CRAC 2025

The Eighth Workshop on Computational Models of Reference, Anaphora and Coreference

Proceedings of the Workshop

November 9, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-342-5

Message from the Program Chairs

If you are one of the staunch supporters of CRAC, you should probably know that this is the 8th edition of CRAC (and the 10th edition if you also count the two CORBON workshops), Following the CRAC tradition, we requested that CRAC 2025 be co-located with the EMNLP conference. At the time of proposal submission, the location of EMNLP 2025 had not been finalized. It was only after we were notified of the acceptance of the proposal that we knew that EMNLP would take place in China, which was certainly a pleasant surprise to us. This will be the third time CRAC takes place in Asia, after CRAC 2022 in South Korea and CRAC 2023 in Singapore.

What is special about this year's workshop is that this is the first time CRAC is held jointly with CODI (Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences) despite the fact that the two workshops organized two shared tasks together in 2021 and 2022. The organizing committee of the joint workshop is composed of the organizers from CRAC and CODI, who worked on the timeline for the workshop, the call for papers, the list of potential invited speakers, and the program schedule. The two workshops, however, had separate program committees, submission sites, and proceedings, and made acceptance decisions independently of each other.

This year, CRAC received 15 submissions, including nine research papers and six shared task papers. Each research paper was rigorously reviewed by three program committee members, and each shared task paper was reviewed by two. Based on their recommendations, we accepted all of the shared task papers, and among the nine research papers, we accepted six, conditionally accepted two, and rejected one. The two conditionally accepted papers were eventually accepted to the workshop after we made sure that the authors adequately addressed the reviewers' comments in the final camera-ready version.

This year we continued to partner with our colleagues at Charles University, Prague and hosted the shared task on Multilingual Coreference Resolution for the fourth time at CRAC. The shared task allowed researchers who did not participate in the workshop to disseminate their work to a smaller and more focused audience which should promote interesting discussions. Following what we did last year, we similarly merged the shared task proceedings with the CRAC workshop proceedings this year. In other words, you can enjoy both the workshop papers and the shared task papers in this proceedings.

As you can imagine, fitting two invited talks, two shared tasks (the Multilingual Coreference Resolution shared task and the DISRPT shared task), and a large number of presentations of papers accepted to CODI and CRAC to a one-day program is by no means an easy task. In the end, the organizing committee decided to have two poster sessions (one in the morning and one in the afternoon) where the majority of the papers will be presented, selecting only a small number of papers for oral presentations. Even so, it has been logistically challenging for us to arrange for virtual paper presentations during the poster sessions.

We are grateful to the following people, without whom we could not have assembled an interesting program for the joint workshop. First, we are indebted to the CRAC program committee members. This year the average reviewing load was the equivalent of two long papers per reviewer. All of our program committee members did the incredible job of completing their reviews in a short reviewing period. Second, we thank Tanya Goyal and Nancy F. Chen, who are established researchers in Discourse, for accepting our invitation to be this year's invited speakers. Finally, we would like to thank the workshop participants for joining us.

We hope you will enjoy the workshop and Suzhou as much as we do!

— Maciej Ogrodniczuk, Michal Novák, Massimo Poesio, Sameer Pradhan, and Vincent Ng

Organizers

Organizing Committee:

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, Poland
Michal Novák, Charles University in Prague, Czechia
Massimo Poesio, Queen Mary University of London, UK
Sameer Pradhan, University of Pennsylvania and cemantix.org, USA
Vincent Ng, University of Texas at Dallas, USA

Program Committee:

Jackie Chi Kit Cheung, Mila / McGill University, Canada
Loic De Langhe, Ghent University, Belgium
Elisa Ferracane, Abridge AI, Inc., USA
Yulia Grishina, Amazon, USA
Lars Hellan, Norwegian University of Science and Technology, Norway
Veronique Hoste, Ghent University, Belgium
Ekaterina Lapshinova-Koltunski, University of Hildesheim, Germany
Sharid Loáiciga, University of Gothenburg, Sweden.
Costanza Navaretta, University of Copenhagen, Denmark
Michal Novák, Charles University in Prague, Czechia
Constantin Orasan, University of Surrey, UK
Massimo Poesio, Queen Mary University of London, UK
Ian Porada, Mila – Quebec Artificial Intelligence Institute, Canada
Bonnie Webber, University of Edinburgh, UK
Juntao Yu, Queen Mary University of London, UK
Yilun Zhu, Georgetown University, USA
Heike Zinsmeister, University of Hamburg, Germany

Invited Talk 1

From Speech to Sense: The Art of Listening in Artificial Intelligence

Nancy F. Chen

Abstract

Unlike sight, which we can shut off with a blink, sound is inescapable. We are always listening, even when we wish not to. Hearing comes naturally, but understanding what we hear requires learning, knowledge, focus, and interpretation. Yet it is sound — be it the quiet drone of an air conditioner, a lover's tender whisper, or the distant rush of a waterfall — that anchors us to our physical surroundings, social connections, and the present moment.

In this talk, I will share our experience in modelling the audio signal in multimodal generative AI to drive translational impact across domain applications. In particular, we exploit the audio modality to strengthen contextualization, reasoning, and grounding. Cultural nuances and multilingual peculiarities add another layer of complexity in understanding verbal interactions. Examples include our generative AI efforts in Singapore's National Multimodal Large Language Model Programme has led to MERaLiON (Multimodal Empathetic Reasoning and Learning In One Network), the first multimodal large language model developed for Southeast Asia context. Such endeavors complement North American centric models to make generative AI more widely deployable for localized needs. Another case in point is SingaKids AI Tutor, which enables young children to learn ethnic languages such as Malay, Mandarin and Tamil. We are currently expanding applications to embodied agentic AI, aviation, and healthcare.

Speaker Bio

Nancy F. Chen is an ISCA Fellow (2025), AAIA Fellow (2025), and A*STAR Fellow (2023), and was recognized with the Asian Women Tech Leaders Award (2025). She is also a 2025 inductee of IEEE Eta Kappa Nu (HKN), the honor society of IEEE recognizing outstanding engineers.

At A*STAR, Dr. Chen leads the Multimodal Generative AI group and the AI for Education Programme. Dr. Chen is a serial best paper award winner across major conferences - including ICASSP, ACL, EMNLP, MICAAI, COLING, APSIPA, SIGDIAL and EACL – her research spans applications in education, healthcare, neuroscience, social media, security and forensics. Dr. Chen's multimodal, multilingual technologies have led to commercial spin-offs and adoption by Singapore's Ministry of Education.

Invited Talk 2

Climbing the Right Hill: On Benchmarking Progress in Long-Form Text Processing

Tanya Goyal

Abstract

Large Language Models (LLMs) are now functionally capable of ingesting very long documents as input, but can they truly process and reason over these massive contexts? In this talk, I will discuss our efforts at answering this question through the lens of long narrative summarization, a setting that naturally requires information synthesis and reasoning over long range dependencies. In the first part, I will describe our work highlighting shortcomings of current models along two key summary quality axes - coherence and factuality - and discuss challenges in automating their evaluation. Next, I will present NoCha, our methodology for constructing realistic and uncontaminated benchmarks for long context narrative reasoning. I will discuss results that show that NoCha is challenging for frontier LLMs; GPT-5 reports <30% worse performance compared to humans, and provide a recipe for building the next generation of robust long context benchmarks.

Speaker Bio

Tanya Goyal is an assistant professor in the Computer Science department at Cornell University.

Her research interests include building reliable and sustainable evaluation frameworks for large language models (LLMs) as well as understanding LLM behaviors as a function of training data and/or alignment strategies. Previously, she was a postdoctoral scholar at Princeton Language and Intelligence Center (2023-2024). Tanya completed her Ph.D. in Computer Science at UT Austin in 2023 where her thesis was awarded UTCS's Bert Kay Dissertation award. Her research is supported by NSF and a gift from Google.

Table of Contents

Workshop Program

Sunday, November 9, 2025

Opening Remarks

09:00–09:10 *Opening and Welcome*
Maciej Ogrodniczuk and Michael Strube

Invited Talk 1

09:10–10:00 *From Speech to Sense: The Art of Listening in Artificial Intelligence*
Nancy F. Chen

Shared Tasks Overview

10:00–10:15 *Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches?*
Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský and Daniel Zeman

10:15–10:30 *The DISRPT 2025 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification*
Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu and Philippe Muller

10:30–11:00 *Coffee Break*

11:00–12:00 **Poster Session 1**

DisCuT and DiscReT: MELODI at DISRPT 2025 Multilingual discourse segmentation, connective tagging and relation classification
Robin Pujol, Firmin Rousseau, Philippe Muller and Chloé Braud

CLaC at DISRPT 2025: Hierarchical Adapters for Cross-Framework & Multilingual Discourse Relation Classification
Nawar Turk, Daniele Comitogianni and Leila Kosseim

DeDisCo at the DISRPT 2025 Shared Task: A System for Discourse Relation Classification
Zhuoxuan Ju, Jingni Wu, Abhishek Purushothama and Amir Zeldes

HITS at DISRPT 2025: Discourse Segmentation, Connective Detection, and Relation Classification
Souvik Banerjee, YI FAN and Michael Strube

SeCoRel: Multilingual Discourse Analysis in DISRPT 2025
Sobha Lalitha Devi, Pattabhi RK Rao and Vijay Sundar Ram

11:00–12:00 **Poster Session 1** (continued)

Few-Shot Coreference Resolution with Semantic Difficulty Metrics and In-Context Learning
Nguyen Xuan Phuc and Dang Van Thin

Fine-Tuned Llama for Multilingual Text-to-Text Coreference Resolution
Jakub Hejman, Ondrej Prazak and Miloslav Konopík

Few-Shot Multilingual Coreference Resolution Using Long-Context Large Language Models
Moiz Sajid, Muhammad Fraz, Seemab Latif and Zuhair Zafar

CorPipe at CRAC 2025: Evaluating Multilingual Encoders for Multilingual Coreference Resolution
Milan Straka

Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production
Debasmita Bhattacharya, Anxin Yi, Siying Ding and Julia Hirschberg

Discourse Relation Recognition with Language Models Under Different Data Availability
Shuhaib Mehri, Chuyuan Li and Giuseppe Carenini

Where Frameworks Disagree: A Study of Discourse Segmentation
Maciej Ogrodniczuk, Anna Latusek, Karolina Saputa, Alina Wróblewska, Daniel Ziembicki, Bartosz Żuk, Martyna Lewandowska, Adam Okrasiński, Paulina Rosalska, Anna Śliwicka, Aleksandra Tomaszewska and Sebastian Żurowski

Information-Theoretic and Prompt-Based Evaluation of Discourse Connective Edits in Instructional Text Revisions
Berfin Aktas and Michael Roth

Joint Modeling of Entities and Discourse Relations for Coherence Assessment
Wei Liu and Michael Strube

Towards Adding Arabic to CorefUD
Dima Taji and Daniel Zeman

Exploring Coreference Resolution in Glosses of German Sign Language
Yuzheng Bao and Haixia Chai

12:00–13:30 *Lunch Break*

Invited Talk 2

13:30–14:20 *Climbing the Right Hill: On Benchmarking Progress in Long-Form Text Processing*
Tanya Goyal

14:20–15:30 **Poster Session 2**

Long Context Benchmark for the Russian Language
Igor Churin, Murat Apishev, Maria Tikhonova, Denis Shevelev, Aydar S. Bulatov, Yuri Kuratov, Sergei A. Averkiev and Alena Fenogenova

Enhancing the Automatic Classification of Metadiscourse in Low-Proficiency Learners' Spoken and Written English Texts Using XLNet
Wenwen Guan, Marijn Alta and Jelke Bloem

Stance Detection on Nigerian 2023 Election Tweets Using BERT: A Low-Resource Transformer-Based Approach
Mahmoud Said Ahmad and Habeebah A. Kakudi

"Otherwise" in Context: Exploring Discourse Functions with Language Models
Guifu Liu, Bonnie Webber and Hannah Rohde

On the Role of Context for Discourse Relation Classification in Scientific Writing
Stephen Wan, Wei Liu and Michael Strube

Automated Conspiracy Narrative Detection Across Social Media Platforms
Calvin Yixiang Cheng and Mohsen Mosleh

Zero-Shot Belief: A Hard Problem for LLMs
John Murzaku and Owen Rambow

Probing the Limits of Multilingual Language Understanding: Nepali Proverbs as LLM Benchmark for AI Wisdom
Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak and Usman Naseem

Measuring Sexism in US Elections: A Comparative Analysis of X Discourse from 2020 to 2024
Anna Fuchs, Elisa Noltenius, Caroline Weinzierl, Bolei Ma and Anna-Carolina Haensch

EmbiText: Embracing Ambiguity by Annotation, Recognition and Generation of Pronominal Reference with Event-Entity Ambiguity
Amna Sheikh and Christian Hardmeier

Human and LLM-based Assessment of Teaching Acts in Expert-led Explanatory Dialogues
Aliki Anagnostopoulou, Nils Feldhus, Yi-Sheng Hsu, Milad Alshomary, Henning Wachsmuth and Daniel Sonntag

Multi-token Mask-filling and Implicit Discourse Relations
Meinan Liu, Yunfang Dong, Xixian Liao and Bonnie Webber

Consistent Discourse-level Temporal Relation Extraction Using Large Language Models
Yi Fan and Michael Strube

Coreference in simplified German: Linguistic features and challenges of automatic annotation
Sarah Jablotschkin, Ekaterina Lapshinova-Koltunski and Heike Zinsmeister

Mention detection with LLMs in pair-programming dialogue
Cecilia Domingo, Paul Piwek, Svetlana Stoyanchev and Michel Wermelinger

The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works
Antoine Bourgois and Thierry Poibeau

14:20–15:30 **Poster Session 2** (continued)

Referential ambiguity and clarification requests: comparing human and LLM behaviour
Chris Madge, Matthew Purver and Massimo Poesio

Revisiting the Givenness Hierarchy. A Corpus-Based Evaluation
Christian Chiarcos

15:30–16:00 *Coffee break*

Oral Session

16:00–16:15 *Unpacking Ambiguity: The Interaction of Polysemous Discourse Markers and Non-DM Signals*
Jingni Wu and Amir Zeldes

16:15–16:30 *Impact of ASR Transcriptions on French Spoken Coreference Resolution*
Kirill Milintsevich

16:30–16:45 *GLaRef@CRAC2025: Should we transform coreference resolution into a text generation task?*
Olga Seminck, Antoine Bourgois, Yoann Dupont, Mathieu Dehouck and Marine Delaborde

16:45–17:00 *Entity Tracking in Small Language Models: An Attention-Based Study of Parameter-Efficient Fine-Tuning*
Sungho Jeon and Michael Strube

17:00–17:15 *Corpus-Oriented Stance Target Extraction*
Benjamin David Steel and Derek Ruths

17:15–17:30 *Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser*
Elena Chistova

Closing Remarks

17:30–17:45 *Closing the workshop with Best Paper Awards*
Maciej Ogrodniczuk, Michal Novák, Michael Strube and Janet Liu

Referential ambiguity and clarification requests: comparing human and LLM behaviour

Chris Madge, Matthew Purver and Massimo Poesio
Queen Mary University of London
{c.j.madge,m.purver,m.poesio}@qmul.ac.uk

Abstract

In this work we examine LLMs' ability to ask clarification questions in task-oriented dialogues that follow the asynchronous instruction-giver/instruction-follower format. We present a new corpus that combines two existing annotations of the Minecraft Dialogue Corpus — one for reference and ambiguity in reference, and one for SDRT including clarifications — into a single common format providing the necessary information to experiment with clarifications and their relation to ambiguity. With this corpus we compare LLM actions with original human-generated clarification questions, examining how both humans and LLMs act in the case of ambiguity. We find that there is only a weak link between ambiguity and humans producing clarification questions in these dialogues, and low correlation between humans and LLMs. Humans hardly ever produce clarification questions for referential ambiguity, but often do so for task-based uncertainty. Conversely, LLMs produce more clarification questions for referential ambiguity, but less so for task uncertainty. We question if LLMs' ability to ask clarification questions is predicated on their recent ability to simulate reasoning, and test this with different reasoning approaches, finding that reasoning does appear to increase question frequency and relevancy.

1 Introduction

Large Language Models (LLM) are much maligned for their tendency to act presumptively, "hallucinating" in the absence of knowledge. Until the recent advent of reasoning orientated LLMs (i.e. models deliberately fine tuned with reasoning as an objective such as DeepSeek-R1 (Guo et al., 2025)), models struggled asking clarification questions and rarely proactively sought missing information (Deng et al., 2023; Li et al., 2022, 2024). Prior works have tested how LLMs respond to uncertainty, and proposed benchmarks (Zhang et al., 2024). However, this remains a challenge.

This is perhaps further complicated as clarification is a conversational strategy applied sparingly by humans (Purver et al., 2003; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005). Certain situations promote greater clarification question usage; for example, situations in which information is asymmetric, and which concerns a task requiring information seeking. This happens to be a popular paradigm for tasks created with the objective of soliciting dialogue (sometimes referred to as instruction giver/instruction follower) and also, in recent years, for studying clarification questions (Chi et al., 2020; Madureira and Schlangen, 2023a; Testoni and Fernández, 2024; Shen and Lourentzou, 2023). In this work we look at the Minecraft Dialogue Corpus (Narayan-Chen et al., 2019), a task orientated, grounded corpus that follows this paradigm. We select this corpus as it has benefited from multiple separate annotation efforts (Thompson et al., 2024; Madge et al., 2025) that extend its already richly structured offering, with useful supplemental information that can inform the experiments undertaken in this work.

We focus on one particular area of linguistic uncertainty, referential ambiguity. There has been a long standing interest in reference, with ambiguity featuring as an interest in the first popular corpora (Pradhan et al., 2012). Our first contribution of this work is combining prior annotation efforts providing annotations for clarification questions (Thompson et al., 2024) and reference (Madge et al., 2025) into a single aligned corpus in the MMAX format (described in Section 3).

This annotation supports our next contribution, a comparison of how LLMs and humans resolve uncertainty. We ask, *"does referential ambiguity really trigger clarification requests from humans, and is this different for LLMs?"*. We look at both the annotated instances of linguistic ambiguity, and the original clarification questions, as posed by human interlocutors for correlation. We test both

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 1–11
Suzhou, China, November 9, 2025.

against different LLM based approaches.

One proposed approach to improving clarification questions with LLMs is through adding further reasoning capabilities with variations on the Chain of Thought approach (Deng et al., 2023). Despite extensive testing in epistemic, aleatoric, linguistic uncertainty (Ortega-Martín et al., 2023) and proposed benchmarks (Zhang et al., 2024), it remains somewhat unclear how effective LLMs are in identifying uncertainty and even more so how LLMs may consistently generate the relevant questions to address it.

For our second research question, we ask if the ability to ask a useful clarification question, or indeed judge when to ask a question, is based on a model's ability to simulate reasoning. We test this hypothesis with an experiment, comparing models that were trained to include reasoning and prompt engineering strategies for inducing reasoning at test time against ordinary models/methods.

For our final contribution, we look further into human reasoning and its constituent parts, with a discussion on how these may affect LLMs seemingly emergent ability to ask some clarification questions, and liken this to human reasoning.

2 Related Work

2.1 Clarification Questions

There has been a great interest in clarification questions in the literature on dialogue systems going back at least twenty years (Purver et al., 2003; Schlangen, 2004; Gabsdil, 2003). More recently, there has been extensive work in modern Natural Language Processing modelling clarification question generation or indeed when to ask them (Majumder et al., 2021; Aliannejadi et al., 2019; Kiseleva et al., 2022). This section will primarily focus on prior works that target clarification question with overlap to our specific goals (i.e. task orientated dialogue with situated and/or embodied agents). Previous works have gathered or annotated datasets in situated dialogue with clarification questions. For example, (Gervits et al., 2021) gather a corpus (HuRDL - Human-Robot Dialogue Learning) and annotate clarification questions in a dialogue gathered from human participants in a robot situated tool gathering task. (Gella et al., 2022) annotate dialogue acts in the TEACh (Task-driven Embodied Agents that Chat) dataset (Padmakumar et al., 2022); the product of a task that has human participants collaborating to perform household tasks in a virtual house environment.

A particularly popular task/dataset for this is CoDraw (Kim et al., 2017). The CoDraw task (Kim et al., 2017) is similar to the previously discussed Minecraft task, in that an instruction giver communicates with an instruction follower to collaboratively reach a goal. As opposed to constructing a 3D voxel based structure, they recreate a scene formed of clipart images. (Madureira and Schlangen, 2023b) annotate this dataset with clarification questions.

Previous works have also compared when humans and models would ask clarification questions (Testoni and Fernández, 2024) use the aforementioned CoDraw dataset to investigate this the relationship between model uncertainty, and human clarification questions based on task properties (e.g. size, orientation, position etc.). The presence of a clarification question is used as the measure of measure uncertainty, and they use logistic regression to see if they can predict this.

The Minecraft Dialogue Corpus (Narayan-Chen et al., 2019) used in this work is different, in that rather than referencing direct objects, continuously changing abstract shapes are created and manipulated during the dialogue. We expand on this further in the following section. There has been other work using Minecraft-like environments as a test-bed for the study of clarification questions in dialogue. However, this was prior to LLMs and looked at clarification question production as a task of ranking available clarification questions, rather than their generation (Kiseleva et al., 2022)

Several works have investigated the use of LLMs for clarification question generation, with methods including: fine tuning on question data (Andukuri et al., 2024); uncertainty estimation over multiple samples (Pang et al., 2024; Zhang and Choi, 2023) and multi turn prompting strategies (Kuhn et al., 2022; Li et al., 2023). To our knowledge, none of these are primarily concerned with reference or situated dialogue settings. There is however evidence to suggest LLMs can successfully resolve reference with performance similar to, or in some cases superior to, reference specific models (Hicke and Mimno, 2024; Le and Ritter, 2023).

2.2 MDC and its extensions

The Minecraft Dialogue Corpus (Narayan-Chen et al., 2019) is a collection of conversations among human participants performing the Minecraft Collaborative Building Task.This follows the typical

instruction giver, instruction follower paradigm, where *the Architect*, who has full observability over the target environment but is unable to act, instructs *the Builder*, to manipulate the environment to meet that target structure. The world is a 3D voxel based $11 \times 9 \times 11$ Minecraft like world, originally provided by project Malmo (Johnson et al., 2016). This results in a 509 multi turn situated dialogues with rich linguistic phenomena including reference and clarification.

Various annotation efforts have extended MDC, including variations of AMR (Bonn et al., 2020; Bonial et al., 2021), reference (Madge et al., 2025) and Segmented Discourse Representation Theory (SDRT, Thompson et al., 2024). We focus on the latter two as they are directly used in this work.

SDRT provides a macrostructure of interconnected logical discourse forms, linking narrative arcs and discourse relations (e.g. clarification questions, corrections, confirmations, acknowledgements etc., see Asher and Lascarides, 2003; Lascarides and Asher, 2007). Thompson et al. (2024) exhaustively annotated MDC with SDRT in their Minecraft Structured Dialogue Corpus (MSDC).

MDC-R (Madge et al., 2025) consists of a subset of 100 dialogues from MDC with reference expert annotated according to the ARRAU guidelines (Poesio et al., 2024). The dynamically changing environment and instruction based two-party dialogue gives rise to various types of reference, much of which, beyond the discourse, is linked directly to the objects in the virtual world. This results in some interesting and challenging examples of ambiguity for a dialogue system to resolve.

2.3 Reasoning in Large Language Models

Chain of Thought (COT, Wei et al., 2022) simulates reasoning at inference time by encouraging the model to think through the answer step by step. In implementation, this can take one of two common forms. The model is either provided an example of thinking through a problem step by step as part of a one-shot/few-shot prompt, or a zero-shot approach that simply prefaces the prompt with something like "Let's think this through, step by step...". The core benefits of the COT approach are given to be: problem decomposition; some explainability/insight into how results are reached; logical problem solving/symbolic manipulation and ease of application to existing models. Previous work has observed an improvement in applying this method when addressing Minecraft orientated tasks (Madge and Poesio, 2024). Following Chain of Thought, several models have been trained or aligned explicitly to follow this process (e.g. Gemini 2.5; Version 3 of the Qwen model (Yang et al., 2024); DeepSeek-R1 (Guo et al., 2025)).

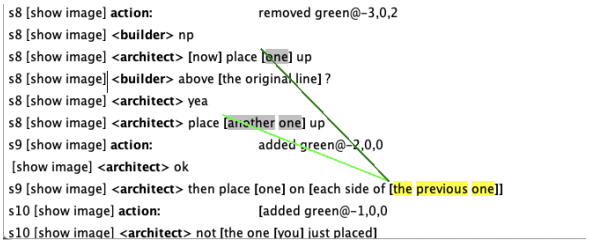

Figure 1: Referential Ambiguity Annotation Example in MMAX

3 Adding MSDC information to MDC-R

In this section we will motivate and describe our effort to add MSDC information to MDC-R to produce a new version of the corpus combining both types of annotation.

We identify two types of utterances or phrases may provoke clarification requests. Firstly, utterances that have been annotated as the subject of a clarification or confirmation request in the dialogue (typically related to task orientated uncertainty , and secondly instances of referential ambiguity.

To support our experimentation, and investigation of any possible relationship between the two, we present a corpus that merges two existing corpora that identifies these. This is a combination of the previously discussed MDC-R (Madge et al., 2025) (providing reference annotations for MDC), and MSDC corpora (Thompson et al., 2024) (providing clarification questions for MDC), permitting convenient examination of reference and more specifically types of referential ambiguity aligned with clarification questions. We automated the merge of these two corpora, through use of a script that operated at token level to produce a common method of addressing and aligning the respective segments in each.

We add a new MMAX[1] layer, referred to as SDRT. Each MMAX markable in this layer represents an Elementary Discourse Unit and the relations between those markables are represented by a *to* attribute on each markable, with the related markables unique ID. Crucially, these relationships

[1] https://mmax2.net/

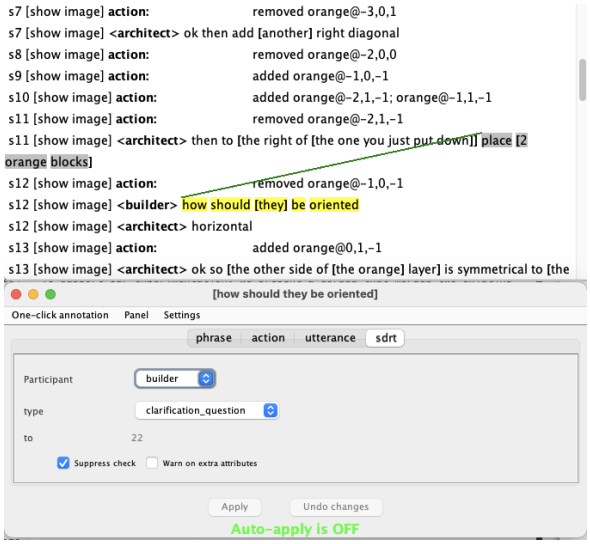

Figure 2: Clarification Question Annotation Example in MMAX

describe links between utterances and their clarification questions.

This combined format allows parsing and examination of adjacent reference in MMAX (shown in Figure 1) and clarification (shown in Figure 2).

Reference annotation has many parameters, with each relationship holding many attributes. Whilst the SDRT annotations exhaustively cover all of the original MDC dataset (Narayan-Chen et al., 2019), MDC-R covers a subset of 100 dialogues (some more detailed descriptive statistics taken from MDC-R (Madge et al., 2025) are given in Table 1). As such, our corpus will be limited to the same 100 dialogues.

Statistic	Count	Statistic	Count
Documents	101	Tokens	29,174
Utterances	3,343	Actions	5,793
Markables	7,600	Discourse old	1960
Bridging	1,053	Object	500
Plural	24	Ambiguous	149

Table 1: MDC-R Corpus Statistics (Madge et al., 2025)

Using the combined corpus, we have counted instances of utterances exhibiting certain attributes that may motivate a question. Firstly, using annotations originating from MSDC, we look at the counts of *confirmation questions* and *clarification questions* occurring in the original human dialogue. Secondly, we count specific instances of referential ambiguity, using annotations originating from MDC-R. These are instances of discourse deixis - relating to parts of the discourse (e.g. "as I said earlier"), and spatial deixis in a real world space - typically our voxel world environment (e.g. "next to that block").

Table 2 shows the frequency of these different instances of utterances that may motivate a question as a percentage of all utterances in the selected subset of the corpus.

Type	Instances	%
confirmation question	218	3.7%
clarification question	182	3.1%
discourse deixis	24	0.4%
spatial deixis	16	0.3%

Table 2: Frequency of instances as a percentage of utterances

77% of dialogues contain a clarification question and 75% a confirmation question.

The SDRT annotations of the complete corpus found in MSDC had 999 confirmation questions and 960 clarification questions over 547 dialogues. We can see from the relative quantity of questions, that the selected dialogues do appear to be representative of the corpus as a whole, with respect to question quantity.

To give some overview, the most common phrases for discourse deixis are: *"that"* (8); *"this"* (7) and *"it"* (3). Spatial deixis has 3 instances of "this", but referents while still ambiguous, tend do be more literal (e.g. *"the red end"*). We expand further on this in a discussion of reoccurring patterns in Section 5.1.

The corpus is available at `https://github.com/arciduca-project/MDC-R/tree/sdrt`.

4 Methodology

To test how LLMs perform clarification questions, we first discover points in the conversation that may require clarification, then we sample from various LLMs with different approaches, feeding the context of the conversation up to the appropriate point. Examples of our prompts are given in the Appendix (see Section A.1). A system prompt describes the nature of environment and it's constraints. This differs slightly between architects and builders, in that architects can see the target structure, and the builder's system prompt specifies the required JSON response format necessary to encode their resultant actions or expect the world state (this experiment is text

only, with the world state/actions encoded in JSON - no images are used). For the chain-of-thought treatments, we supplement these system prompts as follows. The zero-shot chain-of-thought approach simply adds, "Think step by step" to the system prompt. The one-shot approach follows the system prompt with an example exchange that incorporates thinking.These approaches are deliberately selected to compare reasoning based models (e.g. `llama3.2`, `DeepSeek-R1` (Guo et al., 2025)) and sampling methods (`llama2:13b-COTZERO`, `llama2:13b-COTONE`) vs. non-reasoning (`llama2:13b`). We compare the approach taken by LLMs to the approach originally taken by humans.

There are three characteristics of the instances that we identify to test against. Firstly, the subject of any clarification and confirmation question as originally annotated in SDRT. Second and thirdly, linguistic ambiguities. We select any referent that has two or more antecedents, whether they be discourse deixis (part of the discourse), or spatial deixis (in reference to objects in the environment. That is to say, in the MDC-R MMAX format, for discourse deixis the phrase in question, would have the following attributes specified `segment_phrase_antecedent_2`, and for spatial deixis, `object2` specified.

We also check for correlation between clarification questions, and the instances of linguistic ambiguity, as permitted by using our new merged corpus (described in Section 3),

We measure the tendency of different approaches to ask a question and use a single human coder to examine specific instances of ambiguity and their responses to attempt to quantify the number of relevant questions asked when applying each method. Prior to quantifying responses to instances we conduct a thorough example driven investigation.

5 Results

In the first section we look at how LLMs and humans respond to instances of referential linguistic ambiguity, including whether humans pose clarification questions when they encounter linguistic ambiguity. In the second section we look at the original clarification questions as posed by humans, to see how LLMs respond. Finally, we perform a quantitative evaluation, counting the tendency of different approaches to ask questions, and a count of question relevancy under one specific condition.

5.1 Linguistic Ambiguities

Do humans ask clarification questions when there is ambiguity? In many cases, where linguistic ambiguity exists, it appears deliberately underspecified in the interest of brevity, and where expanding more literally or verbosely would not have any further positive effect on task completion. Here, we see two common patterns emerge.

The first common pattern is in the communication of approval. The exact target of approval is unspecified, and arguably redundant if correct, as no further intervention is required. This appears to be used as a signal to indicate objective completion and generally precede new instruction. Examples include: *"oh that is magnificent builder"*; *"... that is great"*; *"...that's perfect"*; *"...let 's see what we can do about that"*; *"yeah, that's fine"*; *"good job"*, *"yes like that"*.

