represents an important step forward, MTEB-NL should still be seen as an intermediate step. In particular, parts of the benchmark rely on machine- and human-translated resources, which may overlook linguistic nuances, domain-specific variation, and cultural context specific to Dutch. In addition, several task categories are represented by only one or two datasets, limiting the benchmark's ability to provide a broad and robust assessment across embedding tasks. With MTEB-NL (v2), which is currently a work in progress, we aim to further develop this effort into a broader, more representative, and community-driven benchmark for Dutch text embeddings. We invite researchers and practitioners, working with Dutch language data to contribute to MTEB-NL (v2). By developing this resource together, we hope to provide a stronger foundation for evaluating and developing Dutch text embedding models.

[1] Banar, N., Lotfi, E., Van Nooten, J., Arhiliuc, C., Kliocaite, M., & Daelemans, W. (2025). MTEB-NL and E5-NL: Embedding Benchmark and Models for Dutch. arXiv preprint arXiv:2509.12340.

KEYWORDS NLP, Dutch, Text embedding models and resources, Benchmarking, MTEB

POSTER 10

LiNT-II: readability formula and readability tool for Dutch

Jenia Kim, Henk Pander Maat, Antal P.J. van den Bosch, Stefan Leijnen, Marieke M.M. Peeters & Antske Fokkens

HU University of Applied Sciences Utrecht, Utrecht University, Vrije Universiteit Amsterdam

In the Netherlands, 14–24% of adults are estimated to be low-literate (Netherlands Court of Audit, 2016; Gubbels et al., 2019), creating urgent demand for readability assessment tools that support fair access to information. LiNT (Pander Maat et al., 2023) is a Dutch readability formula, based on four linguistic features: word frequency, proportion of concrete nouns, syntactic dependency length, and content words per clause. It is the only empirically validated readability tool for Dutch; the features

were causally linked to reading comprehension via cloze-test data from 2,753 secondary-school students (Kleijn 2018). However, LiNT relies on T-Scan, an NLP pipeline with legacy dependencies that is slow and hard to install; in practice, its use is limited to analysis of small text batches via a web interface. We present LiNT-II, a modern reimplement for both research and production. The core change is replacing T-Scan with spaCy, which simplifies installation and substantially accelerates processing while preserving the validated feature set. We refit the linear regression on the original comprehension data. As part of the optimization process, we (a) compared two frequencies corpora: SUBTLEX-NL (Keuleers et al., 2010) and wordfreq (Speer 2022), (b) compared the effect of using decomposed versus whole-word frequencies for closed compounds, and (c) revised and extended our semantic classification list. The refitted model reaches an adjusted R^2 of 0.74, same as the original LiNT. The applied contribution is twofold. First, LiNT-II makes a validated Dutch readability instrument practically usable in modern LLM-based workflows, as a prompting signal, reinforcement reward, and evaluation metric. Second, LiNT-II is transparent: alongside a readability score it outputs all underlying feature values, enabling linguistically informed feedback in automated pipelines and user-facing writing assistants. The tool is open-source: https://github.com/vanboefer/lint_ii. References: Gubbels, J., Van Langen, A., Maassen, N. & Meelissen M. (2019). Resultaten PISA-2018 in vogelvlucht (Results PISA-2018 - An overview). Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, 42(3):643–650. Kleijn, S. (2018). Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing. Ph.D. thesis, Utrecht University. Netherlands Court of Audit (2016). *Aanpak van laaggeletterdheid (Tackling low literacy)*. Pander Maat, H., Kleijn, S., & Frissen, S. (2023). LiNT: een leesbaarheidsformule en een leesbaarheidsinstrument. *Tijdschrift voor Taalbeheersing*, 45(1):2–39. Speer, R. (2022). *rspeer/wordfreq*: v3.0.

KEYWORDS readability assessment, readability formula, Dutch

POSTER 11

Validating the LiNT readability metric on various text categories

Anneroo van Diermen, Natalia Amat Lefort & Wessel Kraaij

Leiden University

Measuring the accessibility of written text is an important task in the domains of education and formal communication. Among the current preferred methods for assessing difficulty of Dutch Text is the LiNT tool (Pander Maat, Kleijn, Frissen, 2023: <https://doi.org/10.5117/TVT2023.3.002.MAAT>). This readability formula uses the textual features word frequency, content words per clause, concrete words, and maximum syntactic dependency length. Our study examines the external validity of the LiNT tool, by testing whether the LiNT algorithmic predictions match real-world human reading difficulty for text genres beyond those on which the LiNT tool was created. To measure this real-world reading difficulty, we performed a ‘cloze’ study with 110 secondary school students each performing two cloze tests randomly chosen from 15 texts across 3 different text genres, namely Fiction (specifically novels), Terms & Conditions, and texts from the experiments used in the original LiNT study, in order to measure how well our experimental setup can reproduce the original experiment by Suzanne Kleijn. Cloze tests are texts where content words are deleted and replaced by gaps. Subjects in the study were asked to fill in the gaps, and their average performance is a measure for readability. Replicating the original Kleijn study, these cloze texts were ‘semantically graded’. This means that the

experimenter grades the words filled in by the subjects as plausible or not (binary grading). This step is necessary, but also introduces a subjective element. Therefore, we also ran an experiment to quantify this subjectivity by computing the inter annotator reliability of the semantic grading by having 15 independent raters score a total of 45 cloze gaps, 3 per text. The findings of the cloze scores compared with the LiNT analysis shows that LiNT successfully ranks relative difficulty between text genres, but differs significantly from actually obtained cloze scores per text. There was a non-significant medium correlation between the two scores. LiNT proved more accurate for the formal, legal texts of the Terms & Conditions, than for the more creative texts in the other two categories. Inter-rater reliability metrics, like Cohen's Kappa, were below standard thresholds, which shows that semantic grading of cloze texts is a very difficult task with subjective elements. A limitation of our inter reliability study is that the 15 raters did only judge the sentence with the gap, without full document context, in order to limit assessment time. Our study suggests that LiNT can be a valuable tool for comparisons, but does not give a perfect estimation of reading difficulty. Furthermore, measuring human reading comprehension remains difficult both from the aspect of validity as well as reliability. Measuring human reading comprehension remains difficult and time consuming for proper assessment of validity and reliability.

KEYWORDS readability measurement, validity, reliability, cloze tests

POSTER 12

FlemChecker: in search of linguistically and culturally Flemish data in Dutch corpora

Florian Debaene, Aaron Maladry, Walter Daelemans, Veronique Hoste

LT3, Language and Translation Technology Team, Universiteit Gent; CLiPS, Universiteit Antwerpen

Dutch language technology predominantly reflects Netherlandish norms. Encoder models such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020, 2022), decoder models such as GEITje (Rijgersberg & Lucassen, 2023) and the recent GPT-NL initiative (Van Oort et al., 2026), and benchmarks such as DUMB (de Vries et al. 2023) are built almost exclusively on Netherlandish data. This Netherlandish bias is further compounded by Anglocentrism in multilingual systems (Søgaard, 2022; Adilazuarda et al., 2024). This type of cultural bias in models leads to measurable degradation in tasks requiring cultural and pragmatic sensitivity such as emotion recognition (De Bruyne, 2023), irony detection (Maladry et al., 2025), and named entity disambiguation when applied to Flemish content. Yet Flemish Dutch is a recognized natiolect with its own phonological (Van De Velde et al., 1997), lexical (Bakema et al., 2004), and morphosyntactic (De Troij et al., 2023) conventions, shaped by distinct public institutions and sociolinguistic norms (De Caluwe, 2013, 2017; Dhondt et al., 2024). Without tools to identify and foreground Flemish data in Dutch corpora and model generations, evaluating or mitigating this bias remains difficult.

We present FlemChecker, a framework for detecting natiolectal variation in Dutch corpora, targeting three classes: Flemish Dutch, Netherlandish Dutch, and unmarked Dutch or sentences whose surface form is consistent with either variety. We conceptualize the unmarked class as the intersection of both natiolects, capturing the substantial shared core of written standard Dutch while preserving each variety's distinctive signal. We train and evaluate classifiers on two large labeled corpora of spoken Dutch (CGN) and written Dutch (SoNar), combining transformer-based models with interpretable n-gram approaches. The latter have previously demonstrated strong performance on subtitle-based variety classification (Van der Lee and Van den Bosch, 2017) and offer computational advantages for

large-scale corpus mining. (Beyond classification, FlemChecker incorporates a lexicon-based component for named entity recognition, enabling culturally grounded analysis of corpus content.) We validate the framework by applying it to unlabeled Dutch social media data from nLT-Tweets (Debaene et al., 2025) and to TRIC, a labeled Dutch irony corpus (Maladry et al., 2025), and use the resulting natiolectal labels to demonstrate the usefulness of framing the cultural representativeness of these datasets.

FlemChecker thus provides a practical, lightweight pipeline for mining Flemish data from large Dutch corpora or evaluating generated system outputs, supporting downstream Flemish-oriented NLP tasks that require linguistic and cultural specificity. By making natiolectal variation tractable at scale, FlemChecker addresses a foundational gap in Dutch language technology and offers a replicable methodology for similar low(er)-resourced natiolectal contexts.

KEYWORDS machine learning, variant classification, sociolinguistics

POSTER 13

Treebank Querying with Universal Dependencies

Mijail Kabadjov, Vincent Prins, Jan Niestadt, Koen Mertens, Jesse de Does, Vincent Vandeghinste

KU Leuven, Instituut voor de Nederlandse Taal

We introduce TrUDy, a workbench for querying Universal Dependency (UD) treebanks. TrUDy builds on concepts developed by GrETEL, a proponent of example-based corpus querying of Dutch treebanks (Augustinus et al., 2012), and to that end harnesses two other systems: BlackLab (de Does et al., 2017), a mature and flexible search engine for corpus indexing and token- and relation-based querying; and TextLens, a web-based platform for automated linguistic annotation for researchers in digital humanities, linguistics and translation studies (Van Hee et al., 2026). GrETEL (Greedy Extraction of Trees for Empirical Linguistics) is designed to make searching within treebanks (users can upload, parse and query their own corpora) accessible to linguists without knowledge of the grammar formalism or formal languages like XPath. The search starts with an example sentence entered by the user, which is then parsed and a query is automatically generated. TextLens is based on GaLAHaD (Depuydt & de Does, 2025), originally developed for historical Dutch texts. Both systems are hosted by the Instituut voor de Nederlandse Taal (INT). BlackLab is the backbone of the new workbench and the emphasis is on the querying aspect. The main goal of this work is to reuse the example-driven, user-friendly approach put forward by GrETEL and repurpose it in the space of Universal Dependencies (UDs), whilst also bridging with transformer-based representations, which encode deeper semantic relationships within text (Reimers & Gurevych, 2019). As a UD treebank query engine, BlackLab is related to PML Tree Query (Pajas & Stepanek, 2009), the main difference being the latter's backend is a relational database, whereas a Lucene-based search engine like BlackLab is better suited for incorporating transformer-based embeddings. The purpose of the system is to cover all use cases covered by GrETEL - such as, verb clusters and copular constructions in Dutch (Augustinus et al., 2017). The key difference, however, is that instead of XPath, BlackLab makes use of BCQL, an implementation of the Corpus Query Language. To illustrate, consider the following example: "Het vredegerrecht van Genk, dat nu een tijdelijk onderkomen heeft bij het vredegerrecht in Bilzen (Bilzen-Hoeselt), keert terug naar Genk." (The Genk Justice of the Peace Court, which is currently temporarily housed at the Justice of the Peace Court in Bilzen (Bilzen-Hoeselt), is returning to Genk). Based on the above, a query like 'who returns' can be encoded using a BCQL query for relations, which translates as

follows: "keert" -nsubj-> _, whereby we first specify the relationship nsubj (i.e., subject-verb relationships), we leave the subject slot open (see usage of the wildcard `\_` on the right-hand side of the relation), and finally anchor on the verb `keert`/terugkeren (to return). As part of this work, we are also investigating the feasibility of generating BCQL queries automatically using LLMs.

KEYWORDS Treebanks, Universal Dependencies, Parsing as a Service, Linguistic Infrastructure

POSTER 14

A Dutch corpus for studying emotion verbalization based on visual prompts

Hannah S. Rognan & Luna De Bruyne

University of Antwerp

Research on emotions in natural language processing relies heavily on datasets with known limitations: emotion categories are often predefined by researchers, annotations are provided by third parties rather than the people who experience the emotions, texts are frequently translated rather than written natively, and the compilation process provides limited opportunity for cross-corpus comparability. These choices introduce systematic biases into the data and constrain what kinds of emotional language patterns can be studied. We present a new emotion corpus designed to address these limitations through two methodological choices: using visual stimuli as prompts to elicit narratives without presupposing specific emotion categories, and allowing native-speaker participants to provide all affective measures themselves.

Fifty Dutch speaking participants each responded to 4 photographs selected to vary in emotional content, from the Geneva Affective PicturE Database (Dan-Glauser & Scherer, 2011). For each picture, participants wrote a short personal narrative, supplied free-form emotion labels, and provided their own ratings of valence and arousal at several points in the process: at the start of the survey as a baseline measure, upon seeing the image, and after completing the narrative. This produced a corpus of 200 Dutch texts with rich, participant-generated metadata including affect ratings, self-assigned emotion labels, and demographic information. Additionally, our use of the experimental tool PsychoPy allowed us to extract highly detailed behavioral measures, including the completion time for every step of the survey such as the time spent considering each image and the text writing duration.

In this work we demonstrate the range of research questions this corpus enables. We apply dictionary-based methods, including the NRC Emotion Lexicon and LIWC, to examine the lexical profile of narratives, and use SpaCy to explore syntactic features of emotion verbalization across both the narrative texts and the participant-supplied labels to understand language-specific features of emotion verbalization. We also examine the relationship between participant-provided affect ratings and the linguistic choices made in their narratives. Preliminary findings suggest associations between valence and arousal ratings and features such as token counts, density of different parts of speech and lexical diversity. Additional analyses include topic modeling to identify recurring themes across texts and embedding-based methods to extract emotion-relevant lexical and sentential content.

We aim to use this corpus to study the verbalization of emotions at the lexical, syntactic, and broader discourse levels, with implications for research on emotions in NLP, linguistics and psychology. Our corpus offers a foundation for studying how emotions are verbalized in Dutch and a design template replicable across other languages, supporting future cross-linguistic comparison.

KEYWORDS language resources, corpus construction, emotions, sentiment analysis

POSTER 15

Automatic Detection and Corpus Creation for Dutch Code-Mixing

Aaron Maladry, Loic De Langhe, Florian Debaene & Veronique Hoste

IT3, Language and Translation Technology Team, Ghent University

Code mixing is the linguistic phenomenon where a person uses two languages within the same sentence. This is most commonly observed in spoken language or on social media, where multilingual communities frequently alternate between languages (Aguilar et al., 2020). This phenomenon appears to be increasingly visible in online communication, particularly through the influence of English-language media and social networking platforms. This is supported by corpus-based studies of many languages, including Dutch and Flemish online discourse. Additionally, borrowing is not a simple “copy-and-paste” process, as English elements are often adapted graphematically, morphologically, and semantically within the target language, which adds an interesting layer of complexity to this phenomenon (De Decker et al., 2013). Although code mixing is receiving growing attention, the data used for pre-training large language models is often subject to language identification and filtering procedures that favour monolingual content. As a result, language models are not explicitly optimized for processing mixed-language input. Larger code-mixed varieties, such as Spanglish (Spanish-English) and Hinglish (Hindi-English), have become established research topics and are represented in dedicated benchmarks and evaluations like LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020). In contrast, Dutch-English code mixing has received comparatively little attention in computational linguistics and has primarily been studied from a linguistic perspective through experimental and theoretical analyses of bilingual speech (Vanen Wyngaerd, 2020). Consequently, large-scale corpus-based studies of Dutch-English code mixing remain scarce. As a first step towards addressing this gap, we propose an automatic method for detecting code-mixed data in Dutch social media corpora. Our methodology combines n-gram statistics derived from large monolingual English (COCA, GloWbE) and Dutch corpora (CGN, Lassy, Alpino, Sonar) with lexical information to identify code-mixed instances in existing datasets. Through manual evaluation, we assess the validity of our approach on social media data from both Reddit and Twitter, where code mixing is particularly prevalent. The main goal of this work is to facilitate future analyses of downstream tasks such as hate speech detection, argument mining, and emotion detection in Dutch-English code-mixed settings.

KEYWORDS Code Mixing, Dutch, Language Resources, Corpus Creation and Analysis**Sign language processing**

POSTER 16

SignNet, WordNet and the hurdle of natiolects and polysemy

Ineke Schuurman, Mijail Kabadjov, Caro Brosens, Kris Heylen

KU Leuven, Vlaams Gebarentaalcentrum, Instituut voor de Nederlandse Taal

For most sign languages the amount of annotated data is still rather small, meaning that up-to-date, AI-based approaches to Machine Translation between sign languages (SLs) and spoken languages (SpLs) are not yet available for most of them. One way to get more data is to extend the number of SpL words that can be linked to SL signs/concepts. Another is trying to make publically available online events, which include live sign-language dubbing. For SpLs a resource like Open Multilingual WordNet

might be useful to do so, for SLs we are working on a similar resource, SignNet, this time focussing on signs and the concepts expressed by these.

An essential issue is that concepts in SL are often broader than the 'related' ones in SpL or just smaller, as iconic aspects are rather important (like involving a horizontal vs a vertical movement) or because a specific concept does not exist with the intended meaning in SpLs (VGT "APPLAUS-DOOF" vs "APPLAUS-HOREND"). Moreover, once SL corpora are widely available, Resnik's semantic similarity based on Information Content can be harnessed to model broader to more specific concepts. When searching for synonyms, hyponyms or hypernyms of SL concepts in WordNet, the written language spoken in the surrounding is used. For both VGT (Flemish Sign Language) and NGT (Sign Language of the Netherlands) that will be Dutch. We are working on the input for a SignNet for VGT, that should in se be applicable to other languages as well, especially NGT. SpL synsets for such new concepts are temporarily added to our own interim version of Open Dutch/Multilingual WordNet. Nowadays Dutch is officially to be treated as a pluricentric language with natiolects (standard variants) such as Dutch Dutch (NN) and Belgian Dutch (BN). This is to be taken into account, especially when building resources that could be relevant to train (AI) NLP tools (not just SL-related ones). We are thus reworking concepts in both Word- and SignNets using updated versions of IVDNT dictionaries, words used by national media, starting with items in the VGT dictionary. ALL words should in se be marked as NN, BN or AN (Algemeen Nederlands). Currently word(meanings) are often marked 'Belgian', almost never 'Netherlandic' (although such words really do exist!). Quite often people in both regions are not familiar with the words used in the other region, or with polysemy issues (the main meaning of 'lopen', 'namiddag' or 'gracht' not being the same). Note that for SL regiolects (schools) are of importance as well. Of course all adaptations are to be discussed with a) the deaf community and b) the linguists behind Open Dutch WordNet as WordNets in general are still ignoring natiolects.

Schuurman et al. (2023) Are there just WordNets or also SignNets? 13th GWC 2023. Resnik (1995), Using Information Content..., IJCAI. De Caluwe (2013). Nederland en Vlaanderen, (a)symmetrisch pluricentrisme ... Internat. Neerlandistiek

KEYWORDS WordNet, SignNet, Natiolects, Polysemy, Pluricentrism

POSTER 17

Applying Sign Language Processing Techniques to Automatically Find Discourse Markers and Discourse Units on the LSFB Corpus

Adélaïde Couplet, Benoît Frénay & Laurence Meurant

Université de Namur

Sign language processing (SLP) is a field at the crossroads between natural language processing (NLP) and computer vision (CV). Nowadays, models for SLP are developed by reusing techniques from CV and NLP such as self-supervised pre-training [1], masking [2], contrastive learning [3], etc., and the goal of these models is often to translate sign languages (SLs) sequences sign by sign into text. Our approach tends to take the same issue from another point of view. We want to add more linguistic information to the process. Our goal is to train a model to automatically recognize discourse markers (DMs). Why DMs? Because according to the research of Johnston [4] and Gabarró-López & Meurant [5], DMs are markers that structure and delimit syntactic units, called "clause-like units" (CLUs) and "basic discourse units" (BDUs). We want to automatically find CLUs and BDUs to study those units

quantitatively and see if we can find recurrent patterns in SLs. This work starts by training a model on a precise DM, the palm-up (PU). It is a polysemous sign that can mean “I don’t know”, “I’m done talking” and can also be a boundary between two syntactic units. We are working with the LSFb-corpus [6] and parts of the data are already manually annotated with several occurrences of PU which allows us to train a small fully-supervised model. This work will make it possible to develop models to automatically annotate SL occurrences and ease this time-consuming task for the linguists who work on SLs and it will also help to find recurrent patterns in SLs.

[1] H. Hu, W. Zhao, W. Zhou, et H. Li, « SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no 9, p. 11221-11239, 2023. [2] W. Zhao, H. Hu, W. Zhou, J. Shi, et H. Li, « BEST: BERT Pre-training for Sign Language Recognition with Coupling Tokenization », in *Proc. AAAI*, vol. 37, no 3, 3597-3605, 2023. [3] R. Wong, N. C. Camgoz, et R. Bowden, « Learnt Contrastive Concept Embeddings for Sign Recognition », in *ICCVW*, Paris, France: 2023, p. 1937-1946. [4] T. A. Johnston, « Clause constituents, arguments and the question of grammatical relations in Auslan (Australian Sign Language): A corpus-based study », *SL*, vol. 43, no 4, p. 941-996, 2019. [5] S. Gabarró-López et L. Meurant, « Slicing your SL data into Basic Discourse Units (BDUs). Adapting the BDU model (syntax+ prosody) to Signed Discourse », in *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, 2016. [6] L. Meurant, « Corpus lsfb. Un corpus informatisé en libre accès de vidéos et d’annotations de la langue des signes de Belgique francophone (lsfb) », Namur, 2015.

KEYWORDS Sign Language Processing, Sign Languages, Discourse Markers

Speech, assessment & interaction

POSTER 18

Utterance-Level Methods for Identifying Reliable ASR-Output in Child Speech

Gus Lathouwers, Lingyun Gao, Catia Cucchiariini & Helmer Strik

Radboud Universiteit

Automatic Speech Recognition (ASR) is increasingly used in applications involving child speech, such as language learning and literacy acquisition (Bhardwaj et al., 2022). However, the effectiveness of such applications is limited by high ASR error rates, which are typically encountered in child speech recordings e.g. due high levels of noise (Dutta et al., 2022). The negative effects can be mitigated by identifying in advance which ASR-outputs are reliable. Use cases of such technologies lie in for example automatically transcribing large new children's speech corpora, which would reduce the amount of manual human verification required.

This work aims to develop a novel framework for selecting reliable ASR-output at the utterance level, one for selecting reliable read speech and one for dialogue speech material. Evaluations were done on an English dataset (CSLU; n=3534 utterances) and a Dutch dataset (JASMIN; n=2129 utterances). For each language, two Whisper models were applied, namely a baseline condition (Whisper-V2 Large) and a model finetuned to the respective language (Whisper-Medium Finetuned). For selecting reliable read material, an approach that compared ASR transcripts with the original intended prompt was used. For selecting reliable dialogue material, an approach employing LLM classification was used. For

the English language, because the original speech files comprised longer audio recordings, a pipeline was developed to parse the children's speech into smaller transcript utterances.

For read material, results show that the optimal strategy combining model agreement and a comparison with the original prompt achieves high precision rates for both Dutch and English ($P > 98.2$). For dialogue material, the optimal strategy combining model agreement and LLM-classification similarly resulted in high precision rates for both datasets ($P > 97.3$). Using the strategies applied in the current research would allow 21.0% to 55.9% of dialogue/read speech datasets to be automatically selected with low (UER of < 2.6) error rates.

References:

V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022

S. Dutta, S. A. Tao, J. C. Reyna, R. E. Hacker, D. W. Irvin, J. F. Buzhardt, and J. H. Hansen, "Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments," in *Proceedings of Interspeech 2022*, 2022, pp. 4322–4326

KEYWORDS asr child speech

POSTER 19

SAGE: Advancing Automated Scoring of Spoken Assessments

Bianca Ciobanica, Anaïs Tack

Itec - imec research group, KU Leuven

Despite growing use of spoken language assessments, technologies to help assessors during the evaluation process remain underdeveloped. Concerns about generative AI in writing have increased interest in spoken assessments (Desai, 2025), yet these rely on manual scoring, making them costly, slow, and sometimes inconsistent. There is a societal and technological need for more reliable and efficient automated solutions (Zechner & Evanini, 2019).

The SAGE (Spoken Assessments Guided by Enhanced technologies) project aims to address these challenges. SAGE is a collaborative research project funded by VLAIO and imec that brings together three industry partners (namely, Televic, Linguineo and Sensotec), and two research laboratories (IDLab UGent and ITEC, an imec research group at KU Leuven). Within this project, our work focuses on applying natural language processing and pre-trained language models (PLMs) to automatically calibrate questions and evaluate open-ended spoken responses, both on the language and the content.

In our study, we address two core challenges: the underperformance of PLMs on understanding informal learner speech patterns (Vajjala et al., 2025), and the need for transparent, scalable modeling. To this end, we apply open-source small language models (SLMs), namely Gemma4:8b, Gpt-oss:20b, and Phi4:14b, on both open (Knill, 2025) and proprietary datasets. We evaluate model performance using standard classification and regression metrics.

Initial results regarding the automated scoring of spoken assessments indicate that the transcription type (verbatim, grammatically corrected, without disfluencies, Whisper-generated) influences language proficiency prediction, with effects varying across speaking tasks. Additionally, our findings

reveal a systematic bias in SLMs: they tend to overestimate proficiency at lower levels while underestimating it at higher levels.

Sources: Desai, H. (2025, May 19). What's worth measuring? The future of assessment in the AI age. UNESCO. Retrieved January 30, 2026, from <https://www.unesco.org/en/articles/whats-worth-measuring-future-assessment-ai-age>

Knill, K., Nicholls, D., Gales, M. J. F., Qian, M., & Stroinski, P. (2025). The speak & improve corpus 2025: An L2 english speech corpus for language assessment and feedback. <https://doi.org/10.17863/CAM.114333>

Vajjala, S., Alhafni, B., Bannò, S., Maurya, K. K., & Kochmar, E. (2025). Opportunities and Challenges of LLMs in Education: An NLP Perspective (arXiv:2507.22753). arXiv. <https://doi.org/10.48550/arXiv.2507.22753>

Zechner, K., & Evanini, K. (Eds.). (2019). Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech. Routledge. <https://doi.org/10.4324/9781315165103>

KEYWORDS Automated Spoken Assessments, Natural Language Processing, Language Models

POSTER 20

Decoding Pragmatic Cues: Evaluating Audio-Enabled Large Language Models in Detecting Gricean Maxim Flouting

Alexis Liao

Leiden University

Traditional text-based communicative AI inherently loses paralinguistic features from user input, relying solely on linguistic data. In speech-enabled systems, modeling non-lexical vocal cues—such as pitch, intonation, speech rate, and pauses—is critical to rendering interactions natural and avoiding pragmatic mismatches. This study investigates the intersection of acoustic paralinguistics and pragmatic machinery by evaluating whether an audio-native large language model (LLM) can distinguish literal meaning from conversational implicature when guided by explicit paralinguistic cues. Grounded in Grice's Cooperative Principle, the experiment tests model sensitivity to the deliberate flouting of three conversational maxims: Quality, Quantity, and Manner.

Using the Qwen2-Audio-7B-Instruct model, we deployed maxim-specific prompts designed to focus the system's attention on acoustic and temporal structures. The evaluation dataset comprised 36 spoken utterances across 18 unique linguistic frames. Each frame was recorded in two distinct conditions: a neutral baseline (control group) and a paralinguistically over-marked version featuring overt vocal flouting (e.g., exaggerated rising-falling pitch for Quality, clipped delivery for Quantity, and artificial pauses/fillers for Manner).

The results reveal a clear hierarchy in the model's diagnostic capability: Quality > Manner > Quantity. The model demonstrated exceptional performance in detecting Quality flouting (100% true positives), reliably leveraging exaggerated pitch contours, contrastive intonation, and hesitation hedges to identify sarcasm and irony. Performance on Manner flouting was moderate but inconsistent (67% true positives); the model successfully flagged surface markers like fillers (um/uh) and vague phrasing but frequently conflated intentional conversational obscurity with general prosodic anxiety.

or disfluency . Conversely, performance on Quantity flouting was the weakest (50% true positives) . The model struggled to identify under-informativeness as an intentional flout, heavily conflating rapid speech rates with over-information and occasionally failing to incorporate obvious paralinguistic cues into its reasoning .

Furthermore, a systemic conservative bias was observed across all categories, yielding elevated false-positive rates on control items (56% overall control classification), driven partly by prompt-induced over-interpretation and occasional output artifacts . These findings suggest that while audio-enabled LLMs possess a robust capacity for mapping explicit tonal variations to non-literal meaning (Quality), they lack the discourse-level pragmatic reasoning required to evaluate context-dependent brevity (Quantity) . We conclude that future iterations of speech-empowered AI must better integrate discourse context with acoustic analysis to accurately calibrate communicative intent .

KEYWORDS Gricean Maxims, Paralinguistics, LLMs

POSTER 21

Towards automatic dental report generation: an exploratory analysis of clinical reporting practices

Colin Swaelens, Valentin Vervack, Guillaume De Moyer, Veronique Hoste & Els Lefever

Language & Translation Technology Team, Universiteit Gent

Automatic generation of clinical reports requires a thorough understanding of existing reporting practices. This paper presents a first exploratory analysis of clinical intake reports in a dental practice as groundwork for future report generation systems. Such systems may help alleviate the growing shortage of dentists in Flanders (Belgian Chamber of Representatives, Written Question No. 56-136, 2025) by assisting clinicians in drafting patient reports. In the envisioned workflow, findings derived from medical imaging, intra-oral scans, and clinical examinations are combined to generate a draft report that is subsequently reviewed and validated by a dentist. Similar human-in-the-loop workflows have already been explored in radiology (Sloan et al., 2024). The reports produced in current clinical practice constitute the target output of such systems; consequently, understanding their structure and variability is a necessary first step. The dataset comprises approximately 12,000 patient reports, covering either clinical intake consultations or implant treatment advice. Surgical reports are currently excluded, as the long-term objective is to generate reports from information available at the time of the initial examination. To obtain a first understanding of reporting practices, we conducted a qualitative and quantitative analysis of a balanced subcorpus of 104 reports produced by four dentists, comprising 13 intake reports and 13 implant advice reports per clinician. After removing headings and salutations, average report length ranged from 145 to 210 words, while the average number of unique tokens ranged from 98 to 142 per report. Quantitative analysis further revealed moderate but consistent inter-clinician variation: reports written by the same clinician exhibited higher lexical similarity than reports written by different clinicians (mean TF-IDF cosine similarity: 0.623 vs. 0.538). Qualitative analysis showed that, despite the use of a shared report template, clinicians regularly add, remove, or modify report sections and differ in their verbalisation of clinical findings. In particular, structured clinical measurements are often expressed through non-standardised qualitative descriptions rather than the underlying numerical values (e.g., Nibali et al., 2017; Tobias & Spanier, 2020), introducing ambiguity and reducing consistency across reports. These findings suggest that dental reports occupy a middle ground between rigid templates and free-text narratives. While

clinicians largely follow a shared reporting structure, substantial variation remains in the way clinical observations are verbalised. Future work will therefore focus on analysing the relationship between examination data and the resulting reports, providing the foundation for data-to-text generation models capable of handling both structured input and natural variation in clinical language.