These seem closely related to the second commonly appearing pattern in sources of linguistic ambiguity, which occur in clarification questions themselves and appear to be used as a polite conversational device to invite architect interruption to general correction and indicate the builder believes the prior instruction to be addressed. Examples include: "*like this?*"; "*is this good?*"; "*this right?*", where "this" and "that" do not have clear referrents. Again, exact specification of what "this" refers to is somewhat redundant, as any corrective instruction may follow. Examining the context an interpreting more explicitly, these would appear to be more semantically equivalent to: "I believe I've completed all instructions, I'm ready for the next instruction or correction".

The first finding of this paper is that linguistic ambiguities are rarely the target of clarification questions directly in our corpus. We only discovered a single instance in which a discourse deixis referential ambiguity is in the vicinity of the subject of a latter clarification question. In the case of linguistic ambiguity, it seems humans prefer to act presumptively, in this setting.

Do LLMs ask clarification questions when there is ambiguity? Unresolved linguistic ambiguities, left unchecked, do propagate, resulting in the necessity for clarification or confirmation later in the conversation. Whilst human participants do not address these immediately, LLMs sometimes do. We show two examples here. In the following example, "the previous one" has an ambiguous referent in the context of the conversation and current world state.

This is not immediately resolved at the time by the human builder, who acts presumptively, and takes action. Having taken the incorrect action, and then been required to perform a revision, the builder then uses a confirmation question, *"like that?"*. In contrast, the LLM however, does ask a question to directly disambiguate, *"the previous one"* by offering the two possible interpretations (see Figure 3).

Architect	then place one on each side of the previous one
Builder (model vs. original human dialogue):	
Model	Are the blocks to be placed on each side of the green block at [-2,0,0] or in relation to the vertical stacking process near the center?
Human	*takes action*
Architect	not the one you just placed
Builder	*revisits with corrective action*
Builder	like that?

Figure 3: Excerpt from B4-A19-C4-1522882189483 demonstrating propagation of referential ambiguity

Similarly, in Figure 4, the human builder does not clarify the ambiguous statement, "the same thing", but proceeds to take action, to discover a lot of work must then be mirrored. Conversely, an LLM asks, about which plane the blocks should be mirrored, to first discover the correct approach.

Builder	*builder takes action*
Architect	now the same thing on the other side
Builder (model vs. original human dialogue):	
Model	Does 'the same thing on the other side' mean mirroring the blocks symmetrically with respect to the y-z plane at x=0?
Human	*builder takes further action*
Architect	that but on the other side
Builder	*builder carries out extensive revision*
Architect	there

Figure 4: Excerpt from B35-A34-C118-1524004322117 demonstrating direct and immediate clarification of linguistic ambiguity from LLM

We show here in this section that whilst linguistic ambiguities rarely trigger direct human clarification, despite the fact they may translate to uneconomical task action, LLMs can and do sometimes respond with targeted clarification questions.

5.2 Clarification/Confirmation Questions

Having shown in the previous section, there is not a strong correlation between linguistic ambiguity and clarification questions, this section looks, at under what circumstances clarification questions are asked by human participants, and how this behaviour compares with how LLMs respond.

A common repeated pattern by human builders is the use of questions such as, *"like that?"*, as previously discussed, and others that are potentially more specific, such *"is this right?"* and *"how's that"*, with contextually evident reference to specific actions. This is not a behaviour employed by the LLMs tested here. As such, these clarification questions are unique to human participants.

Another clarification question approach exhibited uniquely by humans, is that of interrupting a set of instructions to proactively anticipate a possible pattern that they have observed to potentially reduce further instruction. For example: *"Are we repeating the structure?"*; *"Are we repeating the pattern with orange and yellow?"*; *"both sides or just this one?"*; *"are we going to repeat what we did on the other side?"*; *"same design? just blue?"*; *"repeat here?"*.

Clarification questions typically occur not directly around linguistic ambiguity, but rather ambiguity in relation to the task objectives (e.g. properties). In the following example, the architect specifies colour and quantity, but does not specify orientation, leading to similar, albeit more concise on part of the human, questions from human and LLM (Figure 5). Similarly, for individual block position in Figure 6.

Architect	then to the right of the one you just put down place 2 orange blocks
Builder (model vs. original human dialogue):	
Model	What is the current orientation of the blocks that have been placed so far? Are they forming a diagonal line from north to south or from east to west?
Human	how should they be oriented

Figure 5: Excerpt from B19-A36-C118-1524093668453 showing the subject of a clarification request in relation to orientation uncertainty

Architect	great now one orange below that one
Builder (model vs. original human dialogue):	
Model	Is it correct to assume 'now one orange below that one' means in the same diagonal direction as 'now one red block down along the diagonal
Human	diagonally or nah?

Figure 6: Excerpt from B16-A29-C144-1524067263550 showing the subject of a clarification request in relation to individual position uncertainty

There are several examples in relation to collective, individual, absolute and relative property. LLMs seem to also approach these well. Another common example is colour (shown in Figure 7).

Again, the LLM reaches a fair clarification question.

Architect	on the yellow block touching the orange , add two blocks to either side of it . making a t.
Builder (model vs. original human dialogue):	
Model	What color are the new blocks to form the T shape and do you want me to remove or replace any existing blocks at [-4,1,0], [-3,0,0], [-2,1,0]?
Human	what color blocks ?

Figure 7: Excerpt from B37-A23-C102-1523629957597 showing the subject of a clarification request in relation to the colour property

5.3 Quantitative Analysis

In this section we aim to quantify the differences between approaches and how they respond to situations that may benefit from clarification. Primarily, we are interested in each individual approaches tendency to generate a question (Table 3), particularly a relevant questions. (Table 4)

Table 3 looks at the number of questions asked in each reply. In this first case, we do not assess question relevance, but simply the tendency to reply with a question. We see a lower tendency to ask a question in the previous generation models that did not feature reasoning (i.e. `llama2`). However, later models that do feature reasoning (e.g. `llama3` and `deepseek-r1`), or previous generation models supplemented with chain-of-thought prompts (e.g. `llama2:13b-COTZERO` and `llama2:13b-COTONE`) that attempt to simulate reasoning, perform comparably with a greater tendency to question.

Approach	Questions Asked
llama2:13b	276
llama2:13b-COTZERO	303
llama2:13b-COTONE	327
deepseek-r1:8b	354
llama3.2:3b	278
llama3.3:70b	383

Table 3: Number of questions asked in different approaches

To provide some notion of relevancy, Table 4 looks specifically at the number of questions asked that target the instance of expert annotated ambiguity compared with the actions of the original human participant in the conversation. These are counted solely for instances of spatial deixis, as these are very literal and therefore the easiest to objectively assess. For example, in *"now a tower of five oranges on top of the red end"*, Llama3.3:70b's response of *"Which end of the red blocks is considered the 'red end', the one at coordinates [3, 0, 1] or [0, 0, 1]?"*, is considered to target the ambiguity.

We see that the previous generation models, not trained for reasoning, but with Chain of Thought prompting, perform comparably to modern reasoning orientated models, in this regard. Human participants did not choose to disambiguate these phrases direct at the time, but rather acted presumptively. This approach does not measure the final utility of asking the question, or any impact it may have on the conversation.

Approach	Spatial Deixis
Human participant	0
llama2:13b	0
llama2:13b-COTZERO	1
llama2:13b-COTONE	3
deepseek-r1:8b	3
llama3.2:3b	2
llama3.3:70b	5

Table 4: Number of relevant spatial deixis questions asked by approach

Another limitation to our experiment. It's challenging to communicate a sense of perspective to the LLMs. As a consequence, some instructions do not make sense, e.g.: *"A: can you come to the side of the structure so you have a side view"*; *"B: left or right"*; *"B: forward to my right or in front of me?"*; *"B: this perspective ?"*.

6 Reasoning and Clarification

Our experiments would appear to show that the recent advent of reasoning in models has the emergent benefit of allowing models to ask clarification questions. In this section, following our hypothesis that clarification is dependent on an ability to reason, we look at clarification questions in relation to reasoning through the lens of human psychology, and where available, assessments of LLM abilities to perform the required components of related human reasoning abilities

Examining the role of reasoning in clarification, knowing when to ask a question, requires reflection on the gaps in ones own knowledge, or a higher order of thought, referred to in human psychology as **Metacognition** (Flavell, 1979). LLMs originally lacked any awareness of gaps in their knowl-

edge, acting presumptively, leading to "hallucination" where, albeit often grammatically valid, a model's output would be factually incorrect or possibly nonsensical. This is most commonly due to a lack of knowledge (Zhang et al., 2023), or perhaps the inability to reason when knowledge is absent. Metacognition is the ability to reflect on held knowledge by self questioning. The Metacognitive capabilities of LLMs have been explored in previous works in relation to reasoning (Didolkar et al., 2024). Metacognitive prompting that, self questions to enhance reasoning (Wang and Zhao, 2023) has been explored in LLMs, as has self-questioning with the goal of reducing hallucination (Dhuliawala et al., 2023). A model knowing whether it has the applicable knowledge or skill to proceed, or whether to direct its process to clarification, could be seen as metacognitive regulation.

When an agent is working in collaboration with other interlocutors, clarification may be dependent on discovering other parties knowledge, abilities or attitudes and approach to a task. This crucial component, in psychology, is an aspect of reasoning known as **Theory of Mind** (Premack and Woodruff, 1978), which relates to reasoning about other participants belief states. There is some evidence to suggest, as a consequence of simulated reasoning ability, LLMs may now be able to simulate this also (Kosinski, 2023). This has been explored with LLMs in Mindcraft, which is a collaborative task in which the players have separate skills and must negotiate to reach a common goal (Bara et al., 2021). This is particularly important for LLMs in referential communication (Sidera et al., 2018)

7 Conclusion

To conclude, we find that in our conversations human participants do not commonly ask a clarification question when language is ambiguous. In the majority of these cases of linguistic ambiguity, the ambiguity appear deliberate in the interest of conversational efficiency. Consequently, there may be little utility to asking a question in many of those situations. (The resultant utility of a clarification question is not examined in this work, and may be the subject of a future work.) We did however identify one situation in which that uncertainty did propagate to create future issues. Regardless of conversational efficiency, we do find that LLMs, particularly reasoning orientated ones, are capable of asking relevant clarification questions under those circumstances.

On the topic of the originally posed human clarification questions, we found they largely followed a specific pattern and strategy not adopted by the LLMs. That was, to perform actions then use a clarification question to verify they were correct. These instances of human clarification questions largely relate to task based ambiguity rather than linguistic ambiguity.

Across all instances, we find a greater tendency of reasoning orientated approaches to pose clarification questions and find that this can be somewhat matched at test time, with methods such as COT.

Acknowledgements

This research was funded by ARCIDUCA, EPSRC EP/W001632/1

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. *arXiv preprint arXiv:2109.06275*.

Claire Bonial, Mitchell Abrams, David Traum, and Clare Voss. 2021. Builder, we have done it: evaluating & extending dialogue-amr nlu pipeline for two collaborative domains. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 173–183.

Julia Bonn, Martha Palmer, Jon Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded minecraft corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),*.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2459–2466.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Spandana Gella, Aishwarya Padmakumar, Patrick Lange, and Dilek Hakkani-Tur. 2022. Dialog acts for task-driven embodied agents. *arXiv preprint arXiv:2209.12953*.

Felix Gervits, Antonio Roque, Gordon Briggs, Matthias Scheutz, and Matthew Marge. 2021. How should agents ask questions for situated learning? an annotated dialogue corpus. *arXiv preprint arXiv:2106.06504*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rebecca MM Hicke and David Mimno. 2024. [lions: 1] and [tigers: 2] and [bears: 3], oh my! literary coreference annotation with llms. *arXiv preprint arXiv:2401.17922*.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *IJCAI*, volume 16, pages 4246–4247.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2017. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *arXiv preprint arXiv:1712.05558*.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, and 1 others. 2022. Interactive grounded language understanding in a collaborative environment: IGLU 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *Preprint*, arXiv:2305.14489.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Haau-Sing Li, Mohsen Mesgar, André FT Martins, and Iryna Gurevych. 2022. Asking clarification questions for code generation in general-purpose programming language. *arXiv preprint arXiv:2212.09885*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Chris Madge, Maris Camilleri, Paloma Carretero Garcia, Mladen Karan, Juexi Shao, Prashant Jayannavar, Julian Hough, Benjamin Roth, and Massimo Poesio. 2025. Mdc-r: The minecraft dialogue corpus with reference. *arXiv preprint arXiv:2506.22062*.

Chris Madge and Massimo Poesio. 2024. A llm benchmark based on the minecraft builder dialog agent task. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Brielen Madureira and David Schlangen. 2023a. " are you telling me to put glasses on the dog?"content-grounded annotation of instruction clarification requests in the codraw dataset. *arXiv preprint arXiv:2306.02377*.

Brielen Madureira and David Schlangen. 2023b. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *arXiv preprint arXiv:2302.14406*.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *arXiv preprint arXiv:2104.06828*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Jing-Cheng Pang, Heng-Bo Fan, Pengyuan Wang, Jia-Hao Xiao, Nan Tang, Si-Hang Yang, Chengxing Jia, Sheng-Jun Huang, and Yang Yu. 2024. Empowering language models with active inquiry for deeper understanding. *arXiv preprint arXiv:2402.03719*.

Massimo Poesio, Maris Camilleri, Paloma Carretero Garcia, and Ron Artstein. 2024. *The ARRAU 3 Annotation Manual*, v. 1.1 edition. Queen Mary University of London.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 239–246, Ann Arbor.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 101–108, Barcelona.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143.

Ying Shen and Ismini Lourentzou. 2023. Learning by asking for embodied visual navigation and task completion. *arXiv preprint arXiv:2302.04865*.

Francesc Sidera, Georgina Perpiñà, Jèssica Serrano, and Carles Rostan. 2018. Why is theory of mind important for referential communication? *Current Psychology*, 37:82–97.

Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. *arXiv preprint arXiv:2402.06509*.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024. Discourse structure for the minecraft corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *arXiv preprint arXiv:2405.12063*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Appendix

A.1 System Prompts

Architect You are an agent in a voxel world, where the most northernly point is 0,0,-5; the most westerly point -5,0,0; the most eastern point is 5,0,0; the most southern 0,0,5 and the y-axis is

up and down, with y=0 being the minimum. Your task is deliver instructions and respond to questions where necessary, to achieve a target world state []. The current world state is []. They should be easy to understand instructions, Do not ask for specific coordinates and do not accept them. Omit any pleasantries. You may ask questions

Builder You are an agent in a voxel world, using the following coordinate system, where the center on the ground is [0,0,0], [0,9,0] is the highest point in the air, the most northernly point is [0,0,-5]; the most westerly point [-5,0,0]; the most eastern point is [5,0,0]; the most southern [0,0,5]. Describe the coordinates of the blocks their colours (must be one of: blue, yellow, green, orange, purple, red) and whether the action is to add or remove them. You MUST include a question if the instruction is potentially unclear or ambiguous and you MUST follow the final JSON format: "add": [[x,y,z,"color"], ...], "remove": [[x,y,z,"color"], ...], "confidence": 0.0, "question": "...". The current world state is []. Give your final response in JSON only, without additional dialog. You may ask questions

A.2 COTONE

user

- builder: hi, let's start
- architect: place three blue blocks in a row in the center. Place a red block on top of the blue block

assistant

- architect: **Thinking:** let's think. we know three blue blocks need to go in the center. The center is [0,0,0], so to place three blue blocks would be [-1,0,0,"blue"],[0,0,0,"blue"],[1,0,0,"blue"]. Now we need to place a red block on top of the blue block. Wait, they didn't tell us which blue block, there are three. We must ask a question also. "Which of the three blue blocks would you like the red block placed on?". We need to formulate this response in the given JSON format.
- **Response:** builder: {"add": [[-1, 0, 0, "blue"], [0, 0, 0, "blue"], [1, 0, 0, "blue"]], "confidence": 0.5, "question": "Which of the three blue blocks would you like the red block placed on?"}

Coreference in simplified German: Linguistic features and challenges of automatic annotation

Sarah Jablotschkin[1] **Ekaterina Lapshinova-Koltunski**[2] **Heike Zinsmeister**[1]
[1]University of Hamburg, [2]University of Hildesheim
[1]sarah.jablotschkin,heike.zinsmeister@uni-hamburg.de,
[2]lapshinovakoltun@uni-hildesheim.de

Abstract

In this paper, we analyse coreference annotation of the German language, focussing on the phenomenon of simplification, that is, the tendency to use words and constructions that are assumed to be easier perceived, understood, or produced. Simplification is one of the tools used by language users in order to optimise communication effectively. We are interested in how simplification is reflected in coreference in two different language products exposed to the phenomena of simplification: simultaneous interpreting and Easy German. For this, we automatically annotate simplified texts with coreference. We then evaluate the outputs of automatic annotation. In addition, we also look into quantitative distributions of some coreference features. Our findings show that although the language products under analysis diverge in terms of simplification driving factors, they share some specific coreference features. We also show that this specificity may cause annotation errors in simplified language, e.g. in non-nominal or split antecedents.

1 Introduction

This paper focuses on coreference phenomena in different variants of simplified German. Simplification is one of the means used by language users in order to optimise communication effectively. Concentrating on linguistic means, we understand simplification as the tendency to use words and constructions that are assumed to be 'easier'.[1] In particular, we analyse Easy German (e.g. Maaß et al., 2021) and simultaneous interpreting (e.g. He et al., 2016). Although both language products are known to be simplified, the driving forces of the optimisation process differ: Easy German (hereafter EG) is simplified to be better perceived and understood by the target audience, i.e. the receiver side. At the same time, simultaneous interpreting (hereafter SI) is simplified due to the production constraints on the producer side, i.e. the interpreter who optimises the output to reduce their own cognitive load.

Following a linguistic approach, coreference describes the reference of two or more expressions (i.e., mentions) to one and the same entity in the extralinguistic context, also called discourse entity in contrast to extralinguistic entities (e.g., Jurafsky and Martin, 2025, chap. 23).[2] Accordingly, the goal of coreference resolution is to identify coreferent mentions and explicitely link them, so that they can be interpreted as equivalent irrespective of their surface forms, thereby forming an equivalence set. In example (1), four mentions are underlined that all refer to the same discourse entity, a plural entity that consists of two events taking place in the city of Hamburg. In addition to marking the mention spans, the syntactic heads are printed in bold face, which are relevant for further analyses.

(1) *Two major **events** are taking place in Hamburg this weekend. **These** are a music festival and a sporting event. **Both** are taking place in St. Pauli. A lot of people are coming to these two **events**.*[3]

Annotating the mentions in example (1) results in the equivalence set (2).

(2) {Two major events, These, Both, these two events}

We are interested in how coreference is expressed

[1]Non-linguistic means of simplification are, for example, layout and formatting for written communication, gestures and body languages for spoken communication.

[2]We emphasise the conceptual discourse space as reference point, to distinguish coreference resolution from the related task of entity linking. The latter task goes beyond textual-conceptual resolution by mapping mentions of named entities to real world entities encoded in knowledge bases such as Wikidata (https://www.wikidata.org/).

[3]The example is fictitious, created by 'normalizing' the Easy German example (4) with DeepLwrite (https://www.deepl.com/de/write) and translating it into English with DeepL (https://www.deepl.com/).

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 12–23
Suzhou, China, November 9, 2025.

in the introduced simplified variants of German and aim to detect coreference features specific to these variants. For this, we annotate simplified (and Standard) texts with coreference using an automatic tool and evaluate a small output sample of the simplified texts manually. In this paper, we address the following research questions (RQs):

RQ1 What challenges occur in automatic and manual annotation of coreference in simplified German?

RQ2 How are the annotation divergences linked to the linguistic specificities of simplified German?

RQ3 Which coreference phenomena are specific for the evaluated variants of simplified German?

We report on the problems of both automatic and manual annotation, employing a state-of-the-art coreference resolver and manual correction of sample annotations. In a qualitative error analysis of a small sample set, we explore to what extent annotation divergences can be explained by features of simplified German. Finally, we present quantitative distributions of coreference features in two corpora of simplified German, which are preliminary because they are biased by the challenges of automatic annotation that we detected in the state-of-the-art coreference resolver and analysed in the qualitative study.

We believe that our findings are helpful for developing better annotation tools for simplified languages, and also for deriving linguistic hypotheses about the expression of mentions and coreference relations in different variants of simplified German.

2 Background

2.1 Notion of coreference

Coreference is an important component of discourse coherence and contributes to comprehensibility and readability of texts. As introduced in Section 1, coreference is achieved by linguistic means that represent the same discourse entity in a text. These mentions can be realised by a variety of linguistic devices such as proper names or typically definite noun phrases (*the event*), pronouns (*it*) and adverbs (*there*) if they pick up an already introduced entity, or indefinite noun phrases, verb phrases or even sentences if they introduce a new entity. There are also language-specific means, such as pronominal adverbs (*dabei* 'at it') in German.[4] For comprehensive descriptions of different mention types see, e.g., Hirst (1981); Mitkov (2002); Ng (2010); Poesio et al. (2016); Kolhatkar et al. (2018).

2.2 Coreference and Cognition

At least for humans, it is assumed that the surface form of a mention serves as processing signal for the reader to facilitate identification of referents (e.g. Ariel, 2019; Kunz, 2010; Gundel et al., 1993). For example in English and German, definite noun phrases typically signal that the referent has already been introduced into the discourse or is inferrable from the linguistic or extra-linguistic context. Some forms, such as pronouns, can only refer to highly accessible referents which are very prominent in the current discourse. In addition, there are cues for discourse newness such as indefinite noun phrases. In terms of an accessibility hierarchy these referents are least accessible. In interaction with mention form, syntactic function, such as subject, object and nominal modifier is assumed to influence the probability of a mention being taken up again as an antecedent or being part of a coreference chain in general, cf., e.g., Centering Theory for English (Grosz et al., 1995). For German, this effect is less pronounced: Strube and Hahn (1999) found stronger influence of the mentions' information status, i.e. their familiarity, than their syntactic function. However, also in German subjects are highly prominent and preferred as coreferential antecedents at least for personal pronoun mentions (Portele and Bader, 2016).

2.3 Coreference resolution

The task of automatic coreference resolution consists of two subtasks, which are often done holistically in recent end-to-end approaches (Jurafsky and Martin, 2025, chap. 23): first mention detection and, second, coreference linking. In the step of mention detection, all referring candidate strings are marked. In the step of coreference linking, all mentions that refer to the same entity are grouped together.

There are various metrics for coreference resolution evaluation. The official score of the CoNLL-2011/2012 shared tasks on coreference resolution,

[4] Coreference can also be implicit in languages that do not require to express verbal arguments in the text, which is modeled as 'zero anaphora'. This phenomenon goes beyond the scope of our paper.

which was intended to provide a standardised evaluation metric, is the arithmetic mean of three other scores: MUC, B^3 and $CEAF_e$ (Pradhan et al., 2014, 30). However, as discussed in Moosavi and Strube (2016), each of these metrics has their shortcomings: While MUC is the least discriminative one, B^3, CEAF (as well as BLANC) show a mention identification effect, meaning that the score improves notably if a mention is identified and the detection of coreference links has a much smaller impact on the scores. To overcome this bias in favour of mention detection, Moosavi and Strube (2016) propose LEA, a link-based entity-aware metric.

2.4 Annotated resources

There exist a number of corpora for German annotated with coreference chains. However, none of them contains simultaneous interpreting, and only the small LeiKo corpus (Jablotschkin and Zinsmeister, 2020, 2024) contains Easy German with an annotation layer for coreference. For the former, resources containing spoken language production could be of relevance. For instance, ParCorFull (Lapshinova-Koltunski et al., 2018) and ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) contain transcripts of TED talks annotated with coreference chains. The German texts in this corpus represent written translations of the English transcripts. The corpus GECCo (Kunz et al., 2021) contains coreference annotation of spoken parts as well, e.g. academic speeches, transcribed interviews, transcripts of TV talkshows in English and German. They are all original spoken text production and there is interpreting available. Studies show that interpreting possesses a number of linguistic characteristics that set it apart as a unique language product, different from other types of spoken production (see e.g. Lapshinova-Koltunski et al., 2021). German translations annotated with coreference are also contained in the corpus described in Grishina and Stede (2015). However, no simplified texts are included. Further corpora with coreference annotation of written Standard German include the richly annotated Potsdam Commentary Corpus (PCC, Stede and Neumann, 2014; Bourgonje and Stede, 2020) containing newspaper commentaries and the TüBa-DZ treebank (Hinrichs et al., 2004; Naumann, 2006) based on diverse newspaper articles from *die tageszeitung*.

2.5 Variation in coreference features

As already introduced above, coreference chains contain linked mentions of the same entities throughout a text. These mentions are realised by nominal phrases, pronouns and other linguistic means. Previous studies show that various text types—genres or registers—may have preferences for certain types of linguistic devices (Lapshinova-Koltunski and Kunz, 2020). Knowledge on these preferences is important, as they may impact performance of coreference resolution systems.

There are studies showing that register and mode have an impact on anaphora prediction models (Zeldes, 2018). Hence, knowledge of genre- or mode-dependent differences in coreference phenomena is useful for coreference resolution that requires domain adaptation (Poesio et al., 2024; Roesiger and Teufel, 2014; Uryupina and Poesio, 2012; Yang et al., 2012; Apostolova et al., 2012).

Dealing with specific types of texts, we expect to identify specific coreference features typical for either Easy German or simultaneous interpreting. Additionally, we expect to find coreference features that are common in both of these language products as they are both prone to linguistic simplification.

2.6 Coreference in simplified German

There is only very limited work on coreference in simplified language. Wilkens and Todirascu (2020) and Wilkens et al. (2020) analyse coreference in simplified French texts. They report a rich set of corpus statistics on a small parallel corpus of French narrative texts simplified for dyslexic children. An important finding is that the simplified texts have more coreference chains with lexical noun phrases than with pronouns (p. 96).[5]

Switching to simultaneous interpreting in German, as can be seen from example (3), the English source contains the chain *the practice of sandblasting – which – jeans sandblasted* with mentions filled with a relative pronoun and a full lexical phrase. At the same time, the interpreting into German contains a demonstrative pronoun (*das*) and an adverb (*so*) instead. From the lexical point of view, the means of referring are simpler in the interpreted output. In contrast, the coreference chain in the Easy German example (4) contains no pro-

[5]Based on their analyses, the authors write simplification guidelines and create another corpus with manual simplifications, on which they then evaluate a rule-based system. We would like to thank the reviewer who pointed out this relevant work to us.

forms, but lexical repetitions as a simplification strategy. In addition, the anaphors are highlighted by being positioned sentence-initially.

(3) **Original**: *In particular, I want to draw attention to the practice of sandblasting of jeans which happens more in Bangladesh than anywhere else in the world. Up to one hundred million pairs of jeans sandblasted a year being export from Bangladesh.*
SI: *Aber was dort in Bangladesch passiert, ist weiter eine Bedrohung für die Gesundheit der Arbeitnehmer, insbesondere die* ***Sandstrahlmethode*** *für Jeans.* ***Das*** *wird in Bangladesch vor allen Dingen durchgeführt. Einhundert Millionen Jeans werden* ***so*** *hergestellt und exportiert pro Jahr.*

(4) *In Hamburg sind am Wochen·ende 2 große* ***Veranstaltungen****. Diese 2 großen* ***Veranstaltungen*** *sind: • Ein Musik·fest. • Und eine Sport·veranstaltung. Die 2 großen* **Veranstaltungen** *sind in St. Pauli. [. . .] Und die 2 großen* ***Veranstaltungen*** *sind [. . .] Zu diesen 2 großen* ***Veranstaltungen*** *kommen sehr viele Menschen.* (There are 2 big events in Hamburg this weekend. These 2 big events are: - A music festival. - And a sports event. The 2 big events are in St. Pauli. [. . .] And the 2 big events are [. . .] A lot of people come to these 2 big events.)

Overall, we expect to find more accessible forms, i.e., subjects and direct objects, as well as demonstrative pronouns in both simplified variants of German if compared to Standard German. However, we also expect to find differences across the two variants. In Easy German, we expect to find fewer personal pronouns and more lexicalised subjects, due to the achievement of ease in perception for the readers (e.g. Bock and Pappert, 2023; Bredel and Maaß, 2016; Netzwerk Leichte Sprache, 2022). In simultaneous interpreting, instead, we expect to find more personal pronouns and fewer lexicalised subjects, as pronouns are shorter and easier to produce (e.g. He et al., 2016).

3 Methodology

3.1 Data

For our analyses, we use two different sets of data: first, texts in Easy German and Standard German from DE-Lite v1 (Jablotschkin et al., 2024), which covers a number of online text genres. Second, transcribed texts of (spoken) German that were simultaneously interpreted from English into German extracted from EPIC-UdS (Przybyl et al., 2022), a multilingual parallel and comparable corpus of simultaneous interpreting of political speeches held by members of the European Parliament. The interpreted speeches were manually transcribed.

The automatic annotation was performed on a sample of about 4,700 texts from DE-Lite v1. This subcorpus contains about 1.2 million tokens. Moreover, DE-Lite also comprises comparable texts in Standard German, about 800 texts and 1.1 million tokens. For simultaneous interpreting, we used a sample of 137 texts of German interpreting from English extracted from EPIC-UdS.

For the manual correction, we identified eight automatically annotated texts of similar length (four text in Easy German from different genres and four text of simultaneous interpreting).

3.2 Automatic annotation

To analyse coreference, we annotated the data with the state-of-the-art coreference resolver CorPipe (Straka, 2023a) that won the CRAC 2023 shared task on multilingual coreference resolution (Žabokrtský et al., 2023). CorPipe is a system for multilingual coreference resolution that was trained on all corpora available in CorefUD 1.1 (Nedoluzhko et al., 2022). The underlying training data for German include two corpora: ParCorFull (Lapshinova-Koltunski et al., 2022) and PCC (Bourgonje and Stede, 2020) (see also Section 2.4). Both corpora contain manual annotations of coreference chains. However, their annotations differ in their definitions of certain structures. One striking difference is the definition of the mention span. While the PCC implements the principle of maximum mention span, which includes, for example, leading prepositions such as in the span *wegen seiner Situation* ('because of his situation'),[6] ParCorFull restricts the mention span in such cases

[6]This is a residual of the syntactic annotation in the Tiger corpus which opted for 'flat' prepositional phrases without a hierarchically embedded noun phrase, see e.g. the discussion in Dipper and Kübler (2017).

to the nominal core (here: *seiner Situation*) (see Nedoluzhko et al. 2021 for a detailed description of the annotation regimes of all CorefUD 1.1 corpora). CorPipe performs coreference resolution in two steps: mention detection and coreference linking. Unlike end-to-end resolution systems, this approach makes it possible to detect singletons (Straka, 2023b, 41).

Since CorPipe requires tokenization and morpho-syntax supplied by UDPipe 2 (Straka, 2018), the outputs in our corpus contain annotations of not only coreference chains, but also syntactic functions and parts of speech based on Universal Dependencies (UD) and Universal POS tags (UPOS), see Nivre et al. (2020) for more details. We use this information to analyse linguistic features of annotated mentions that are members of entity sets in both simplified variants and also in a subcorpus of written Standard German, see Section 4.3.

3.3 Annotation study

In order to get insights into the quality of automatic annotation and into typical features of simplified language products that are not captured satisfactorily by automatic annotation, we conducted an annotation study:[7] Two student annotators performed manual correction on a small subset of the automatically annotated data (four Easy German texts and four interpreted texts, cf. Section 3.1) using CorefAnnotator (Reiter, 2018), a tool for manual coreference annotation. Correction steps included adjustment, deletion or addition of mention spans and reorganisation of mentions into appropriate equivalence sets (i.e., entities). We analysed annotation divergences with respect to both annotators, focusing on mention detection and saving a deeper investigation of coreference linking for future work.

We also calculated inter-annotator agreement (IAA, see details in Section 4.1 below) which will help us analyse annotation problems and some of the specificity of simplified language. We leverage these system measures also to report human inter-annotator agreement. We do not state chance-corrected values, such as weighted α (Passonneau, 2006), since their interpretability and comparability for measuring agreement of manual coreference annotation are open to doubt (cf. Paun et al., 2022, pp. 66-70).

[7] Our annotation guidelines in German are available here: https://www.fdr.uni-hamburg.de/record/17944.