KEYWORDS Clinical NLP, report generation, data-to-text generation, clinical documentation

Metaphor & figurative language

POSTER 22

Can Large Language Models Identify Lexical Metaphor in Wine Reviews? A MIP-Based Simple Benchmark Study

Léonieke Ariaans¹, Iris Hendrickx², Ilja Croijmans^{2,3}

1) Radboud University Nijmegen 2) Centre for Language Studies, Radboud University, Nijmegen 3) Donders Institute for Brain, Cognition and Behavior

Natural-language metaphors can be regarded as manifestations of core cognitive mechanisms, particularly processes, such as analogical inference, and when Large Language Models (LLMs) fail, implied meaning in practical applications may be overlooked (Tong et al, 2024). While lexical metaphors are typically used to share sensory experiences in wine reviews, systematic comparisons between LLMs concerning metaphor-identification have not yet been conducted to our knowledge. Recent research in computer science has explored the use of LLMs to identify and interpret conceptual metaphors at a larger scale than manual annotation would allow, creating the requirement of establishing if LLMs can support or replace manual annotation while maintaining linguistic rigor. However, their results suggest that current LLMs still face reliability limitations and do not yet consistently approximate human metaphor interpretation (Kim et al., 2023; Dmitrijev et al., 2024; Tian et al., 2024; Pedersen et al., 2025; Moa et al., 2024). As a methodological research gap has been identified concerning the application of Metaphor Identification Procedure (MIP) to analyze lexical metaphors with different LLMs, including both locally hosted and cloud-based models, and no studies have compared LLMs that meet transparency requirements for academic research to date, this study examined if LLMs can reliably be used to analyse lexical metaphors in wine reviews with results comparable to manual annotation in terms of inter-rater reliability (Ariaans, 2025). To this end, a single, standardised prompt was designed to evaluate the reliability and accuracy of each model in classifying lexical units as either lexical metaphors or non-metaphors [i.e. a binary classification] in accordance with MIP. The simple benchmark included each of the six models receiving both the wine review text and the part-of-speech tags, and relevant individual content words were analyzed, a strategy recommended in a newer manual annotation MIP, i.e. MIPVU. Consequently, all LLMs analysed an identical dataset containing non-expert wine reviews (i.e. Vivino) and expert wine reviews (i.e. Wine Enthusiast). Although the LLMs struggled consistently classifying binary data, all models generally represented contextual and basic meanings accurately. The reliability scores obtained: OLMo02:7b (Cohen's κ = .10), OLMo02:13b (Cohen's κ = .12), Tülu03:8b (Cohen's κ = .05), DeepSeek-r01:8b (Cohen's κ = .06), Llama03.1:8b (Cohen's κ = .11), and ChatGPT-05 (Cohen's κ = .44). Only ChatGPT-05 borders on moderate reliability obtained. Given that the tested models were base models rather than systems fine-tuned for metaphor identification, their performance is nevertheless promising, indicating that LLMs show significant potential for metaphor identification in wine

writing, regardless of expert-level, substantial challenges remain, concerning contextual sensitivity, interpretive transparency, and cross-cultural reliability.

KEYWORDS metaphor, metaphor identification procedure, Large Language Models, simple benchmark study, reliability analysis

POSTER 23

MethHOPE: A Severity-Based Error Annotation Framework for Evaluating Machine Translation of Metaphor

Jiahui Liang, Lifeng Han

Leiden University

Metaphors pose challenges for both machine translation (MT) and broader natural language processing (NLP) tasks. Recent advances in neural MT (NMT) and large language models (LLMs) have substantially improved translation quality, with some systems achieving performance comparable to human translators on general translation benchmarks (Kocmi et al., 2025). However, such improvements do not necessarily extend to metaphor translation (Han et al., 2026). Karakanta et al. (2025) report metaphor translation accuracy rates of only 64-80%, while Wang et al. (2024) find that around 20% of metaphorical expressions remain non-equivalent in translation. To better understand the gap between general MT performance and metaphor translation performance, it is necessary to systematically analyse metaphor translation errors. Existing studies mainly focus on translation strategies (Pedersen, 2017; Zajdel, 2022; Li & Chen, 2025) or translation quality, such as equivalence, fluency, emotional effect, and authenticity (Wang et al., 2024). However, fine-grained error analysis remains limited. To address this gap, this study adapts the HOPE framework (Gladkoff & Han, 2022) for metaphor translation evaluation. Originally developed as a lightweight version to Multidimensional Quality Metrics (MQM) (Lommel et al., 2014, 2024), HOPE reduces annotation complexity through a smaller set of error categories and a severity-based scoring scheme. Building on this design, we develop a metaphor-oriented annotation framework, MetaHOPE, that enables the systematic identification and severity assessment of metaphor translation errors. MetaHOPE consists of five error categories: Impact, Style, Mistranslation, Required Adaptation Missing, and Proofreading Error, together with a five-level severity scale (minor, medium, major, severe, critical). Using this framework, the study investigates: What types of metaphor translation errors are produced by different SOTA MT systems? How do the frequency and severity of errors vary across systems and translation directions (EN-ZH and ZH-EN)? We compare three MT systems: Google Translate, GPT-5.4, and Hunyuan-7B, an open-source LLM fine-tuned for translation tasks. The evaluation data are drawn from the news sections of two publicly available metaphor corpora: the VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al. 2010) for English and the Peking University Chinese Metaphor Corpus (PSUCMC) (Lu & Wang, 2017). From each corpus, 200 sentences containing metaphorical expressions were sampled, yielding 565 English metaphors and 368 Chinese metaphors. In addition, a pilot dataset of 20 sentences per translation direction was created to refine the annotation framework and assess annotation feasibility. Preliminary results show that our human annotators' agreement levels for [GoogleMT, GPT-5.4, Hunyuan-LLM-7B] are [0.536, 0.726, 0.333] for Pearson's correlation, and [76.9%, 70.8%, 61.5%] for exact agreement.

KEYWORDS Metaphor Translation Assessment, LLMs, Neural Machine Translation, Translation Annotation Framework

POSTER 24

Translating Political Metaphors in U.S.–China Trade Discourse: A Qualitative Comparison of Machine Translation and Large Language Model Outputs

Resul Yalcin, Xiaojuan Tan, Jelke Bloem

University of Amsterdam, University of Lille

Metaphors play an important role in political discourse by shaping interpretation, emotion, and ideological framing. Yet, metaphor translations pose persistent challenges for Machine Translation (MT) platforms and Large Language Models (LLMs). In this study, we evaluate how effectively Google Translate, DeepL, and ChatGPT (GPT-5) translate 300 conventional and novel Metaphor-Related Words (MRWs) from English to Dutch that occur in 80 sentences from the Political Metaphor Corpus at VU Amsterdam (VUPMC). The VUPMC English dataset contains political trade discourse between the United States and China. Each sentence was translated by all three systems, resulting in 240 translated sentences for qualitative analysis. The Metaphor-Related Words (MRWs) in all translated sentences were identified according to MIPVU (Metaphor Identification Procedure Vrije Universiteit Amsterdam) and CNMIP (Conventional and Novel Metaphor Identification Procedure). We rated the translations across four dimensions: translation quality, equivalence, emotional conveyance, and authenticity. The findings indicate that Google Translate, DeepL, and GPT-5 achieve high translation quality, with DeepL performing slightly better overall. Novel metaphors are translated more successfully than conventional metaphors, and most metaphorical expressions retain their conceptual structure and emotional intensity. However, systematic issues persist, including loss of metaphoricality, occasional mistranslations, and reduced emotional force. While contemporary MT platforms and LLMs are capable of translating political metaphors at an acceptable level, they cannot fully replicate the contextual sensitivity, cultural awareness, and interpretive judgment of professional translators.

KEYWORDS Machine Translation, Large Language Models, Metaphor Translation, Translation Evaluation

POSTER 25

Context-sensitive modeling of metaphoric potential with Large Language Models

Shihui Li, Xiaojuan Tan & Jelke Bloem

Universiteit van Amsterdam, Université de Lille

Research on cognitive metaphor suggests that semantic and psycholinguistic variables play a significant role in shaping a word's metaphorical potential. For instance, words with highly imageable meanings and concrete semantic fields are more likely to be used metaphorically, as such meanings tend to serve as the basic meanings of linguistic metaphors or as source domains in conceptual metaphors. Familiarity and iconicity may also influence metaphorical usage by affecting how readily these basic meanings or source domains are accessed and processed. Despite this theoretical interest, few studies have empirically examined how these variables interact in predicting a word's metaphoric potential. Moreover, existing work has almost exclusively operationalized psycholinguistic variables as static, lemma-level ratings, ignoring how meaning shifts with context.

Recent research indicates that Large Language Models (LLMs) can reliably predict psycholinguistic properties such as concreteness. Building on this, our work leverages LLMs to extend psycholinguistic feature ratings both out of context (lemma-level) and in context (token-level), enabling a more accurate investigation of how these variables influence word-level metaphoric potential. To this end,

we annotated 600 metaphor-annotated sentences from the VU Amsterdam Metaphor Corpus (VUAMC) with in-context human ratings for concreteness, imageability, familiarity, and iconicity. We then evaluate several LLMs for their ability to predict these psycholinguistic variables at both the lemma level (without context) and the token level (within context) against human ratings. This allows us to add LLM-estimated ratings to potentially metaphoric words in the entire VUAMC, which in turn makes it possible to quantify the extent to which each of our variables predicts metaphoric use. This provides empirical and quantitative support for theories of metaphor use and development.

KEYWORDS metaphor, computational psycholinguistics, lexical semantics

Multilingual & cross-lingual NLP

POSTER 26

Improving machine translation with named-entity awareness

Simon Bosch & Wafaa Mohammed

Universiteit van Amsterdam

This study investigates whether prompt engineering can enhance the MT quality by improving the quality of named-entity translation using open-source large language models (LLMs) on the XC-Translate dataset (Conia et al., 2024). Five LLMs of 7B-12B parameters in size (Tower Plus 9B (Rei et al., 2025), EuroLLM 9B (Martins et al., 2025) Gemma3 12B (Team et al., 2025) , Qwen 2.5 7B (Qwen et al., 2025) and Llama3.1 8B (Grattafiori et al., 2024)) are evaluated on the XC-Translate benchmark which contains then language pairs of English to X and manually-curated named-entity translations. Five prompt types are used to assess the effect on translation performance. These are baseline, zero-shot, few-shot, chain-of-thought and output-improvement. The MT quality is evaluated using the BLEU (Papineni et al., 2002; Post, 2018), CHRF (Popovic, 2015; Post, 2018) and COMET (Rei et al., 2020) scores, while the named-entity translation quality is assessed using the m-ETA score (Conia et al., 2024). The results show that prompt engineering improves MT and named-entity translation for almost every model-prompt pair. Few-shot prompting achieved the highest performance, while output-improvement achieved the worst. For Qwen2.5 7B and Llama3.1 8B even worse than the baseline prompt. The improvement of the few-shot prompt is largely due to the improvement of the English-Chinese and English-Japanese language pairs. However, performance varies across models: EuroLLM 9B, Tower Plus 9B, and Gemma3 12B show consistent improvements, but Qwen2.5 7B performs significantly worse overall. Several limitations are identified, including the inconsistencies in target translations for some language pairs and differences in model output formats. This study concludes that prompt engineering is an effective method for improving machine and named-entity translation with open-source LLMs, but further work such as statistical significance testing, dataset refinement, experiments with larger models and translation improvement can help to draw more generalizable conclusions and possibly improve machine translation even more.

KEYWORDS Machine translation, Named-entity recognition, Prompt engineering, Large Language Models

POSTER 27**Prompt-based Japanese lexical normalization with large language models**

Yefei Zhou & Jelke Bloem

Universiteit van Amsterdam

Withdrawn by the authors.

POSTER 28**Cross-lingual transfer of concreteness ratings from English to Arabic**

Salma Amrah, Jelke Bloem

Universiteit van Amsterdam

Concreteness is a lexical-semantic property that describes how easily a word can be experienced through the senses. The word "book" is highly concrete, while "freedom" is highly abstract. There is increasing research interest in using language models to predict lexical-semantic properties of words (Trott, 2024), but this work is mostly limited to English. Large concreteness datasets exist for English but most other resource rich languages, including Arabic, have very limited availability of annotated data.

The Kalimah dataset (Alzahrani et al., 2025), which contains ratings for 2,467 Modern Standard Arabic words, is the first large-scale Arabic concreteness dataset which makes a systematic comparison of computational methods for Arabic possible. This study asks whether concreteness ratings for Arabic words can be predicted using models trained on English data.

We compare three approaches: regression on static embeddings, regression on contextual embeddings and zero-shot prompting of a large language model. All models are trained on the English norms by Brysbaert et al. (2014) and tested on the Kalimah dataset. We perform an error analysis to see where each method succeeds and fails. Performance is measured using Pearson and Spearman correlations, along with the ARLUE benchmark (Abdul-Mageed et al., 2021).

KEYWORDS concreteness, Arabic, cross-lingual**POSTER 29****Benchmarking LLMs for multilingual aspect-based sentiment analysis in the hospitality domain**

Marie Dewulf, Orphée De Clercq, Mariet Raedts & Luna De Bruyne

Universiteit Antwerpen, Universiteit Gent

In the current digital era, online reviews are highly trusted sources of information that shape potential customers' decision-making processes and perceptions of a hotel (Leung et al., 2013). While hotel managers' responses can influence customer satisfaction and loyalty, the impact of different response strategies remains underexplored (Diouf et al., 2025). To better understand customer-hotel interactions, we extend the analysis of customer reviews by examining corresponding managerial responses. We employ aspect-based sentiment analysis (ABSA) to identify the topics that are mentioned in hotel reviews and may be addressed in hotel responses. This fine-grained analysis also supports our broader goal of automatically generating an appropriate response to a review.

ABSA breaks a review down into aspects, such as service or facilities, and assigns a sentiment polarity to each aspect individually (Zhang et al., 2023). As with various other NLP tasks, LLMs have now become a dominant approach for ABSA. However, most existing ABSA datasets are English-centric, which hinders robust multilingual evaluation. We address this gap by evaluating the capabilities of LLMs for multilingual ABSA in the hotel domain. We also extend the standard ABSA task by estimating the importance of a mentioned topic.

We have access to a dataset of hotel reviews in six different languages, retrieved from Booking.com. For each language, a small sample has been enriched with manual ABSA annotations and was subsequently used to benchmark state-of-the-art commercial and open-source LLMs for instruction fine-tuning and zero- and few-shot learning.

We will present these benchmark results on the separate ABSA subtasks and also report performance of a unified model that jointly addresses multiple ABSA subtasks, including 1) topic detection, 2) aspect sentiment classification and 3) aspect importance. In this respect, we will pay specific attention to possible differences between languages and perform an extensive error analysis. Ultimately, this work outlines the practical value and strategic implications of applying LLMs to hospitality analytics.