4 Results

4.1 RQ1: Evaluation of Coreference Annotation

We start by analysing the manual correction of a small subset of eight automatically annotated texts by two student annotators (four EG texts and four SI texts with about 48.5 mentions per text detected by CorPipe on average [median]), see also section 3.1, and calculating inter-annotator agreement.

Table 1 and Table 2 report automatic annotation quality in terms of mention detection as well as all of the scores mention in Section 2.3 against the manual corrections of annotator 1 and annotator 2, respectively. Instead of performing adjudication and creating a gold annotation we scored the automatic annotation against both manual annotation sets individually. The scores were calculated with the Reference Coreference Scorer (Pradhan et al., 2014) and the CoVal Scorer (Moosavi et al., 2019). Both scorers take CoNLL files as input, not CoNLL-U files. This means that no syntactic information, such as head token, is stored in the evaluated files and only exact matches are considered matches.

	Recall	Precision	F1
mentions	74.95	90.18	81.87
MUC	63.75	87.43	73.73
B^3	64.73	87.11	74.27
$CEAF_m$	69.82	84.01	76.26
$CEAF_e$	71.26	77.13	74.08
BLANC	55.07	84.86	66.77
LEA	59.61	77.97	67.56
CoNLL score	–	–	74.03

Table 1: CorPipe vs. annotator 1

In Table 1 all F1 scores are higher than in Table 2. This shows that the automatic annotation set and

	Recall	Precision	F1
mentions	66.98	81.05	73.35
MUC	57.20	80.33	66.82
B^3	55.55	76.03	64.20
$CEAF_m$	60.45	73.28	66.25
$CEAF_e$	58.86	63.24	60.97
BLANC	48.33	72.53	57.96
LEA	49.56	65.72	56.51
CoNLL score	–	–	64.00

Table 2: CorPipe vs. annotator 2

the manually corrected version of annotator 1 are more similar to each other than the automatic annotation and the manually corrected version of annotator 2. Moreover, in all of the scores Precision is higher than Recall, meaning that the annotators added more mentions and links than they deleted. We therefore provide a qualitative analysis of False Negative mentions in Section 4.2.

In the CRAC 2023 shared task, CorPipe achieved a CoNLL score of 72.12 % for the ParCorFull test set and 71.09 % for the PCC test set (Straka, 2023b). These results are slightly lower than our result compared to annotator 1, which was 74.03 % (see Table 1). However, the original results are much better than our result compared to annotator 2 (64.00 %, see Table 2). To understand the quality of the coreference linking and its relation to phenomena of simplification, further analysis is required.

While Table 1 and Table 2 provide overall scores for the annotated data, Figure 1 allows a more fine-grained analysis: By calculating F1 scores of mention identification for the individual texts, we show that automatic annotation quality differs considerably across texts. While the score lies between 76 (text 2) and 86 (text 5) for annotator 1, the span for annotator 2 is even greater and lies between 60 for text 4 and 88 for text 1. Inter-annotator agreement in terms of F1 of mention identification is lowest for text 6 (63) and highest for text 1 (92). Numbers for comparing annotator 1 and annotator 2 directly can be found in Table 3 and Table 4 in the appendix.

Some of the deviations between human annotators and model performance is due to the fact that the model was trained on different corpora based on partly diverging guidelines. Hence the model's training data consisted of diverging annotations which leads to seemingly inconsistent model decisions, e.g., with respect to the yield of mentions by including/excluding prepositions (cf. Section 3.2).

4.2 RQ2: Specific Annotation Problems of Simplified German

By qualitatively analysing divergences between automatic and manually corrected annotations, we were able to identify some coreference features that are related to simplification and that are not (systematically) captured by automatic coreference resolution. Even though we did not perform adjudication to create a gold annotation set, the categories presented in this section repeatedly appeared in the two annotators' corrections. Due to the costly process of manual data creation and the resulting small sample size, the findings presented here can barely be analysed statistically and mainly serve as a basis for further hypotheses and exploration. Furthermore, since annotators added more mentions than they deleted (see Section 4.1), we focus on False Negatives of the automatic annotation.

First of all, we detected a frequent use of demonstrative pronouns and pronominal adverbs in both simplified language products. While they are primarily used to reduce syntactic complexity in Easy German, they are also a typical feature of interpreted language. They have a short form and allow for packing and wrapping larger information pieces into smaller units, and are therefore frequent in interpreted texts (cf. example (3)). In our data, a relevant proportion of automatically detected singletons constituted demonstrative pronouns or pronominal adverbs. In the course of manual correction, the annotators identified corresponding antecedents which often were verb phrases or whole sentences but could also be (complex) nominal phrases. Non-nominal antecedents are one of the most frequent categories of False Negatives in our automatically annotated data.

(3) *Genau wie die Vorsitzende des Ausschusses habe ich große Sympathie für Kommissar Kovács [wegen seiner schwierigen Situation]$_1$. [Er versucht jetzt diese Besteuerungsregelung da durch die Maschinerie der Gemeinschaft zu bringen]$_1$. Und [dafür]$_1$ braucht er größtmöglichen politisches Taktgefühl wegen der Einstimmigkeit.*

(SI_EN_DE_029: Like the chair person of our committee I would like to sympathise with the Commissioner with Commissioner Kovács [because of his difficult situation]$_1$. [He is now trying to get this taxation regulation through the Community machinery]$_1$. And [for this]$_1$ he needs the greatest possible political tact because of the unanimity.)

(4) *[Einige Züge]$_1$ fallen am späten Abend aus.*

Und [einige Züge]$_2$ fallen am frühen Morgen aus.

Wir wissen nicht: Wie lange fallen [die Züge]$_{1,2}$ aus?

(m_3045_easy: [Some trains]$_1$ late at night are being cancelled. And [some trains]$_2$ early in the morning are being cancelled. We don't know: How long are [the trains]$_{1,2}$ being cancelled?)

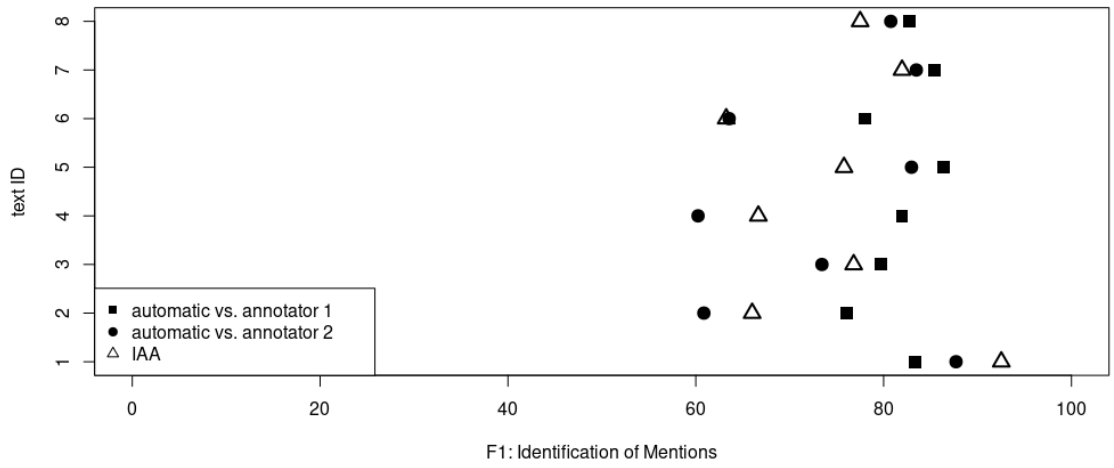

Figure 1: Agreement scores (F1) of mention detection per manually corrected text (Text 1-4: EG; Text 5-8: SI)

(5) *Und da müssen wir uns jetzt wirklich einmal darauf konzentrieren auch [diese praktische Hilfe]*$_{1,2,3}$ *zu leisten. [Geld]*$_1$ *[Trinkwasser]*$_2$ *[Strom]*$_3$.

(SI_EN_DE_114: And we really have to focus there now on also giving [this practical help]$_{1,2,3}$. [money]$_1$ [drinking water]$_2$ [electricity]$_3$.)

(6) *27 Länder in Europa haben sich [zu einer Gruppe]*$_1$ *zusammen-geschlossen.*

[Die Gruppe]$_1$ *heißt: Europäische Union.*

[Dazu]$_1$ *kann man auch kurz sagen: EU.*

Die Länder [in der EU]$_1$ *wollen zusammen politisch und wirtschaftlich stärker sein.*

(p_806_easy: 27 countries in Europe have formed [a group]$_1$. [The group]$_1$ is called: European Union. A shorter name [for this]$_1$ is: EU. Together, the countries [in the EU]$_1$ want to be politically and economically stronger.)

(7) *Deshalb brauchen wir auch mehr [Auszubildende]*$_1$.

[Auszubildende]$_1$ *sind junge Menschen.*

[Diese jungen Menschen]$_1$ *lernen einen Beruf.*

(m_3045_easy: That is why we also need more [trainees]$_1$. [Trainees]$_1$ are young people. [These young people]$_1$ learn a profession.)

The automatic annotation also didn't capture split antecedents, which are frequent in Easy German text because they allow for syntactic and content simplification. In example (4), the sentence segment *Einige Züge fallen [...] aus* is repeated. This way, it's possible to avoid coordination ellipsis. At the same time, the discourse model of the reader is only slowly enriched with information and the meaning of the discourse segment is made more explicit. However, reference resolution becomes more complex because in the last sentence, two expressions that originally establish reference separately from each other have to be subsumed under the broader reference of the expression *die Züge*.

Split antecedents also occur in the interpreting data where they have an argumentative function. In example (5), the speaker emphasises the necessity to provide practical help by splitting up the vague term *help* into smaller and more concrete measures in order to specify what kind of help is needed. The help measures are listed asyndetically, which is another source of false negatives in our data: In lists, CorPipe usually overlooks some or even all mention spans. Lists are frequent in Easy German as well where they are typically used to present information in a syntactically simple way or to specify concepts that are not considered part of the readers' background knowledge.

In example (6), different terms are used in order to refer to the same concept, namely the European Union. For Easy German, it is usually recommended to avoid using different terms for the same concept. However, as can be seen in example (6), lexical substitution also occurs as part of concept

explanations. The same holds true for example (7), where both *Auszubildende* and *diese jungen Menschen* refer to trainees. Concept explanations are necessary in Easy German because potential readers are not expected to have large background and world knowledge, and consequentially we hypothesise lexical substitution to be frequent in Easy German as well. However, lexical substitution, which occurs in our interpreted as well as Easy German data, often isn't detected by automatic coreference resolution.

Example (7) presents another typical feature of Easy German which is neither captured by most coreference guidelines for German nor by automatic coreference resolution: The quantified nominal phrase *mehr Auszubildende* and the bare plural *Auszubildende* do not refer to an individual referent. Instead, as part of a concept explanation, they generically refer to a category of referents. According to the annotation guidelines, generic expressions should not be annotated as referring expressions However, since we were aware that they play an important role in Easy German texts, we specifically instructed the annotators to look out for them. Even though not all indefinite noun phrases are generic, in our data they constitute a frequent category of false negatives.

4.3 RQ3: Coreference features in simplified German

Finally, based on the syntactic functions of their heads, we selected a subset of automatically detected mentions in 4,700 Easy German texts, 800 Standard German texts, and 137 simultaneously interpreted texts (see Section 3.1). We analysed the proportions of the following UD labels (Nivre et al., 2020) with regard to all mentions (see Figure 2): `nsubj`, `nsubj:pass`, `nmod`, `obj`, `obl` and `obl:arg`. These labels were selected based on the assumption that their proportions reflect differences in the expression of accessibility in the respective subcorpora (see Section 2.2). In a second step, we analysed the distribution of POS labels (STTS, Schiller et al., 1999) among the pre-selected dependency labels (see Figure 3).

As seen from Figures 2 and 3, we observed similar tendencies in the distribution of both syntactic functions and parts-of-speech in both variants of simplified German. Interestingly, both simplified variants use more subjects as compared to standard German (see Figure 2). However, they differ in terms of the form of the subject mention: While Easy German prefers to use common nouns, personal pronouns are used in simultaneous interpreting. These findings confirm our assumptions about the distribution of more accessible forms, see Section 2.6 above.

However, our manual pilot analysis in Section 4.2 revealed that certain types of mentions remain undetected by CorPipe. That is why the quantitative distributions of syntactic functions and parts of speech among the syntactic heads of mentions only provide a first glimpse of morphosyntactic features of mentions in the respective subcorpora. Based on our manual analysis, we assume that there is also a considerable amount of verb phrases or even larger units of text that constitute antecedents of demonstrative pronouns (cf. ex. (3)) and that the numbers of mentions with nominal head must be even higher than depicted in Figure 3 due to structures like split antecedents, lexical substitution and generic coreference that are not captured by CorPipe (see examples (4) to (7)).

In addition, since we only analysed mention heads, we cannot make conclusive remarks about the accessibility of mentions, which is often determined by modifiers, articles or attributes (see Section 2.2). For example, mentions with a nominal head (`NN`) preceded by a definite article are more accessible than mentions with a nominal head preceded by an indefinite or no article.

5 Conclusion and Discussion

In this study, we use automatic coreference annotation to detect coreference chains in two variants of simplified German. Although these language products diverge in terms of simplification driving factors (producer's vs. receiver's perspective), we find some similarities in their linguistic features. We also show that the specificity of simplified language may cause annotation errors, especially in case of non-nominal antecedents and split reference. We manually explore these errors to find out that the main reason for their occurrence is the linguistic specificity of simplified texts: most frequent errors are observed in the linguistic constructions that are typical for both Easy German and interpreting. This points to the need for dedicated resources that are specifically trained on simplified German. This underlines the need for domain adaptation in coreference resolution on the data for less-ressourced and less-researched language products.

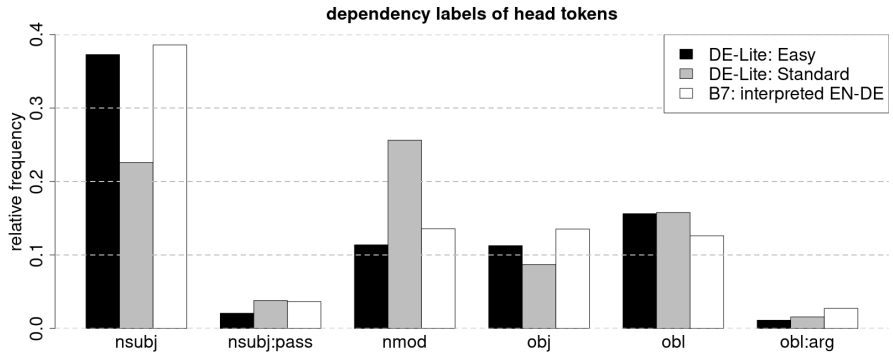

Figure 2: Syntactic functions (UD labels) of mention heads. Total: All automatically detected mentions

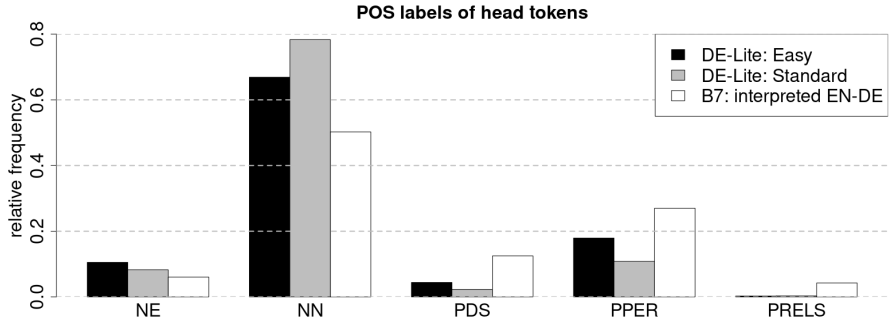

Figure 3: POS (STTS) of mention heads. Total: Automatically detected mentions with one of the following UD labels: nsubj, nsubj:pass, nmod, obj, obl, obl:arg

Limitations

One of the limitation of our study is that the we look into the annotation errors rather exploratively without providing statistical analysis on the errors types. Also, the overall dataset for simultaneous interpreting is relatively small. We also understand that the two variants of simplified German are not entirely comparable. So, we are aware of the genre effect that may have an impact on our results. To validate this, we would need to compare our results with the distributions in spoken German too, which remains beyond the scope of this paper. Also, we do not perform any comparison with annotation errors in standard German. However, we know that CorPipe performs slightly better on commentary texts than on spoken data, as reported by Straka (2024, 2023a), which may have an impact on our results too. Another problem is that we have some errors in the pre-processing that also impact the automatic divergences. For instance, segmentation errors in the automatic pipeline that introduce erroneous sentence boundaries affect mention detection, because mention spans never cross a sentence boundary.

Acknowledgments

This research is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We would like to thank Nele Benz and Anastasiia Stulen for their thorough annotation work and Florian Schneider from the HCDS Hamburg for his assistance in parsing the corpora with CorPipe. Additionally, we would like to thank the reviewers for their insightful comments and for directing us to additional relevant literature.

References

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat, and Dina Demner-Fushman. 2012. Domain adaptation of coreference resolution for radiology reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 118–121, Montréal, Canada. Association for Computational Linguistics.

Mira Ariel. 2019. *Accessing Noun-Phrase Antecedents*. Routledge, London.

Bettina M. Bock and Sandra Pappert. 2023. *Leichte Sprache, einfache Sprache, verständliche Sprache*.

Narr Studienbücher. Narr Francke Attempto, Tübingen.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Dudenverlag, Berlin.

Stefanie Dipper and Sandra Kübler. 2017. German Treebanks: TIGER and TüBa-D/Z. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 595–639. Springer Netherlands, Dordrecht.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 14–22, Beijing, China. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. 69(2):274–307.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. 2004. Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT)*, pages 51–62.

Graeme Hirst. 1981. *Anaphora in Natural Language Understanding: A Survey*, volume 119 of *Lecture Notes in Computer Science*. Springer.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-lite - a new corpus of easy German: Compilation, exploration, analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.

Sarah Jablotschkin and Heike Zinsmeister. 2020. LeiKo: A corpus of easy-to-read German. https://zenodo.org/record/3923917.

Sarah Jablotschkin and Heike Zinsmeister. 2024. LeiKo — ein umfassend annotiertes Korpus mit Texten in Leichter und Einfacher Sprache. *KorDaf*, 4(2):256–263. Medium: application/pdf,application/xml Publisher: Universitäts- und Landesbibliothek Darmstadt.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. https://web.stanford.edu/~jurafsky/slp3/.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora With Non-Nominal Antecedents in Computational Linguistics: A Survey. *Computational Linguistics*, 44(3):547–612.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, José Manuel Martínez Martínez, Katrin Menzel, and Erich Steiner. 2021. *GECCo - German-English Contrasts in Cohesion*. De Gruyter Mouton, Berlin, Boston.

Kerstin Anna Kunz. 2010. *Variation in English and German Nominal Coreference: A Study of Political Essays*. Number 21 in Saarbrücker Beiträge Zur Sprach-Und Translationswissenschaft. Lang, Frankfurt, M. [u.a.].

Ekaterina Lapshinova-Koltunski, Yuri Bizzoni, Heike Przybyl, and Elke Teich. 2021. Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 82–90, online. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. ParCorFull2.0: a parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813, Marseille, France. European Language Resources Association.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. 2020. Exploring coreference features in heterogeneous data. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 53–64, Online. Association for Computational Linguistics.

Christiane Maaß, Isabel Rink, and Silvia Hansen-Schirra. 2021. Easy language in Germany. In Camilla Lindholm and Ulla Vanhatalo, editors, *Handbook of Easy languages in Europe*, volume 8, pages 191–218. Frank & Timme.

Ruslan Mitkov. 2002. *Anaphora Resolution*, 1 edition edition. Routledge, London; New York.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642. Association for Computational Linguistics.

Karin Naumann. 2006. Annotation of Referential Relations. Technical report, University of Tübingen. Annotation guidelines.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdenek Zabokrtskỳ, and Daniel Zeman. 2021. Coreference meets universal dependencies–a pilot experiment on harmonizing coreference datasets for 11 languages. Technical report, ÚFAL MFF UK, Praha, Czechia. https://ufal.mff.cuni.cz/techrep/tr66.pdf.

Netzwerk Leichte Sprache. 2022. Die Regeln für Leichte Sprache vom Netzwerk Leichte Sprache.

Vincent Ng. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical methods for annotation analysis*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Massimo Poesio, Maris Camilleri, Paloma Carretero Garcia, Juntao Yu, and Mark-Christoph Müller. 2024. The ARRAU 3.0 corpus. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 127–138, St. Julians, Malta. Association for Computational Linguistics.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution. Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer.

Yvonne Portele and Markus Bader. 2016. Accessibility and referential choice: Personal pronouns and d-pronouns in written german. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (18).

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35. Association for Computational Linguistics.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022. EPIC UdS - creation and applications of a simultaneous interpreting corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1193–1200, Marseille, France. European Language Resources Association.

Nils Reiter. 2018. CorefAnnotator: a new annotation tool for entity references. Doi: 10.18419/OPUS-10144.

Ina Roesiger and Simone Teufel. 2014. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung, Seminar für Sprachwissenschaft, Stuttgart, Tübingen.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik*, pages 925–929.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL*

2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2023a. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Milan Straka. 2023b. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Michael Strube and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25:309–344.

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 187–191, Istanbul, Turkey. European Language Resources Association (ELRA).

Rodrigo Wilkens, Bruno Oberle, and Amalia Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100, Marseille, France. European Language Resources Association.

Rodrigo Wilkens and Amalia Todirascu. 2020. Simplifying coreference chains for dyslexic children. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1142–1151, Marseille, France. European Language Resources Association.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor Wai-Hung Tsang, Kian Ming Adam Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 744–753, Jeju Island, Korea. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Amir Zeldes. 2018. A predictive model for notional anaphora in English. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

	Recall	Precision	F1
mentions	74.57	74.15	74.36
MUC	67.33	66.15	66.73
B^3	64.64	64.55	64.59
$CEAF_m$	67.87	67.35	67.61
$CEAF_e$	62.45	62.91	62.68
BLANC	58.92	60.54	59.71
LEA	57.61	57.09	57.35
CoNLL score	–	–	64.67

Table 3: Annotator 1 vs. Annotator 2

	Recall	Precision	F1
mentions	74.15	74.57	74.36
MUC	66.15	67.33	66.73
B^3	64.55	64.64	64.59
$CEAF_m$	67.35	67.87	67.61
$CEAF_e$	62.91	62.45	62.68
BLANC	60.54	58.92	59.71
LEA	57.09	57.61	57.35
CoNLL score	–	–	64.67

Table 4: Annotator 2 vs. Annotator 1

Revisiting the Givenness Hierarchy. A Corpus-Based Evaluation

Christian Chiarcos
Applied Computational Linguistics (ACoLi)
University of Augsburg, Germany

Abstract

The Givenness Hierarchy (GH) models how speakers signal cognitive statuses of referents in discourse, playing a key role in computational models of situated communication and in applied linguistics. We present an empirical method to evaluate the Givenness Hierarchy using large corpora with coreference annotations. Our findings confirm predicted associations between cognitive statuses and referring expressions across multiple languages, while also highlighting limitations, notably difficulties to approximate the cognitive status UNIQUE and to account for demonstrative noun phrases. Additionally, we demonstrate how coreference data can be used to bootstrap GH annotations, facilitating automatic labelling of cognitive statuses and advancing discourse-aware NLP. Finally, we provide conversion scripts to transform Japanese and Korean corpora into CorefUD-compatible formats, supporting broader multilingual research despite current annotation and licensing constraints. Our work bridges theoretical linguistics and practical computational methods, offering a scalable framework to study givenness across diverse languages.

1 Background and Motivation

Effective reference – whether by humans or dialogue systems – requires contextually appropriate expressions. As illustrated by variations in translation, language offers multiple ways to express the same referent: While Challoner (1749-1752 (revision) translated 1Ki, 11,28 with pronoun and elision (*Solomon seeing him (...), Ø made him chief ...*), Darby (1890) used a definite NP and a pronoun (*Solomon saw the young man (...), and he made him ruler ...*). To account for the functional dimension of this flexibility, various theories of information status (Prince, 1981; Givón, 1983; Ariel, 1990; Chafe, 1994) posit hierarchies or scales mapping referential forms to degrees of 'salience', 'accessibility', or 'givenness', but while they agree on broad trends – pronouns denote high degrees of givenness, full NPs low – they differ in terminology and granularity. From the set of prominent theories, we adopt the Givenness Hierarchy (GH) by Gundel et al. (1993) for three reasons: (1) It is a relatively detailed theory in that it accounts not only for pronominal, nominal, definite and indefinite forms, but also for different qualities of demonstrative pronouns and demonstrative NPs, as well as for possible deviations from the expected encoding of the statuses it proposes; (2) unlike any other of the aforementioned theories, it comes with explicit, and practical annotation guidelines;[1] and (3) the theory and its annotation guidelines have been applied to a considerable number of typologically diverse languages.[2] Moreover, the GH is while widely cited in technical contexts (Han et al., 2022; Pal et al., 2021; Spevak et al., 2022; Higger and Williams, 2024; Daigler et al., 2024),[3] as well as in applied linguistics (Gundel and John-

[1] There are other annotation guidelines for information status, e.g., Nissim et al. (2004); Ritz et al. (2008); Baumann and Riester (2013); Dyer et al. (2024), but these aim to generalize over multiple theories and are thus not directly comparable. In particular, they cannot be directly used to evaluate specific claims of Gundel et al.'s theory if the underlying theories did not share the same predictions, esp., regarding the use of demonstratives.

[2] Aside from the major languages considered here and by Gundel et al. (1990, 1993), this also includes Breton (Hedberg and Schapansky, 1996), Yapese (Ballantyne, 2004), Kumyk (Humnick, 2005), Irish (Mulkern, 2008), Kaqchikel Maya (Hedberg, 2010), Eegimaa, Ojibwe (Gundel et al., 2010), Farsi (Khormaee and Skrouchi, 2015), Luo (Omondi, 2016) and American Sign Language (Swabey, 2011), among others.

[3] In the era of LLMs, many of the challenges that theories of information status such as the Givenness Hierarchy account for – reordering constituents, anaphor resolution and generation, prediction and interpretation of non-canonical structures, lexicalizing frames, and handling grammatical voice – may be solved to some extent in practice, but only for major languages and but without any insight into the underlying processes and their actual effects on the interlocutors, thus not directly applicable for low-resource languages or in controlled and vulnerable settings such as in human-robot dialog.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 24–41
Suzhou, China, November 9, 2025.

son, 2013; Kim, 2016; Velnic, 2018; Krüger, 2018). At the same time, however, its empirical basis remains limited. Early support came from elicitation experiments (Gundel et al., 1990, 1993), but few annotated corpora are publicly available, the limiting replicability and application. Here, we provide a technical operationalization of the Givenness Hierarchy on the basis of existing corpora with coreference annotation for all languages originally considered by Gundel et al. (1990, 1993) to motivate and develop their theory (English, Arabic, Chinese, Japanese, Korean, Russian, and Spanish), with the goal of to evaluate the theory, to test and confirm its predictions, and to develop a method for bootstrapping language-specific givenness hierarchies for other languages from existing coreference annotation.

Gundel et al. (1993) postulate a hierarchy of six cognitive statuses and their alignment with prototypical expressions. Below, these ranked from most to least accessible

1. **in focus**: referent is the current focus of attention and highly prominent in the local context ($\sim$ *he*, *she*).
2. **activated**: referent is present in the local context ($\sim$ *this/that*, *this man*).
3. **familiar**: referent is known to both speaker and hearer from prior discourse ($\sim$ *that man*).
4. **unique**: the referent is uniquely identifiable to hearer and speaker ($\sim$ *the man*).
5. **referential**: the speaker refers to a specific but possibly unknown entity ($\sim$ *this guy*).
6. **type** (type identifiable): hearer can identify the category of a referent ($\sim$ *a man*).

Unlike other theories, this is an implicative hierarchy, i.e., a status higher in the hierarchy entails all lower statuses, so, it can be referred to with their forms – and speakers may use such deviations to convey implicatures (e.g., using a definite NP for an in-focus entity to emphasize contrast). Gundel et al. (1990,1993) provide predictions for English, Arabic, Chinese, Japanese, Korean, Russian, and Spanish and evaluate these in elicitation experiments. However, no annotations seem to be publicly available, and the numbers they (and many later papers) report do not always reach the levels of statistical significance. Table 1 replicates their English data and adds Pearson correlation and binary χ^2 significance scores for each pairing of referring expressions and statuses, and this reassessment confirms key GH assumptions: significant positive correlation between **in focus** and pronouns, **unique** and definite NPs, **type** and indefinite NPs. It also shows negative correlations where expected. However, data on demonstratives and emphatic pronouns remain sparse, indicating a potential weak spot in theory – but, ironically, these are the very predictions that distinguish GH from competing theories: While the general pattern of pronoun $>$ definite NP $>$ indefinites is generally accepted, its fine-grained distinctions, especially regarding demonstratives, remain controversial. Siddharthan et al. (2011) argue that GH conflates dimensions, and psycholinguistic experiments such as Xu and Xiang (2021) failed to confirm some predicted effects. The status of demonstratives is particularly contested: In direct opposition to Gundel et al. (1993), Sgall et al. (1986) claimed that demonstrative pronouns rank higher than personal pronouns, and Ariel (1991) saw demonstrative NPs as lower than definites.

To address data sparsity and controversies – but also potential biases of annotators who are aware that certain forms indicate certain categories –, we propose a new approach: extrapolating cognitive status from existing corpora with coreference annotation, which substantially exceed the traditional elicitation experiments in scale. In the last years, this approach has become feasible due to the increased availability of corpora with coreference annotation, covering now all original GH languages. In comparison with earlier elicitation methods, these offer higher coverage – hundreds or thousands of tokens per referential form – and reduce circularity risks. We replicate the original GH setup by focusing on the same referential expressions,[4] to the extent they are annotated.[5] By grounding the evaluation of the Givenness Hierarchy in independently created coreference corpora, we aim to reassess its predictions and offer a scalable, reproducible methodology to support or revise its theoretical foundations. We are specifically interested in debated GH claims that are not suffi-

[4]This is particularly important for calculating totals. In particular, we do not evaluate against *all* referring expressions, but only against (the total of) those considered by Gundel et al., so that certain categories, e.g., first- and second-person anaphora, pronominal adverbs, quantified NPs and proper names are excluded.

[5]Some corpora have an annotation bias, e.g., we have no annotation of zero anaphora (Ø) for Korean and Russian, and some corpora, in particular, OntoNotes (Pradhan and Xue, 2009), only provide annotations for specific referents, effectively neglecting the **type** category. No corpus we worked with has annotations of event anaphora.

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
it	0,895++	-0,373++	-0,21++	-0,311++	-0,183+	-0,212++
HE	-0,03 n/a	0,072 n/a	-0,012 n/a	-0,017 n/a	-0,01 n/a	-0,012 n/a
this	-0,119 n.s	0,281 n/a	-0,046 n/a	-0,068 n/a	-0,04 n/a	-0,046 n/a
that	-0,111+	0,286 n/a	-0,05 n/a	-0,075 n/a	-0,044 n/a	-0,051 n/a
this N	-0,082(+)	0,224 n/a	-0,041 n/a	-0,061 n/a	-0,036 n/a	-0,041 n/a
that N	-0,127(+)	0,14 n/a	0,195 n/a	-0,073 n/a	-0,043 n/a	-0,049 n/a
the N	-0,479++	0,227++	0,268++	0,514++	-0,226++	-0,262++
indef. this N	-0,03 n/a	-0,021 n/a	-0,012 n/a	-0,017 n/a	0,149 n/a	-0,012 n/a
a N	-0,479++	0,227++	0,268++	0,514++	-0,226++	-0,262++

1.0 > r > 0.5 +/++
0.5 > r > 0.1 +/++
0.1 > r > 0 +/++
n.s / (+) / n.a
0 > r > -0.1 +/++
-0.1 > r > -0.5 +/++
-0.5 > r > -1.0 +/++

n/a	χ^2 not applicable
n.s.	not significant
(+)	marginal, p<=.05
+	significant, p<=.01
++	highly significant, p<=.001

Table 1: Cognitive statuses and referential expressions in English (Gundel et al., 1993), with correlation and χ^2 significance, Gundel et al.'s absolute numbers are provided in Tab. 7

ciently substantiated by earlier empirical analyses, in particular, the relative givenness of demonstratives in comparison to pronouns and definite NPs.

2 Experimental Setup

We provide an operationalization of the Givenness Hierarchy on the basis of the original annotation guidelines (Gundel et al., 2006). These instruct annotators to assign the highest applicable cognitive category according to the following overview. In addition to the original instructions, we indicate whether a criterion is directly implementable (✓), can be heuristically approximated ((✓)), or not operationalizable (?). We also provide approximation criteria, but only operate with those that do not introduce dependencies from surface forms. These are marked by (✓*).

Given a referring expression e and referent r:

- annotate **in focus** if
  - ✓ r is subject of the preceding utterance (∼ `nsubj`)
  - ✓ r mentioned earlier in same utterance
  - ✓ r mentioned in both of the two previous utterances
  - (✓) r is the event of the preceding utterance (∼ neuter weak pronouns without antecedent)
  - ? r is a discourse topic inferred but not overtly mentioned
- **activated** (if not **in focus** and):
  - ✓ r is mentioned in one of the two previous utterances
  - (✓*) r evoked by gesture or gaze (n/a, we operate with written text)
  - ? r is an associated proposition or speech act
- **familiar** (if not **activated** or **in focus** and):
  - ✓ r previously mentioned
  - ? r known from shared background
- **unique** (if not **familiar**, etc):
  - (✓*) e contains sufficient lexical material to create a unique referent (∼ by the use of more than 3 content words)
  - (✓) r linked via lexical association to activated referent (∼ possessive pronouns)
- **referential** (if not **unique**, etc.):
  - ✓ r mentioned later in discourse
  - (✓) r linguistically marked for discourse prominence
- **type** (if not **referential**, etc.):
  - (✓*) e encodes interpretable conceptual content (∼ anything subject to coreference annotation)

Out of 15 criteria, 6 are directly implemented, 6 can be approximated, and 3 not covered. By including surface criteria, we can cover up to 80% of the original protocol – although, here, we decided to remain agnostic about surface forms to avoid circular reasoning and operate only with (approxiations for) 9 criteria (60%). While this introduces some noise, we assume it will not preclude meaningful generalizations if statistically significant patterns emerge. Additional design decisions include equating 'utterance' with sentences (based on provided parse or produced by a parser), and the definition of 'mentioned in' as 'having an anaphor/antecedent annotated in'. For pro-drop languages without Ø annotation (Russian, Korean), this leads to an underrepresentation of **in focus** and **activated**. As none of our corpora systematically annotates event anaphora; such cases may be wrongly treated as discourse-new.

Aside from Gundel et al. (2006), there are alternative GH operationalizations that reflect language-specific needs or annotation trade-offs,[6] but mostly

[6]Alternative operationalizations of the GH include Henschel et al. (2000), who redefine **in focus** as subject of last

represent simplifications. We operate with Gundel et al. (2006) to order to follow Gundel et al.'s original six-way distinction.

3 Empirical Evaluation

Based on corpora with coreference annotation, and using the heuristics described above, we compute Pearson's r and assess correlation significance using the χ^2 test for each pairing of cognitive status and the type of referring expression. This section provides aggregate results, with full data in the Appendix.

Table 2 gives an overview over the corpora considered, using OntoNotes v5.0 (Pradhan and Xue, 2009), NTC 1.5 (Iida et al., 2017), KoCoNovel (Kim et al., 2024), and the CorefUD 1.3 (Nedoluzhko et al., 2022) editions of AnCora (Recasens and Martí, 2010), ECMT, GUM (Zeldes, 2017), LitBank (Dyer et al., 2024), ParCorFull (Lapshinova-Koltunski et al., 2018), and RuCor (Ju et al., 2014).[7] These corpora vary in scope, genre, and annotation practices. Not all provide full coreference coverage – OntoNotes only annotates specific, referential entities, while KoCoNovel restricts to protagonists, omitting inanimates. Event anaphora are generally excluded. Because of the resulting noise, it is thus important to interpret them in conjunction with data from Gundel et al. (1990) and later papers that is smaller in scale but produced under more controlled conditions. As for CorefUD corpora, we use the existing UD annotations for classifying referring expressions and extracting grammatical features (esp. `nsubj`), for other corpora, we operate with automated parses obtained from spaCy, resp., for Arabic, spaCy-udpipe (i.e., UDpipe 2.5). We operate with sentence boundaries as provided and use those predicted by the parsers where these are lacking.

We distinguish the following types of referring expressions (cf. Tab. 6 for other languages) :

Ø zero anaphor (if annotated)
pron third-person pronoun (e.g., *it*)
dem.prox proximal dem. pronoun (e.g., *this*)
dem.med medial dem. pronoun (e.g., Span. *ese*)
dem.dist distal dem. pronoun (e.g., *that*)
dem... N demonstrative NP (e.g., *this house*)
def N definite NP (e.g., *the house*)
Ø N bare NP (e.g., Russian *dom*)
ind N indefinite NP (e.g., *a house*)

These categories are language-specific and functionally not always equivalent. For instance, medial demonstratives exist only in Spanish, Japanese, and Korean. Zero anaphors vary in distribution and constraints; e.g., Spanish allows them for subjects, only, but Japanese also for objects. The category `Ø N` also has different functions languages, depending on (the lack of) a grammatical opposition with indefinite or definite NPs.

The Givenness Hierarchy postulates three principles to account for deviations from the expected associations between forms and cognitive statuses: (1) all statuses can be expressed with forms for lower statuses (implicative hierarchy), (2) deviations are used to trigger Gricean quantity implicatures (and thus, rarer than non-deviations), and (3) these deviations are monodirectional (downward only). Statistically, we thus expect positive correlations between statuses and their prototypical form, absence of low statuses encoded with higher-status forms, and noise from our heuristic-based status approximations, with a possible overrepresentation of **type** (and, possibly, **referential**) for actual cases of event anaphora. Tables 3 and 4 summarize our results with aggregate correlation data for English and other languages, respectively. Overall, the reported correlations are statistically significant, but low, at times, reflecting both imperfections in the annotation-based cognitive status approximations and noise in the data.

In all English corpora (Tab. 3), `pron` correlates positively with **in focus** and **activated**, and negatively with lower statuses, consistent with Gundel et al. Demonstrative pronouns correlate positively with **activated** and negatively with **in focus**, supporting their distinct status from personal pronouns. Unexpected positive correlations with **referential** and **type** (in OntoNotes) may result from event anaphora.

Proximal demonstrative NPs exhibit negative

utterance plus *all* discourse-old entities, with an accompanying salience ranking based on Grosz et al. (1995). Although Gundel (1998) also explore this link, they reject scalar models, preferring categorical distinctions. Henschel et al.'s simplification thus collapses several GH categories into the equivalent of **familiar**. Another simplification is Traugott et al.'s (2008) use of only three GH statuses (**familiar**, **unique**, **referential**) for Old English. For Spanish, (Blackwell and Quesada, 2012) merged **referential** and **type**, and further distinguish between recoverable and non-recoverable **activated** referents. However, it is unclear how to technically operationalize their notion of recoverability on the basis of coreference annotations.

[7] Note that for Arabic, OntoNotes reflects Modern Standard Arabic, which may differ from the spoken varieties used by (Gundel et al., 1990, undocumented variety) and (Gundel et al., 2010, Tunesian Arabic).

	OntoNotes	LitBank	GUM	AnCora	NTC	KoCoNovel	ECMT	RuCor
version	5.0	CU 1.3	CU 1.3	CU 1.3	1.5	—	CU 1.3	CU 1.3
language	ar / en / zh	en	en	es	ja	ko	ko	ru
modality	written	written	written/spoken	written	written	written	written	written
genre	news, web, lit	literature	diverse	news	news	literature	news	diverse
tokens (K)	325 / 1,750 / 235	190	170	429	1,000	165	439	145

Table 2: Coreference corpora considered (CU = CorefUD)

correlation with **in focus** and the expected positive correlation with **activated**, but also with **familiar**, contrary to Gundel et al. This suggests a possible reclassification aligning them with **familiar**. Isolated positive correlations with **referential** (in OntoNotes) are in line with predictions for indefinite *this*, but may be due to the incomplete nature of coreference annotation. Distal demonstrative NPs correlate positively with **activated** and **unique** in OntoNotes, and with **referential** and **type** in LitBank. Their inconsistent behavior suggests Gundel et al.'s hierarchy may not fully explain their use. Other functions appear to be likely, e.g., the use of distal demonstratives in comparisons with referents referred to with proximal demonstratives. Overall, proximal forms are more frequent (e.g., OntoNotes: 3743 vs. 1904; GUM: 982 vs. 374), possibly due to their broader use contexts, and these (but only these), seem to adhere to the expected distribution.

Definite NPs correlate strongly with **familiar** and **unique**. Correlations with lower statuses may reflect limitations of the approximation of **unique** by lexical richness. Still, they tend to encode lower statuses than demonstratives. As for indefinite NPs, these are negatively correlated with previous mention (**familiar** or higher) and positively correlated with the lack thereof (**unique** or lower). Again, the differentiation between **unique** and lower statuses may be insufficient to delineate the narrower scope of indefinite NPs.

As for the cognitive statuses themselves, we see good evidence for **in focus** (positively correlated with the use of third-person pronouns, and only these) and **activated**) (positively correlated with demonstrative pronouns. and only these), as well as for the differentiation between **familiar** (negatively correlated with indefinite NPs) and lower statuses (positively correlated with indefinite NPs). The **unique** status can probably not be approximated from coreference annotations that would be sufficiently reliable to be used in a meaningful way in this evaluation. This is different for **referential**, which can be easily identified from coreference annotation. Yet, in the corpora of (mostly) written language considered here, there are no referring expressions that seem to require a differentiation between **referential** and **type**.

For other languages (Tab. 4), pronouns and Ø correlate with **in focus**, as predicted. The positive correlations between Ø and **activated** in Japanese and Chinese may be due to the fact that these languages do not limit zero anaphors to subject antecedents and can have more than one Ø as argument. The positive correlations between third-person pronouns and ACTIVATED for Russian, Japanese and Chinese may reflect that in these languages, pronouns can be more easily replaced by zero anaphors, so that overt pronouns are more likely to take on characteristics of stressed pronouns ... that Gundel et al. (1993) associate with **activated**. Demonstrative pronouns correlate with **activated** but not with **in focus** or with lower statuses, with the possible exception of Korean. Some **type** correlations may stem from exophoric or event-based reference, especially in dialogue.

Similar to English, demonstrative NPs appear heterogeneous and hard to interpret. Their negative correlation with **in focus** is in line with Gundel et al., and they seem to be associated with cognitive statuses at the same level or below demonstrative pronouns. Noteworthy is the systematic association between proximal demonstrative NPs and **activated**, which is predicted by Gundel et al. Indeed, English seems to be exceptional in this regard in extending the scope of *this*-NPs to **familiar**. For medial and distal demonstrative NPs, there seems to be no clear positive correlation with any cognitive status. This, again, indicates that the functions of demonstrative NPs (other than proximal demonstrative NPs) may involve other functions than givenness marking. Definite NPs correlate with previous mention (i.e., **familiar**, for Spanish), but aside from their negative correlation with **in focus**, they can be used for any status at the level of **unique** or higher. That we also find correlations with **referential** and **type** may be due to the insufficiencies of our approximation of **unique**, as this differentiation, indeed, was already statistically

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	4:0	2:1	0:4	0:3	0:4	0:2
dem.prox	0:1	2:0		0:2	1:1	0:1
dem.dist	0:1	2:0	0:1	0:2	0:1	1:1
dem.prox N	0:4	3:0	3:0	0:1	1:1	0:1
dem.dist N	0:2	2:0	0:1	1:0		
the N	0:4	1:2	4:0	3:0	4:0	2:0
a N	0:4	0:3	0:3	3:0	2:0	2:0

4:0
3:0
2:0, 3:1
1:0, 2:1
1:1
0:1, 1:2
0:2, 1:3
0:3
0:4

Table 3: Aggregate correlations for four English corpora; green indicates significant positive correlations, red negative ($p \leq 0.01$).

	IN FOCUS		ACTIVATED		FAMILIAR		UNIQUE		REFERENTIAL		TYPE
Ø	es,ja, zh,ar		ja,zh	es,ar		ar	es,ja, zh,ar		es,ja, zh,ar		es,ja
pron	ru,es,ja, zh,ar		ru,ja,zh	es,ar	ru,es,zh, ar (ja)		ru,es, zh,ar		ru,es, zh,ar		
dem.prox	ko1	ja	es,ja,zh		ko2,zh	es,ko1	es,zh		es,ko1, zh		ko
dem.med	ko	ja	es,ko,ja		ko1	es, ko2 (ja)			es,ko		ko
dem.dist	ko2	zh	zh								es
dem.prox N	ko2	es, ja,zh	ru,es, ko1,ja,zh		zh	ko1 (ja)	es	zh	ru,es, ko1,zh		
dem.med N		ko1,ja		ko1	(ja)	es,ko1	es		es,ko1		ko1
dem.dist N		zh	zh		zh				ru		
the N		es,ar	es	ar	es		es	ar	es	ar	es
N		ru,ko, ja,zh		ru,ko, ja,zh	ru,ko2, zh (ja)		ru, ja,zh	ko1	ru,ko, ja,zh		ko, ja
a N		es,zh, ar		es,zh, ar		es,zh,ar	zh,ar		es,zh,ar		

ar	Arabic (OntoNotes)
es	Spanish (AnCora)
ja	Japanese (NTC)
ko1	Korean (ECMT only)
ko2	Korean (KoCoNovel only)
ko	Korean (ECMT=KoCoNovel)
ru	Russian (RuCor)
zh	Chinese (OntoNotes)

4:0, 5:0
3:0
2:0, 3:1
1:0, 2:1, 3:2
1:1, 2:2
0:1, 1:2, 2:3
0:2, 1:3
0:3
0:4, 0:5

Table 4: Aggregate correlations for Arabic, Chinese, Japanese, Korean, Russian, and Spanish. Significant correlations ($p \leq 0.01$): positive on the left, negative on the right of each status cell.

significant in Gundel et al.'s original data. As for Ø NPs, these are negatively correlated with **in focus** and **activated**, and positively with **familiar** and lower statuses. This underlines the relevance of the differentiation between **activated** and **familiar**, but also that Ø NPs cover functions otherwise adopted by definite and indefinite NPs and are thus applicable to *any* cognitive status, if the conditions are met. The correlations of indefinite NPs underline the importance to differentiating previous mention (**familiar**) and the lack thereof (**unique** or lower).

As for the evaluation of cognitive statuses, the distribution differences support the differentiation of **in focus** (for Ø and pron), **activated** (for stressed pronouns and demonstrative pronouns), **familiar** (for Ø NPs in languages without grammaticalized determiners) and statuses lower than **familiar** (for indefinite NPs). Again, **unique** cannot be reliably differentiated from lower statuses with the heuristics adopted here, but whereas **referential** can, it does not seem to be necessary to account for any of the major classes of referring expressions.

Similarly as for the case of English, we thus find that pronouns mark higher givenness than demonstratives, demonstrative pronouns and *proximal* demonstrative NPs rank above definites and bare NPs. Medial/distal demonstrative NPs resist straightforward classification and may involve other pragmatic functions beyond those captured by the Givenness Hierarchy.

4 Consolidation, Inference and Revision

We would like to combine our findings with those of Gundel et al. (1993) – who report statistically significant differences between **unique** and **referential** when accounting for definite NPs –, and

overall suggest to reconsider the status of **referential**. While absent from the original proposal (Gundel et al., 1990), this was introduced by Gundel et al. (1993) specifically to account for indefinite *this* in English, but have failed to demonstrate its (statistical) significance, and neither do our correlations call for a differentiation between **referential** and **type identifiable**.

For this reduced hierarchy of *five* cognitive statuses, we find robust evidence both in our data and Gundel et al.'s: correlations between **in focus** and pronouns/Ø; between **activated** and demonstratives/proximal demonstrative NPs; between **familiar** and the avoidance of indefinite NPs; between **unique** and definite NPs (per Gundel et al.); and between lower statuses and indefinite NPs. However, medial and distal demonstrative NPs remain difficult to classify: neither our data nor Gundel et al.'s show significant correlations, suggesting these are sensitive to factors other than givenness.

Further, the approach enables bootstrapping Givenness Hierarchies in other languages or for phenomena that Gundel et al. did not originally consider. For instance, Mulkern (2011) investigates the referential properties of proper names, distinguishing between full names (e.g., given name plus surname) and single names (e.g., family name alone or nicknames). She suggests that full names align with **unique** and single names with at least **familiar**.

Using the correlation analysis described above, we can now verify these claims in an empirical quantitative manner. We approximate the notion of single name by single token proper names, full names by multi-token proper names and perform the evaluation against the OntoNotes corpus, as it is by far the largest corpus in our sample. The results (Tab. 5) confirm, indeed, that proper names are associated with the middle segment of the Givenness Hierarchy (thus negative correlation with **in focus**), and moreover, that short names, or, at least, single-token names, tend to be associated with higher cognitive statuses than full, resp., multi-token names, as these differ in their correlation with **activated**. As mentioned before, the approximation of **unique** from coreference annotations is insufficient, but on conceptual grounds – as pointed out by Mulkern –, any element that a hearer can recognize as a proper name is by definition **unique**. An interesting observation is, however, that the cognitive statuses that full and short names seem to be associated with are not **familiar** and **unique**, as postulated by Mulkern, but, rather, **activated** and **unique**. However, this may be an artifact of the heuristic approximation of single names by single tokens and full names by multi-word expressions, as a considerable number of single tokens will indeed just represent the complete name of, say, a country, and these might behave differently from person names.

We can confirm the general pattern of short names associated with higher givenness and long names associated with lower givenness, but it also seems that further differentiations within the larger group of proper nouns are likely, with their own alignments with cognitive statuses, and that these align with, but complement the distinction of different kinds of proper names studied by Mulkern. With this in mind, future research may now explore more fine-grained distinctions of referring expressions in an empirical fashion, and, potentially, revise the Givenness Hierarchy.

For future studies of the Givenness Hierarchy with coreference-annotated corpora, we suggest a to operate with a simplified model where the current **referential** category is abandoned. However, unlike (Traugott and Pintzuk, 2008), (Blackwell and Quesada, 2012), and (Abisambra Miccheli and Quesada, 2023), we do not suggest to merge it with **type**, but, instead with **unique**, as, according to Gundel et al. (1993), it is a superset of **unique**, and it provides clear, verifiable criteria for its distinction from **type**: **referential** can be inferred from subsequent anaphora. To avoid ambiguity, we propose renaming this unified category to **XREF** (extended referential). While our data could not distinguish **referential** from **type** conclusively, future research may uncover such distinctions for **XREF**. With truly **unique** referents accumulating in this category, we would expect that some of the effects (Gundel et al., 1993) and later studies found for **unique** are detectable in this broader category.

5 Results and Perspectives

We describe the empirical verification of the Givenness Hierarchy (Gundel et al. 1990, 1993) using coreference-annotated corpora for the languages for which this theory has been originally formulated. Unlike the original, small-scale manual annotations, our approach relies on publicly available corpora with coreference annotations, allowing for statistically significant analyses.

For English and across languages, we confirmed strong associations between **in focus** and the use

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
single-token names							
frequency	5433	6839	6444	n/a	4851	5	23572
correlation	-0,119+++	0,034+++	0,126+++		0,11+++	-0,004 n.s	
multi-token names							
frequency	3582	5006	5531	4389	4765	3	23276
correlation	-0,188+++	-0,042+++	0,083+++	0,147+++	0,108+++	-0,006(+)	
other							
frequency	41882	24654	11895	8758	8058	48	95295
total	50897	36499	23870	13147	17674	56	142143

Table 5: Distribution of names and approximative cognitive statuses in the OntoNotes corpus compared to other referring expressions. Colors are used in accordance with Tab. 1.

of pronouns and zero anaphora, whereas all other types of referring expressions are negatively correlated with **in focus**. Demonstrative pronouns are associated with locally evoked (**activated**) referents; demonstrative NPs showed similar trends, but medial and distal forms lacked consistent associations with cognitive statuses. This suggests that such forms serve specialized discourse functions (e.g., deixis, exophora, contrast) beyond simple givenness marking and should be analyzed separately from the hierarchy. Definite NPs (or Ø NPs in languages lacking definite determiners) tend to occupy the middle-to-lower part of the hierarchy. However, the inability to reliably distinguish **unique** from **referential** and **type** using coreference data makes it difficult to confirm whether definite NPs require **unique** status. While Gundel et al. confirmed this link experimentally, it cannot be directly replicated using anaphoric annotation alone. At the lower end of the hierarchy, indefinite NPs (or Ø NPs in relevant languages) dominate. The absence of previous mention (the primary criterion for **familiar**) helps distinguish them from higher statuses.

Aside from issues with **unique** and medial/distal demonstrative NPs, we confirm GH predictions for Arabic, Chinese, English, Japanese, Korean, Russian, and Spanish. Despite variation in corpus formats, genres, and annotation schemes, and a considerable noise arising from the incomplete nature of coreference annotations in comparison to Gundel et al. (2006), we observed correlations sufficiently strong to yield significant results also for aspects of the Givenness Hierarchy previously described with insufficient amounts of data, only.

Because of difficulties surrounding the approximation of **unique**, we recommend simplifying the GH by collapsing it with **referential** into a novel combined category: **XREF** (extended referential), encompassing both **unique** and **referential** referents as defined by the original GH manual, with the primary criterion for their identification drawn from the GH definition of **referential**. With **XREF**, future studies may better capture the transitional space between **familiar** (hearer-old) and **type identifiable** ($\sim$ hearer-new).

Beyond evaluating the Givenness Hierarchy, we showed how phenomena not originally covered by the Givenness Hierarchy (the givenness of full and short names) can be investigated with this methodology, leading to insights consistent with previous qualitative analyses (Mulkern, 2011). This methodology can thus *infer and extend* language-specific givenness hierarchies. Also, our ability to approximate givenness from coreference annotations is practically relevant: The procedure introduced in Section 2 can bootstrap givenness annotations and thus yield the first available training data for GH annotations. Such data can support the development of automated taggers and serve as evaluation material for future methods of inducing or predicting givenness – potentially even in the absence of coreference annotation. Note that labelling cognitive statuses appears simpler and more robust than full anaphora resolution, while still providing valuable discourse-level insights. Thus, extrapolating GH annotations from coreference annotation may serve as a useful intermediary task for applications requiring discourse-aware processing.

We would like to emphasize that – at this stage – we do not aim to evaluate the theory per se. Although many of the categorizations put forward by Gundel et al. seem plausibile from a cognitive-linguistic perspective, and the factors they consider (proximity, previous mention, assumed hearer knowledge, intention to refer to a specific entity) certainly play a role, it is not uncontested that information is, in fact, categorial by nature (Ariel, 1990), and how these categories are differentiated,

cf. Chiarcos (2010, 2011b,c) for overview, discussion and criticism, and (Poesio and Modjeska, 2008), (Chiarcos, 2011a) or (Hou, 2021) (among others) for possible alternatives and their operationalization. Instead, our primary goal is to demonstrate the potential for using coreference-annotated corpora for bootstrapping Givenness Hierarchies for other languages, and these, in turn, may be useful for the empirical, cross-linguistic evaluation of the theory, potentially, along with other theories of reference, accessibility and information status. We are aware that the approximations suggested in this paper are, to a large extent, imperfect, but we would argue that these heuristics nevertheless capture prototypical examples at sufficient numbers, and that statistically significant patterns observed over these allow us to gain insights into the underlying theoretical models. At this stage, we thus conclude that the theory is to a large extent verifiable with coreference-annotated corpora, that there are limitations in the heuristic identification of UNIQUE referents (for which we suggest a simplification for empirical studies that is based on the implicative nature of the Givenness Hierarchy), but also that not all the claims of the Givenness Hierarchy could be confirmed, especially regarding demonstratives, that motivates more extensive research into other languages, as well, and that this research can be conducted with the bootstrapping methodology suggested here.

As a secondary contribution, we converted two corpora – Japanese NTC 1.5 and Korean KoCoNovel – into CorefUD-compatible formats. The conversion scripts and accompanying materials for these and all other corpora considered here are available under an Apache v.2 license from `https://github.com/acoli-repo/givenness-hierarchy`.

Limitations

This study presents an approximative operationalization of the Givenness Hierarchy using coreference annotations. While most cognitive statuses can be approximated reliably, the status **unique** could not be accurately derived from the available data. For this reason, we propose merging **unique** and **referential** into a single category, **XREF**, in future GH implementations.

Note that to facilitate comparability between our numbers and those of Gundel et al. (1990,1993), we limit our analysis to referring expressions studied by Gundel et al., which restricts coverage and means that totals are not calculated over the full set of referring expressions annotated in a corpus, but only to those that fall into categories also considered by Gundel et al. In particular, proper nouns, quantified nouns, pronominal adverbs and possessive NPs were not considered in Sect. 3. With the extension to proper nouns and the replication of Mulkern (2011), the totals in Sect. 4 were extended to cover anaphors and antecedents annotated as `PROPN`, as well.

Also, we are restricted to referring expressions that we can reliably identify in our data, so, while Gundel et al. (1993) distinguished stressed and unstressed pronouns (available in spoken data), the (primarily) written text we operate with does not provide such cues. Likewise, we did not disambiguate between indefinite *this* and proximal demonstratives, because these are identical in form. In particular, because of limitations in the reliable detection of **unique** referents, we could not naively identify non-**unique** *this*-NPs with 'indefinite *this*' and Gundel et al.'s **referential** category. Similarly, our handling of proper names diverges from Mulkern (2011). Whereas Mulkern seems to focus on person names, exclusively, we evaluated all referring expressions annotated as `PROPN`. Our data thus includes organizations, dates, and locations, as well.

A major limitation lies in the comparability of the corpora. They differ in genre, size, annotation design, and coverage. Crucially, none provide annotations for event anaphora, resulting in underrepresentation of certain referential types (e.g., demonstrative pronouns for events). Annotations for bridging or other models of information status, which could also inform GH annotations, are available for a subset of corpora considered and have been excluded, so that cross-linguistic comparison could be established. Several corpora are also affected by biases in their annotation. For example, OntoNotes and ParCorFull only annotate specific (i.e. **referential** or higher) referents, thus systematically excluding **type identifiable** expressions. Similarly, the KoCoNovel corpus annotates protagonists, only. None of the corpora provide annotations for event anaphora. Russian and Korean corpora lack annotations for zero anaphora (Ø), causing misclassification of referents that should be **in focus** or **activated** as lower-status categories like **familiar**. In Japanese (NTC 1.5), missing text boundaries can lead to erroneous coreference

chains. In the Chinese OntoNotes corpus, an extraction error led to misclassifying a group of proper names as bare nominals.

In order to categorize referring expressions and grammatical features in a systematic way, we rely on Universal Dependencies (UD). While such annotations were available for subset of 5 corpora drawn from the CorefUD collection, they had to be created automatically for Arabic, Chinese, Japanese, the Korean KoCoNovel and the English OntoNotes corpus. We used spaCy (resp., for Arabic, spaCy-udpipe, i.e., UDPipe 2.5) to create them automatically, introducing potential parsing errors as an additional source of noise. For identifying **in focus** referents on the basis of their realization in the preceding sentence (i.e., referents not mentioned in the penultimate sentence, as well), we rely on the UD label `nsubj`, whereas the original definition by Gundel et al. (2006) would also include morphosyntactic topic and focus markers as present (but not annotated) in Korean and Japanese.

Overall, while these limitations introduce variability, our methodology provides a scalable, corpus-based framework for exploring the Givenness Hierarchy across languages and modalities. Despite the aforementioned issues, our analysis is strengthened by recurring patterns observed across multiple languages and multiple independently annotated corpora. We report only statistically significant findings and emphasize robust patterns across datasets, rather than isolated results.

References

Ingrid Abisambra Miccheli and Margaret Lubbers Quesada. 2023. Cognitive status and subject reference in spanish written discourse. *Studies in Hispanic and Lusophone Linguistics*, 16(1):5–33.

M. Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of pragmatics*, 16(5):443–463.

Keira Gebbie Ballantyne. 2004. Givenness as a ranking criterion in centering theory: evidence from yapese. *Oceanic Linguistics*, 43(1):49–72.

Stefan Baumann and Arndt Riester. 2013. Coreference, lexical givenness and prosody in german. *Lingua*, 136:16–37.

Sarah E Blackwell and Margaret Lubbers Quesada. 2012. Third-person subjects in native speakers' and l2 learners' narratives: Testing (and revising) the givenness hierarchy for spanish. In *Selected proceedings of the 14th Hispanic linguistics symposium*, pages 142–164. Cascadilla Proceedings Project Somerville, MA.

Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Richard Challoner. 1749-1752 (revision). Douay-Rheims Bible Online. https://www.drbo.org/, accessed 2025-08-05. Official Catholic Version with Search.

Christian Chiarcos. 2010. *Mental Salience and Grammatical Form*. Ph.D. thesis, Potsdam: Universität Potsdam.

Christian Chiarcos. 2011a. Evaluating salience metrics for the context-adequate realization of discourse referents. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 32–43.

Christian Chiarcos. 2011b. The Mental Salience Framework: Context-adequate generation of referring expressions. In Christian Chiarcos, Berry Claus, and Michael Grabski, editors, *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, volume 227, pages 105–140. Walter de Gruyter.

Christian Chiarcos. 2011c. On the dimensions of discourse salience. *Bochumer Linguistische Arbeitsberichte*, 3:31–44.

Logan Daigler, Mark Higger, Terran Mott, and Tom Williams. 2024. Challenges in annotating gesture-based cognitive status in human-robot collaboration datasets. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 364–368.

John Nelson Darby. 1890. Bible in English. Darby Translation. https://kingdomjc.com/Bible_EN.htm, accessed 2025-08-05.

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska, Syahidah Asma Umniyati, and Helena Rodrigues Menezes de Oliveira Vaz. 2024. A multilingual parallel corpus for coreference resolution and information status in the literary domain. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 55–64, Hamburg,Germany. Association for Computational Linguistics.

Talmy Givón. 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study*, volume 3. John Benjamins Publishing.

Barbara J Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Jeanette Gundel, Nancy Hedberg, Ron Zacharski, Ann Mulkern, Tonya Custis, Bonnie Swierzbin, Amel Khalfoui, Linda Humnick, Bryan Gordon, Mamadou Bassene, and Shana Watters. 2006. Coding protocol for statuses on the givenness hierarchy (gundel, hedberg and zacharski 1993). Technical report, University of Minnesota.

Jeanette K Gundel. 1998. Centering theory and the givenness hierarchy: Towards a synthesis. *Centering theory in discourse*, pages 183–198.

Jeanette K Gundel, Mamadou Bassene, Bryan Gordon, Linda Humnick, and Amel Khalfaoui. 2010. Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, 42(7):1770–1785.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1990. Givenness, implicature, and the form of referring expressions. In *Annual Meeting of the Berkeley Linguistics Society*, pages 442–453.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307. Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society (1990), pp. 442-453.

Jeanette K Gundel and Kaitlin Johnson. 2013. Children's use of referring expressions in spontaneous discourse: Implications for theory of mind development. *Journal of Pragmatics*, 56:43–57.

Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th international conference on natural language generation*, pages 1–14.

Nancy Hedberg. 2010. Centering and noun phrase realization in kaqchikel mayan. *Journal of Pragmatics*, 42(7):1829–1841.

Nancy Hedberg and Nathalie Schapansky. 1996. A referentiality constraint on preverbal nps in breton. In *annual meeting of the LSA, San Diego, Jan*, pages 3–7.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Mark Higger and Tom Williams. 2024. Gaia: A givenness hierarchy theoretic model of situated referring expression generation. In *Proceedings of the annual meeting of the cognitive science society*, volume 46.

Yufang Hou. 2021. End-to-end neural information status classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1377–1388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linda Humnick. 2005. Pronominal references and the encoding of cognitive status in kumyk. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 41, pages 149–163. Chicago Linguistic Society.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2017. Naist text corpus: Annotating predicate-argument and coreference relations in japanese. In *Handbook of linguistic annotation*, pages 1177–1196. Springer.

Toldova S Ju, A Roytberg, AA Ladygina, MD Vasilyeva, IL Azerkovich, M Kurzukov, G Sim, DV Gorshkov, A Ivanova, A Nedoluzhko, and 1 others. 2014. Rueval-2014: Evaluating anaphora and coreference resolution for russian. *Computational Linguistics and Intellectual Technologies*, 13:681–694.