Diouf, S., Nakouri, H., & Ménélas, B.-A. J. (2025). Impact of hotel responses to online reviews on customer loyalty and acquisition: A longitudinal sentiment analysis with booking. *International Journal of Data Science and Analytics*, 20(8), 7257–7279. <https://doi.org/10.1007/s41060-025-00870-4>

Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. *Journal of Travel & Tourism Marketing*, 30(1–2), 3–22. <https://doi.org/10.1080/10548408.2013.750919>

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2023). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11), 11019–11038. <https://doi.org/10.1109/TKDE.2022.3230975>

KEYWORDS large language models, aspect-based sentiment analysis, hospitality

POSTER 30

The needle in the haystack: Retracing meaning among part-words in Turkish semantic similarity data

Nebi Eren Aygün & Giovanni Cassani

Tilburg Research Center for Cognitive Science and Artificial Intelligence

SotA language models' tokenizers encode words as the combination of sub-lexical tokens, which need not correspond to morphemes [1], especially in morphologically rich languages, multilingual models and lower-resource languages, such as Turkish. We evaluate different token composition strategies to produce static embeddings [2] against human similarity ratings in Turkish, to identify which tokens carry the bulk of the lexical meaning. We benchmark the following strategies: considering only first token, only the last token, averaging all tokens, or applying max pooling. Word representations were extracted from three transformer models: BERTurk (monolingual [3]) and two multilingual models, mBERT [4] and XLM-R [5]. Representations were sampled at three layers (1, 7, and 12), using two strategies to derive static, type-level embeddings from contextualised embedding models: isolated words and averaged contextualised embeddings of a target words across multiple sentence contexts.

Cosine similarities between the word embeddings were correlated with human judgments from the AnlamVer [6] using Spearman rank correlation. We ran paired bootstrap tests to establish statistical significance. Considering the first token produced the strongest correlations. BERTurk outperformed other models. In line with previous evidence [3], averaging contextualised representations and extracting embeddings at layer 7 worked best. Correlations were lower when words were tokenized more aggressively. Results from a odd-one-out test including triplets where a target word shared the stem with one word and the affix with another confirmed that considering the first token favored shared stems, suggesting that this strategy produced stronger correlations with human ratings because it better preserved the meaning of the stem.

[1] Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. In 2012 IEEE international conference on acoustics, speech and signal processing. IEEE. [2] Apidianaki, M. (2023). From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2). [3] Schweter, S. (2020). BERTurk: BERT models for Turkish [Data set]. [4] Devlin, J., Chang, M.-W., 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. [5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th annual meeting of the association for computational linguistics. [6] Ercan, G., & Yıldız, O. T. (2018). AnlamVer: Semantic model evaluation dataset for Turkish—word similarity and relatedness. In Proceedings of the 27th international conference on computational linguistics.

KEYWORDS Semantic similarity, tokenisation, word embeddings

POSTER 31

Investigating reasoning processes in large language models in a multilingual context

Mara C. Spadon, Roos M. Bakker, Daan L. Di Scala & Piek T.J.M. Vossen

Vrije Universiteit Amsterdam, TNO

This research investigates whether Large Language Models (LLMs) perform genuine multi-step reasoning or primarily rely on superficial statistical patterns when solving Machine Reading Comprehension (MRC) tasks that require logical reasoning. We examine this in a multilingual setting by comparing model behaviour in English and Dutch and by analysing whether intermediate reasoning traces align with the logical structure of a task rather than merely producing correct answers.

The ReClor benchmark [1] was selected as the primary evaluation dataset since it is a native English human-written multiple-choice MRC benchmark specifically designed to assess logical reasoning. As no Dutch equivalent existed, this research introduces ReClor-NL, a Dutch counterpart created by translating 4,500 ReClor training instances using DeepL and subsequently refining them through an automated LLM-assisted review pipeline. To support more detailed analyses, question-type labels were assigned using a classifier trained on the original labelled data, refined through LLM-assisted validation. To our knowledge, ReClor-NL is the first Dutch logical MRC benchmark.

Experiments compare GPT-5.2 and Mistral Medium 3.5 under both direct and chain-of-thought (CoT) prompting [2]. CoT prompting encourages the model to express its reasoning in intermediate steps,

decomposing the reasoning process and enabling step-level analysis. Beyond measuring accuracy, reasoning quality is evaluated through LLM-based assessment of intermediate reasoning steps and manual analysis of a subset of incorrect predictions. Additional experiments examine prompt-language effects, temperature sensitivity, robustness of answer and question-order shuffling, answer-label modifications, and performance on culturally specific questions.

The main experimental results indicate that GPT-5.2 consistently outperforms Mistral Medium 3.5, and that Dutch remains slightly more challenging than English across both prompting strategies. In addition, CoT prompting consistently improves accuracy by approximately two percentage points compared to direct prompting. Robustness experiments reveal minor performance changes under shuffling and relabelling conditions, suggesting that performance is driven by semantic reasoning rather than superficial answer-selection cues.

Overall, our research contributes a new Dutch benchmark for logical reasoning MRC, and a framework for analysing (multilingual) reasoning traces. Our work highlights the importance of evaluating LLM reasoning processes by looking further than final-answer accuracy.

[1] W. Yu, Z. Jiang, Y. Dong, and J. Feng, “ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning,” in International Conference on Learning Representations (ICLR), 2020 [2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv preprint arXiv:2201.11903, 2022

KEYWORDS Large Language Models (LLMs), Machine Reading Comprehension (MRC), Logical Reasoning, Multilingual NLP, Chain-of-Thought (CoT) Prompting

POSTER 32

Seeing the unseen: visual similarity for pixel language model adaptation

Ran Zhang, Miryam de Lhoneux & Wessel Poelman

KU Leuven

Pixel-based language models (LMs) replace traditional tokenizers by processing rendered images of text (Rust et al., 2023), making cross-lingual transfer heavily dependent on the visual and structural properties of writing systems (Rahman et al., 2023). However, the dynamics of adapting these models to low-resource languages with complex morphology and written in unique scripts are not yet explored. Using Tibetan as a case study, we analyze how continued pre-training of pixel-based LMs is influenced by data scale, initial script exposure, and cross-lingual transfer from languages written in other Brahmic scripts. We introduce four rendering-level metrics to quantify visual script similarity. We evaluate downstream performance across three tasks. Our results show that higher orthographic proximity enhances semantic transfer, even under severe data constraints. Additionally, we find a performance asymmetry based on the pre-training starting point: while multilingual pre-training PIXEL-M4 (Kesen et al., 2025) has stronger initial performance, its capacity for subsequent adaptation seems to be constrained, whereas adapting a monolingual model PIXEL (Rust et al., 2023) with mixed scripts yields more gains on sentence-level tasks. Our metrics and case study offer empirical observations that could help inform data selection and script adaptation choices when working with pixel-based models in similar low-resource settings.

reference: Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual Pretraining for Pixel Language Models. In Proceedings of the 2025 Conference on Empirical Methods

in *Natural Language Processing*, pages 29594–29611. Association for Computational Linguistics

Md Mushfiqur Rahman, Fardin Ahsan Sakib, Fahim Faisal, and Antonios Anastasopoulos. 2023. To token or not to token: A Comparative Study of Text Representations for Cross-Lingual Transfer. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 67–84. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language Modelling with Pixels. In *The Eleventh International Conference on Learning Representations*.

KEYWORDS Pixel language models, cross-lingual transfer, visual similarity, low-resource NLP, Tibetan

POSTER 33

Diagnosing Cross-Lingual Transfer in Bilingual Models & L2 Learners of Dutch

Aaricia Herygers & Lisa Beinborn

University of Göttingen

Large language models exhibit cross-lingual structural knowledge [1], but which grammatical phenomena transfer between languages, and to what extent, remains understudied. Phenomena of cross-lingual transfer have often been diagnosed in second language acquisition research [2], but comparison across studies remains difficult due to a lack of standardized definitions of transfer effects and quantifiable evaluation strategies for fine-grained phenomena.

We present ongoing work to develop a diagnostic dataset dedicated to explicitly measure transfer effects in Dutch. We use BLiMP-style [3, 4] minimal pairs and focus on specific transfer patterns from 11 source languages. We consider four transfer phenomena that are often observed in second language acquisition: word order, article assignment, grammatical gender, and negation placement. The dataset thus enables targeted evaluation of specific contrasts rather than aggregate cross-lingual performance.

We train 11 bilingual GPT-2 small models, each with a varying first language (L1) across a range of language families and Dutch as second language (L2). We measure both the overall transfer and the per-phenomenon transfer as the difference in accuracy compared to a monolingual Dutch baseline, allowing us to isolate the contribution of each typological contrast.

For the interpretation of the results, we encounter two main challenges. One challenge is distinguishing transfer effects from typological proximity: a model trained on a structurally similar language may perform better on Dutch not because knowledge actively transfers, but simply because fewer L1-L2 conflicts exist. To address this, we correlate transfer effect magnitudes with typological distances derived from WALS [5] features, allowing us to assess to what extent observed transfer effects are predicted by typological distance. Secondly, linguistic phenomena are not fully independent — e.g., word order and negation placement interact in ways that may produce compounding errors in minimal pair evaluations. We account for this interaction through careful paradigm design and cross-phenomenon analysis.

[1] Eronen, J., Ptaszynski, M., & Masui, F. (2023). Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3), 103250.

[2] Odlin, T. (1989). *Language transfer*. Cambridge, UK: Cambridge.

[3] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377-392.

[4] Suijkerbuijk, M., Prins, Z., Kloots, M. D. H., Zuidema, W., & Frank, S. L. (2025). BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4), 1267-1301.

[5] Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. *WALS Online* (v2020.4) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13950591>

KEYWORDS cross-lingual transfer, bilingual models, L2 Dutch, diagnostic dataset

POSTER 34

Language-Aware Structured Pruning for Multilingual Large Language Models

Jutte Vijverberg

Universiteit van Amsterdam, SURF

As multilingual large language models grow in size, compression methods such as pruning become increasingly important for reducing model size while maintaining performance. Unlike unstructured pruning, which removes individual weights regardless of structure, structured pruning removes entire components such as attention heads and MLP neurons, enabling hardware-friendly speedups. Yet the field remains heavily English-centric. While prior work has explored the role of calibration data language in unstructured pruning (Kurz et al., 2025), its effect on structured pruning methods remains unclear.

We address this gap by systematically analyzing how calibration language affects structured pruning across two models (Llama-3.1-8B and Qwen2.5-7B) and two pruning methods: LLM-Pruner (Ma et al., 2023) and 2SSP (Sandri et al., 2025). We further investigate whether instruction-tuning after pruning can recover lost performance, and how the choice of recovery language interacts with calibration language. Pruned models are evaluated on multilingual benchmarks spanning next-token prediction, linguistic, and reasoning tasks.

We find that English generally performs worst among the tested calibration languages. Furthermore, language-matched calibration yields the best results on perplexity and linguistic tasks. However, this advantage does not extend to multiple-choice reasoning, where calibration language has a weaker and less consistent effect. When performing recovery, we find that nearly all calibration and fine-tuning language combinations improve over no recovery, suggesting cross-lingual transfer benefits. Fine-tuning substantially reduces, but does not eliminate, calibration language differences. While language-matched recovery is optimal for some languages, multilingual recovery is a strong alternative that generalizes to unseen languages. Overall, our results show that calibration language is an important factor when pruning multilingual models, and that English is a poor default choice.

Kurz, S., Chen, J.-J., Flek, L., & Zhao, Z. (2025). On the Limitations of Language Targeted Pruning: Investigating the Calibration Language Impact in Multilingual LLM Pruning. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2408.14398>

Ma, X., Fang, G., & Wang, X. (2023). LLM-Pruner: On the Structural Pruning of Large Language Models. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2305.11627>

Sandri, F., Cunegatti, E., & Iacca, G. (2025). 2SSP: A Two-Stage Framework for Structured Pruning of LLMs. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2501.17771>

KEYWORDS Multilingual large language models, model compression, structured pruning, calibration language, recovery

Poster session 2

12:30 – 14:00 · Q building: Nelson Mandela Hall

Reasoning, puzzles & distillation

POSTER 1

Knowledge distillation for cryptic crossword solving

Rey Rashid, Antske Fokkens & Janneke van der Zwaan

Vrije Universiteit Amsterdam, Sopra Steria Netherlands

Cryptic crossword solving requires complex linguistic task that requires multi-step reasoning. Each clue encodes an answer through a definition component and a wordplay mechanism (such as an anagram, homophone, reversal, or hidden word). The clues have a misleading surface reading. Prior work has shown that even state-of-the-art large language models struggle with this task, particularly at the level of wordplay type identification and solution explanation. This work investigates whether reasoning ability can be transferred into a smaller, more cost-efficient model through knowledge distillation. We use GPT-5.4 as a teacher model to generate structured five-step chain-of-thought (CoT) reasoning traces for cryptic clues, including identifying the definition, wordplay type, indicator, executing the wordplay and reaching an answer. A student model (GPT-4.1-mini) is then fine-tuned on these traces on Azure AI Foundry. The teacher model is provided with both the correct answer and wordplay type and asked to explain the reasoning. This ensures better coverage of all wordplay types including rare and difficult ones. To maintain trace quality, we filter for traces that contain all predefined reasoning steps in the expected structure. Evaluation goes beyond standard exact match accuracy to assess whether distillation improves the quality of the student model's reasoning itself. We evaluate across four dimensions: exact match accuracy stratified by wordplay type, structural completeness of generated reasoning traces, wordplay type correctness, and answer-trace consistency. Questions addressed include: -Can reasoning ability be distilled into a smaller model through reasoning trace supervision? -Does fine-tuning improve not only answer accuracy but the quality of the student's reasoning? -Which wordplay types benefit most from distillation? This work contributes both a task-specific reasoning distillation pipeline and a multi-dimensional trace evaluation framework that can be used for tasks other than cryptic crossword puzzles.

KEYWORDS knowledge distillation, chain-of-thought reasoning, cryptic crossword solving, supervised fine-tuning, reasoning evaluation

POSTER 2

Optimising chain-of-thought reasoning for cryptic clue solving using LLMs as in-context teachers

Keze Hu, Pauline van Nies & Antske Fokkens

Vrije Universiteit Amsterdam, Sopra Steria NL

Cryptic crossword clue solving remains challenging to Large Language Models (LLMs) (Sadallah et al., 2025), since it requires wordplay expertise, world knowledge, and creativity for coherent, multi-step

reasoning to disambiguate and construct the final answer (Efrat et al., 2021). Despite the centrality of reasoning in the solution process, prior studies have largely emphasised final-answer accuracy, providing limited insights into the quality and optimisation of the reasoning traces.

To address this research gap, we adapt the Trace-of-Thought framework (McDonalds and Emami, 2024) to investigate Chain-of-Thought prompting with LLM-generated reasoning demonstrations. In this framework, a teacher LLM with stronger capabilities pre-generates stepwise reasoning traces for a small set of clues stratified by wordplay types. These traces are evaluated through a fine-grained LLM-as-a-Judge mechanism on final accuracy, decomposition quality, inference soundness, and structural coherence, thereby forming a curated pool of high-quality and reusable traces per wordplay to guide a student LLM in cryptic clue solving.

Based on this research design, the study answers the overarching question: To what extent can LLM-generated reasoning traces, curated as in-context expert-style demonstrations, improve another LLM's reasoning performance in cryptic crossword clue solving? Two sub-questions help operationalise this: 1) How does prompt-based optimisation with curated reasoning traces improve performance compared to standard CoT prompting across different wordplay types? 2) Which reasoning structures or strategies emerge in student LLMs' reasoning traces when guided by externally curated demonstrations? By foregrounding reasoning traces rather than accuracy alone, this work contributes a rigorous process-based prompting paradigm for supervising and optimising LLM reasoning for complex linguistic tasks.

References: Abdelrahman Sadallah, Daria Kotova, and Ekaterina Kochmar. 2025. What Makes Cryptic Crosswords Challenging for LLMs?. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5102–5114, Abu Dhabi, UAE. Association for Computational Linguistics.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A Cryptic Crossword Benchmark for Extreme Ambiguity in Language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tyler McDonald and Ali Emami. 2024. Trace-of-Thought Prompting: Investigating Prompt-Based Knowledge Distillation Through Question Decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 293–306, Bangkok, Thailand. Association for Computational Linguistics.