Alireza Khormaee and Elina Skrouchi. 2015. A study of discourse anaphora in persian language based on gundel, hedberg and zacharski's givenness hierarchy. *Language Science*, 3(4):154–111.

Hyunwoo Kim. 2016. The effects of accessibility of english reference in korean efl learners' sentence processing. *ITL-International Journal of Applied Linguistics*, 167(1):78–97.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Koconovel: Annotated dataset of character coreference in korean novels. *Preprint*, arXiv:2404.01140.

Martina Krüger. 2018. *Prosodic decoding and encoding of referential givenness in adults with autism spectrum disorders*. Ph.D. thesis, Universität zu Köln.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. Parcorfull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ann E Mulkern. 2008. Knowing who's important: Relative discourse salience and irish pronominal forms. In *The Grammar–Pragmatics Interface: Essays in honor of Jeanette K. Gundel*, pages 113–142. John Benjamins Publishing Company.

Ann E Mulkern. 2011. The game of the name. In *Reference and referent accessibility*, pages 235–250. John Benjamins Publishing Company.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Achuodho V Omondi. 2016. *Topic and Focus in Dholuo Oral Speeches: Case of Givenness Hierarchy Framework.* Ph.D. thesis, University of Nairobi.

Poulomi Pal, Grace Clark, and Tom Williams. 2021. Givenness hierarchy theoretic referential choice in situated contexts. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Massimo Poesio and Natalia N Modjeska. 2008. Focus, activation, and this-noun phrases: An empirical study. In *Anaphora processing: Linguistic, cognitive and computational modelling*, pages 429–449. John Benjamins Publishing Company.

Sameer S Pradhan and Nianwen Xue. 2009. Ontonotes: the 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44(4):315–345.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: an evaluation across different types of texts. In *LREC*.

Petr Sgall, Eva Hajicová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Kevin Spevak, Zhao Han, Tom Williams, and Neil T Dantam. 2022. Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6109–6115. IEEE.

Laurie A Swabey. 2011. Referring expressions in asl discourse. *Discourse in signed languages*, pages 96–120.

Elizabeth Closs Traugott and Susan Pintzuk. 2008. Coding the york-toronto-helsinki parsed corpus of old english prose to investigate the syntax-pragmatics interface. *Studies in the History of the English Language IV. Empirical and Analytical Advances in the Study of English Language Change. Berlin/New York: Mouton de Gruyter*, pages 61–80.

Marta Velnic. 2018. *Ditransitive structures in Croatian adult and child language: The role of animacy and givenness*. Ph.D. thesis, UiT Norges arktiske universitet.

Weijie Xu and Ming Xiang. 2021. Is there a predictability hierarchy in reference resolution? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Appendix

In this appendix, we provide the detailed list of referring expressions (Tab. 6) considered by us and Gundel et al. (1990, 1993), as well as raw frequencies and correlation analyses for the individual corpora. Table 3 aggregates over Tables 8, 9, 10 and 11 for English. Table 4 aggregates over Tab. 12 for Arabic, Tab. 13 for Chinese, Tab. 14 for Japanese, Tables 15 and 16 for Korean, Tab. 17 for Russian and Tab. 18 for Spanish. As for the color codes, we use the same schema as Tab. 1.

	Arabic	Chinese	English	Japanese	Korean	Russian	Spanish
Ø	Ø	Ø	[0]	Ø	(Ø)[1]	(Ø)	Ø
pron	*هُوَ ، هِيَ ، هُم*	*他, 他们, 她*	*he, she, it*	*彼, 彼女*	[2]	*он, она, они*	*él*
Gundel&al label	hua[3]	tā	it	kare		on	él
dem.prox	*هٰذِي ، هٰذَيْنِ*	*这, 这个*	*this, these*	*これ*	*이*	*это*	*éste*
Gundel&al label	haaðaa[3]	zhè	this	kore	i[3]	èto	éste
dem.med				*それ*	*그*		*todo ese*
Gundel&al label				sore	kū[3]		ése
dem.dist	*أُولَئِكَ ، تِلْكَ ، تِيكَ*	*那, 那个*	*that, those*	*あれ*	*저*	*то*	*todo aquel*
Gundel&al label	ðaalika[3]	nèi	that	are	cə[3]	to	aquél
dem.prox N	*-هٰؤُلاءِ المُعارِضِينَ*	*这次会议*	*these people*	*この事件*	*이런 식으로*	*этот процесс*	*este situación*
Gundel&al label	haaðaa N[3]	zhè N	this N	kono N	i N[3]	èto N	este N
dem.med N				*その責任*	*그 곳에서*		*ese día*
Gundel&al label				sono N	kū N[3]		ese N
dem.dist N	*أُولَئِكَ الحِيتانِ وَ- الدينا صورات-*	*那句话*	*that day*	*あの戦争*	*저 양반*	*тот явление*	*aquel país*
Gundel&al label	ðaalika N[3]	nèi N	that N	ano N	cə N[3]	to N	aquel N
def N	*اللَجْنَةِ*		*the company*				*el Gobierno*
Gundel&al label	al N[3]		the N				el N
Ø N[4]	Ø N	Ø N		Ø N	Ø N	Ø N	Ø N
a (one) N		*一个人*	*a year*				*uno año*
Gundel&al label		yi N	a N				un N

[0] Non-applicable categories rendered in gray
[1] Referring expressions in brackets are not covered by the annotation.
[2] We follow Gundel et al. (1990) in considering Korean 그 a proximal demonstrative, not a personal pronoun.
[3] According to Gundel et al. (1990). Unmarked labels follow Gundel et al. (1993).
[4] Only for languages for which *Ø N* was also considered by Gundel et al. (1990, 1993).

Table 6: Inventories of referring expressions considered by Gundel et al. (1990, 1993) and in this study, listing type of referring expression, original label used by Gundel et al. and corpus examples

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
it	214	1					215
HE		1					1
this		15					15
that	1	17					18
this N	1	11					12
that N		10	7				17
the N	30	95	47	108			280
indef. this N					1		1
a N					41	55	96
total	246	150	54	108	42	55	655

Table 7: Cognitive statuses and referential expressions in English as reported by Gundel et al. (1993), absolute numbers, for correlation and binary χ^2 significance see Tab. 1

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	33056	10695	1983		812	24	46570
dem.prox	325	881	199		204	1	1610
dem.dist	606	1236	157		144	6	2149
dem.prox N	688	1421	871	226	532	5	3743
dem.dist N	450	750	238	315	151		1904
the N	6217	9435	8211	5630	4888	11	34392
a N	540	236	236	2587	1327	1	4927
total	41882	24654	11895	8758	8058	48	95295

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,532+++	-0,065+++	-0,243+++	-0,311+++	-0,236+++	0,001 n.s
dem.prox	-0,063+++	0,086+++	0 n.s	-0,042+++	0,02+++	0,001 n.s
dem.dist	-0,048+++	0,11+++	-0,024+++	-0,048+++	-0,01+	0,015+++
dem.prox N	-0,104+++	0,056+++	0,066+++	-0,022+++	0,042+++	0,007(+)
dem.dist N	-0,058+++	0,044+++	0 n.s	0,036+++	-0,003 n.s	-0,003 n.s
the N	-0,392+++	0,027+++	0,259+++	0,187+++	0,155+++	-0,006 n.s
a N	-0,155+++	-0,112+++	-0,054+++	0,35+++	0,155+++	-0,003 n.s

Table 8: Referring expressions in OntoNotes 5.0, English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	3066	949	258	7	323	117	4720
dem.prox	105	162	38		72	39	416
dem.dist	117	174	42	1	35	34	403
dem.prox N	155	216	156	130	109	216	982
dem.dist N	101	76	19	54	34	90	374
the N	1046	910	1082	1169	1021	2921	8149
a N	691	130	123	731	416	1628	3719
total	5281	2617	1718	2092	2010	5045	18763

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,475+++	0,103+++	-0,074+++	-0,203+++	-0,073+++	-0,319+++
dem.prox	-0,01 n.s	0,109+++	0 n.s	-0,053+	0,032+++	-0,059+++
dem.dist	0,003 n.s	0,125+++	0,007 n.s	-0,051+++	-0,01 n.s	-0,062+++
dem.prox N	-0,065+++	0,055+++	0,055+++	0,016(+)	0,003 n.s	-0,026++
dem.dist N	-0,004 n.s	0,026++	-0,02+	0,015(+)	-0,007 n.s	-0,009 n.s
the N	-0,298+++	-0,07+++	0,125+++	0,089+++	0,051+++	0,177+++
a N	-0,106+++	-0,15+++	-0,101+++	0,134+++	0,008 n.s	0,189+++

Table 9: Referring expressions in the GUM corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	6488	1233	298		132	90	8241
dem.prox		1	2		1	1	5
dem.dist					1	1	2
dem.prox N	35	53	43	6	25	41	203
dem.dist N	43	22	16	14	17	36	148
the N	375	409	710	82	456	1329	3361
a N	110	43	41	80	225	883	1382
total	7051	1761	1110	182	857	2381	13342

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,659+++	0,066+++	-0,216+++	-0,149+++	-0,25+++	-0,556+++
dem.prox	-0,02n/a	0,004n/a	0,022n/a	-0,002n/a	0,011n/a	0,001 n/a
dem.dist	-0,013 n/a	-0,005n/a	-0,004n/a	-0,001n/a	0,022n/a	0,01n/a
dem.prox N	-0,089+++	0,047+++	0,058+++	0,017(+)	0,03++	0,008 n.s
dem.dist N	-0,05+++	0,005 n.s	0,01 n.s	0,074 n/a	0,022(+)	0,018(+)
the N	-0,485+++	-0,018(+)	0,269+++	0,054+++	0,169+++	0,329+++
a N	-0,306+++	-0,101+++	-0,066+++	0,13+++	0,137+++	0,409+++

Table 10: Referring expressions in the LitBank corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	195	66	3		1	11
dem.prox	6	9	2			1
dem.dist	12	4				
dem.prox N	3	8	1		2	
dem.dist N		1		1	2	
the N	14	17	9	1	29	1
a N		1	1	1	12	
total	230	106	16	3	46	13

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,43+++	-0,055 n.s	-0,204+++	-0,121 n/a	-0,484+++	0,069 n/a
dem.prox	-0,095 n.s	0,119(+)	0,08 n/a	-0,018 n/a	-0,075 n/a	0,03 n/a
dem.dist	0,078 n.s	-0,003 n/a	-0,04 n/a	-0,017 n/a	-0,071 n/a	-0,036 n/a
dem.prox N	-0,128+	0,135 n/a	0,032 n/a	-0,016 n/a	0,019 n/a	-0,034 n/a
dem.dist N	-0,11 n/a	-0,001 n/a	-0,02 n/a	0,283 n/a	0,122 n/a	-0,018 n/a
the N	-0,328+++	-0,017 n.s	0,208+++	0,037 n/a	0,43+++	-0,045 n/a
a N	-0,217++	-0,084 n/a	0,028 n/a	0,136 n/a	0,425 n/a	-0,035 n/a

Table 11: Referring expressions in the ParCorFull corpus (UDcoref 1.3), English

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	3651	684	99	3	53	1	4491
pron	521	34	4		7		566
dem.prox	1	1			1		3
dem.dist					1		1
dem.prox N			1		1		2
dem.dist N				3			3
the N	1780	2168	1379	1391	2714	1	9433
N	401	785	636	1780	1761		5363
total	6354	3672	2119	3177	4538	2	14499

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,506+++	-0,156+++	-0,235+++	-0,354+++	-0,435+++	0,005 n/a
pron	0,196+++	-0,09+++	-0,079+++	-0,107+++	-0,131+++	-0,002 n/a
dem.prox	-0,003 n/a	0,003 n/a	-0,006 n/a	-0,008 n/a	0,001 n/a	0 n/a
dem.dist	-0,007 n/a	-0,005 n/a	-0,003 n/a	-0,004 n/a	0,012 n/a	0 n/a
dem.prox N	-0,01 n/a	-0,007 n/a	0,012 n/a	-0,006 n/a	0,005 n/a	0 n/a
dem.dist N	-0,013 n/a	-0,008 n/a	-0,006 n/a	0,027 n/a	-0,01 n/a	0 n/a
the N	-0,686+++	-0,074+++	0 n.s	-0,236+++	-0,074+++	-0,004 n/a
N	-0,561+++	-0,188+++	-0,06+++	0,209+++	0,025+	-0,009 n/a

Table 12: Referring expressions in OntoNotes 5.0, Arabic

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	11395	3403	1221		1121		17140
pron	7509	2416	688	179	437	1	11230
dem.prox	659	576	278				1513
dem.dist	184	147	87	83	112		613
dem.prox N	719	703	351	78	209		2060
dem.dist N	78	78	63	37	52		308
N	34344	17768	14701	12512	20002	7	99334
a N	209	210	173	900	630		2122
total	55097	25301	17562	13789	22563	8	132198

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,194+++	0,007+	-0,07+++	-0,132+++	-0,108+++	-0,003 n/a
pron	0,156+++	0,018+++	-0,064+++	-0,088+++	-0,107+++	0,001 n/a
dem.prox	0,004 n.s	0,052+++	0,016+++	-0,037+++	-0,049+++	-0,001 n/a
dem.dist	-0,016+++	0,008+	0,002 n.s	0,007(+)	0,002 n.s	-0,001 n/a
dem.prox N	-0,017+++	0,048+++	0,014+++	-0,027+++	-0,023+++	-0,001 n/a
dem.dist N	-0,016+++	0,008+	0,01++	0,003 n.s	0 n.s	0 n/a
N	-0,25+++	-0,055+++	0,078+++	0,123+++	0,142+++	0,002 n/a
a N	-0,082+++	-0,03+++	-0,019+++	0,134+++	0,043+++	-0,001 n/a

Table 13: Referring expressions in OntoNotes 5.0, Chinese

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	25516	7240	4790		99	1	37646
pron	42	36	15			1	94
dem.prox	72	213	414		10	2	711
dem.med	121	178	406		20	5	730
dem.dist		6	1				7
dem.prox N	150	367	557	1	10	3	1088
dem.med N	69	108	578		18	5	778
dem.dist N	2	5	19		1		27
N	12600	13951	100305	138	3034	927	130955
total	38572	22104	107085	139	3192	944	172036

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,576+++	0,101+++	-0,541+++	-0,015+	-0,062+++	-0,039+++
pron	0,012+++	0,018+++	-0,022+++	-0,001 n/a	-0,003 n/a	0,002 n/a
dem.prox	-0,019+++	0,033+++	-0,005(+)	-0,002 n/a	-0,002 n.s	-0,002 n/a
dem.med	-0,009++	0,023+++	-0,009++	-0,002 n/a	0,004 n.s	0,001 n/a
dem.dist	-0,003 n/a	0,014 n/a	-0,006 n/a	0 n/a	-0,001 n/a	0 n/a
dem.prox N	-0,017+++	0,05+++	-0,018+++	0 n/a	-0,006(+)	-0,003 n.s
dem.med N	-0,022+++	0,002 n.s	0,017+++	-0,002 n/a	0,002 n.s	0,001 n/a
dem.dist N	-0,005 n.s	0,002 n/a	0,002 n.s	0 n/a	0,002 n/a	-0,001 n/a
N	-0,548+++	-0,117+++	0,528+++	0,015+++	0,061+++	0,038+++

Table 14: Referring expressions in NTC 1.5, Japanese

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
dem.prox	106	58	21			565	750
dem.med	1391	897	365			312	2965
dem.dist			1			1	2
dem.prox N	68	87	23	18		336	532
dem.med N	34	21	8	30	1	376	470
dem.dist N						1	1
N	3465	3396	3385	178	4782	25642	40848
subtotal	5064	4459	3803	226	4783	27233	45568

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
dem.prox	0,012+	-0,009 n.s	-0,026+++	-0,009 n/a	-0,044+	0,041+++
dem.med	0,301+++	0,182+++	0,038+++	-0,019 n.s	-0,09+++	-0,265+++
dem.dist	-0,002 n/a	-0,002 n/a	0,01 n/a	0 n/a	-0,002 n/a	-0,001 n/a
dem.prox N	0,006 n.s	0,024+++	-0,016++	0,045 n/a	-0,037+	0,008 n.s
dem.med N	-0,013+	-0,018+++	-0,025+++	0,086 n/a	-0,034+++	0,042+++
dem.dist N	-0,002 n/a	-0,002 n/a	-0,001 n/a	0 n/a	-0,002 n/a	0,004 n/a
N	-0,246+++	-0,146+++	-0,006 n.s	-0,025+++	0,116+++	0,181+++

Table 15: Referring expressions in the ECMT corpus (UDcoref 1.3), Korean

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
dem.prox		4	10		4	6	24
dem.med	631	710	255		167	77	1840
dem.dist	42	21	12		10	1	86
dem.prox N	34	19	12		8	7	80
dem.med N	43	32	22		20	15	132
dem.dist N	5	6	2		2	2	17
N	1812	2391	1670	1	1357	825	8056
total	2567	3183	1983	1	1568	933	10235

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
dem.prox	-0,028 n.s	-0,015 n.s	0,027+	0 n/a	0,002 n.s	0,027+
dem.med	0,103+++	0,076+++	-0,066+++	-0,005 n/a	-0,082+++	-0,081+++
dem.dist	0,05+++	-0,013 n.s	-0,013 n.s	-0,001n/a	-0,009 n.s	-0,025(+)
dem.prox N	0,036++	-0,014 n.s	-0,01 n.s	-0,001 n/a	-0,013 n.s	-0,001 n.s
dem.med N	0,02(+)	-0,017 n.s	-0,008 n.s	-0,001 n/a	-0,001 n.s	0,009 n.s
dem.dist N	0,004 n/a	0,004 n.s	-0,008 n/a	0 n/a	-0,004 n/a	0,004 n/a
N	-0,108+++	-0,062+++	0,065+++	0,005 n/a	0,081+++	0,072+++

Table 16: Referring expressions in the KoCoNovel corpus, Korean

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
pron	1253	543	71		30	13	1910
dem.prox	2	1			1		4
dem.dist							
dem.prox N	86	129	39		42	7	303
dem.dist N		6	5		30		41
N	848	950	1062	26	1825	60	4771
total	2189	1629	1177	26	1928	80	7029

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
pron	0,455+++	0,076+++	-0,213+++	-0,037 n.s	-0,354+++	-0,026(+)
dem.prox	0,01 n/a	0,001 n/a	-0,011 n/a	-0,001 n/a	-0,001 n/a	-0,003 n/a
dem.prox N	-0,013 n.s	0,098+++	-0,022 n.s	-0,013 n.s	-0,065+++	0,023 n/a
dem.dist N	-0,052(+)	-0,016 n.s	-0,009 n.s	-0,005 n/a	0,079+++	-0,008 n/a
N	-0,42+++	-0,112+++	0,215+++	0,042++	0,353+++	0,016 n.s

Table 17: Referring expressions in RuCor (UDcoref 1.3), Russian

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE	total
Ø	6098	851	197	8	215	24	7393
pron	1536	313	42	25	129	14	2059
dem.prox	100	79	10	2	20	3	214
dem.med	65	72	4	1	5		147
dem.dist	4	3	1		2	1	11
dem.prox N	471	444	153	88	105	14	1275
dem.med N	214	181	32	39	22	3	491
dem.dist N	23	7	2	5	4		41
the N	5503	4256	3098	1285	4691	159	18992
a N	270	85	64	37	328	8	792
total	14014	6206	3539	1453	5193	218	30623

	IN FOCUS	ACTIVATED	FAMILIAR	UNIQUE	REFERENTIAL	TYPE
Ø	0,416+++	-0,123+++	-0,157+++	-0,123+++	-0,211+++	-0,026+++
pron	0,155+++	-0,034+++	-0,08+++	-0,045+++	-0,077+++	-0,001 n.s
dem.prox	0,002 n/a	0,035+++	-0,018++	-0,015++	-0,017++	0,007 n/a
dem.med	-0,002 n.s	0,05+++	-0,019++	-0,013(+)	-0,025+++	-0,006 n/a
dem.dist	0,019 n.s	0,003 n/a	-0,001 n/a	-0,004 n/a	0,001 n/a	0,019++
dem.prox N	-0,037+++	0,075+++	0,003 n.s	0,021++	-0,048+++	0,01 n.s
dem.med N	-0,006 n.s	0,053+++	-0,02++	0,019++	-0,042+++	-0,002 n/a
dem.dist N	0,008 n.s	-0,003 n.s	-0,008 n.s	0,013(+)	-0,007 n.s	-0,003 n.s
the N	-0,431+++	0,068+++	0,19+++	0,121+++	0,264+++	0,019++
a N	-0,038+++	-0,039+++	-0,018+	-0,001 n.s	0,106+++	0,006 n.s

Table 18: Referring expressions in AnCora (UDcoref 1.3), Spanish

Mention detection with LLMs in pair-programming dialogue

Cecilia Domingo, Paul Piwek, Michel Wermelinger
The Open University
Milton Keynes, England
cecilia.domingo-merino, paul.piwek, and michel.wermelinger (@open.ac.uk)
and **Svetlana Stoyanchev**
Toshiba Europe
Cambridge, England
svetlana.stoyanchev@toshiba.eu

Abstract

We tackle the task of mention detection for pair-programming dialogue, a setting which adds several challenges to the task due to the characteristics of natural dialogue, the dynamic environment of the dialogue task, and the domain-specific vocabulary and structures. We compare recent variants of the Llama and GPT families and explore different prompt and context engineering approaches. While aspects like hesitations and references to read-out code and variable names made the task challenging, GPT 4.1 approximated human performance when we provided few-shot examples similar to the inference text and corrected formatting errors.

1 Introduction

Pair programming is a collaboration technique which has received a lot of scholarly attention due to the numerous benefits it can lead to, such as improved confidence and code quality (Hawlitschek et al., 2023). It involves two programmers working together on the same piece of code. The setting may vary (e.g., co-located or distributed pair programming); the pair dynamics may also vary: e.g., scholars mostly observe a navigator and a driver role, but these may switch variably during the session, and some scholars also observe different roles (Hanks et al., 2011). However, one aspect remains constant: dialogue drives the task. Dialogue complicates NLP tasks by introducing new challenges not found in the more traditionally studied written genres, and the idiosyncrasies of pair-programming dialogue further add to those challenges.

Below we present a short excerpt from our dataset to illustrate the type of dialogue that we are working with. In this excerpt, we can see some general characteristics of dialogue, such as hesitations (e.g., the repetition of determiners on the first line or the numerous filler sounds on the last line) or incomplete sentences (e.g., the turn in line 2 ends abruptly). We also observe some characteristics more unique to our type of setting, such as the use of domain terminology (e.g., here 'a string' is not thin rope) and references to unrealised entities (e.g., the speakers keep mentioning a string but they only type it into the code with the name 'text' on the fifth turn. This is frequent in this type of dialogue because the collaborative setting makes it necessary to discuss ideas with one's partner before deciding what to put into practice.).

A: Can we, uh, I don't know, define a, a string, maybe the, the so-cool string.
B: Uh... Yeah, that seems like a good place to start. And then we can kind of maybe try and split it up into the.
A: Yeah. Yeah. So should I start defining these, this string?
B: Yeah, sure. Sounds good.
A: Um. Uh, how should I, uh, call it? Uh... Just. Um, sentence. [B types 'text'; the name 'sentence' is discarded and entity becomes realised as 'text'] Oh, text. Yeah, text

In this work we focus on mention detection, the basic pillar of work on reference (e.g., this importance has been described in terms such as 'The performance of mention detection is to this day one of the most important factors in anaphora resolution' (Poesio et al., 2023, p. 571)). In simple terms, it consists on extracting all text spans that refer to some entity in the world, be it physical or abstract, or a broader element of the discourse in the case of discourse deixis. We use 'entity' to mean anything that exists, whether concrete or abstract; thus, mentions will always refer to an entity, and sometimes they may also be linked to other mentions if they all refer to the same entity. A mention that refers to an entity that is only referred to once in the discourse is called a singleton mention. Although the basic definition of the mention detection task is rather simple, researchers often differ in the

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 42–54
Suzhou, China, November 9, 2025.

specification of the concrete types of mentions (and references that these make) that they consider (e.g., see Zeldes (2022) for a critique of the commonly used OntoNotes schema, which includes its omission of singletons, predication, generic mentions, and nested mentions). While a full discussion of our annotation scheme and process is beyond the scope of the current paper, we do provide some key details in Section 3.1. In relation to the topic of what is considered a mention, we shall note that we included singletons (single mentions to an entity not mentioned any other time), predication (mentions equated to each other through a copula verb, like in 'Sahil is a lecturer'), generic mentions (e.g., 'Good children eat their vegetables and then Santa brings presents to good children'), and nested mentions (e.g., 'the book on the table' would be labelled as the book on the table and the table). We did however not include bridging anaphora (Clark, 1975) in our annotations[1], as we considered it much different from other types of anaphora — e.g., it is distinct enough to warrant a separate task in the CODI-CRAC task (Khosla et al., 2021; Li et al., 2021). Nonetheless, at the mention detection stage, the anaphors in bridging anaphora are still considered mentions; we simply consider them first mentions at the coreference resolution stage, instead of linking them to an existing antecedent through bridging anaphora. As we are working with LLMs, we shall rely of their vast training data to supply the schemata needed to interpret bridging anaphora — we however encourage further work in this area upon the release of our dataset[2]. With regard to discourse deixis, this is also considered a sufficiently distinct type of anaphora resolution to warrant its own task (Khosla et al., 2021; Li et al., 2021). As such, we did not include any discourse antecedents in our annotation of mentions, though for our later annotation of coreference we did add them separately after noting that discourse deixis was too frequent not to be included in the interpretation of references in our data.

[1]In bridging anaphora, the anaphor (i.e., the mention) is linked to a referent that it is not equivalent to, but from which it is inferred through shared common ground. For example, in 'I went to a Spanish restaurant. The waiter was from Cuenca', *the waiter* would be linked to *a Spanish restaurant* via bridging — it is the waiter's first appearance in the discourse, but we could already infer his existence from our knowledge of restaurants.

[2]Due to ongoing work, we are currently unable to release the dataset, but have scheduled its release for the beginning of 2026. Data will only made available upon request to avoid data contamination.

In order to analyse the characteristics of pair-programming dialogue and observe how they may impact NLP tasks related to reference, we collected and annotated a dataset of pair-programming sessions. We describe the collection and annotation procedure in Section 3.1. An analysis of our data confirmed the relevance of references and shed light on their characteristics in this domain. We then used this data to experiment with state-of-the art LLMs and measure their performance on this task, paying special attention to how the observed characteristics of pair-programming dialogue impact it. We describe our experimental methodology in Section 3.2, and then present and discuss our results in Sections 4 and 5. Our work is motivated by the ulterior goal of facilitating the development of AI agents that can act as pair-programming partners in an educational setting when no suitable human partner is available for the student to benefit from this practice, as suggested in the work of Robe and Kuttal (Kuttal et al., 2020; Robe et al., 2020; Kuttal et al., 2021; Robe, 2021; Robe and Kuttal, 2022). This influenced both the design of our data collection and experiment design: we use LLMs as the most accessible tools for dialogue agent design under the new LLM-based paradigm (Jurafsky and Martin, 2025). Our results have important repercussions for research not only on mention detection, but also on other tasks related to referring acts, as they build upon mention detection; we discuss this impact in Section 5, but it is first contextualised through the body of research we present now in Section 2.

2 Related work

Although a lot of research on reference has focused on written genres, an increasing body of research has been developed in dialogue as well, with more available datasets (Khosla et al., 2021; Li et al., 2021; Poesio et al., 2023). With the recent paradigm shift introduced by the popularisation of LLMs, research in this area has also been facilitated. With their vast training datasets and their optimisation for dialogue, these tools offer great promise for NLP tasks related to reference, even in dialogue settings. Nonetheless, initial research on coreference resolution using these types of models shows that they do not always surpass previous approaches (Mitkov and An Ha, 2024), but they offer great generalisability in unsupervised settings (Le and Ritter, 2023). The models' vast inherent knowl-

edge and their capacity for in-context learning have also been successfully harnessed for entity linking (Liu et al., 2024). One area in which knowledge is still much lacking, though, is mention detection (e.g., the works mentioned above rely on ground-truth mentions for the successful results).

Mention detection, however, cannot be taken for granted, as it is the basic task upon which all other reference tasks are built (Li et al., 2021). This task has been shown to be challenging and, therefore, attempts have been made at simplifying it. Manikantan *et al*. (2024) proposed a task that focuses on the major entities. i.e., the most frequent ones — that task is useful in their literary setting, where the main characters of a story are known, but such major entities are not so easily extracted in an online setting with a dynamic environment. Even with that simplified task, the approach also had to be broken down into steps for the models to achieve satisfactory performance; in this case, the grammatical heads of the mentions were first extracted before they were expanded into the full span of each mention. However, the full mention detection task (i.e., working on all mentions, not only those referring to the most frequent entities) still poses a big challenge. This is specially true in the domain of situated dialogue with a dynamic environment, where the system has no prior information about the entities that may be mentioned. Some evidence of these challenges are already observed in the work of Madge *et al*. (2025), who tested coreference resolution in one such environments and corroborated that performance was significantly lower than in other simpler settings — their experiments included the extracted mentions as part of the input, thus not reflecting mention detection performance, but we can expect the challenges of the dialogue setting to similarly affect mention detection.

3 Methodology

3.1 Data collection and annotation

We collected a dataset of 22 distributed (remote) pair-programming dialogues between students at our institution. We recorded a total of 25 dialogues, though two were discarded for technical reasons and a participant's withdrawal; a further dialogue is excluded, as it was used only for training annotators. Each session lasted around 30 minutes, and communication took place only via voice call and a shared programming interface. We recorded several data sources: dialogue (audio and transcription[3]), keylog records[4], video and screenshots of the programming interface, and files registering all changes to the code. The keylog records were incorporated into the json files containing the dialogue transcripts through their timestamps; however, that level of context was not used in this task, as the human annotators did not use it either for mention labelling. The keylog records and the separate visual information are employed instead for other tasks in our project for which a richer context is needed. Further details about the data recording can be found in (Domingo et al., 2024).

The dataset was then annotated by a team of 7 people trained specifically for the task; before training, they had to demonstrate the necessary linguistic and programming skills through a test or relevant qualifications. The majority of the team worked on annotating coreference chains and linking code references to code files. The task of locating mentions was carried out by the two team members with expert knowledge of Linguistics using LabelStudio.[5] The interface was configured so that no unit smaller than a word could be captured to avoid human errors, and any adjoining punctuation marks (e.g., a comma at the end of a mention) was removed during post-processing. The annotation scheme and guidelines[6] were developed through discussion among the research team validated through three rounds of the two experts double coding sections of two dialogues and discussing the process as a team with the main researcher, who performed both a quantitative and qualitative analysis of the output. We thus combined a traditional iterative development approach (Fuoli, 2018) with a socialisation-based approach (Godwin and Piwek, 2016) for improved efficiency and reliability. The annotators who labelled the mentions also classified them into the linguistic categories outlined in Appendix B. As we have discussed, a full discussion of the annotation scheme is beyond the scope of this article, though more details can be found in the supplementary materials. One important piece of information is that we in-

[3]Dialogues were transcribed using Whisper (Radford et al., 2022) and revised manually. Manual revision was necessary due to the tool's inability at the time to successfully handle disfluencies and overlapping speech, the imprecisions in audio segmentation, and the challenging domain terminology.

[4]Keylog records were obtained using a custom tool or RUI, depending on compatibility with participants' computer.

[5]https://huggingface.co/LabelStudio

[6]The section of the guidelines concerning mention detection is available as additional materials and a summary can be found on Appendix B.

cluded singletons — we want the data to be usable as training/testing for an online system, where it is not possible to know if a singleton is a singleton or part of a coreference chain until the dialogue is over.

Table 1 shows the key details of our dataset. An analysis of our data shows that around a third of all the words in the corpus correspond to mentions: the average number of mentions in our dialogues is 692, with an average number of 3445 words per dialogue and an average length of 1.6 words per mention.

Number of:	Average (per dialogue)	Total (22 dialogues)
turns	385 (SD = 72)	8468
words	3445 (SD = 768)	75797
mentions	692 (SD = 189)	15222
mentions in chains	289 (SD = 59)	6365

Table 1: Dataset details

For our experiments, we used one dialogue as our development set, and the remaining 21 for evaluation. The development set was chosen semi-randomly to ensure usefulness: we selected it randomly from the top-ten most 'average' dialogues in terms of the percentage of multimodal mentions, mentions to abstract programming concepts, names, read-out sections of code, and 'intensional' objects (borrowing the terminology from (Madge et al., 2025) to denote references to planned task outputs that have not been produced yet, or ever).

3.2 Prompt and context engineering and output processing

In recent years, numerous LLMs have been released, including many trained on programming languages in addition to natural languages (Jiang et al., 2024). It is therefore a futile attempt to try to carry out a comprehensive performance test of all possible LLMs, nor even of the most recent ones, given the rapid developments in the field. Instead, we chose representative examples to illustrate how the challenging aspects of our domain may be tackled with an LLM approach. We thus limited our experiments to recent variants of the Llama and GPT families. With our choice of models we strove to select frequently used ones — e.g., these are the families used too by Le & Ritter (2023), and they represent both proprietary and open-weights models. Our model selection was further motivated by the availability of API services that offer sufficient data protection safeguards. With regard to the model parameters, throughout our experiments, we have used a constant temperature of 0, for more deterministic, replicable results.

Prompting makes running the models easy in principle (Sarkar, 2024); however, results are highly dependent on the type of prompt used (White et al., 2023). Bearing this in mind, we tested different prompting approaches. Our initial prompt refinement was based on a qualitative analysis of 20 random outputs from each prompting approach, considering task completion, format adherence, and task accuracy. We are aware of possible hallucinations, especially with regard to numerical values, so we quickly discarded any approach reliant on index numbers or any kind of numerical identifier. Instead, we obtained more consistent results with simple XML tags (<M></M>). Previous work with LLMs (Domingo et al., in press) showed us the effectiveness of a persona-based prompt: instead of providing many details about the task we expect the model to complete, we describe a persona for it to adopt and rely on its vast training data to supply the definition of what that persona entails. A non-human persona showed the best results — the prompts can be found on Appendix A.1. In our pursuit of consistency, we did not perform many experiments with temperature parameters, selecting a temperature of 0 most of the time for consistent, replicable results. Based on the work by Manikantan et al. (2024), we also tested splitting the task into the two subtasks that they identify: mention heads are detected first, and then the second task consists one expanding them, which can be done with the same model or using SpaCy[7].

In addition to the prompt, the context also requires 'engineering'. Recent models are capable of processing long inputs, and it is sometimes the case that exploiting this capacity by adding long contexts improves performance. However, long contexts can also introduce noise and draw the model's attention away from the main instructions. Therefore, our context engineering efforts concerned not only context length, but also quality. We experimented with different few-shot settings where we provided a varying number of example dialogue turns with ground-truth labels (from 1 to 10). The

[7] https://spacy.io/

examples were randomly selected, or fabricated by us aiming to exemplify the main difficulties of our setting, or manually selected to be the most representative overall, or mixed. We also tested extracting pool of turns[8] from which the best one was retrieved for each inference turn based on sentence similarity (also using SpaCy[9]).

We paired our quantitative evaluation metrics (F1) with continuous small qualitative analyses to better understand the performance of each approach. We observed some consistent errors that could be corrected through simple rules (e.g., when the models added spaces or prefaced the output with an arrow), so we added an automated post-processing step to our pipeline. Of special interest were some inconsistencies in the models' processing of contracted verbs: e.g., we'll was sometimes returned as <M>we'll</M> and sometimes as <M>we</M>'ll; we corrected the output to follow the latter format, in line with our ground truth. The use of dialogue data added another difficulty: sometimes the models struggled with mentions broken by disfluencies. We corrected the cases where a determiner was repeated, sometimes with a filler sound in between (e.g., 'the, uh, the string'), ensuring that the mention labels grouped the two determiners in the same mention.

We carried out our initial experiments with smaller models for cost/time efficiency: Llama 3 8B and GPT 4o mini. Both models' release date is only a few months apart, and both are claimed to have a similar size (Abacha et al., 2025), though the GPT model is distilled from a larger one. Nonetheless, these two options offer the highest comparability among the ones available to us. After analysing our quantitative results, we tested the best approaches on bigger versions of the models: Llama 3 70B and GPT4o, as well as GPT 4.1. We then tested the generalisability of the approaches on the evaluation set. As our human performance ceiling we use the agreement between our annotators during the validation stages of the annotation scheme development: 82.21% to 90.39%.

[8]For our preliminary tests, we extracted a random pool as a training set. For our final experiments with the evaluation set, we were instead able to use a whole dialogue (the development data) as a more complete training set that we could expect to contain a variety of turns that could always allow to find sufficiently similar examples to the turn used for inference.

[9]Martino Mensio's Github

4 Results

Here we present the results over the evaluation data: i.e., the 21 dialogues that are not dialogue 032x028, which was used as development data The naming structure \d\d\dx\d\d\d reflects the code assigned to each speaker in the pair during anonymisation. The numbers represent the order in which people interested in participating signed the consent forms. The experiments with this data allow us to have a clearer view of model performance without a single dialogue biasing results. The design of our experiments was based on our preliminary work with the development data (dialogue 032x028)[10]. Based on preliminary work with the development data, we concluded that the most successful approach was providing few-shot examples that were similar to the turn being parsed. We used the development dialogue 032x028 as the pool from which to retrieve the few-shot examples using sentence similarity; using a whole dialogue would allow us to have a rich pool of possible examples. We also tested a zero-shot approach to have a clear view of how the few-shot examples contribute to the task. For our final experiments, we used the whole range of models available to us, both big and small: GPT 4o, GPT 4o mini, GPT 4.1, GPT 4.1 nano, Llama 3 70B, and Llama 3 8B. Under the few-shot condition, we used three few-shot examples, which had proven to be sufficient with the GPT models. However, as we had also observed that the Llama models were more sensitive to the amount of few-shot examples, we also used six few-shot examples with Llama 3 8B — given the amount of data we were testing on, we did not test with a larger number of examples, and we only used the six examples on the smaller Llama model. Also drawing on the insights from our preliminary experiments, we expected the Llama models to perform below the GPT models under any condition. Thus, with the evaluation tests we did not attempt to boost their performance closer to the GPT models; we only wished to determine to which extent increasing the number of few-shot examples boosts performance in these more context-sensitive models

Table 2 shows the GPT models' performance under the approaches we've described; Table 3 shows the performance for the Llama models. As we tested two few-shot conditions, we did not test a zero-shot approach — our preliminary experiments

[10]More details about our preliminary work can be found in Appendix A

Model	Zero-shot average			Few-shot average		
	F1pp	Ppp	Rpp	F1pp	Ppp	Rpp
GPT 4.1	0.64	0.80	0.53	0.80	0.84	0.76
GPT 4.1 nano	0.25	0.48	0.17	0.57	0.60	0.54
GPT 4o	0.27	0.53	0.18	0.70	0.72	0.68
GPT 4o mini	0.22	053	0.14	0.73	0.78	0.69

Table 2: Average performance of the GPT models across the 21 evaluation dialogues. Few-shot performance involves examples selected based on sentence similarity; in this case we use three examples. The pp suffix means that the score was obtained after post-processing the output.

made it evident that the Llama models rely to a greater extent on the few-shot examples; thus, our focus was on seeing the effect of increasing the number of examples instead.

From these tables we can observe that, as was the case with the development dialogue, the GPT models perform better across all dialogues. The best-performing model is GPT 4.1., which is meant to be suitable for very complex text tasks. Surprisingly, GPT4o mini's performance is not much lower, though it relies heavily on the post-processing of the output. As expected, the Llama models perform below the GPT models in general. We can also see that increasing the number of few-shot examples does improve performance to some extent. Here we have presented the F1 scores; however, we have also made some observations about precision and recall. For instance, the average zero-shot precision for GPT 4.1 was 0.80, but recall was 0.53; with the few-shot approach, precision increased noticeably to 0.84, but recall increased even more remarkably to 0.76. We see this tendency in the other models through both the final and preliminary experiments, where precision is higher than recall, and the difference is larger in the worse-performing approaches.

Additionally, we re-evaluated the final results using a more lenient metric, based on the work by Moosavi et al. (2019). They point out that mention detection can be evaluated based on minimum span match, instead of requiring a system to define the mention boundaries in exactly the same way as the ground truth data. They develop an algorithm for automatically extracting minimum spans from mentions without the need of additional manual annotations. Both their algorithm and a simpler method based on head extraction correlate highly with human annotations in the few datasets that include such minimum span annotations. Here we use the head extraction method for simpler implementation. Table 4 shows the adjusted average scores.

Model	FS	Few-shot average		
		F1pp	Ppp	Rpp
Llama 3 70B	3	0.59	0.62	0.56
Llama 3 8B	6	0.49	0.50	0.48
Llama 3 8B	3	0.46	0.48	0.44

Table 3: Average performance of the GPT models across the 21 evaluation dialogues. Few-shot performance involves examples selected based on sentence similarity. FS stands for the number of few-shot examples. The pp suffix means that the score was obtained after post-processing the output.

Model	Precision-pp	Recall-pp	F1-pp
GPT 4.1	0.89	0.80	0.84
GPT 4.1 nano	0.67	0.61	0.64
GPT 4o	0.76	0.72	0.74
GPT 4o mini	0.84	0.74	0.78
Llama 3 70B	0.70	0.62	0.66
Llama 3 8B	0.59	0.55	0.57

Table 4: Scores measuring minimum-span matches on the similarity-based few-shot example approach with 3 examples. The pp suffix indicates that we evaluated the output after post-processing it.

4.1 Error analysis

As our preliminary experiments covered only one dialogue, with the evaluation data we were interested in seeing how the characteristics of each dialogue affected task performance — all dialogues in this setting have some broader characteristics in common, but we already saw in previous analyses that these are displayed to different extents in each dialogue. Looking at performance across different models, we concluded that the 'easiest' dialogue was 040x054 (with a maximum F1 of 0.84 and a minimum of 0.52), and the 'hardest' was 062x059 (with a maximum F1 of 0.67 and a minimum of 0.45). Table 5 shows the main distinctive characteristics of mentions in this setting for each of these dialogues. We see that these two dialogues are in distant sides of the spectrum with regard to the percentage of mentions that are proper nouns and the mentions that are read-out code — we must also bear in mind that, in this domain, both cate-

gories are linked, as proper nouns are often variable names. To better understand performance differences, we performed a brief error analysis of these two dialogues by looking at their false negatives and their false positives; we analysed 20 mentions for each error type for each dialogue. With regard to the false negatives, in the easy dialogue, 8 of the missed mentions were direct references to code (e.g., 'for x in grade'), while this figure was 14 for the hard dialogue. In this small sample, we did not find any names in the easy dialogue's false negatives, but we found two for the hard dialogue: 'student student, test scores' and (BLEEP). These examples illustrate how names (and mentions in general) can be challenging in this domain. The first name includes a hesitation, and the second one is anonymised — though the pool of few-shot examples featured this as well. Looking at the false positives, we can observe how this task is also challenging and ambiguous for humans, as some of the false positives could be considered valid positives or even an error in the ground truth (e.g., in 'Right. I'm doing this wrong, aren't i? That's it.' the ground truth missed 'this'). In some other cases, however, the false positives cannot be considered mentions even if we apply the annotation scheme very flexibly (e.g., in the previous example, 'That's it' was returned as a mention, thus a truly false positive). We observe such verb phrases or verbs treated as mentions even when the turn shows no discourse deixis that would justify labelling a whole verb phrase as a referent; we find four cases in the easy dialogue, and two in the hard one. The main cause of false positives that we observe in both dialogues is a mismatch between the ground truth span and the output span, where the model misses parts of a mention when it is a complex noun phrase — e.g., one turn says 'Um and then we need it to output the percentage of students who passed and then output the grade', and the ground truth extracts 'the percentage of students who passed', whereas the model only extracts the main part of the phrase, 'the percentage of students'. We find four such cases in the easy dialogue, and six in the hard one. Lastly, one important source of false positives related to this is the presence of hesitations, which are common in our dataset as part of the nature of spoken dialogue; in such cases, the model can miss part of the mention or interpret a repetition as two separate mentions (e.g., in 'Yeah um maybe uh student student, test scores', the ground truth extracts 'student student, test scores' as a mention to the variable storing the student test scores, but the model extracts three separate mentions: 'student', 'student', and 'test scores').

Type of mention (percentage range)	Value for 040X054 (easy)	Value for 062x059 (hard)
Multimodality (2.82-21.36)	5.87	8.24
Abstract Mentions (0-19.67))	0.78	1.90
Names (3.84-24.29)	**8.92**	**15.96**
Read-dictate (1.39-15.35)	**6.22**	**10.47**
Intensional mentions (1.27-31.85)	5.60	3.95
Total mentions (350-1154)	482	783

Table 5: Main characteristics of the 'easiest' (040x054) and 'hardest' (062x059) dialogues. We show the types of mentions as a percentage of the total. In parentheses we show the value range across all dialogues to contextualise this data.

5 Discussion

As we mentioned in Section 3.2, agreement between our annotators for this task ranged between 82.21% and 90.39% for the three double-coding validation rounds that we ran. This shows that, even though this was one of the 'easier' tasks in our work on reference, even trained human experts sometimes disagreed. We observed some human errors in the ground truth in our error analysis (Section 4.1), but our validation work for the annotation scheme involved qualitative analyses and discussions with the annotators to ensure that there were no misalignments in their internalisation of the scheme; therefore, most disagreements at the final validation stage can be primarily attributed to inherent ambiguities of the task — while this task is simpler than the subsequent task of coreference resolution, for this latter task Poesio *et al.* (2023) noted that there is a great degree of ambiguity in references, with figures of up to 40% in dialogues annotated for discourse deixis. Bearing this in mind, we cannot expect any model to surpass the human scores; that would indicate both overfitting

to one annotator's perspective and an imbalance in our data split — we are preventing this by testing on all our dialogues, thus balancing data from both annotators. To be able to compare the models' performance with the human annotators, we calculated the models' agreement with the ground truth. For GPT4.1, this ranges between 71.86% and 88.14%; GPT 4o and GPT4o mini reach maximums of 79.86% and 81.99% respectively, but their minimums are much lower (38.08% and 37.22%). We can therefore conclude that, under the best approach, some recent models approximated human performance, but only GPT 4.1 did so consistently.

Due to the recent advances in language models, few studies exist to this date on mention detection using LLMs. However, Manikantan *et al.* (2024) and Le & Ritter (2023) offer some comparable results. The latter observed unsatisfactory performance with Instruct GPT, which yielded an F1 score of 46.5. The former, however, obtained F1 scores ranging between 77.1 and 85.5 using GPT 4 with their best approach (having the LLM extract the nucleus of the main mentions, and using SpaCy to expand the mention span). Our results are unsurprisingly higher than Le & Ritter's, possibly due to primarily their use of a less powerful model — in fact their results are similar to what we obtained with our poorest model, Llama 3 8B. Manikantan *et al.*'s better results are more similar to our best results, despite them performing a simplified task on literary texts that lack the challenges of pair-programming dialogue, probably due to our use of the latest, most powerful models. We find evidence of this in the fact that performance only reached this high range of F1 scores between 0.75 and 0.86 with GPT 4.1.

In addition to overall model performance, we have made some other key observations. We have observed through models' zero-shot performance that their base knowledge allows them to detect mentions with precision, but that in-context learning is needed for them to detect a broader range of mentions. Additionally, as we expected from analysing the characteristics of our dataset, one key issue that made mention detection difficult in this domain is the mentions to code, especially variables with their flexible form unlike that of names in other domains. Additionally, the fact that we are dealing with spoken dialogue resulted in hesitations, which also pose a significant challenge.

6 Conclusions

Through this work, we set out to explore the challenges that a pair-programming dialogue setting presents for work on reference, starting with the base task of mention detection — there can be no good coreference resolution without very good mention detection. We used LLMs for this exploration as the most recent and accessible tools for this task, imagining the kinds of tools that might available for an online pair-programming agent.

We have looked at the different experimental settings that may improve model performance in a few-shot setting: prompt and context engineering, carefully crafting suitable prompts and selecting the optimal type and amount of few-shot examples. We have observed that GPT 4.1 is close to human performance, so it could potentially replace human annotators with adequate prompt and context-engineering on texts that are not exceedingly complicated. As we have discussed throughout this work, LLMs are powerful tools that can detect mentions to some extent. However, just as humans require annotation schemes and there is often debate about which types of mentions should be included (Zeldes, 2022), the models require few-shot examples to capture the whole range of mentions required. Pair-programming dialogue presents some additional challenges that are not found in other text types, primarily hesitations stemming from it being spoken dialogue, and references to code, which involve terminology much unlike that of other domains.

Mention detection is the basic task upon which work in reference resolution is built. Therefore, insights into it are important, as outside of research scenarios coreference resolution cannot rely on having gold-standard mentions as input. We have seen some limitations from LLMs performing this task, but we have also discussed where some of these come from and how performance can be improved in many cases. These insights will inform our future work in reference resolution, and we hope it can also prove useful to the broader NLP community. As we discussed in 3.1, we recorded several types of data, and our annotation work went beyond mention detection into reference resolution and phrase grounding. We will thus work on those tasks with the outputs and insights obtained at this first stage of mention detection.

Limitations

As we have discussed in Section 3.2, we did not strive to do a comprehensive analysis of LLMs' performance that included all possible kinds of such models, nor is that feasible with the rapid developments in this area. We also did not perform a comparative analysis with other neural or rule-based approaches: our focus is on LLMs as an example of the SotA in NLP for numerous tasks, to examine to what degree the idiosyncrasies of our dialogue setting pose significant challenges. Another significant limitation of our work is the conversion of spoken dialogue into text. We used transcriptions generated with Whisper (Radford et al., 2022) and revised by us, thus having very robust verbatim transcripts. Future work with multimodal LLMs will need to determine whether the same performance is achieved when the models process the audio directly — this could hinder performance through transcription errors, or it could even simplify the task through the removal of hesitations in the initial steps of speech processing. Additionally, although our dataset is not exceedingly small, as shown on Table 1, it is very far from the size of popular datasets such as OntoNotes (Weischedel et al., 2013), which must be taken into account when interpreting our results/

Ethical considerations

This research project has been reviewed by, and received a favourable opinion from, The Open University's Human Research Ethics Committee. Participants gave informed consent for the use of their data, which has been anonymised. They were informed of their right to withdraw from the study. While participation was voluntary, participants received a voucher as a token of gratitude.

Acknowledgments

This work has financial support from EPSRC Training Grant DTP 2020-2021 Open University and Toshiba Europe Limited.
We thank our annotators for their valuable work.

References

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes. *arXiv preprint*. ArXiv:2412.19260 [cs].

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Cecilia Domingo, Paul Piwek, Michel Wermelinger, and Svetlana Stoyanchev. 2024. Annotation Needs for Referring Expressions in Pair-Programming Dialogue. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Trento, Italy. SEMDIAL.

Cecilia Domingo, Paul Piwek, Michel Wermelinger, Svetlana Stoyanchev, Rama Doddipatla, and Kaustubh Adhikari. Human ratings of LLM response generation in pair-programming dialogue. In *Proceedings of the 18th International Natural Language Generation Conference*, Hanoi, Vietnam. Status: in press.

Matteo Fuoli. 2018. A stepwise method for annotating appraisal. *Functions of Language*, 25(2):229–258.

Keith Godwin and Paul Piwek. 2016. Collecting Reliable Human Judgements on Machine-Generated Language: The Case of the QG-STEC Data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 212–216, Edinburgh, UK. Association for Computational Linguistics.

Brian Hanks, Sue Fitzgerald, Renée McCauley, Laurie Murphy, and Carol Zander. 2011. Pair programming in education: a literature review. *Computer Science Education*, 21(2):135–173.

Anja Hawlitschek, Sarah Berndt, and Sandra Schulz. 2023. Empirical research on pair programming in higher education: a literature review. *Computer Science Education*, 33(3):400–428.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating Code-Switching Translation with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A Survey on Large Language Models for Code Generation. *arXiv preprint*. ArXiv:2406.00515 [cs].

Daniel Jurafsky and James Martin. 2025. Chatbots & Dialogue Systems. In *Speech and Language Processing*, pages 1–39. Stanford University.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandeep Kaur Kuttal, Jarow Myers, Sam Gurka, David Magar, David Piorkowski, and Rachel Bellamy. 2020. Towards Designing Conversational Agents for Pair Programming: Accounting for Creativity Strategies and Conversational Styles. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–11, Dunedin, New Zealand. IEEE.

Sandeep Kaur Kuttal, Bali Ong, Kate Kwasny, and Peter Robe. 2021. Trade-offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Yokohama Japan. ACM.

Nghia T. Le and Alan Ritter. 2023. Are Large Language Models Robust Coreference Resolvers? *arXiv preprint*. ArXiv:2305.14489 [cs].

Shengjie Li, Hideo Kobayashi, and Vincent Ng. 2021. The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution in Dialogue: A Cross-Team Analysis. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 71–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13634–13651, Miami, Florida, USA. Association for Computational Linguistics.

Chris Madge, Maris Camilleri, Paloma Carretero Garcia, Mladen Karan, Juexi Shao, Prashant Jayannavar, Julian Hough, Benjamin Roth, and Massimo Poesio. 2025. MDC-R: The Minecraft Dialogue Corpus with Reference. *arXiv preprint*. Version Number: 1.

Kawshik Manikantan, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. 2024. Major Entity Identification: A Generalizable Alternative to Coreference Resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11679–11695, Miami, Florida, USA. Association for Computational Linguistics.

Ruslan Mitkov and Le An Ha. 2024. Are rule-based approaches a thing of the past? The case of anaphora resolution. *Procesamiento del Lenguaje Natural*, 73(0):15–27.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using Automatically Extracted Minimum Spans to Disentangle Coreference Evaluation from Boundary Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational Models of Anaphora. *Annual Review of Linguistics*, 9(1):561–587.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*. ArXiv:2212.04356 [cs, eess].

Peter Robe. 2021. Designing a Pair Programming Conversational Agent. Master's thesis, University of Tulsa, Tulsa, Oklahoma.

Peter Robe, Sandeep Kaur Kuttal, Yunfeng Zhang, and Rachel Bellamy. 2020. Can Machine Learning Facilitate Remote Pair Programming? Challenges, Insights & Implications. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–11, Dunedin, New Zealand. IEEE.

Peter Robe and Sandeep Kaur Kuttal. 2022. Designing PairBuddy—A Conversational Agent for Pair Programming. *ACM Transactions on Computer-Human Interaction*, 29(4):1–44.

Advait Sarkar. 2024. Intention Is All You Need. In *Proceedings of the 35th Annual Conference of the Psychology of Programming Interest Group*, Liverpool.

Ralph Weischedel, Martha Palmer, Marcus Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. 2013. OntoNotes Release 5.0. Artwork Size: 2806280 KB Pages: 2806280 KB.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint*. ArXiv:2302.11382 [cs].

Amir Zeldes. 2022. Opinion Piece: Can we Fix the Scope for Coreference?: Problems and Solutions for Benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

A Appendix: Prompt and context engineering

Table 6 shows the compared performance of different models tested under the same simple setting (3 randomly selected few-shot examples; we evaluated over five folds with repeated random subsampling validation to account for variations resulting from the different random examples.). Given the size of the development set, we cannot read

Model	P	R	F1	F1pp
GPT 4.1	0.80	0.75	0.77	**0.79**
GPT 4.1 nano	0.55	0.57	0.56	**0.60**
GPT 4o	0.62	0.61	0.62	**0.66**
GPT 4o mini	0.73	0.70	0.72	**0.80**
Llama 3 8B	0.00	0.00	0.00	**0.44**

Table 6: Model comparison using 3 random few-shot examples. Initial results on the development set for P(recision,), R(ecall) and F1 after post-processing.

much into the results, but they allow us to exemplify three basic observations from our broader set of preliminary experiments: the GPT models perform markedly better than Llama, smaller distilled models are competitive with their full version, and postprocessing significantly improves results. This last point is especially true for Llama, which was consistently inconsistent in its formatting of the output.

Our initial qualitative analysis allowed us to refine our base prompt, as we discussed in Section 3.2. However, considering that the models that we are using are optimised for dialogue, we also tested a prompt that focuses more on this dialogue setting instead of a parser persona — we called this prompt a 'chatty' prompt. The different prompts are shown below. Table 7 shows the different F1 scores using our base prompt and two 'chatty' prompt versions, comparing two small models and different numbers of randomly-selected few-shot examples. The first 'chatty' prompt also follows the persona approach, encouraging the model to respond like a dialogue partner to a student; the second one introduces the setting as a dialogue, but without asking the model to adopt any human-like behaviour. As we can see, performance with the first of the 'chatty' prompts decreases noticeably, but with the second one it is similar to the base prompt. The table also allows us to exemplify the effect of the number of few-shot examples. We see that, with the GPT model, performance is stably high with few examples, but decreases when we use more than a couple of examples; for Llama, however, a higher number of examples increases performance. In addition to the number of few-shot examples, we must also consider the type (e.g., Huzaifah *et al.* 2024 observed the benefits of carefully selecting examples, with the ones most closely related to the input content being most useful in clarifying the task to the model). We therefore compared the use of random examples against the use of specifically chosen examples. One approach we followed was manufacturing our own examples that imitated the style of the dialogues but concentrated in a few sentences the main phenomena that could be challenging in our data; we call these examples 'ideal' examples. We also tested a similar approach where we instead selected the most representative examples from real data; we call these examples 'real'. Finally, we also tested an approach where we used real examples from the data but we did not manually select the best; instead, we used sentence similarity[11] to adapt what the best were for each case. We separated a pool of dialogue turns; before inference, each input turn was compared against the pool of turns and the top n most similar turns were retrieved as few-shot examples; we call this approach 'Similar'. Table 8 shows a comparison of the performance under the different types of few-shot examples. Again, using the development data limits the interpretability of the results, but they do suggest that a careful selection of the few-shot examples is far from irrelevant.

Model	FS	Base prompt F1	Chatty prompt 1 F1	Chatty prompt 2 F1
GPT 4o mini	2	0.79	0.56	0.70
GPT 4o mini	3	0.80	0.59	0.72
GPT 4o mini	6	0.81	0.57	0.76
Llama 3 8B	1	0.29	0.11	0.27
Llama 3 8B	2	0.45	0.33	0.28
Llama 3 8B	3	0.44	0.39	0.35
Llama 3 8B	10	0.47	0.47	0.51

Table 7: Performance over development data comparing two prompt styles, two small models, and three amounts of randomly-selected few-shot examples. The tests cover the whole development split (dialogue 032x028), excluding the 1-10 turns used for few-shot learning. FS refers to the number of few-shot examples. The F1 scores represent the value after post-processing the output.

Following the work by Manikantan *et al.* (2024), we also tested whether dividing the task into two simpler tasks improved performance. For the first task, we ask the model to only label the heads of mentions, expecting that this will reduce the burden of having to decide which determiners and modifiers to include — we leave that for the second task, where the heads are expanded into the full mention span. For the second step, as it relies purely on

[11] Martino Mensio's Github

Model	Type of FS	FS	F1pp
GPT4omini	Similar	3	0.81
L3-8b-v1	Similar	3	0.47
GPT4omini	Ideal	3	0.75
L3-8b-v1	Ideal	3	0.42
GPT4omini	3 ideal + n random	1	0.79
GPT4omini	3 ideal + n random	2	0.79
GPT4omini	3 ideal + n random	3	0.78
L3-8b-v1	3 ideal + n random	1	0.45
L3-8b-v1	3 ideal + n random	2	0.43
L3-8b-v1	3 ideal + n random	3	0.50
GPT4omini	Real	3	0.77
L3-8b-v1	Real	3	0.37
GPT4omini	3 ideal + n real	3	0.79
L3-8b-v1	3 ideal + n real	3	0.44

Table 8: Performance with different types and numbers of few-shot examples (FS). We present F1 scores after the output was post-processed (F1pp).

Model	Approach	FS	F1pp
GPT4o mini	2 prompts	3	0.67
GPT4o mini	1 prompt + SpaCy	3	0.63
GPT4o mini	1 prompt	3	0.80
Llama 3 8B	2 prompts	3	0.33
Llama 3 8B	1 prompt + SpaCy	3	0.37
Llama 3 8B	1 prompt	3	0.44
GPT4o mini	2 prompts	6	0.67
GPT4o mini	1 prompt + SpaCy	6	0.64
GPT4o mini	1 prompt	6	0.81
Llama 3 8B	2 prompts	6	0.36
Llama 3 8B	1 prompt + SpaCy	6	0.43
Llama 3 8B	1 prompt	6	0.49

Table 9: Performance comparison with one prompt or splitting the task into mention-head detection and mention span expansion (with a second prompt or with SpaCy). We present F1 scores after post-processing the output (F1pp).

identifying dependencies, we test it both with an LLM and with a simple script using SpaCy. Table 9 shows performance under these conditions, with a single-prompt approach for comparison. We can observe that the two-prompt approach never surpasses the single-prompt approach, and that using SpaCy to expand the heads does not improve performance but actually hinders it.

A.1 Prompts

Zero-shot prompt

You are an NLP parser specialised on extracting mentions from text. Your output is fed to a coreference resolution system and an entity linking system. Therefore, your output should respect format restrictions and not add any comments; if there are no mentions, just return the original text. You will be given a text and you should extract all the mentions in it. You should return the text with mention opening tags <M> and closing tags </M>.

Base prompt

You are an NLP parser specialised on extracting mentions from text. Your output is fed to a coreference resolution system and an entity linking system. Therefore, your output should respect format restrictions and not add any comments; if there are no mentions, just return the original text. You will be given a text and you should extract all the mentions in it. You should return the text with mention opening and closing tags as in the examples.

'Chatty' persona-based prompt

You are a university student pair-programming with a partner, who is also a university student. Your partner says something and you need to understand it to respond. To show that you've understood your partner, you need to label the things that your partner has mentioned, to later think about what each of those mentions refers to.

'Chatty' dialogue-based prompt

You are going to see a bit of text from a dialogue partner. Think of all the objects, concrete or abstract, that they mention in their text. Return the text with <M> </M >tags framing the objects. Return nothing else. Here are some examples of how you've done this before.

B Appendix: Mention classification labels

The guidelines provided to annotators include numerous images and examples, spanning 28 pages, so here we only provide a brief summary.

The annotation units are mentions, be it to ele-

ments in the dialogue (e.g., a previous word now referred to with a pronoun), in the context (e.g., the participants, programming concepts, elements of the code being created, etc.), or both. These mentions can be one or more words, and they may be split by punctuation; if the words are part of the same mention, we label them as one broad span that includes the words and the spaces (and possibly commas or apostrophes) between them. If we did not highlight the space in between the words as part of the annotation, that would create separate mentions. If there is a filled pause between parts of one mention, the pause can be included as part of the mention. We also include repetition within the same unit; for example, in many cases the speakers will repeat an article – we annotate them all as part of the mention. After labelling the mention spans, the annotators also classified the mentions according to their grammatical number and the linguistic categories summarised below:

- Pronoun - Personal
- Pronoun - Demonstrative
- Pronoun - Other
- NP - Definite
- NP - Indefinite
- NP - Meta (this category was used for mentions referring to words as abstract concepts, not the meaning represented by the word)
- NP - Read-dictate (this category was used for read-out or dictated code)
- Name
- Name variation (this category was used for variable names that were not reproduced exactly, e.g., when the 'len()' function is mentioned as 'length')
- Location adverb
- Incomplete (this label was added independently of the others to mentions that spanned more than one turn, so that we could later join the segments)

The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works

Antoine Bourgois and Thierry Poibeau
Lattice (CNRS & ENS-PSL & Université Sorbonne Nouvelle), Montrouge, France
antoine.bourgois@protonmail.com
thierry.poibeau@ens.psl.eu

Abstract

While coreference resolution is attracting more interest than ever from computational literature researchers, representative datasets of fully annotated long documents remain surprisingly scarce. In this paper, we introduce a new annotated corpus of three full-length French novels, totaling over 285,000 tokens. Unlike previous datasets focused on shorter texts, our corpus addresses the challenges posed by long, complex literary works, enabling evaluation of coreference models in the context of long reference chains. We present a modular coreference resolution pipeline that allows for fine-grained error analysis. We show that our approach is competitive and scales effectively to long documents. Finally, we demonstrate its usefulness to infer the gender of fictional characters, showcasing its relevance for both literary analysis and downstream NLP tasks.