KEYWORDS Chain-of-Thought, reasoning optimisation, prompt engineering, cryptic crossword clues

POSTER 3

Evaluating Levels of Structured Guidance for LLM Reasoning in Cryptic Puzzles

Rico Nicolaas

Vrije Universiteit Amsterdam, Sopra Steria

Artificial intelligence has become increasingly useful across a wide range of applications, demonstrating strong performance on many language and reasoning tasks. However, current systems still struggle with multi-step reasoning. In cryptic puzzle solving, clues often require the solver to combine information from multiple knowledge domains to reach a solution. Research has shown that

decomposing a complex task into smaller subtasks through prompting techniques such as Chain-of-Thought (Wei et al., 2022) and Tree-of-Thought (Yao et al., 2023) prompting can significantly improve the performance of LLMs. However, previous studies on cryptic puzzle solving have focused exclusively on cryptic crosswords, which inherently provide a structured template that guides the reasoning process of LLMs. Constraints such as a predefined grid and letter counts substantially reduce the space of possible solutions, thereby simplifying the task.

Therefore, this paper aims to investigate whether the current reasoning capabilities of state-of-the-art language models are advanced enough to solve extremely complex cryptic clues which do not follow a conventional linguistic formula comprised of wordplay and definition indicators (Andrews and Witteveen, 2025). To simulate the thinking steps of a human solver, we augment LLMs with access to external knowledge sources and structural tools. Using linguistic puzzle entries from the AIVD Kerstpuzzel, an annual collection of cryptic puzzles published by the Dutch General Intelligence and Security Service (Algemene Inlichtingen- en Veiligheidsdienst, 2025), this study specifically explores what level of structured guidance these models require to successfully reach an answer. Understanding how and when structured guidance improves performance helps clarify the limitations of current models and indicates what kinds of methodological developments are needed to enhance their reasoning abilities. Finally, we examine whether these guidance strategies are transferable to similar but slightly different complex puzzles.

Bibliography:

Algemene Inlichtingen- en Veiligheidsdienst. 2025. AIVD Kerstpuzzel 2025. Technical report, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, Den Haag.

Martin Andrews and Sam Witteveen. 2025. A Reasoning-Based Approach to Cryptic Crossword Clue Solving. arXiv:2506.04824 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

KEYWORDS Multi-step reasoning, Large language models, Cryptic puzzle solving

POSTER 4

Does Teacher Matching Matter for Warm Starts in On-Policy Distillation?

Shaozhen Shi, Huiyuan Lai, Yftah Ziser, Yevgen Matuskevych, Malvina Nissim

University of Groningen

In natural language processing, knowledge distillation is a setup where a smaller language model learns from a larger one, making it cheaper and faster while retaining part of the larger model's ability. In standard distillation, the student learns from fixed examples prepared in advance. On-policy distillation (OPD) is different: the student generates its own responses, and the teacher provides guidance on those responses. At every point, the student has to continue from what it has already

generated, not from ideal examples prepared in advance. Using the student's own responses makes learning closer to actual use but also creates a problem: if the student starts from weak responses, the teacher's guidance may be less useful. We therefore ask whether basic reasoning support before OPD is enough, or whether a better warm start is needed, to first make the student closer to the teacher (Agarwal et al., 2024; Gu et al., 2024; Yang et al., 2025). We study this question in mathematical reasoning using a Qwen3-8B teacher and a Qwen3-0.6B student. We compare three ways of preparing the student before OPD: no extra preparation, preparation on filtered OpenThoughts-3 step-by-step math solutions, and fresh step-by-step solutions written by our own teacher for the same prompts. For the two data-based settings, we use either supervised fine-tuning or knowledge distillation. All students are then trained with OPD under matched settings and are evaluated on three commonly used data sets, AIME24/25 and GSM8K. Our preliminary results suggest that OPD depends not only on how well a warm start performs before OPD, but also on how closely it aligns the student with the teacher in terms of their token-level predictions on student-generated responses. In particular, warm starts based on solutions written by our own teacher seem to work better than warm starts based on public solutions. References Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos Garea, S., Geist, M., and Bachem, O. (2024). On-policy distillation of language models: Learning from self-generated mistakes. In Proceedings of ICLR 2024. Gu, Y., Dong, L., Wei, F., and Huang, M. (2024). MiniLLM: Knowledge distillation of large language models. In Proceedings of ICLR 2024. Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. arXiv:2505.09388. Guha, E. K., Ghalebikesabi, S., Yao, Y., et al. (2025). OpenThoughts: Data recipes for reasoning models. arXiv:2506.04178.

KEYWORDS on-policy distillation, knowledge distillation, reasoning models, warm start, mathematical reasoning

POSTER 5

Explainable Multi-Agent Reasoning through Explicit Agent Mental Models

Elke Vandermeersch, Tinne De Laet, Tim Van de Cruys

KU Leuven

Multi-agent systems based on large language models have shown promise for improving reasoning through debate, collaboration, and role specialization. However, many existing approaches rely on static coordination mechanisms such as majority voting or fixed agent roles, while lacking explicit modelling of agent reliability, expertise, and behavioural dynamics. This research presents the design and implementation of a teacher-centred multi-agent framework in which a dedicated meta-agent constructs and continuously updates mental models of participating language agents during interaction. The teacher agent maintains representations of their domain expertise, calibration, reasoning quality, recurrent failure patterns, and behavioural tendencies. These mental models are used to dynamically weight agent contributions, guide aggregation decisions, and adapt coordination strategies over time. The proposed framework is inspired by concepts from cognitive science, including theory-of-mind modelling, meta-cognition, and socially distributed reasoning. The teacher agent functions as a meta-reasoning entity that forms and revises internal representations of the other agents, analogous to how humans construct mental models of collaborators during cooperative problem solving. A central objective of the framework is to improve the interpretability of multi-agent reasoning. The teacher agent's decisions are transparently grounded in both the observable contributions of the participating agents and the teacher's explicit mental models of their respective strengths and weaknesses. This makes the aggregation process more inspectable than opaque voting

or selection mechanisms, while enabling analysis of why certain agents are trusted, weighted, or overruled in specific contexts. This architecture establishes connections between language-agent orchestration, cognitive modelling, explainable AI, and intelligent tutoring systems, while remaining compatible with open-source small language models in resource-constrained settings. The system will be evaluated in collaborative reasoning tasks to investigate whether explicit meta-reasoning and adaptive agent modelling improve the quality, robustness and transparency of multi-agent reasoning.

KEYWORDS Open SLM-based Multi agent systems, slm-mental model, explainable multi-agent coordination

POSTER 6

Modal sense disambiguation in language models: The role of context, training domain, and input frequency

Ezgi Başar, Yevgen Matuskevych, Annemarie van Dooren & Arianna Bisazza

University of Groningen

Modal verbs such as *may* and *must* can take on different meanings depending on discourse context, alternating between root interpretations (e.g., permission, obligation) and epistemic interpretations (e.g., possibility, belief). While modal expressions are known to be notoriously difficult for children to acquire due to their context-dependency and lack of physical grounding (Cournane & Pérez-Leroux, 2020; van Dooren et al., 2022), limited research exists on whether language models (LMs) can reliably distinguish between modal meanings. To address this gap, we conduct computational tests using two different methods across seven LMs that vary in size, architecture, and training data type. In addition to LMs pre-trained on large amounts of web data, we also employ smaller LMs pre-trained on child-directed language (CDL) to provide a more cognitively plausible reference point. Our first method probes vector representations of modal verbs to investigate whether LMs have distinct representations for each meaning. The second uses surprisal measures to assess whether LMs show a preference for one modal meaning over the other. Across both methods, we evaluate LMs on three different datasets while manipulating the availability of additional conversational context. We find that larger LMs successfully use contextual information to distinguish between root and epistemic meanings. In the absence of context, they exhibit a preference for the meaning most frequent in their training data. Smaller LMs trained on CDL or other limited input fail to show consistent sensitivity to modal meaning distinctions and do not benefit from context to the same degree. Our findings suggest that two factors thought to influence children's acquisition of modals, namely input frequency and the ability to make use of contextual cues, also guide how LMs disambiguate modal meanings.

References

Cournane, Ailís & Ana Teresa Pérez-Leroux (2020), 'Leaving obligations behind: epistemic incrementation in preschool English'. *Language Learning and Development* 16: 270–91.

Tulling, Maxime, Ryan Law, Ailís Cournane & Liina Pykkänen (2020), 'Neural Correlates of Modal Displacement and Discourse-Updating under (Un)Certainty'. *eNeuro* 8(1): 1-19.

van Dooren, Annemarie, Anouk Dieuleveut, Ailís Cournane & Valentine Hacquard (2022), 'Figuring out root and epistemic uses of modals: The role of the input'. *Journal of Semantics*, 1-36

KEYWORDS language acquisition, modals, modal sense disambiguation, language models, child-directed language

Model behaviour, evaluation & efficiency

POSTER 7

How Small Can We Go? Benchmarking Open-Weight Models for Sustainable Stance Detection

Nityaa Kalra, Dimitar Shterionov & Jisk Attema

Tilburg University

Language models are increasingly used for a wide range of Natural Language Processing (NLP) tasks and one such application is stance detection: the task of determining whether a text expresses support, opposition or no clear position towards a given target. However, comparisons with traditional baselines often rely on closed, commercial models. Smaller open-weight models remain underexplored, and existing evaluations tend to prioritize predictive performance over resource dimensions such as environmental cost. This is especially important in light of Green AI, which argues that computational cost and environmental impact should be treated as first-class evaluation criteria rather than afterthoughts.

Motivated by this gap, we ask whether smaller open-weight models can offer a better balance between predictive performance, carbon efficiency, and throughput for stance detection. We further examine how far model size can be reduced before performance begins to meaningfully degrade.

To answer this, we benchmark closed API models, including GPT-5.2, GPT-5-mini, and Claude-Haiku-4.5, against open-weight models from the Qwen3, Phi, DeepSeek, Llama, and Nemotron families, ranging from 1.7B to 14B parameters. We evaluate all models on three stance detection datasets: SemEval 2016, EZStance, and MTCSD. To ensure a controlled comparison, all open-weight models are served with vLLM on a single NVIDIA H100 NVL at temperature 0. We compare multiple prompting strategies and report average three-class Macro F1, prompt-level variance, CO₂eq emissions, and inference throughput.

Our results show that although closed models achieve the highest Macro F1, models around the 4B scale form a strong efficiency frontier: they remain competitive with larger counterparts while producing lower CO₂eq emissions and higher-throughput inference. We also observe that performance does not increase consistently with parameter count. Instead, it depends strongly on model family and post-training choices. For example, Qwen3-4B instruction-tuned outperforms larger Qwen3 variants, and Phi models offer one of the strongest overall trade-offs between performance. At the same time, models below 4B degrade noticeably, suggesting a lower bound beyond which aggressive downsizing compromises stance detection quality.

This positions the 4B range as a favourable trade-off point: small enough to reduce environmental and computational cost, but large enough to preserve reliable stance detection performance. For practitioners with limited compute budgets or restricted access to proprietary systems, these findings show that smaller open-weight models can support more accessible and sustainable stance detection while keeping the performance competitive.

KEYWORDS Stance detection, Small language models, Green AI, Computational efficiency

POSTER 8**Where Do the Errors Go? Class-Level Trade-Offs in Language Model Stance Detection**

Nityaa Kalra, Dimitar Shterionov & Jisk Attema

Tilburg University

Stance detection determines whether a text expresses support, opposition, or no clear position toward a given target. Despite its multi-class nature, many evaluation frameworks focus primarily on FAVOR/AGAINST performance, either by excluding the neutral class from the main metric or treating it as secondary. Recent stance detection studies have reported language models matching or even surpassing supervised baselines on these benchmarks. However, when a whole class is excluded or downplayed, it remains unclear whether these gains reflect better stance detection or simply a redistribution of errors.

We present a class-level analysis of predictions across closed commercial LLMs and smaller, open language models on three stance detection datasets: SemEval-2016 target stance, EZ-STANCE mixed target/claim stance, and MT-CSD conversational stance. Alongside aggregate Macro F1, we examine class-level performance across different model families (such as Qwen3, Phi, DeepSeek, Llama, and Nemotron) and prompting strategies. To ensure a controlled comparison, all models are served with vLLM on a single NVIDIA H100 NVL at temperature 0.

Our results show that while supervised baselines inherit biases from their training data, language models exhibit class-specific prediction preferences that vary across model families, likely reflecting differences in their post-training alignment. These preferences frequently shift ambiguous cases towards either NONE or FAVOR classes. We also find that aggregate gains often come from correcting one class while degrading another, rather than uniform improvements across all three classes. For example, a high NONE recall may indicate accurate neutral class detection, or an over-assignment of that label at the expense of FAVOR and AGAINST.

Therefore, we argue that stance detection should treat class-level performance as core evidence, not secondary consideration, especially with the increased use of language models.

KEYWORDS Stance detection, LLM prediction preferences, post-training alignment

POSTER 9**LoRA vs. DAPT in low-resource military Swedish domain adaptation**

Anton Wallin & Jelke Bloem

Universiteit van Amsterdam

Domain adaptation of language models allows for adjustment to different statistical language use in specialised linguistic domains. Domain Adaptive Pre-training (DAPT) is the primary method for this adjustment but often requires large amounts of data and computational resources, which can complicate deployment. This study investigates the possibility of utilising Low-Rank Adaptation (LoRA) as a low-resource substitute for DAPT. It does so by evaluating adaptation of the Swedish KB-BERT model to the domain of military Swedish. A corpus of a total of 692k tokens was sourced from publicly available Swedish Defence Force (Försvarsmakten) manuals and models were trained under varied data availability conditions (20%, 50%, 100%) and evaluated on PPPL alongside a qualitative analysis of first-layer embeddings. Lexical comparison showed a vocabulary overlap of 44,3% between

military Swedish and a Wikipedia sourced-corpus, affirming the distinctiveness of the domain. Both DAPT and LoRA models showed a statistically significant decrease in PPPL compared to the base KB-BERT. However, DAPT consistently outperformed LoRA, yielding a lower PPPL (7.93) under the 20% condition in a fraction of the training time compared to the 100% condition LoRA model (10.92). The results suggest that while LoRA can partially mimic the domain adaptation performance of DAPT, the former remains inferior with regards to both raw performance and resource efficiency at our dataset scale. Lastly, the qualitative semantic analysis displayed a realignment in the embeddings of sample words toward a militarily-coded language space, suggesting that Swedish military domain adaptation is likely beneficial for NLP applications.

KEYWORDS Domain Adaptation, BERT, Swedish

POSTER 10

Is Memorization Context-Sensitive? Prefix-Based Extraction Beyond Isolated Prefixes

Ali Satvaty, Narjes Sharafi, Jirui Qi, Suzan Verberne, Fatih Turkmen

University of Groningen, University of Leiden

Large language models (LLMs) can expose memorized training sequences under prefix-based extraction: given a prefix from a training example, the model may assign high probability to the original continuation. In deployed systems, however, prefixes are rarely evaluated in isolation. They often appear together with instructions, retrieved documents, or other task-specific context, as in retrieval-augmented generation (RAG). This motivates examining whether contextual conditioning mitigates memorization or merely changes the set of memorized samples that become extractable. We investigate this issue through paired item-level measurements of probabilistic suffix extraction. For each prefix-suffix pair, we score the target suffix under an empty prompt and under retrieved contexts of varying relevance, across three open-weight instruction-tuned models. We find that context does not simply erase memorization. Instead, extractable memorization consists of a context-robust core and a context-sensitive boundary. Many samples that are extractable without context remain extractable under the retrieved context, especially as the prefix length increases. At the same time, context mainly affects marginal samples near the extraction threshold: it suppresses some exposures, but also enables new ones that are missed by prefix-only evaluation. These findings qualify the view that RAG reduces memorization risk. Context can lower aggregate extraction by suppressing boundary cases, yet robustly extractable samples persist, and context-enabled extractability remains security-relevant.

KEYWORDS memorization, privacy, RAG

POSTER 11

Steering sycophantic drift at inference within a multi-turn conversation setting

Lennard Froma, Max van Duijn & Maaike H. T. de Boer

Universiteit Leiden, TNO

Sycophantic behaviour is the tendency of an LLM to conform to the user's beliefs rather than remain factual. This manifests, among other things, as excessive agreement with incorrect premises or unwarranted validation of flawed reasoning. As LLMs are increasingly deployed in collaborative

settings, this behaviour poses a concrete threat to the quality of human-AI collaboration as users exposed to sycophantic models often distrust these models, hindering collaboration..

A growing line of work addresses sycophancy and related alignment failures through activation steering, a technique that modifies a model's internal representations. Steering vectors are typically derived by computing the difference in residual-stream activations between contrastive prompt pairs that elicit or suppress the target behaviour. The resulting vector is then added to the hidden state at a chosen layer during the forward pass. This approach has been applied to suppress sycophantic behaviour, as well as to improve truthfulness, reduce refusals, and modulate a range of other model properties. But steering can also degrade the model's performance on inputs that do not require correction.

Existing methods are, however, predominantly evaluated on single question-answer pairs. This is not representative of real-world usage, where LLMs are deployed in extended, multi-turn conversations in which sycophantic drift may accumulate gradually across turns. This work addresses both multi-turn sycophancy and real-time, conditional intervention. We propose a method that trains a linear probe on the residual stream activations to detect impending sycophantic drift within a multi-turn conversation, and couples this probe with a steering vector that is applied conditionally, only when the probe fires, to suppress the drift before it manifests in the model's output. Our method is evaluated on the SYCON-Bench, in which we compare no intervention, always-on steering, probe-triggered steering and our method. This design isolates the contribution of conditional detection and contrasts inference-time with parameter-level intervention, providing a systematic account of when and how real-time steering can improve the honesty of LLMs in conversational settings.

KEYWORDS Sycophancy, LLM, Factuality, SYCON-Bench, probe steering, multi-turn conversation, drift

POSTER 12

Explainability-based token replacement on LLM-generated text

Hadi Mohammadi, Anastasia Giachanou, Daniel L. Oberski & Robert A. Bagheri

Utrecht University

Large language models (LLMs) can generate fluent texts that closely resemble human writing, but they often leave subtle linguistic patterns that make their outputs detectable. At the same time, rewriting or paraphrasing can reduce the reliability of AI-generated text detectors. This raises an important question for computational linguistics and NLP: can explainability methods help us understand, manipulate, and improve the detection of AI-generated text?

In this work, we investigate explainability-based token replacement as both an adversarial strategy and an analytical tool for AI-generated text detection. We train several classifiers, including XGBoost, BERT, DistilBERT, XLM-RoBERTa, and an ensemble model, to distinguish between human-written and AI-generated texts. The experiments are conducted on English and Dutch data across different domains, including news, reviews, and social media texts. The main dataset is based on the CLIN33 shared task corpus, which makes the work especially relevant for the CLIN community.

We apply SHAP and LIME to identify the tokens that most strongly influence AI-detection decisions. Based on these explanations, we propose several token replacement strategies, including human-preferred similar word replacement, part-of-speech-constrained replacement, GPT-based replacement,

and GPT-based replacement with genre context. We evaluate how these strategies affect detector performance and textual fidelity using classification metrics and text-similarity measures such as BLEU and ROUGE.

Our results show that explainability-guided token replacement can substantially reduce the performance of individual detectors, especially when influential tokens are targeted. However, the ensemble model remains more robust across languages and domains. A small human evaluation further suggests that rewritten AI-generated texts can also become harder for human readers to identify. Overall, the study shows the dual role of explainability in AI-generated text detection: it can reveal model reasoning, but it can also expose weaknesses that make detectors easier to evade.

KEYWORDS AI-generated text detection, explainable NLP, large language models, token replacement, multilingual NLP

POSTER 13

Evaluating Ontology Concept Placement Using LLMs: Not all placements are the same

Upal Bhattacharya, Maaïke de Boer & Sergey Sosnovsky

Utrecht University, TNO

Ontologies are semantic representations of knowledge in a domain and are widely used in various applications. Incorporating new conceptual information in an ontology is one of the most important tasks when updating ontologies requiring care to avoid introducing inconsistencies. However, updating ontologies is a time and resource-intensive process. Recently, LLMs have proven their capabilities in performing several knowledge-intensive language processing tasks. Consequently, there has been increasing interest in applying LLMs to various semantic tasks, including updating ontology taxonomies. However, most research in adding new concepts to an ontology (concept placement) view the objective as a monolithic task and do not account for the varying complexity in placing concepts in an ontology. In this study, we aim to investigate the ability of LLMs to place new concepts within an ontology at three levels of difficulty: as a leaf concept, as a new top-level concept and, as an intermediate concept. This work investigates the capabilities of LLMs to perform concept placement using advanced LLM strategies like Retrieval-Augmented-Generation (RAG) on the FoodOn ontology and analyzes the variation in performance across the three defined categories of placement. Utilizing different forms of context in the retrieval database, the approach analyzes: whether LLMs struggle with placing more 'abstract' top-level concepts as opposed to leaf and intermediate concepts, the disparity in predicting parents and children and, if retrieving more items leads to better performance. Initial experimentation suggests that LLMs are capable of placing intermediate and leaf concepts with better accuracy than top-level concepts. Our experimentation also highlights that LLMs are generally better at predicting parents of concepts than children. Retrieving more items from the RAG database (top_k from 10 to 300) does not lead to any performance benefits. From the initial observations, we aim to devise better strategies that incorporate semantically grounded information of ontology taxonomies when retrieving relevant items before predicting placements and contrast the performance with conventional embedding similarity-based RAG approaches that incorporate different forms of ontological information.

KEYWORDS ontology learning, concept placement, Semantic Web, LLM evaluation, RAG

POSTER 14

SoftwareCiter: Agentic Generation of Verifiable Software Citations with a Metadata-Aware Benchmark

Congfeng Cao, Yixian Shen, Jelke Bloem

Universiteit van Amsterdam

Research software is a fundamental component of modern scientific workflows, yet it is often mentioned informally in papers rather than cited as a first-class scholarly object, making software use difficult to trace, verify, reproduce, and credit. Existing work primarily studies software mention detection or local attribute extraction, leaving open the harder problem of transforming incomplete software mentions into complete, citation-ready, and evidence-grounded software references. We introduce SoftwareCiter, an agentic framework for end-to-end software citation generation from scientific publications. SoftwareCiter decomposes citation generation into publication parsing, software metadata extraction, missing-field planning, external metadata enrichment, FORCE11-compliant citation construction, and citation verification, enabling iterative refinement when citations are incomplete or inconsistent. We further introduce SoftwareCiteBench, a metadata-aware benchmark derived from the gold-standard SoMeSci dataset, comprising 1,217 publications and 2,605 manually enriched software citations across multiple metadata fields. We design a multi-level evaluation protocol that measures software detection, FORCE11 principle compliance, citation correctness, and fine-grained metadata correctness. Experiments show that SoftwareCiter substantially outperforms a single-call LLM baseline, improving overall metadata correctness from 0.435 to 0.704 and citation correctness from 0.345 to 0.717. We publicly release SoftwareCiter, SoftwareCiteBench, and the annotation resources.

KEYWORDS Software Citation, LLM Agents, Scientific Publications**Bias, society & political discourse**

POSTER 15

WinoQueer-NL: Assessing Bias in Dutch Language Models toward LGBTQ+ Identities

Jiska Beuk, Gerasimos Spanakis

Maastricht University

While English language models have been widely examined for anti-queer bias, Dutch models remain understudied. To address this gap, we developed a culturally and linguistically adapted Dutch dataset based on the English WinoQueer benchmark. The dataset contains pairs of stereotypical and counter-stereotypical sentences. To validate and expand it, we conducted an online survey with 43 Dutch queer participants, who evaluated 171 stereotypes. Of these, 145 were confirmed as culturally relevant and 22 new biases were identified through free-text responses. The final dataset, comprising 42,906 sentences, was evaluated using a range of Dutch-specific and multilingual models, including both masked language models (MLMs) and autoregressive language models (ARLMs) and is made available to the community. Bias was measured via a bias score, which compares log-likelihoods of stereotypical versus counter-stereotypical sentences. While the mean bias score across models appeared neutral ($\approx 50\%$), closer analysis revealed significant disparities. Some models favored stereotypical sentences up to 97% of the time for transgender identities, while producing them only 6% of the time for gay-

related pairs. Transgender and non-binary identities consistently received the highest bias scores, exposing uneven treatment despite overall neutrality. Our findings highlight the importance of culturally grounded datasets for evaluating and mitigating biases that disproportionately impact marginalized groups in Dutch language models.

KEYWORDS bias in LLMs, culturally adapted benchmarks, Dutch queer representation

POSTER 16

How good is frontier AI at spotting democratic decay? a logical audit of political elites' social media rhetoric

Husnain Raza, Nicolo Pennucci, J  r  my Dodeigne

University of Namur

Political hostility in digital spaces has intensified significantly in recent years, with extensive scholarship demonstrating the role of social media platforms in amplifying this surge. However, existing computational studies remain severely constrained in their ability to differentiate between standard partisan conflict and the specific forms of hostility that actively endanger democratic practices. The current work-in-progress paper addresses this gap by operationalising the Political Hostility Scale (PHOS) to systematically distinguish between democratically compatible friction (Agonism) and norm-eroding delegitimation (Antagonism) in elite political rhetoric. In compliance with recent social media data use policies and restrictions, which mandate the use of alternative approaches for classifying big data rather than training task-specific models, we design an empirical experiment to identify the optimal computational approach for this task. Specifically, we evaluate and compare the performance of multiple frontier Large Language Models (LLMs), encompassing both prominent open-source and commercial architectures, and diverse prompting strategies, including shots of zero, one, few and many against a gold-standard baseline of human annotations. We systematically investigate the capacity of these generative models to follow a logical approach when navigating the deeply contextual boundaries of hostile discourse. We used a diverse corpus of tweets from European politicians and the former US president, and our analysis evaluates how different prompting strategies and model scales affect classification accuracy in differentiating the boundaries between agonist and antagonist hostility. Furthermore, we analyse the systematic errors made by the models by mapping where algorithmic heuristics diverge from human logic, particularly when encountering complex edge cases, target ambiguities, and structural sarcasm. We provide actionable insights to improve model instruction guidelines and advance our understanding of how automated systems handle highly sophisticated normative distinctions in political hostility.

KEYWORDS Hostility Detection, Social Media, LLMs, Logical Flow

POSTER 17

Evaluating open-source LLMs for community-specific content moderation on Reddit

Jonathan Cowley & Jelke Bloem

Amsterdam University College, University of Amsterdam

Content moderation on Reddit is community-specific: each subreddit enforces its own norms, so the same comment can be acceptable in one community and removed in another. It is not well understood whether current open-weight language models can reproduce these community-specific decisions, or

what automating them would imply. We study 15 subreddits sampled from a 2025 to 2026 archive, comparing moderator-removed comments against approved ones.

Open-weight LLMs can replicate per-community moderation. A Qwen 3 14B model fine-tuned separately on each community reaches 79% balanced agreement with human moderators (Cohen's $\kappa = 0.573$), approaching the inter-moderator agreement ceiling of roughly 0.72 implied by reported disagreement rates. The same model used zero-shot is no better than chance on one community ($\kappa = 0.05$), which isolates fine-tuning rather than model scale as the decisive factor. A single adapter trained on all 15 communities matches the per-community models ($\Delta\kappa = +0.013$), so one shared model is sufficient, while prompted commercial systems are weaker on every subreddit, with Gemini 2.5 Flash at $\kappa = 0.209$ and Claude Sonnet 4.6 at $\kappa = 0.267$.

Replicating these decisions faithfully, however, also replicates what they encode. Estimating each comment author's political leaning and each community's moderators' leaning from cross-community participation, we find that human moderators remove comments more politically distant from them at higher rates, an effect concentrated in the most politically engaged communities and absent in apolitical ones. The fine-tuned models reproduce this asymmetry on held-out comments, removing politically distant comments at higher rates much as the moderators do. Automating a community's moderation by fine-tuning on its own removals therefore inherits its political asymmetry rather than neutralizing it.

KEYWORDS content moderation, large language models, fine-tuning, Reddit, political bias

POSTER 18

Semantics versus syntax: adjusting microportrait extraction by implementing semantic role labeling for Dutch

Sanne van der Wal, Antske Fokkens & Pia Sommerauer

Vrije Universiteit Amsterdam

The subjectivity of constituting what a stereotype is, makes stereotype detection a difficult task within Natural Language Processing (NLP). Designing annotation tasks to create data for training machine learning models then becomes challenging to avoid subjective annotations. The Microportraits pipeline avoids this subjectivity by extracting summaries about target entities directly from natural language itself (Fokkens et al., 2018). The use case for developing the pipeline is to compare how Muslims are represented in Dutch news articles compared to Dutch people. Each portrait of an entity contains the labels that are used to refer to a social category, the properties that describe an entity and their roles and behaviors. Biases can occur when speakers choose such labels, properties and roles to describe a certain entity or social group, which come to the surface through the patterns that exist within the extracted microportraits. However, the roles and behaviors as they are extracted in the original design of the Microportrait pipeline are based on syntactic dependencies, which we substitute in this study with a semantic role labeling (SRL) system for Dutch. To do this, we finetuned an XLM-RoBERTa model with Dutch and English data from the Universal Propositions bank to supplement the pipeline with SRL. The model has been evaluated using standard metrics, and yielded F1 scores for the agent and patient roles above 0.80. The SRL information is included in the Microportraits pipeline to be able to account for the semantic roles predicted by the SRL model. This information should provide more accurate semantic representations of the roles and behaviors assigned to specific entities, as well

as create microportraits that contain more semantic associations rather than syntactic associations. The output of the SRL-based pipeline and the syntax-based pipeline are compared on a quantitative and qualitative level. The quantitative level explores PMI scores with a minimum frequency threshold to investigate how the associations with target words changes between the syntax- and semantics-based portraits. The qualitative level evaluates how the quality of the extracted portraits changes in terms of how informative the extracted portraits are, and how the assigned roles and behaviors change. This is done by comparing both outputs to a manually annotated gold subset. The initial results show less noise in the SRL-based portraits, as the properties and roles capture semantic associations better compared to extracting all syntactic associations. Furthermore, the semantic arguments seem promising as they provide more accurate semantic roles than the syntactic-based heuristics do. This study would be a promising improvement of the Microportraits pipeline to research stereotypes as they occur in natural language. Fokkens, A., Ruigrok, N., Beukeboom, C., Gagestein, S., & van Atteveldt, W. (2018). Studying Muslim Stereotyping through Microportrait Extraction.

KEYWORDS Semantic Role Labeling, Stereotype Detection

POSTER 19

Quantifying different conceptualizations of sustainability in the New Space economy

Sanne van der Wal, Xiaojuan Tan, Diliara Valeeva & Jelke Bloem

Universiteit van Amsterdam, Université de Lille

In the rapidly developing commercial space industry, associated corporate activities carry significant environmental risks, such as orbital debris and atmospheric pollution. In response to emerging sustainability governance, companies conceptualize and communicate sustainability and environmental responsibilities in varying ways in their communications to the general public. In this study, we aim to quantify the different ways in which commercial space companies conceptualize sustainability in their website content across three regions, the US, Japan and the EU.