1 Introduction

Coreference Resolution (CR)—the task of identifying and grouping textual mentions that refer to the same entity (e.g., a person, an organization, a place)—is a fundamental component of natural language processing (NLP). It underpins downstream applications such as information extraction (Yao et al., 2019), text summarization (Liu et al., 2021), and machine translation (Vu et al., 2024). Over the past decades, significant progress has been made in CR, evolving from rule-based multi-sieve systems to end-to-end neural models, encoder-decoder architectures, and large language models based approaches, all contributing to improvements on benchmark datasets (Porada et al., 2024).

These models have long been trained and evaluated solely on generic datasets such as OntoNotes (Hovy et al., 2006). As CR drew attention in other fields, it became evident that models trained on general datasets underperformed when applied to domain-specific tasks. To address this flaw, dedicated datasets have been developed, covering areas such as biomedical (Lu and Poesio, 2021) and encyclopedic data (Ghaddar and Langlais, 2016).

Driven by the availability of extensive digitized collections, literary texts have emerged as a key subject of digital humanities (Moretti, 2013). A large part of such research focuses on characters, considered a fundamental aspect of fiction works. The study of characters is essential for analyzing narrative structures, plot development or conducting diachronic studies. CR is crucial for applications such as quote attribution (Vishnubhotla et al., 2023), character archetypes inference (Bamman et al., 2014), and social networks extraction (Elson et al., 2010). Additionally, it has been employed to study the representation and behavior of characters according to their gender (van Zundert et al., 2023).

As outlined by Roesiger et al. (2018), literary texts present unique challenges for CR, including character evolution throughout the narrative and the prevalence of dialogues involving multiple participants. They also contain a high proportion of pronouns and nested mentions. Complex narrative structures—such as letters, flashbacks, and sudden narrator interventions—further complicate the task. Additionally, authors often rely on readers' contextual understanding rather than explicit statements, creating ambiguities when linking mentions.

To address these challenges, annotated datasets have been developed, covering multiple languages and genres, from classical novels and fantasy tales to contemporary literature. These resources enable training and evaluating in-domain coreference resolution models, leading to steady performance improvements (Martinelli et al., 2024). Despite visible progress on benchmarks, current state-of-the-art CR models still struggle with full-scale literary texts, limiting usefulness for downstream applications (Vishnubhotla et al., 2023).

A key factor contributing to this limitation lies in the scarcity of fully annotated long documents. Most existing datasets consist of short excerpts or

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 55–69
Suzhou, China, November 9, 2025.

relatively brief texts. Since coreference annotation is labor-intensive and costly, there exists a trade-off between annotating a larger number of short documents or a smaller number of long ones.

We argue that the lack of representative datasets for long literary texts is a major obstacle to effectively scaling CR models. This work aims to bridge this gap, and our contributions are as follows:

- an annotated dataset of character coreference for three full-length French novels spanning three centuries, showcasing the feasibility of combining automatic mention detection with manual coreference annotation.
- A modular CR pipeline scalable to long documents, enabling fine-grained error analysis and achieving competitive performance on benchmark dataset.
- A comprehensive study of the impact of document length on CR performance.
- A case study on character gender inference using CR models.[1].

2 Related Work

2.1 Coreference Models

Coreference resolution has undergone several paradigm shifts (Poesio et al., 2023), evolving from rule-based, linguistically informed models tested on limited examples to data-driven statistical approaches enabled by the creation of large annotated datasets such as those from the Message Understanding Conference (MUC) and the Automatic Content Extraction (ACE) shared tasks (Grishman and Sundheim, 1995; Doddington et al., 2004).

The adoption of neural network-based models, beginning with Wiseman et al. (2015), marked significant progress. The introduction of end-to-end models by Lee et al. (2017, 2018), further advanced CR by jointly detecting mention spans and resolving coreference, eliminating the need for external parsers and handcrafted mention detection models. Building on this foundation, higher-order inference (HOI) strategies and entity-level models were developed to refine entity representations during inference and leverage cluster-level information.

However, as highlighted by Xu and Choi (2020), the performance gains from these strategies have been marginal compared to the substantial improvements achieved by the use of more powerful encoders like ELMo, BERT and DeBERTaV3.

Alternative approaches using encoder-decoder architectures and large language models have been proposed, framing CR as sequence-to-sequence (Hicke and Mimno, 2024) or question-answering (Wu et al., 2020; Gan et al., 2024) tasks. While showing promising results, these methods are computationally intensive and do not scale efficiently to long documents or resource-constrained scenarios.

2.2 Existing Datasets

While MUC and ACE laid the foundation for coreference datasets, OntoNotes has since become the primary benchmark for CR. Published in 2006 (Hovy et al.) and regularly updated, OntoNotes has been used in the CoNLL shared tasks (Pradhan et al., 2011, 2012). Its latest version (Weischedel et al., 2013) spans multiple languages (English, Chinese and Arabic), and genres, including conversations, news, web, and religious texts. The English part contains 1.6M tokens across 3,943 documents, averaging 467 tokens per document. OntoNotes does not contains singleton mentions—those that do not corefer with any other mention.

The growing interest for large literature corpora has driven the development of dedicated annotated datasets. The late 2010s saw the emergence of the first literary CR datasets, beginning with DROC (Krug et al., 2018), including samples from 90 German novels annotated with character coreference chains. With over 393,000 tokens (averaging 4,368 tokens per document), DROC remains the largest literary CR dataset to date. The RiddleCoref dataset (van Cranenburgh, 2019) followed, covering excerpts from 21 contemporary Dutch novels, though it is not publicly available due to copyright restrictions. Bamman et al. (2020) released LitBank, consisting of the first 2,000 tokens from 100 English novels. This dataset covers six entity categories (persons, faculties, locations, geopolitical, organizations and vehicles). Other datasets include FantasyCoref (Han et al., 2021), KoConovel covering 50 full-length Korean short stories (Kim et al., 2024), and LitBank-fr (Mélanie et al., 2024). This last dataset is noteworthy in that it covers longer excerpts of text—averaging 9,834 tokens and up to 30,987 for the longest document.

Despite these resources, extrinsic evaluations re-

[1]All code and data are publicly available at github.com/lattice-8094/propp. The trained coreference resolution pipeline is readily usable through the open-source `propp_fr` Python library.

[2]standardebooks.org

	Lang.	Domain	Doc.	Tokens	Tokens / Doc.	
					Avg.	Max.
Annotated Datasets						
OntoNotes^{en} (Weischedel et al., 2013)	English	Non-literary	3,493	1,600,000	467	4,009
DROC (Krug et al., 2018)	German	Fiction	90	393,164	4,368	15,718
RiddleCoref (van Cranenburgh, 2019)	Dutch	Fiction	21	107,143	5,102	-
LitBank (Bamman et al., 2020)	English	Fiction	100	210,532	2,105	3,419
FantasyCoref (Han et al., 2021)	English	Fantasy	214	367,891	1,719	13,471
KoCoNovel (Kim et al., 2024)	Korean	Fiction	50	178,000	3,578	19,875
LitBank-fr (Mélanie et al., 2024)	French	Fiction	28	275,360	9,834	30,987
Target Datasets						
Standard Ebooks[2]	English	Fiction	770	82,855,210	107,604	1,105,964
Chapitres (Leblond, 2022)	French	Fiction	2,960	240,971,614	81,409	878,645
Contribution						
Ours	French	Fiction	3	285,176	95,058	115,415

Table 1: Comparison of coreference annotation datasets: OntoNotes (English section), fiction datasets, and target datasets across languages.

veal that CR models perform poorly on full-length documents (van Zundert et al., 2023). Studies consistently show that performance degrades with increasing document length (Joshi et al., 2019; Toshniwal et al., 2020; Shridhar et al., 2023). This represents a major challenge given that practical applications involve digitized collections such as Project Gutenberg or Wikisource, where documents frequently exceed 90,000 tokens and can reach up to a million as illustrated in Table 1.

While some initiatives annotate entire books, they often diverge from standard guidelines. He et al. (2013) annotated *Pride and Prejudice* but focused solely on proper mentions. Similarly, van Zundert et al. (2023) labeled character aliases across 170 novels, omitting pronouns and noun phrases. Other datasets, such as QuoteLi3 (Muzny et al., 2017) and PNDC (Vishnubhotla et al., 2022), include coreference annotations for speakers and direct speech but lack broader character coverage.

Until recently, the only coreference resolution results reported on a document of substantial length (37k tokens) came from Guo et al. (2023), though their work omits singletons, plural mentions, and nested entities. Since then, Martinelli et al. (2025) released an extended dataset, BOOKCOREF$_{gold}$, comprising two fully annotated English-language novels averaging 97,140 tokens per document, along with benchmark results, further illustrating the growing interest in long-document CR.

These observations underscore the need for an annotated corpus of full-length literary documents. Such a resource will enable more robust evaluation and improvement of CR models, addressing the gap between current datasets and intended applications.

3 New Dataset

We selected three average-length French novels spanning three centuries, resulting in a total of 285,176 tokens. We chose to annotate coreference for character mentions only for several reasons. First, most downstream tasks in literary NLP focus on characters. Second, previous work shows that characters account for the majority of annotated mentions—83.1% in LitBank. Restricting annotations to character mentions allows us to leverage the 31,570 mentions already annotated in LitBank-fr to train an accurate mention detection model.

For consistency and interoperability, we adhere to the annotation guidelines from Mélanie et al. (2024). We annotate all mentions referring to a character, including pronouns, nominal phrases, proper nouns, singletons and nested entities. Coreference links capture strict identity relations.

On [their]$_1$ way to visit [John]$_2$, [[my]$_3$ parents]$_1$ met [[Mrs. Smith]$_4$ and [[her]$_4$ husband]$_5$]$_6$.

This sentence illustrates some annotation principles:

- Mention types: pronoun (*my*), nominal phrase (*her husband*), and proper noun (*John*);
- Nested entities, including third-level nesting (e.g., *her* within *Mrs. Smith and her husband*);
- Plural mentions (*their*, *my parents*, *Mrs. Smith and her husband*) are treated as distinct coreference chains separate from their individual components;
- Singletons, such as *John*, are annotated even if they are not referenced again.

3.1 Mentions Detection Model

While Mélanie et al. (2024) report strong results for mention detection, we opted to retrain our own model. Our approach builds on a stacked BiLSTM-CRF architecture inspired by Ju et al. (2018), leveraging contextual token embeddings from CamemBERT$_{\text{LARGE}}$ (Martin et al., 2020). When evaluating for exact match with gold annotations, We achieved an improvement of 4.99 in F1-score on the test set from LitBank-fr (Table 2). To assess generalization performance and due to the small number of documents in the dataset, we also conducted a leave-one-out cross-validation (LOOCV). Details of the model architecture and hyperparameters are available in the Appendix A.

Model	P	R	F1	Support
Mélanie et al. (test set)	85.0	92.1	88.4	4,061
Ours (test set)	**91.29**	**95.59**	**93.39**	4,061
Ours (LOOCV)	90.72	93.52	92.05	31,570

Table 2: Mention detection performances.

Coreference annotation is usually carried out in two stages: annotating the mention spans, then linking mentions referring to the same entity together. Given our model's 92.05 F1-score, we consider its performance sufficient to automate the first operation, significantly reducing annotation time.

3.2 Coreference Annotation

Coreference annotation is performed manually, building on the automatically detected mentions. A single annotator reviews the text, assigns entity identifiers to each mention, corrects errors from the mention detection step, deleting spurious mentions, adding missed ones, and adjusting incorrect boundaries. This process yield gold-standard annotations for both mentions and coreference chains.

To assess annotation consistency, we double-annotated a sample from each of the three novels (5,000 tokens per text, 5% of the corpus). Inter-annotator agreement (IAA) was measured for mention spans (F1-score) and coreference chains (MUC, B^3, and $CEAF_e$). Results show high consistency: mention span F1-score of 97.47 (vs. 86.0 in Bamman et al. (2019)), benefiting from our focus on a single, well-defined entity type. Coreference agreement is also high: MUC 96.40, B^3 91.02, and $CEAF_e$ 71.65 (86.36 CoNLL F1). The lower $CEAF_e$ reflects differences in annotator decisions regarding long coreference chains and ambiguous cases such as plural entities leaving room for multiple valid interpretations. These results overall demonstrate the reliability and robustness of our annotations.

To perform annotation we use SACR, an open-source, browser-based interface (Oberle, 2018). This tool meets our requirements, allowing efficient processing of long texts, tracking a large number of entities and handling nested mentions.

Mention detection errors mainly involve difficult cases, such as nested and ambiguous mentions (animals with agentivity, appositions, reflexive pronouns) or other edge cases. It shows the feasibility of leveraging automatic mention detection to accelerate coreference annotation. The manual annotation of a 100k-token text takes around 40 hours.

3.3 Dataset Statistics

Table 3 summarizes statistics from our dataset. The entity spread refers to the distance between the first and the last mention of an entity (Toshniwal et al., 2020). This highlights a key specificity of literary texts, characters can be referred to thousands times over several hundred pages, comprising thousands of tokens.

Average Mentions / Doc.	13,178
Singletons Ratio	1.15%
Coreference Chains / Doc.	159
Average Mentions / Chain	82
Maximum Mentions / Chain	4,932
Average Entity Spread (tokens)	17,529
Maximum Entity Spread (tokens)	115,369
Second-Level Nested Mentions	5.74%
Third-Level Nested Mentions	0.30%
Plural Mentions Ratio	8.13%
Proper Mentions	12.79%
Nominal Mentions	12.26%
Pronominal Mentions	74.95%

Table 3: Dataset statistics summary.

Another important metric for characterizing coreference is the distance to the nearest antecedent (Han et al., 2021). For each mention, we locate the previous mention belonging to the same coreference chain and measure the difference in terms of mention positions. Bamman et al. (2020) analyzed the distribution of distance to nearest antecedent for proper nouns, noun phrases and pronouns. We replicate their experiment and report similar results. While 95% of pronouns appear within 7 mentions of their last antecedent, this distance reach up to 270 mentions for proper nouns and noun phrases.

This observation calls for distinct handling of pronouns, common, and proper nouns during CR. The the last 1% of proper and common noun mentions exhibit a distance of over 1,700 mentions, presenting a significant challenge for CR. See Appendix B for the full distribution of antecedent distances.

3.4 Corpus Merging

Since we followed the guidelines from Mélanie et al. (2024), the newly annotated dataset is fully compatible with the character annotations from the LitBank-fr dataset. It allows us to merge the two datasets, resulting in a combined dataset containing 31 documents and 71,105 character mentions. This decision is motivated by the goal of evaluating generalization across a broader range of texts.

This merged dataset becomes the largest annotated literary coreference dataset in terms of tokens (560,536), average document length (18,081 tokens), and maximum document length (115,415 tokens). Unless otherwise specified, all results presented in this paper pertain to this merged corpus, which we refer to as `Long-LitBank-fr`.

4 Coreference Resolution

Several coreference resolution pipelines are available off-the-shelf, such as the *CoreferenceResolver* module from Spacy[3], Fastcoref (Otmazgin et al., 2022) and AllenNLP (Gardner et al., 2018). BookNLP (Bamman et al., 2020), is a pipeline performing, among other, mentions detection and coreference resolution for English. A French adaptation, BookNLP-fr, was developed by Mélanie et al. (2024) and trained on the LitBank-fr dataset. The BookNLP pipelines implement an end-to-end coreference resolution model (Ju et al., 2018).

Diverging from recent trends of end-to-end architectures, we propose to implement coreference resolution as a modular pipeline, facilitating the study of each component's role and enabling fine-grained error analysis.

Additionally, the use of compact, specialised models ($\sim$15M and $\sim$11M parameters for mention detection and mention scoring models) is motivated by practical end-use considerations: the need to process large literary corpora under limited computational resources. This is further supported by recent critiques of the "bigger-is-better" trend in AI, arguing that simply increasing scale doesn't always lead to better results. Instead, smaller, task-specific models have been shown to offer more sustainable, transparent, and often competitive solutions for domain-specific applications (Varoquaux et al., 2025).

[3] https://spacy.io/api/coref

4.1 Pipeline Description

Our mention-pair-based coreference resolution pipeline is composed of the following modules:

Mention Detection: We employ the mention detection module described in section 3.1, which consists of a stacked BiLSTM-CRF architecture using token-level embeddings from pretrained $\text{CamemBERT}_{\text{LARGE}}$ model as input. We retrained it on the merged corpus, achieving an increase of 2.82 points in F1-score (94.87). As mention detection can impact overall CR performance, we make it possible to bypass the errors introduced by this module by using gold mentions as input to the mention-pair encoder.

Considered Antecedents: To address the quadratic complexity of considering all antecedents, recent approaches introduce hyperparameters to uniformly limit the number of considered antecedents (Thirukovalluru et al., 2021; Wu et al., 2020). Inspired by Bamman et al. (2020) and supported by our observations regarding antecedent distance, we adopt a mention-type-specific approach. We limit the number of antecedents to 30 for pronouns and 300 for proper and common nouns.

Mention Pair Encoder: Mention-pairs are encoded by concatenating the representations of the two mentions with a feature vector that includes attributes such as gender, grammatical person, and the distance between the mentions. For multi-token mentions, the representation is calculated as the average of the first and last tokens embeddings.

Mention Pair Scorer: Encoded mention-pairs are passed into a feedforward neural network trained to predict if two mentions refer to the same entity. Details about the features, model architecture and parameters are provided in the Appendix C.

Antecedent Ranker: Following Wiseman et al. (2015), candidate antecedents are ranked according to their predicted scores. During inference, the highest-scoring antecedent is selected unless all scores fall below 0.5, in which case the null antecedent is assigned.

Entity Clustering: Default strategy for linking mentions into clusters is to scan the document from

left to right, each new mention is either merged into the cluster of its best-ranked antecedent or left as a standalone entity. Coreference chains are defined as the set of mentions in a cluster.

We explore additional strategies to address specific challenges and improve overall performance.

Handling Limited Antecedents: Limiting the number of antecedents can lead to split coreference chains. A common strategy in literary texts is to link all matching proper nouns at the document level, along with their derivatives. While previous works have been using hand-crafted sets of aliases to link proper mentions (Bamman et al., 2020), we leverage local mention-pairs scoring to perform coreference resolution at the document scale. Let's say that all local predictions involving mentions of "Sir Ralph Brown" and "Raphael" are coreferent, we propagate this decision to all mention-pairs at the global scale, bridging the gap between a mention and an antecedent that would otherwise be out of the range of locally considered antecedents.

Leveraging Non-Coreference Predictions: While most mention-pair models focus on coreference links, the cross-entropy loss used during training involves that they are equally trained to predict non-coreference. We propose leveraging high-confidence non-coreference predictions to prevent later incorrect cluster merging. Mention-pairs containing a coordinating conjunction, such as "[Ralph] and [Mr. Delmare]", are a strong indication of non-coreference between these mentions, which can be used to prevent the merging of these entities at document level. This approach is combined with an "easy-first" clustering strategy (Clark and Manning, 2016), which processes mentions in order of confidence rather than left-to-right, thus delaying harder decisions.

The addition of these two strategies is refered to as the *easy-first, global proper mentions coreference approach*. This approach follows a hierarchical iterative process, where high-confidence local mention-pair predictions are resolved first, constraining subsequent decisions at the document level. This post-processing module is not trained.

4.2 Evaluation Metrics

We evaluate CR performance using MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005) scores. For overall performance assessment we report the average F1-score of the three metrics which we refer to as the CoNLL F1-score (Pradhan et al., 2012). We use the scorer implementation by Grobol.[4]

4.3 Document Length

While Poot and van Cranenburgh (2020) investigated the impact of document length on CR by truncating documents to different sizes, we adopt a splitting approach. This allows us to evaluate CR performance on more text excerpts.

Given a target sample size of L tokens, we first select all documents from our corpus that exceed this length. Each document is split into non-overlapping samples, each containing L tokens. CR is performed independently on each sample, and the results are averaged across samples of a given document. The overall CR scores are calculated as the macro-average across all retained documents.

4.4 Coreference Resolution Results

4.4.1 Mention-Pairs Scorer Results

The mention-pairs scorer, evaluated using leave-one-out cross-validation with gold mention spans, achieved an overall accuracy of 88.10%. As shown in Table 4, performance disparities between classes reflect the underlying class imbalance, with significantly higher precision and recall for non-coreferent pairs (class 0). Most errors occurred for mention pairs where the scorer's confidence is low (∼0.5) (Appendix D). As we use the highest ranked antecedent strategy, not all scorer decisions are used during entity clustering, mitigating the number of wrong decisions considered.

Coref.	P	R	F1	Support
0	92.31	93.18	92.74	5.52M (82%)
1	68.49	65.62	67.02	1.25M (18%)

Table 4: Mention-pairs scorer performance on LongLitBank-fr corpus. Precision (P), Recall (R).

4.4.2 Highest Ranked Antecedent

After sorting, the correct antecedent was predicted in 88.05% of cases, highlighting the effectiveness of this approach. Errors occurred for 8,496 mentions (11.95%). In 1,478 cases (2.08%), the range of considered antecedents is too narrow, leaving true antecedents out of reach. For these mentions, the null antecedent is assigned approximately half the time, while an unrelated antecedent is assigned in the other half. In 7,018 cases (9.87%), the true

[4] https://github.com/LoicGrobol/scorch

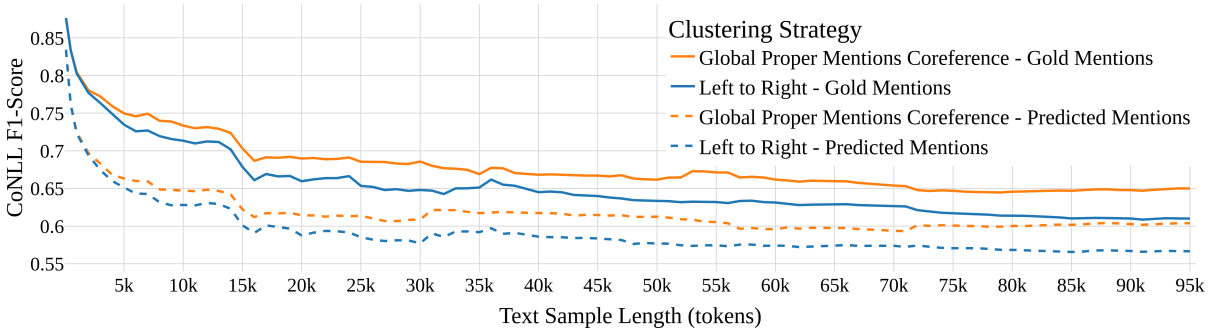

Figure 1: Impact of document length on CR performance for different strategy. Gold and predicted mentions.

antecedent is within reach, but the model incorrectly assigned a different antecedent in nearly 90% of instances. In the remaining 10%, the null antecedent is wrongly predicted.

The additional global proper mentions coreference strategy aims at reducing both types of errors, by bridging the gap between proper mentions and their long distance antecedent, and by limiting clustering of mentions that are believed to be distinct from local mention-pair scores.

4.4.3 Entity Clustering Strategies

The global proper mentions strategy leads to an overall gain in performance measured by CoNLL F1-score of 1.68 points. We observe a slight drop for MUC, but a significant improvement on both B^3 and $CEAF_e$.

Strategy	MUC	B^3	$CEAF_e$	CoNLL
Left to Right	**94.61**	62.95	60.36	72.64
Global Proper CR	94.45	**67.32**	**61.18**	**74.32**

Table 6: Coreference resolution for Long-LitBank-fr corpus. Average F1-scores. Gold mentions.

These scores reflect the overall performance gain of this strategy on the full `Long-LitBank-fr` corpus (averaging 18,081 tokens per document). However, it is best suited to long texts that present both the risk of out-of-reach antecedent, and sufficient local evidence on proper mentions-pairs to propagate document-wide decisions.

4.4.4 Influence of Document Length

When analyzing performance gains as a function of document length, we observe that the MUC score remains relatively stable. For $CEAF_e$ we see a consistent improvement of around 1 point, regardless of document length. The most striking trend is observed on the B^3 score: for documents exceeding 20,000 tokens, the gain from the global proper mentions strategy increases significantly, ranging from 5 to 10 points. See Appendix E.

From Figure 1, we observe that the overall CR performance decreases with document length. Much of the performance loss is observed in the lower range. This might well explain why CR models trained and evaluated on documents of limited length (<10k), have been deceiving when used for downstream tasks on full length documents.

The proper mentions global coreference strategy consistently outperform the vanilla left-to-right method. Performance gains is mostly negligible for short documents (< 2k tokens), but becomes significant and stable beyond, reaching +3 points on the CoNLL F1-score. This shows the effectiveness of our approach for handling CR in longer documents.

Additionally, Figure 1 shows the impact of using predicted mentions as input to the mention-pair encoder, leading to a performance drop of ∼7%, this result is consistent with previous publications.

4.4.5 Comparison to Baseline

For French, our new pipeline consistently outperforms the model proposed by Mélanie et al. (2024) on their test set, setting a new baseline on this specific dataset. We also report average performances on the 3 newly annotated novels for future comparison ; both with gold and predicted mentions.

See Appendix G for cross-dataset and cross-language coreference performance comparison.

While this experiment reveals performance limitations exacerbated by document length, commonly

Corpus (test set)	Model	Mentions	Tokens / Doc	MUC	B^3	CEAFe	CoNLL
LitBank-fr (test-set)	Mélanie et al. 2024	Gold	2,000	88.0	69.2	71.8	76.4
LitBank-fr (test-set)	Ours	Gold	2,000	**92.43**	**70.67**	**75.59**	**79.56**
Long-LitBank-fr (3 docs)	Ours	Gold	93,019	96.64	52.36	46.45	65.15
Long-LitBank-fr (3 docs)	Ours	Predicted	93,019	95.59	45.4	35.95	58.98

Table 5: CR performance on LitBank-fr test-set and on the three fully annotated novels. Gold and predicted mentions.

used CR metrics (MUC, B^3, $CEAF_e$) have been criticised for presenting systematic flaws. Alternative metrics such as LEA (Moosavi and Strube, 2016) and BLANC (Recasens and Hovy, 2011) have been proposed as better aligned with linguistic intuitions. Others argue for extrinsic evaluation (O'Keefe et al., 2013; Vishnubhotla et al., 2023), where CR is assessed based on its contribution to easier to evaluate, downstream tasks.

5 Gender Prediction Case study

As mentioned, studies gravitating around character gender have attracted substantial attention from computational humanities researchers (Underwood et al., 2018). A key challenge is accurately predicting the gender of as many character mentions as possible to ensure representative results.

Early works relied on heuristics to infer gender from explicit clues (he, Mrs, the man), achieving high precision (90%) but lower recall (30-50%), due to the high proportion of ambiguous mentions in literary texts. Recent works leverages CR for broader gender prediction (Vianne et al., 2023).

5.1 Data Preparation

We use the *Long-Litbank-fr* corpus. Starting with all character mentions, we discard singletons (2.74%) and plural mentions (9.84%). We manually annotate the gender of the remaining 62,162 mentions at the entity level. We adopt a binary approach to gender. Works of fiction are subject to play on characters' gender, such as gender revelation or asymmetry of knowledge between characters. To assign character gender we adopt the omniscient perspective (Kim et al., 2024), refering to the knowledge one have at the end of the entire book. We discard chains whose gender cannot be annotated with certainty, leaving us with 804 entities and 61,852 mentions (86.99% of all mentions).

5.2 Prediction Pipeline

To predict the gender of character mentions we implement a multi-stage solution:

Heuristic rules: assign gender based on heuristics from explicit gender clues (pronouns, noun phrases, articles and adjectives).

First-name database: determine the gender of proper mentions using a statistical database of first names given in France since 1900.[5]

[5]French National Institute of Statistics and Economic Studies (*INSEE*).

Coreference propagation: resolve coreference, compute the male/female ratio of processed mentions, and assign the majority gender to all mentions within the coreference chain.

We compare our results with those of Naguib et al. (2022) who used a similar combination of heuristic rules and CR to infer character gender.

5.3 Case Study Results

CR significantly improves recall compared to rule-based methods. While heuristics achieve high precision (>98%), they suffer from low recall (37-47%), reflecting the significant number of mentions whose gender cannot be inferred without additional context. Our approach outperforms the baseline by leveraging sophisticated heuristic rules, a first-names database, and a more effective CR pipeline. Although CR slightly reduces precision—a consequence of clustering errors—the substantial recall gain makes it a robust method overall.

	Male			Female		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline Naguib et al. 2022	95.0	45.0	60.6	97.0	58.0	72.7
Heuristic Rules	**99.8**	37.0	54.0	**98.9**	46.7	63.4
+ First-name data	99.8	38.4	55.4	98.8	47.4	64.1
+ Coreference	95.4	**91.6**	**93.4**	90.4	**93.4**	**91.9**

Table 7: Mentions gender prediction performance (Precision, Recall, F1).

6 Conclusion

We highlight critical limitations in coreference resolution (CR) for literary texts, particularly the scarcity of representative datasets, limiting the possibility to train and evaluate models tailored for literary computational studies. To bridge this gap, we release an annotated corpus of character coreference chains for three full-length French novels spanning three centuries (285,000+ tokens). We introduce a modular CR pipeline tailored for long documents, integrating global coreference propagation for proper nouns and an easy-first clustering approach. After carrying out a detailed error analysis of each component, we study the impact of document length on overall coreference performance. Our approach is competitive with existing state-of-the-art models, demonstrating good performance on longer texts. To demonstrate practical value, we apply it to character gender inference, significantly improving recall over rule-based baselines while maintaining high precision, and outperforming other CR-based approach. This study

underscores the need for robust datasets and well-evaluated models to advance literary CR research.

Limitations

While our dataset is among the largest annotated literary datasets in terms of tokens (285,000), it is limited by the fact that it only contains three documents. This implies that it does not encompass the full diversity of time periods, literary movements, and genres within French literature. This limitation may impact the generalizability of the coreference resolution (CR) models trained on this dataset. The proposed *Long-LitBank-fr* corpus resulting from the concatenation with the *LitBank-fr* dataset mitigates this issue by increasing diversity and improving the potential for model generalization.

Another limitation is that we focused solely on annotating coreference chains for characters. Some downstream applications may require resolving coreference for other entity types (e.g., geographical entities, events). Since our annotations are restricted to characters, a model trained exclusively on this data may not easily transfer to tasks involving other entity types. In such cases, enriching the annotations would be necessary for broader applicability.

Furthermore, our study is limited to French-language texts, and we did not explore cross-lingual generalization of our pipeline. Expanding the dataset to include full documents in other languages could improve its applicability. This could be achieved through annotation transfer or by leveraging multilingual models, which would help reduce the cost of manual annotation.

Finally, while extrinsic evaluation is not the primary focus of this work, we have only begun to assess our pipeline through its application to character gender inference. A more comprehensive evaluation of the models' suitability for full-document literary analysis would require additional extrinsic assessments, such as network extraction or quote attribution.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *International Conference on Computational Linguistics*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018.

AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Dual cache for long document neural coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15272–15285, Toronto, Canada. Association for Computational Linguistics.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], oh my! literary coreference annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Koconovel: Annotated dataset of character coreference in korean novels.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2018. Description of a corpus of character references in german novels - DROC [deutsches ROman corpus].

Aude Leblond. 2022. Corpus chapitres. ANR Chapitres.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 12–23, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot.

2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Giuliano Martinelli, Tommaso Bonomo, Pere-Lluís Huguet Cabot, and Roberto Navigli. 2025. BOOKCOREF: Coreference resolution at book scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24526–24544, Vienna, Austria. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Frédérique Mélanie, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau. 2024. Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. *Journal of Computational Literary Studies*, 3(1):1–34.

Marco Naguib, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, and Olga Seminck. 2022. Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS (male and female novelists : an automatic study of gender of authors and their characters). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 66–77, Avignon, France. ATALA.

Bruno Oberle. 2018. Sacr: A drag-and-drop based tool for coreference annotation. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

Tim O'Keefe, Kellie Webster, James R. Curran, and Irena Koprinska. 2013. Examining the impact of coreference resolution on quote attribution. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 43–52, Brisbane, Australia.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9(Volume 9, 2023):561–587.

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 79–90, Barcelona, Spain (online). Association for Computational Linguistics.