With limited data for computing word associations through traditional corpus methods or by training models from scratch, we explore to what extent pre-trained contextual word and document embedding models can help quantify the way in which the concept of sustainability is framed in each region. We create distinct contextual embedding vectors for words related to sustainability as used in each of the three regions, and observe a statistically significant difference between Japanese companies' use of sustainability terms and both the US and EU companies. Based on a critical discourse analysis of the websites of a few example companies, we then identify potential framings of sustainability concepts, select representative keywords for these framings (such as 'future', 'ecosystem', 'profit') and assess which region's contextual embedding of sustainability is closer to which keyword in embedding space. This analysis also shows keyword-specific differences by region, though not all framings observed in the critical discourse analysis could be quantified in this way.

KEYWORDS lexical semantics, framing, contextual word embeddings, sustainability

POSTER 20

Science worthy of news? Automated detection of newsworthiness and churnalism in university science communication

Luna De Bruyne, Miguel Vissers, Steve Paulussen & Gert-Jan de Bruijn

Universiteit Antwerpen

The growing reliance of science journalists on institutional communication has raised concerns about the increasing influence of university press offices on science news production (Vogler & Schäfer, 2020). Studying these dynamics at scale, however, remains methodologically challenging because linking press releases to news coverage and annotating relevant content characteristics traditionally requires extensive manual coding. This study therefore presents a computational approach for analyzing science communication pipelines.

We collected 792 press releases distributed by all Flemish universities during the COVID-19 pandemic and linked them to related news articles from major Belgian newspapers. To identify news articles that were based on the press releases, we developed a multi-stage filtering pipeline that combines temporal filtering, cosine similarity matching based on sentence-transformer embeddings, and LLM-based relevance classification. A manual check of this filtering method revealed there was an error margin (false positive/negative) of 5.5% throughout the whole dataset, giving us a final dataset of 1,783 relevant news articles.

The manually corrected dataset was then used to investigate the influence of newsworthiness on the uptake of press releases by journalists. Newsworthiness was analysed by identifying the presence or absence of seven news factors: reach, controversy, continuity, surprise, influence, prominence and trigger (Kroon and Schafraad, 2013). After benchmarking multiple open-source models on a manually annotated gold-standard subset, a combination of Gemma 4 and Llama 3.3 was selected (macro F1 between 0.66 and 1). Finally, to assess journalistic reuse of press releases, we employed the Churnalism Index (CHINDEX), a composite similarity measure based on Jaccard similarity, cosine similarity, and Levenshtein distance.

The results show that press releases remain highly influential in science news production: over 60% of research press releases generated news coverage. News factors such as trigger and reach significantly increased the likelihood of news uptake, while substantial levels of textual reuse were observed across the corpus. Although newsworthiness helped explain news selection, it was less predictive of copying behaviour, which was more strongly associated with publication characteristics such as online versus print publication.

References:

Vogler, D., & Schäfer, M. S. (2020). Growing Influence of University PR on Science News Coverage? A Longitudinal Automated Content Analysis of University Media Releases and Newspaper Coverage in Switzerland, *International Journal of Communication*, 14, 3143–3164.

Kroon, A., & Schafraad, P. (2013). Copy-paste of journalistieke verdieping? Een onderzoek naar de manier waarop nieuwsfactoren in universitaire persberichten nieuwsselectie en redactionele bewerkingsprocessen beïnvloeden. *Tijdschrift voor Communicatiewetenschap*, 41(3), 283-303.

KEYWORDS automated content analysis, computational social science, computational journalism, LLM annotation

POSTER 21**Evidence Use in Political News**

Andreea-Gabriela Ion, Federico Pianzola, Sara Nabhani, Khalid Al-Khatib

University of Groningen

News articles do more than report events. They may include claims and interpretations supported by attributed statements, concrete examples, and statistics. In this work, we examine whether the use of these different forms of evidence differs across topics, news outlets, and political orientation. We fine-tune a ModernBERT sequence-labelling model on Webis-Editorials-16 to identify five types of argumentative units: assumption, anecdote, testimony, statistics, and other. Because the model is trained on editorials, we first test how well it transfers to regular news. Results show that relaxed F1 falls from 0.765 on the editorial test set to 0.608 on 45 manually annotated news articles. Testimony transfers most reliably, while the error analysis shows frequent problems with span segmentation, noisy text, missed units, and ambiguity between categories.

We then apply the model to 7,616 articles from NLPCSS-20 and to a second dataset of 660 articles from AllSides, covering 220 events with one left-, centre-, and right-rated article for each event. Across both datasets, assumptions, testimony, and anecdotes are the most common predicted argumentative units. The most apparent differences are related to topic and outlet. Economic reporting contains much more statistical evidence, while reporting on violence, terrorism, and criminal justice contains more testimony and anecdotes. Some differences also appear between articles from left-, centre-, and right-rated outlets in the larger corpus, but the effect sizes are generally small. When the analysis is done at the outlet level, these political differences are no longer statistically significant.

Overall, we find limited evidence for a particular left, centre, or right pattern of evidence use. Topic and outlet show more noticeable variation. The lower performance on news compared with editorials also shows why argument-mining models should be evaluated on the target genre before their predictions are used for large-scale analysis.

KEYWORDS Argument mining; Evidence type classification

POSTER 22**A researcher's role in shaping AI policy and literacy in higher education**

Heike Pauli, Michael Bauwens

UCLL

As generative AI becomes more widely used, organisations need more than ad hoc experimentation: they need clear policy, shared guidance and a realistic understanding of employees' AI literacy. At UCLL University of Applied Sciences, these needs come together in a collaborative approach that connects AI governance with practical support for staff. We present how we have contributed to the recently launched AI policy of UCLL as Language Technology researchers.

Our GenAI Readiness Dashboard is designed as a growth-oriented instrument that helps organisations assess readiness for generative AI across three core dimensions: vision and strategy, AI literacy, and responsible use. Rather than producing a static score, it supports a staged development path from demystification and exploration to application – all based on hands-on experience in training professionals on generative AI. At UCLL, we practise what we preach by using the Readiness

Dashboard to measure and strengthen employees' AI literacy across the institution. In doing so, it supports the operationalisation of the new AI policy and contributes to a broader culture of responsible, human-centred AI adoption. We highlight how policy and literacy measurement can reinforce one another: policy provides structure, expectations and governance, while the dashboard offers a concrete way to monitor readiness, identify support needs and guide professional development.

This combined approach is also timely in light of the EU AI Act, which increases the importance of AI literacy, accountability and appropriate organisational measures. UCLL's experience suggests that AI policy is most effective when it is not treated as a stand-alone compliance document, but as part of a broader learning and change process in which employees are actively supported in developing the knowledge, attitudes, and practices needed for responsible AI use. As the AI policy at UCLL is grounded in the expertise of researchers in Language Technology, this contribution is especially relevant to academic researchers in the field who are exploring how disciplinary expertise can contribute to institutional AI governance and literacy.

Disclosure of use of generative AI: During the preparation of this abstract, the authors used Copilot for Microsoft 365 for the purpose of rephrasing and rewriting of the text. The authors reviewed and edited the content as needed and take full responsibility for the content of the abstract.

KEYWORDS AI policy, AI literacy, generative AI, HR

Emergent communication & cognition

POSTER 23

Compositionality beyond referential games: emergent communication in survival-oriented action-based environments

Luke Kraakman, Phong Le & Raquel G. Alhama

Universiteit van Amsterdam, University of St. Andrews

Emergent communication research investigates how artificial agents develop communication protocols through interaction while solving cooperative tasks. A central question in this field is how agents develop compositional communication systems, in which complex meanings are constructed from reusable parts. Much of the existing literature studies compositionality in referential games, where agents communicate about static objects in controlled environments. However, it remains unclear whether the factors that promote compositional communication in these settings generalize to more realistic environments that require behavioural coordination and context-dependent decision making. This thesis investigates whether established drivers of compositionality from referential games generalize to a survival-oriented communication game in which agents must coordinate actions under partial observability.

A two-agent grid-world environment was developed in which a speaker observes the full environmental state while a listener must select appropriate actions based on its own information and the received message. The environment introduces context-dependent objectives involving food acquisition, danger avoidance, and managing internal hunger states. Experiments examine the effects of communication capacity, environmental complexity, curriculum learning, entropy regularization, communication channel discreteness, and self-play on both task performance and emergent language

structure. Compositionality is evaluated using Topographic Similarity, Positional Disentanglement, and Bag-of-Symbols Disentanglement. In addition, our work introduces Positional Mutual Information (PosMI), a novel information-theoretic framework for analysing how semantic information is distributed across message positions.

The results show that several drivers of compositionality identified in referential games remain relevant in action-based environments, although their effects are often less direct and depend on interactions between communicative pressures and learning dynamics. Intermediate communication capacities produced the most favourable balance between performance and compositionality, while curriculum learning, entropy regularization, and self-play consistently influenced language structure. The experiments further reveal that compositionality and task performance are related but distinct properties, as more compositional languages did not necessarily achieve higher accuracy. PosMI analysis demonstrates that emergent languages frequently rely on distributed and redundant encoding strategies that are not captured by traditional compositionality metrics.

These findings suggest that compositional communication can emerge in survival-oriented environments, but that its emergence is shaped by multiple interacting pressures. The proposed PosMI framework provides a complementary tool for analysing the internal organization of emergent languages in complex multi-agent systems.

KEYWORDS Emergent Communication, Referential Games, Action Games, Compositionality

POSTER 24

Magnitude symbolism without sound: investigating phonetic size-symbolic associations in Large Language Models

Ewelina Kowalczyk, Kiana Shahrasbi & Tessa Verhoef

Universiteit Leiden

Magnitude sound symbolism is a cross-linguistic human cognitive bias linking particular sounds with certain sizes (e.g. high-front vowels with smallness). While cross-modal associations are well-studied in humans, it has recently also become an area of growing academic interest within computational models and artificial intelligence [1]. Alignment of visuo-linguistic representations can facilitate more effective interactions between humans and machines [2]. Although sound symbolism is a multimodal phenomenon, this paper investigates whether size sound-symbolic associations can be learned even without sound, from just textual data. Modern Large Language Models (LLMs) have been found to develop internal representations that capture detailed aspects of our sensory and perceptual world [3], even without embodied experience. The goal of this research is to understand whether Large Language Models (LLMs) encode human-like magnitude sound symbolism. The experiment is based on an existing large-scale cross-linguistic human study [4]. It uses 40 disyllabic nonce words as stimuli for size judgments which are evaluated across three phonetic features (vowel height, vowel backness, and obstruent voicing). The task is performed by three LLMs (gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B). For each word LLMs judge their size on a numerical scale and token log-probabilities are extracted from models' answers to obtain a more detailed insight into their decisions and capture the nuances of sound-symbolic associations. Findings show that all three LLMs display some sound-symbolic associations, however, each model to a different degree. Gpt-4o-mini most accurately reflects human magnitude sound-symbolic biases, almost perfectly replicating all human phonetic mappings.

DeepSeek-V3.1 also reflects human biases accurately, however, only partially. In contrast, the only significant effects exhibited by Qwen3.5-9B involve a reversal of the human vowel height trend. These findings suggest that human-like magnitude sound-symbolic associations can be successfully extracted from textual data, given controlled and well-curated conditions. They contribute to the debate about the capabilities and limitations of computational models in reflecting phenomena grounded in human cognition.

[1] Kouwenhoven, T., Shahrabi, K., & Verhoef, T. 2025 Cross-modal Associations in Vision and Language Models: Revisiting the Bouba-Kiki Effect. *NeurIPS*, 39, pp. 76452-76479. [2] Kouwenhoven T, Verhoef T, de Kleijn R, Raaijmakers S. 2022 Emerging Grounded Shared Vocabularies Between Human and Machine, Inspired by Human Language Evolution. *Front. Artif. Intell.* 5, 886349. [3] Marjeh R, Sucholutsky I, van Rijn P, Jacoby N, Griffiths TL. 2024 Large language models predict human sensory judgments across six modalities. *Sci Rep* 14, 21445. [4] Shinohara K, & Kawahara S. 2010 A cross-linguistic study of sound symbolism: The images of size. In *Annu Meet Berkeley Ling Soc*, pp. 396–410.

KEYWORDS Magnitude sound symbolism, Large Language Models, Cognitive science, Textual grounding

POSTER 25

Tracking the emergence of linguistic structure in self-supervised models learning from speech

Marianne de Heer Kloots, Martijn Bentum, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, Willem Zuidema

University of Amsterdam, Radboud University, Tilburg University

Self-supervised speech models learn effective representations of spoken language, which have been shown to reflect various aspects of linguistic structure. But when does such structure emerge in model training? We study the encoding of a wide range of linguistic structures, across layers and intermediate checkpoints of six new Wav2Vec2 and HuBERT models trained on 831 hours of spoken Dutch. Our analysis suite includes probes at the level of phonetic, syllabic, lexical and syntactic structure. We find that different levels of linguistic structure show notably distinct layerwise patterns as well as learning trajectories, which can partially be explained by differences in their degree of abstraction from the acoustic signal and the timescale at which information from the input is integrated. Moreover, we find that the level at which pre-training objectives are defined strongly affects both the layerwise organization and the learning trajectories of linguistic structures, with greater parallelism induced by higher-order prediction tasks (i.e. iteratively refined pseudo-labels).

Reference: de Heer Kloots, M., Bentum, M., Mohebbi, H., Pouw, C., Shen, G., Zuidema, W. (2026). Tracking the emergence of linguistic structure in self-supervised models learning from speech. <https://arxiv.org/abs/2604.02043>

KEYWORDS speech processing, interpretability, learning dynamics

Stylometry, literature & creativity

POSTER 26

Towards a new writing style: Stylometric drift and authorship in human-AI written texts

Ilinca Rotaru & Giovanni Cassani

Tilburg Research Center for Cognitive Science and Artificial Intelligence

We show that human-AI co-written texts can be identified as a distinct stylistic category and that their linguistic profile changes as AI intervention increases. We sampled 1K human-written texts from the PAN'25 Generative AI Detection Task 2 dataset [1] and created a controlled corpus which consisted of four versions for each source text: three hybrid versions produced through low-, medium-, and high-intensity AI rewriting, and one fully AI-written text generated from extracted topic and text metadata. We compared interpretable stylometric features (function words, PoS tags, syntactic and lexical measures) with TF-IDF-weighted character n-grams in their ability to identify how each text was generated using logistic regression, Support Vector Machines, and Random Forests. Distance-based analysis measured stylometric drift from the original human texts. Results show that human, hybrid, and fully AI-written texts can be classified as distinct categories, with character n-grams producing the strongest performance. Human and low-intensity AI edited texts were more often confused, mid- and high-intensity AI edits were most often confused, whereas fully AI generated texts were hardly ever misclassified, suggesting that the original stylistic features are progressively diluted but still indicate human authorship in a tightly controlled dataset. Accordingly, stylometric drift increased with rewriting intensity: lightly edited texts remained closest to the human baseline, while high-intensity rewrites and fully AI-written texts moved further away. These findings suggest that human-AI co-written text is best understood as a continuum of authorship rather than a binary human/machine distinction. Our work complements current research on AI-assisted writing [2] as well as work on detecting AI-generated texts [3], highlighting that new forms of authorship afforded by LLMs should not be dichotomized as human versus AI generated, but rather characterized along the interplay between the two [4].

[1] Wang, Y., Shelmanov, A., Mansurov, J., Tsvigun, A., Habash, N., Alham Fikri, A., Artemova, E., Xie, Z., Su, J., Xing, R., Gurevych, I., & Nakov, P. (2025). PAN'25 Generative AI Detection (Task 2): Human-AI Collaborative Text Classification [Data set]. Zenodo. [2] O'Sullivan, J. (2025). Stylometric comparisons of human versus AI-generated creative writing. *Humanities and Social Sciences Communications*, 12(1). [3] Richburg, A., Bao, C., & Carpuat, M. (2024). Automatic Authorship Analysis in Human-AI Collaborative Writing. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. [4] Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *CHI Conference on Human Factors in Computing Systems*.

KEYWORDS stylometry; text mining; AI-assisted writing

POSTER 27

Translation as Palimpsest: Authorial and Translatorial Signals in Ken Liu's English Translations of Chinese Science Fiction

Xiaorui Yu

King's College London

English translations of Chinese science fiction (CSF) occupy an ambiguous linguistic position: they circulate as English-language SF while remaining shaped by Chinese source texts, authors, and narrative worlds. This paper asks, from a linguistic perspective, what makes CSF “Chinese” after translation into English. Drawing on Rybicki's metaphor of translation as a palimpsest (2025, 284), it examines whether CSF is pulled towards the linguistic norms of English-language SF, or whether it continues to preserve detectable traces of its Chinese source texts and source authors.

Ken Liu is used as a case study because he is both one of the most visible translators of contemporary CSF and an established Anglophone writer of speculative fiction. The corpus consists of forty CSF translations by Ken Liu, covering Cixin Liu, Jingfang Hao, Qiufan Chen and Xia Jia, amounting to 1,029,217 tokens, alongside 52 of Ken Liu's original SF short stories, totalling 311,219 tokens.

The methodology proceeds in three stages. First, supervised classification examines whether the four CSF authors remain distinguishable within Ken Liu's translated corpus. Ten train-validation splits were generated using two chunk sizes, seven feature settings, and three attribution methods (Support Vector Machines, Burrows's Delta and Naïve Bayes). Following Hill and Tolonen (2021), only combinations achieving 100% validation accuracy were retained. Second, pairwise comparisons test whether Ken Liu's translations of each author can be distinguished from his original SF writing. Third, PCA and close reading of outlying samples are used to identify linguistic features underlying the computational separation.

The preliminary findings suggest that source-author signals remain recoverable within Ken Liu's translations: 2,006 of 4,200 first-stage result files achieved 100% validation accuracy. The separation between Ken Liu's translations and his original writing is even stronger, with perfect-model rates above 97% across all four pairwise comparisons. Close reading of word-bigram outliers suggests that the distinction is shaped by multiple textual dimensions, including dialogue and proper-name patterns in Jingfang Hao, cosmological and technical vocabulary in Cixin Liu, affective first-person narration in Qiufan Chen, relational and temporal patterns in Xia Jia, and different speculative registers in Ken Liu's original fiction.

These results suggest that translated CSF should not be understood simply as the English prose of the translator. Rather, it is a layered linguistic space: written in English and mediated by the translator, but still marked by recoverable patterns associated with Chinese source authors, themes, narrative traditions and imagined worlds. Ken Liu's translatorial practice does not collapse into a single authorial voice; rather, it preserves detectable traces of each source author's stylistic fingerprint.

KEYWORDS computational literary studies, corpus-based translation studies, text classification

POSTER 28**Storing, crafting, poaching: A computational analysis of canon formation on the SCP Wiki**

Harry P. Zhao & Jelke Bloem

Universiteit van Amsterdam

Canons typically evoke ideas about fame, value, circulation, and authority. However, in recent years, they have been drastically recontextualised by critique from counter-hegemonic scholars on the one hand and decentralised media ecosystems on the other. This exploratory project seeks to quantify the characteristics and processes of canon formation taking place on the SCP Wiki, a massive collaborative fiction project, through the use of network and regression analyses. This project employs a tripartite computational framework using PPMI similarity to link entities extracted via NER and ER, hyperlinks and co-citation scores to link documents, and the number of citations and collaborations to link users. We find preliminary evidence that (a) collaborative sites inadvertently replicate traditional processes of preferential attachment alongside real-world cultural biases; (b) encyclopaedic entries and textual entities innovate new diegetic frames and drive differentiation, while narrative stories and core users consolidate those frames and fuse them into the overarching canon; and (c) digital archives give rise to cycles of productive influence where existing works lose and regain their status as possible reference products for future activity.

KEYWORDS named entity recognition, community detection, canon formation, preferential attachment, multilevel modelling

POSTER 29**From AI to ART: the Journey of Woutje at the Watou Arts Festival**

Aaron Maladry, Christophe Scholliers, Maud Vanhauwaert, Jelle Jespers, Veronique Hoste

LT3 (Language and Translation Technology Team), Ghent University

Poetry occupies a special place in computational creativity. Unlike many language-generation tasks, poetic quality cannot be reduced to factual correctness or grammaticality alone, but emerges from the interplay of meaning, imagery, emotion, and form. Most research on automatic poetry generation focuses on explicit formal constraints such as rhyme, meter, or text completion (Oliveira, 2017; Van De Cruys, 2020). Contemporary poetry, however, relies more heavily on semantic associations and emotional resonance, making it substantially harder to generate and evaluate computationally. Moreover, research on poetry generation has primarily focused on high-resource languages such as English, French, and Chinese, leaving Dutch largely unexplored.

To address this research gap, we explore to what extent state-of-the-art AI models can learn to produce contemporary Dutch poetry. To make this possible, we digitize the historical archive of the Watou Arts festival, which dates back to the early editions from 1992 until last year, and create the first contemporary poetry corpus for Dutch. With this corpus we explore a variety of training methods, using both smaller character-based models (built on the same architecture as the initial GPT model) as well as larger state-of-the-art generative models like Llama to build the final poetry generation model we name Woutje.

The evaluation of contemporary poetry is far from trivial, and this only becomes more difficult when AI is involved. To assess the quality of the generated poetry while also accounting for the subjectivity of

the task, we present the outputs of the generated models to a wide audience during the Watou Arts festival, which runs from the 15th of July until the 30th of August. The respondents are presented with a mix of human-written and AI-generated poems and are asked which of the poems they like best. This allows us to assess whether the poems, regardless of whether they are human-like, appeal to human readers and are enjoyable to read. As a second step, the source of the poem (AI-generated or written by a professional poet) is revealed to the audience, and we ask the user how they feel about this origin. With these questions, we explore how important the source of a text is on its perception. As a final evaluation step, the visitors are invited to make the final verdict on the future of Woutje, providing their answer to the question “What can be the source of poetry?”.

References: Oliveira, H. G. (2017). A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In Proceedings of the 10th international conference on NLG Van de Cruys, T. (2020). Automatic poetry generation from prosaic text. In Proceedings of the 58th ACL

KEYWORDS poetry generation, contemporary poetry, language resources and evaluation

POSTER 30

Topics, Tropes, and Televotes: The Story of Seven Decades of Sanremo and Melodifestivalen in Data

Thomas Moerman, Pranaydeep Singh, Alessandra Teresa Cignarella, Joni Kruijsbergen

LT3, Ghent University

The Italian Sanremo Music Festival (officially: Festival della canzone italiana) and the Swedish Melodifestivalen are two of Europe's longest-running song competitions, and each is closely linked to the Eurovision Song Contest: Melodifestivalen chooses Sweden's Eurovision entry, and the winner of Sanremo is offered the chance to represent Italy. The two contests differ in one important rule. Sanremo has required song lyrics to be in Italian throughout its 75-year history. Melodifestivalen required the use of the Swedish language until 1999; after Eurovision dropped its national-language rule, any language was allowed, and English quickly became the norm. This makes the two song festivals a well-matched pair for comparison. We present a diachronic, multi-decade, cross-lingual study that combines song lyrics, competition metadata, jury and televote results, and audio features, and investigate how language, theme, emotion, and sound relate to success in the competitions.

The first part of this study has several components. First, we track how the use of language in the songs changed over time: at Melodifestivalen, the shift after 1999 from Swedish to mostly English and mixed lyrics, and at Sanremo, whether English words and phrases have been entering the lyrics despite the Italian-only rule. We then describe the vocabulary, style, and code-switching patterns in each competition. Second, we use topic modelling to track how the themes of the songs changed across more than seven decades, and ask whether the two countries move together or one follows the other. Third, we map the networks of songwriters and composers to find the most consistently successful creators and see how their collaborations are organised.

The second part of our work focuses on emotion. For each song, we build an emotional arc from the lyrics, measuring how valence and arousal move from verse to chorus, and a matching arc from the music through audio analysis. We then test whether the shape of the arc relates to a song's final placement in the ranking of the competition. We also ask whether this relationship differs between the Italian and the Swedish festival, and whether expert juries reward different emotional profiles

than the general audience, using the published jury and televote split as a guideline. Where data is available, we add Spotify audio features such as ‘valence’, ‘energy’, ‘danceability’, and ‘tempo’ to assess if and how much they contribute to the lyric and audio signal in explaining the results.

Together, these approaches provide a comparative, multimodal picture of how language, theme, and emotion relate to success in two culturally different countries (Italy and Sweden), and they offer a reusable framework for the cross-lingual study of popular song in varying contexts.

KEYWORDS emotions, multimodality, diachronics, topic modelling, code-switching

POSTER 31

Stories for the win: creating and evaluating a synthetic Dutch children’s storytelling dataset

Sabijn Perdijk, Gijs Wijnholds, Suzan Verberne & Max van Duijn

Leiden Universiteit

Narratives have shaped society since the earliest beginnings: they explain, educate, and provide purpose to mankind. They play a constitutive role in child development, boosting acquisition of communicative skills, world knowledge, and Theory of Mind. Training of LLMs also builds upon narratives: novels and short stories are classic elements of pretraining datasets. So far, however, it is largely unclear what exactly the contribution of narrative language is in the LLM pipeline.

To enable further investigation of the role children’s narratives play in pretraining LLMs, we construct a large and realistic synthetic narrative dataset, which we call StoriesfortheWin (StoriesFTW). In light of Wilcox et al.’s plea [1] to use more data that resembles children’s input, we create realistic everyday narratives as told by (and intended for) children. Existing child-narrative datasets either do not match this criterion [2, 3] or are too small in scale [4, 5]. Therefore, we generated a synthetic twin, containing 619,000 stories, of a small Dutch children’s narrative corpus (ChiSCor) [4]. We do so with a systematic approach to prompt design and model selection, guided by metrics of story quality at different levels. We test the alignment of the metrics and the output quality with human judgements.

For the metric alignment, a low correlation between the automated metrics and the human judgements was found, which can be due to faulty automated metrics or inconclusive human judgements on a highly subjective task. For the dataset generation, we found that few-shot prompting with three demonstrations sampled from ChiSCor resulted in narrative language that most closely resembled ChiSCor. We, therefore, used that prompting technique to generate StoriesFTW in combination with two other well-performing prompt techniques to avoid a collapse in dataset diversity. The evaluation of StoriesFTW with automated metrics showed that its language use is more complex and contains stories that are less creative than the stories in ChiSCor. Despite this seemingly underwhelming result, we are cautiously optimistic about the quality of StoriesFTW, as the generated stories could not be distinguished from ChiSCor stories in the human evaluation on all defined categories except for Complexity and Human Likeness.

[1] Wilcox, E.G., et al. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.

[2] Eldan, R. & Li, Y. (2023). Tinstories: How small can language models be and still speak coherent english?

[3] Bens, E. (2021). Children Stories Text Corpus.

[4] van Dijk, B., et al. (2023). ChiSCor: A Corpus of Freely Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science.

[5] Nicolopoulou, A. (2019). Using a storytelling/story-acting practice to promote narrative and other decontextualized language skills in disadvantaged children.

KEYWORDS Narratives, dataset creation, evaluation

Corpora & annotation methods

POSTER 32

Collection-level realism evaluation for synthetic customer-service conversations

Tom Brand, Madelon Molhoek & Han Kruiger

TNO

Within the European LLMs4EU project, we work on synthetic data generation for the telecom use case in collaboration with Orange and KPN. Our long-term goal is to build reusable multilingual customer-service datasets that support research and development while reducing dependence on real conversations, which typically contain personal data and are thus legally difficult to use. To achieve this goal, we use ASR transcripts of customer-service conversations and their related annotations as grounding material for LLM-based generation of synthetic conversation transcripts.

The primary challenge is to make the resulting data realistic enough to replace original data in downstream tasks. In this work, we focus on:

- evaluating realism at the collection level; - iteratively optimizing the generation process; and - supporting multilingual applicability through language-specific metric configuration.

We frame the synthetic data generation task as an iterative optimization problem: we generate a new data collection, quantify its similarity to the source data along several axes, adjust the generation configuration, evaluate again, and repeat the process until improvements plateau.

Quantifying realism is difficult. We therefore implement a selection of both pre-existing and novel collection-level evaluation methods. These compare the empirical distribution of the synthetic data to that of the source data through metric-specific distributional comparisons.

The metrics capture both structural properties, such as speaker turn-taking balance and turn length, and content-based properties, such as the semantic diversity of both conversation participants measured through precision- and recall-based cross-entropy, and language complexity. Since these metrics tend to be language-specific to varying degrees, we tailor the metric configuration to each language in scope of the use-case, e.g. by choosing appropriate embedding models and language-specific readability formulas.

We argue that collection-level evaluation is essential for practical synthetic dialogue generation in multilingual telecom settings. Systematic prompt and generation-setting optimization makes it possible to determine when synthetic data is sufficiently realistic, which enables the creation of larger reusable datasets while allowing us to assess the quality of the synthetic data compared to real data.

Relevant references:

- Géraldine Damnati, Aleksandra Guerraz, and Delphine Charlet. 2016. Web Chat Conversations from Contact Centers: a Descriptive Study. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2017-2021. <https://aclanthology.org/L16-1319/>
- Hannah Venhuizen. 2026. Evaluating Semantic Diversity in Dutch Synthetic Text Data. Forthcoming Master's thesis.

KEYWORDS synthetic data, synthetic data realism, natural language generation, evaluation, multilingual NLP

POSTER 33**Comparing automatic linguistic annotation methods for the study of Spanish differential object marking**

Carla Verwijs-Rodriguez, Maria Tepei & Jelke Bloem

Universiteit van Amsterdam

Differential object marking (DOM) is a cross-linguistically widespread phenomenon in which only a subset of direct objects receives overt morphological marking. For Spanish, previous studies have hypothesised animacy, specificity, definiteness, and some syntactic contexts as the main factors triggering DOM. However, most of these accounts tend to examine potential factors in relative isolation, which often results in an incomplete picture of the phenomenon. We address the question of whether a large-scale, annotated dataset of Spanish DOM can be created and annotated through automatic methods, and whether the resulting annotations are reliable enough for fine-grained subsequent linguistic analysis.

In particular, we compare approaches based on task-specific tools such as dependency parsers, to Large Language Model-based annotation approaches. We develop an annotation pipeline for contemporary Spanish that combines dependency parsing, named entity recognition and lexical resources to extract variables commonly hypothesized to condition DOM, including argument structure, object and subject animacy, and object and subject definiteness. For comparison, we also develop prompts that aim to elicit these same variables from LLMs. To support evaluation, we manually annotate a dataset of news text, wiki text and literary text for these variables.

We observe that the LLM-based approach has high performance on subject and direct object animacy classification as well as DOM detection, but lower performance on subject and object identification. The UD parser-based approach shows better performance for subject and direct object identification, as well as definiteness classification. We find that the optimal approach for constructing an annotated dataset of Spanish DOM, given the currently available tools, is to use an annotation pipeline of specialized tools for some variables and a LLM-based approach for others. In particular, a hybrid approach of using a parser to identify the subject and direct object and passing this information to the LLM for animacy classification is more effective than an end-to-end approach with the LLM.

KEYWORDS Syntax, lexical semantics, linguistic annotation, differential object marking