Ian Porada, Xiyuan Zou, and Jackie Chi Kit Cheung. 2024. A controlled reevaluation of coreference resolution models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 256–263, Torino, Italia. ELRA and ICCL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

M. Recasens and E. Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Nat. Lang. Eng.*, 17(4):485–510.

Ina Roesiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138, Santa Fe, New Mexico. Association for Computational Linguistics.

Kumar Shridhar, Nicholas Monath, Raghuveer Thirukovalluru, Alessandro Stolfo, Manzil Zaheer, Andrew McCallum, and Mrinmaya Sachan. 2023. Longtonotes: OntoNotes with longer coreference chains. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1428–1442, Dubrovnik, Croatia. Association for Computational Linguistics.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3921–3931, Online. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Cultural Analytics*.

Andreas van Cranenburgh. 2019. A dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

Joris van Zundert, Andreas van Cranenburgh, and Roel Smeets. 2023. Putting dutchcoref to the test: Character detection and gender dynamics in contemporary dutch novels. In *Proceedings of the Computational Humanities Research conference 2023*, pages 757–771. CEUR Workshop Proceedings (CEUR-WS.org). Computational Humanities Research Conference ; Conference date: 06-12-2023 Through 08-12-2023.

Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker. 2025. Hype, sustainability, and the price of the bigger-is-better paradigm in ai. *Preprint*, arXiv:2409.14160.

Laurine Vianne, Yoann Dupont, and Jean Barré. 2023. Gender Bias in French Literature. In *Conference on Computational Humanities Research CHR2023*, Paris, France. Ariane and Epita and Humanistica.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*, pages 45–52. ACL.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Context-aware machine translation with source coreference explanation. *Transactions of the Association for Computational Linguistics*, 12:856–874.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

A Mention Detection Model

The mention detection module consists of two stacked BiLSTM-CRF models, each trained on a different nesting level of mentions. During inference, predicted spans from both models are combined. If two mention spans overlap, the span with the lower prediction confidence is discarded.

BERT embeddings: The raw text is split into overlapping segments of length L (the maximum embedding model context window) with an overlap of $L/2$ to maximize the context available for each token. Each segment is passed through the $\text{CamemBERT}_{\text{LARGE}}$ model, and we retrieve the last hidden layer as the token representations (1024 dimensions). The final token embedding is computed as the average from overlapping segments. We do not fine-tune CamemBERT for this task.

BIOES tag prediction: For each sentence, token representations are passed through the BiLSTM-CRF model, which outputs a sequence of BIOES tags: B-PER (Beginning of mention), I-PER (Inside), E-PER (End), S-PER (Single-token mention), and O (Outside).

A.1 Model Architecture

- **Locked Dropout** (0.5) applied to embeddings for regularization.
- **Projection Layer**: Highway network mapping $1024 \rightarrow 2048$ dimensions.
- **BiLSTM Layer**: Single bidirectional LSTM (256 hidden units per direction).
- **Linear Layer**: Maps 512-dimensional BiLSTM outputs to BIOES label scores.
- **CRF Layer**: Enforces structured consistency in predictions.

A.2 Model Training

- **Data Splitting**: Leave-One-Out Cross-Validation (LOOCV) with an 85%/15% train-validation split.
- **Batch Size**: 16 sentences per batch.
- **Optimization**: Adam optimizer (lr = 1.4×10^{-4}, weight decay = 10^{-5}).
- **Learning Rate Scheduling**: ReduceLROnPlateau (factor = 0.5, patience = 2).
- **Average Training Epochs**: 20.
- **Hardware**: Trained on a single 6GB Nvidia RTX 1000 Ada Generation GPU.

B Nearest Antecedent Distribution

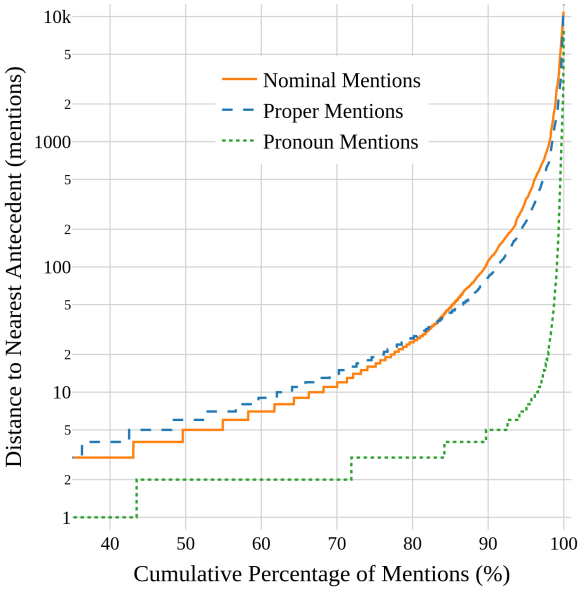

Figure 2: Distance to nearest antecedent for mentions of different type.

C Coreference Resolution Model

C.1 Model Architecture

- **Model Input**: 2,165-dimensional vector, composed of concatenated:
  - **CamemBERT embeddings**: Maximum context embeddings for both mentions ($2 \times 1{,}024 = 2{,}048$ dimensions).
  - **Mention Features** (106 dimensions):
    - Mention length.
    - Position of the mention's start token in the sentence.
    - Grammatical category (pronoun, common noun, proper noun).
    - Dependency relation of the mention's head (one-hot encoded).
    - Gender (one-hot encoded).
    - Number (one-hot encoded).
    - Grammatical person (one-hot encoded).
  - **Mention Pair Features** (11 dimensions):
    - Distance between mention IDs.
    - Distance between start and end tokens of mentions.
    - Sentence and paragraph distance.
    - Difference in nesting levels.
    - Ratio of shared tokens between mentions.
    - Exact text match (binary).
    - Exact match of mention heads (binary).

    - Match of syntactic heads (binary).
    - Match of entity types (binary).
- **Hidden Layers**:
  - Three fully connected layers.
  - 1,900 hidden units per layer with ReLU activation.
  - Dropout rate of 0.6 for regularization.
- **Final Layer**:
  - Linear layer mapping from 1,900 dimensions to a single scalar score.
  - Output: Continuous value between 0 (not coreferent) and 1 (coreferent).

C.2 Model Training

- **Data Splitting**: Leave-One-Out Cross-Validation (LOOCV) with an 85%/15% train-validation split.
- **Batch Size**: 16,000 mention-pairs per batch.
- **Optimization**: Adam optimizer (lr = 1.4×10^{-4}, weight decay = 10^{-5}).
- **Antecedent Candidates**:
  - 30 for pronouns.
  - 300 for common and proper nouns.
- **Hardware**: Trained on a single 6GB Nvidia RTX 1000 Ada Generation GPU.

D Mention-Pairs Scorer Error Distribution

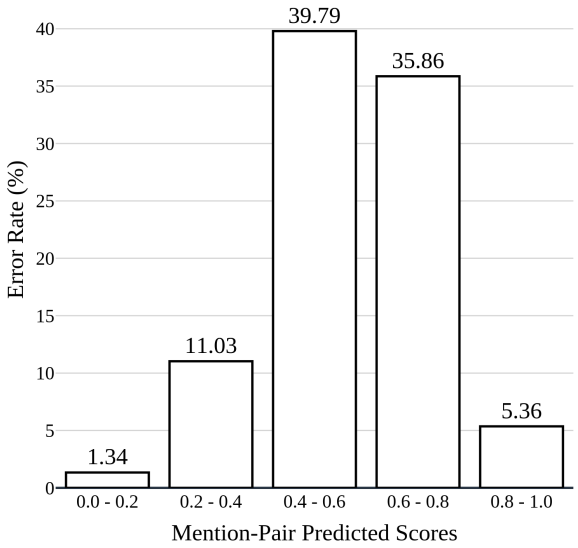

Figure 3: Error Rate by Mention-pair Predicted Score Range.

E Detailed performance gain from clustering strategy

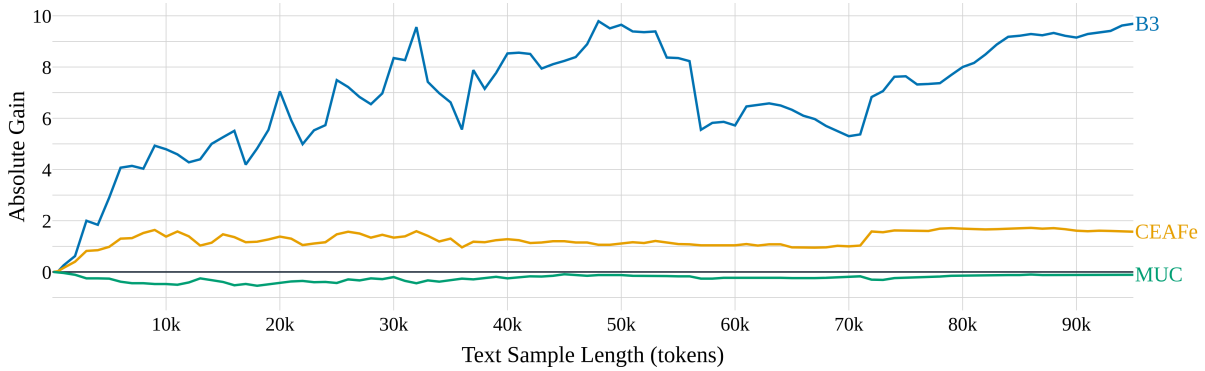

Table 8: Absolute CR performance gain from the global proper mentions clustering strategy over vanilla left-to-right, as a function of document length. Predicted mentions.

F Annotated Dataset Details

Year	Author	Text	Tokens
1731	Antoine-François Prévost	*Manon Lescaut*	71,219
1832	George Sand	*Indiana*	115,415
1923	Delly	*Dans les ruines*	98,542

Table 9: Annotated Dataset Details

G Comparison of CR performance with other datasets and languages

Corpus	Model	Mentions	Tokens / Doc	MUC	B^3	CEAFe	CoNLL
LitBank (*English*)	Bamman et al. 2020	Gold	2,105	88.5	72.6	76.7	79.3
LitBank-fr (LOOCV)	Ours	Gold	2,105	91.93	74.6	75.35	80.63
LitBank (*English*)	Bamman et al. 2020	Predicted	2,105	84.3	62.73	57.3	68.1
LitBank (*English*)	Thirukovalluru et al. 2021	Predicted	2,105	89.50	78.21	67.59	78.44
LitBank-fr (LOOCV)	Ours	Predicted	2,105	84.58	74.77	63.30	73.21
KoCoNovel (*Korean*)	Kim et al. 2024	Predicted	3,578	71.06	57.33	44.19	57.53
Long-LitBank-fr (LOOCV)	Ours	Predicted	3,578	88.31	68.79	47.17	68.09
G. Orwell, *Animal Farm*	Guo et al. 2023	Predicted	37,000	-	-	-	36.3
Long-LitBank-fr (LOOCV)	Ours	Predicted	37,000	92.79	52.35	32.89	59.34
$\text{BookCoref}_{\text{gold}}$	Longdoc	Predicted	76,419	93.5	62.4	45.3	67.0
$\text{BookCoref}_{\text{gold}}$	$\text{Maverick}_{\text{xl}}$	Predicted	76,419	94.3	55.3	33.4	61.0
Long-LitBank-fr (LOOCV)	Ours	Predicted	76,000	94.99	47.51	37.49	60.00

Table 10: Comparison of CR performance with other work on literary coreference with predicted and gold mentions.

Towards Adding Arabic to CorefUD

Dima Taji and Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics (ÚFAL)
Prague, Czechia
{taji, zeman}@ufal.mff.cuni.cz

Abstract

Training models that can perform well on various NLP tasks requires large amounts of data, which becomes even more apparent with more nuanced tasks such as anaphora and coreference resolution. This paper presents the automatic creation of an Arabic CorefUD dataset through the automatic conversion of the existing gold-annotated OntoNotes.

1 Introduction

Coreference resolution is the linguistic task of clustering the different noun phrases that refer to the same entity within a text (Nedoluzhko et al., 2022; Zheng et al., 2011; Elango, 2005). For example, in the sentence "*As Gregor Samsa awoke one morning from uneasy dreams he found himself transformed in his bed into a gigantic insect.*", the mentions *Gregor Samsa*, *he*, *himself*, and *his* all refer to the same entity in the real world.

Recognizing these mentions and clustering them has been shown to improve the performance of different NLP tasks, particularly the tasks that require constructing the meaning of the text, such as opinion target identification (Jakob and Gurevych, 2010), machine translation (Luong and Popescu-Belis, 2016; Miculicich Werlen and Popescu-Belis, 2017), and machine reading comprehension (Huang et al., 2022). Research has also shown that the applications of coreference resolution can extend to tasks in other fields, such as building an ontology in the biomedical domain (Ashury Tahan et al., 2024), improving diversity in rankings (Zhu et al., 2007), and weather forecasting (Belz, 2007).

However, datasets, even for the same language, vary greatly in their format, the phenomena covered, and the way they are annotated. As a result, different state-of-the-art systems are evaluated on different datasets, and as such, evaluations are often not directly comparable (Kobayashi and Ng, 2020). Moreover, using data from different languages to train multilingual systems cannot be achieved without extensive preprocessing of the data to harmonize it.

Multiple efforts have been made to create multilingual coreference-annotated corpora and harmonize existing corpora to follow a unified scheme. Of all these efforts, which will be discussed in Section 2, we are interested in adding an Arabic dataset to the CorefUD (Nedoluzhko et al., 2021, 2022) corpus, following the approach that has been designed to convert the English OntoNotes (Weischedel et al., 2013) dataset to CorefUD.

Due to the morphologically-rich nature of Arabic, we had to modify the existing conversion approach to account for the presence of zero mentions, in addition to converting the annotated data following the same approach used for the conversion of the English data. These decisions are presented in Section 3.

Finally, Section 4 outlines our future plans and concludes.

2 Literature Review

In this section, we present previous efforts that have been made pertaining to the work we are describing in this paper.

2.1 Multilingual Coreference and Anaphora Corpora

Multilingual corpora annotated with coreference and anaphora information are an intuitive solution for creating multilingual and language-agnostic systems. Some of these corpora are limited to languages that are of the same family, such as AnCora (Recasens and Martí, 2010) for Spanish and Catalan, PAWS (Nedoluzhko et al., 2018) for Czech, English, Polish, and Russian, ParCor (Guillou et al., 2014) and ParCorFull (Lapshinova-Koltunski et al., 2022) for English and German, PCEDT (Nedoluzhko et al., 2016) for Czech and English,

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 70–76
Suzhou, China, November 9, 2025.

Wino-X (Emelin and Sennrich, 2021) for English, German, French, and Russian. Others include distant families such as TransMuCoRes (Mishra et al., 2024) for English and 31 South Asian languages, MMC (Zheng et al., 2023) for English, Chinese, and Farsi, and OntoNotes (Weischedel et al., 2011) for English, Chinese, and Arabic.

The issue with this approach is that every corpus follows its own scheme and has a unique set of definitions and annotation guidelines, making the process of adding new languages a costly endeavor in terms of time, effort, and monetary cost. On the other hand, there is a potential for combining these corpora if their schemes and annotations were harmonized, as discussed next.

2.2 Harmonized Schemes

The first effort to a corpus that harmonizes schemes began with the SemEval 2010 shared task on coreference resolution in multiple languages (Recasens et al., 2009), which extracted coreferences from a number of datasets with varying schemes, and represented them in a CoNLL-like format. The MMAX tool (Müller and Strube, 2022) introduced an XML format that can be used to annotate anaphora as well as other linguistic phenomena, and was used for multiple corpora. However, the annotation approaches and the annotated attributes varied greatly between projects. Universal Anaphora (Poesio et al., 2024) proposes a markup scheme for encoding anaphoric information to facilitate the creation of a collection of corpora using the same scheme. Similarly, CorefUD (Nedoluzhko et al., 2021, 2022) addresses the challenges posed by varying data formats and annotation guidelines in existing coreference corpora by creating a unified scheme and format for coreference annotation, facilitating cross-lingual research and development in anaphora and coreference resolution.

Nevertheless, these efforts focus more on unifying the underlying file format, while there is no work being done on the harmonization of linguistic content.

2.3 Arabic Coreference Corpora

Although, as far as we are aware, OntoNotes 5.0 (Weischedel et al., 2013) is the only current multilingual corpus that includes Arabic, there are several coreference corpora for Arabic alone.

Abolohom and Omar (2015) and Abolohom and Omar (2017) use the Quranic corpus, annotated with antecedent references of pronouns. However, since the linguistic structure of Quranic Arabic (QA) is quite distinct from Modern Standard Arabic (MSA), the transfer of the models' knowledge from QA to MSA cannot be directly compared to the experiment results presented in both papers, where their models were evaluated on QA.

Others created their own corpora, which have been used for a limited number of models, such as Mezghani et al. (2009) and Abdul-Mageed (2011).

However, the corpus that made the most sense to be our starting point was OntoNotes (Weischedel et al., 2011, 2013). Since the Arabic portion of the corpus has been used in numerous efforts (Pradhan et al., 2012; Li, 2012; Pradhan et al., 2013; Aloraini et al., 2020; Min, 2021; Aloraini et al., 2022), that indicates that (1) the corpus is popular enough so the momentum of using it to create and test models could be transferred to our new format, and (2) evaluating new systems created with our converted corpus against existing systems would be easy. On the other hand, the downside is that OntoNotes cannot be redistributed freely, which unfortunately affects accessibility of derived works.

2.4 CorefUD

Inspired by the progress achieved by standardizing the labels and annotation guidelines of morphosyntactic labels brought on by Universal Dependencies (Nivre et al., 2020), Nedoluzhko et al. (2022) introduced CorefUD, a collection of corpora with a harmonized scheme that would unify and standardize the annotation of anaphoric and coreference relations.

CorefUD has proved to be beneficial, especially for languages with small training data sets (Pražák et al., 2021; Chai and Strube, 2023), and has been used in four shared tasks focusing on systems for multilingiual coreference resolutions (Žabokrtský et al., 2022, 2023; Novák et al., 2024, 2025). Additionally, the CorefUD format is being used to produce new corpora (Dyer et al., 2024; Jørgensen and Kåsen, 2024). All of these efforts indicate that following this format has the potential to further propel research in the area of anaphora and coreference resolution.

3 Data and Conversion

For this experiment, we used the OntoNotes 5 Arabic dataset (Weischedel et al., 2013). The data set comprises 599 articles with approximately 400K tokens. Since the data is entirely from the Penn

Arabic Treebank, it only contains news articles.

3.1 OntoNotes Annotations and Labels

The annotations contained in OntoNotes are organized in layers, namely treebank, proposition, word sense, ontology, coreference, and named entity. For the purpose of our conversion, and following the approach used by Nedoluzhko et al. (2022), we require the treebank and coreference layers only.

The treebank layer consists of the syntactic annotations of the sentences. These annotations are the parses that are provided by the LDC for the data in the PATB part 3 – v3.1 (Maamouri et al., 2004). This layer is relevant for us because it contains the zero nodes that are discussed in Section 3.2.

For our current purposes, this layer contains the most relevant information. The annotations in this layer connect names, nominal references, and pronouns that refer to the same entity, marking them as coreferents. Similarly, verbs and their equivalent noun phrases are also marked as coreferents. These annotations can span multiple sentences as long as they occur in the same document. Appositions are also marked in this layer. In the Arabic OntoNotes dataset, only 447 articles are annotated for coreference, making a total of 319K annotated tokens.

The OntoNotes tags in the *Coreference* part are the ones that currently appear in our converted files. They include two types; *IDENT* and *APPOS*, and two subtypes; *HEAD* and *ATTRIB*.

- *IDENT* denoting any nominal mentions of the same entity. This is reflected in our dataset by giving the entities the same IDs, without any further elaboration on tags.
- *APPOS* denoting the initial nominal phrase when combined with the *HEAD*, or the referent when combined with the *ATTRIB*.

Figure 1 shows an excerpt from OntoNotes that illustrates the use of these labels.

3.2 Zeros

In pro-drop languages such as Arabic, subject pronouns can be omitted; these omitted subjects are called zero pronouns (Aloraini et al., 2024). However, even when these pronouns are dropped, they can still be part of a coreference chain. As such, in order to identify all the coreference occurrences in a text, zeros must be identified and inserted in their appropriate locations. Additionally, not all zeros need to be part of a coreference chain, and making

```
<COREF ID="64" TYPE="IDENT">
    <COREF ID="66" TYPE="APPOS"
    SUBTYPE="ATTRIB">
```

وزير الدفاع *wzyr AldfAς* ‘Minister of Defense’

```
    </COREF>
    <COREF ID="66" TYPE="APPOS"
    SUBTYPE="HEAD">
```

أنجلو رييس *Ânjlw ryys* ‘Angelo Reyes’

```
    </COREF>
</COREF>
```

Figure 1: An example of OntoNotes’ Coreference annotation showing the tags that are used to identify the nominal mentions. There are two mentions of the same entity connected with an `APPOS` relation and labeled as its `HEAD` and `ATTRIB`, respectively. Additionally, the whole apposition is labeled as a mention in an `IDENT`(ity) coreference relation, whereas `ID="64"` links it to other mentions of that entity elsewhere in the document (not shown here). Arabic transliteration follows the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

this distinction is another task that a coreference resolution model needs to learn.

As the existing conversion approach is based on English, which does not contain zeros, the generated output does not cover this linguistic phenomenon. Fortunately, the PATB annotations included within OntoNotes contain zero nodes, and the subsequent coreference annotations in Arabic OntoNotes take the zeros into consideration.

Per the PATB annotation guidelines (Maamouri et al., 2009), there are five types of zero nodes, four of which appear in the data included in OntoNotes:

- *ICH tag* denoting discontinued constituents, when something interrupts the sentence, without affecting its syntax.
- *T tag* for subjects preceding the verb.
- **0* tag* indicating the existence of a null complementizer or zero WH-pronoun.
- * *tag* indicating the object of a passive verb, the subject of a nominal verb, or an omitted subject of a verb.
- **?* tag* denoting ellipses, which do not appear in the OntoNotes data.

Of the mentioned types of zero node tags that appear in our corpus, the nodes with the * tag are the only ones that represent a coreference relation. The other tags indicate relations that can be realized using Deep UD (Droganova and Zeman, 2019).

ID	Token	Coreference Annotations
3	وزير *wzyr* 'Minister'	Entity=(0001@ann@nw@ar@on_d1sec0c20–1(0001@ann@nw@ar@on_d1sec0c2–1-ATTRIB
4	الدفاع *AldfAς* 'Defense'	Entity=0001@ann@nw@ar@on_d1sec0c2)
5	أنجلو *Ânjlw* 'Angelo'	Entity=(0001@ann@nw@ar@on_d1sec0c2–1-appos:1,HEAD
6	رييس *ryys* 'Reyes'	Entity=0001@ann@nw@ar@on_d1sec0c2)0001@ann@nw@ar@on_d1sec0c20)

Table 1: A segment of the generated CorefUD annotations corresponding to the example shown in Figure 1.

3.3 Output Annotation

Table 1 shows the coreference labels generated by our conversion process that correspond to the example shown in Figure 1. We can see that the tokens belonging to the phrases وزير الدفاع *wzyr AldfAς* 'Minister of Defense' and أنجلو رييس *Ânjlw ryys* 'Angelo Reyes' are annotated with the same entity ID. We can also see that subtypes *ATTRIB* and *HEAD* have been maintained for the heads of each of the phrases.

Table 2 gives a general overview of the size and distribution of clusters and labels in our corpus. The number of unique coreference clusters, i.e. entities with the same Entity identifier, is 12,672, spanning 41,556 tokens. The average cluster size is 3.27, with cluster sizes spanning from 1 to 80 tokens per cluster.

We retained 92% of the zero nodes that appear in the original OntoNotes annotations during our conversion. As previously mention, these are the nodes of the type * which indicate the object of a passive verb, the subject of a nominal verb, or an omitted subject of a verb. The 8% of the zero nodes contained information that we can utilize to provide additional syntactic annotations for the Deep UD treebank (Droganova and Zeman, 2019).

It is worth noting that, according to the OntoNotes Release 5.0 document (Weischedel et al., 2013), the annotated coreferences were limited to only the intra-document occurrences. Nominal mentions, which would be marked with the *IDENT* label excluded all occurrences where the connection between entities can be derived from the use of copula or similar verbs. This is reflected in the rare appearance of this label in the Arabic OntoNotes corpus.

While we cannot directly redistribute the OntoNotes data, our code needed to reproduce the converted output from one's own copy of OntoNotes files will be publicly available.[1]

[1]The conversion code and documentation can be found under our GitHub repository https://tinyurl.com/arabic-corefud

Documents	447
Sentences	30,601
Tokens excluding zeros	299,362
Tokens including zeros	336,735
Tokens including retained zeros	309,631
Tokens part of a coreference cluster	41,556
Coreference clusters	12,672
Minimum cluster size	1
Maximum cluster size	80
Average cluster size	3.27
APPOS labels	1,789
IDENT labels	5
HEAD labels	1,749
ATTRIB labels	1,790

Table 2: Statistics of the converted corpus.

4 Conclusion and Future Work

In this paper, we presented our effort to add Arabic to the CorefUD collection of corpora. We described the decisions we made to modify the existing conversion process to accommodate phenomena that were not in the English corpus, namely the appearance of zero nodes.

Moving forward, we would like to test the quality of multilingual coreference resolution systems when trained on the entirety of CorefUD, including Arabic. Additionally, we plan to prepare a publicly available CorefUD dataset based on UD_Arabic-PADT (Taji et al., 2017). We believe this will be beneficial to furthering research in this area.

Acknowledgments

This work has been supported by the Charles University, project GA UK No. 190125, LINDAT/CLARIAHCZ (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic, and was partially supported by SVV project number 260 821.

References

Muhammad Abdul-Mageed. 2011. Automatic detection of Arabic non-anaphoric pronouns for improving anaphora resolution. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(1):1–11.

Abdullatif Abolohom and Nazlia Omar. 2015. A hybrid approach to pronominal anaphora resolution in Arabic. *Journal of Computer Science*, 11(5):764.

Abdullatif Abolohom and Nazlia Omar. 2017. A computational model for resolving Arabic anaphora using linguistic criteria. *Indian Journal of Science and Technology*, 10(3):1–6.

Abdulrahman Aloraini, Sameer Pradhan, and Massimo Poesio. 2022. Joint coreference resolution for zeros and non-zeros in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 11–21, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdulrahman Aloraini, Juntao Yu, Wateen Aliady, and Massimo Poesio. 2024. A survey of coreference and zeros resolution for Arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Abdulrahman Aloraini, Juntao Yu, and Massimo Poesio. 2020. Neural coreference resolution for Arabic. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Barcelona, Spain (online). Association for Computational Linguistics.

Shir Ashury Tahan, Amir David Nissan Cohen, Nadav Cohen, Yoram Louzoun, and Yoav Goldberg. 2024. Data-driven coreference-based ontology building. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14290–14300, Miami, Florida, USA. Association for Computational Linguistics.

Anja Belz. 2007. Probabilistic generation of weather forecast texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 164–171, Rochester, New York. Association for Computational Linguistics.

Haixia Chai and Michael Strube. 2023. Investigating multilingual coreference resolution by universal annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10010–10024, Singapore. Association for Computational Linguistics.

Kira Droganova and Daniel Zeman. 2019. Towards deep Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152, Paris, France. Association for Computational Linguistics.

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska, Syahidah Asma Umniyati, and Helena Rodrigues Menezes de Oliveira Vaz. 2024. A multilingual parallel corpus for coreference resolution and information status in the literary domain. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 55–64, Hamburg,Germany. Association for Computational Linguistics.

Pradheep Elango. 2005. Coreference resolution: A survey. *University of Wisconsin, Madison, WI*, 1(12):12.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3191–3198, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Baorong Huang, Zhuosheng Zhang, and Hai Zhao. 2022. Tracing origins: Coreference-aware machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1281–1292, Dublin, Ireland. Association for Computational Linguistics.

Niklas Jakob and Iryna Gurevych. 2010. Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 263–268, Uppsala, Sweden. Association for Computational Linguistics.

Tollef Emil Jørgensen and Andre Kåsen. 2024. Aligning the Norwegian UD treebank with entity and coreference information. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 704–710, Torino, Italia. ELRA and ICCL.

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. ParCorFull2.0: a parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813, Marseille, France. European Language Resources Association.

Baoli Li. 2012. Learning to model multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 129–135, Jeju Island, Korea. Association for Computational Linguistics.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Mohamed Maamouri, Ann Bies, Sondos Krouna, Fatma Gaddeche, and Basma Bouziri. 2009. Penn Arabic Treebank Guidelines. *Linguistic Data Consortium*.

Souha Mezghani, Lamia Belguith, and Abdelmajid Ben Hamadou. 2009. Arabic anaphora resolution: Corpora annotation with coreferential links. *Int. Arab J. Inf. Technol.*, 6:480–488.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Bonan Min. 2021. Exploring pre-trained transformers and bilingual transfer learning for Arabic coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 94–99, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ritwik Mishra, Pooja Desur, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2024. Multilingual coreference resolution in low-resource South Asian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11813–11826, Torino, Italia. ELRA and ICCL.

Mark-Christoph Müller and Michael Strube. 2022. Annotating anaphoric and bridging relations with mmax. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. September 1 - 2, 2001, Aalborg, Denmark*, page 6.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, and Maciej Ogrodniczuk. 2018. PAWS: A multi-lingual parallel treebank with anaphoric relations. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 68–76, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. Is one head enough? mention heads in coreference annotations compared with UD-style heads. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 101–114, Sofia, Bulgaria. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025. Findings of the fourth shared task on multilingual coreference resolution: Can llms dethrone traditional approaches? In *Proceedings of the Joint Sixth Workshop on Computational Approaches to Discourse (CODI) and Eigth Computational Models of Reference, Anaphora and Coreference (CRAC)*, Suzhou, China.

Massimo Poesio, Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, Amir Zeldes, Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Universal anaphora: The first three years. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17087–17100, Torino, Italia. ELRA and ICCL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Marta Recasens and M Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, 44:315–345.

Marta Recasens, Antonia Martí, Mariona Delor, Lluís Màrquez, and Emili Sapena. 2009. Semeval-2010 task 1. page 70.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Ralph Weischedel, Hovy Eduard, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Hovy Eduard, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, and Benjamin Van Durme. 2023. Multilingual coreference resolution in multiparty dialogue. *Transactions of the Association for Computational Linguistics*, 11:922–940.

Jiaping Zheng, Wendy W Chapman, Rebecca S Crowley, and Guergana K Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6):1113–1122.

Xiaojin Zhu, Andrew Goldberg, Jurgen Van Gael, and David Andrzejewski. 2007. Improving diversity in ranking using absorbing random walks. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 97–104, Rochester, New York. Association for Computational Linguistics.

Exploring Coreference Resolution in Glosses of German Sign Language

Yuzheng Bao and Haixia Chai
Department of Computing Science
School of Natural and Computing Sciences
University of Aberdeen
{y.bao.21, haixia.chai}@abdn.ac.uk

Abstract

In recent years, research on sign languages has attracted increasing attention in the NLP community and requires more effort from a linguistic perspective. In this paper, we explore coreference resolution in German Sign Language (GSL) primarily through gloss-based analysis. Specifically, in GSL glosses, we conduct a linguistic analysis of coreference, add coreference annotations based on three videos, and evaluate the ability of two large language models to resolve coreference. We gain valuable insights into coreference resolution in GSL, which pave the way for future research.[1]

1 Introduction

Natural language develops naturally for daily communication among humans. As a first language for deaf and hearing impaired individuals, sign languages (e.g., American Sign Language and German Sign Language) are visual-spatial natural languages with their own sophisticated linguistic systems—including lexicon, morphology, phonology, syntax, and pragmatics—not only gestures (Stokoe, 1980). The complexity of signs lies in the fact that they can be articulated through multiple phonological units, including handshape, palm orientation, position, and facial expressions (e.g., eyebrow movement and head motion) within a three-dimensional signing space (Herrmann and Steinbach, 2011; Michael et al., 2011). Compared to spoken language, sign language remains underexplored and demands linguistic insights from the Natural Language Processing community (Yin et al., 2021b).

To record and analyse signs in written form, glosses annotated by experts are used as linear labels that approximate the semantic meaning of each sign, typically using the base form of a corresponding word in spoken language. Table 1 presents an example of glosses used in German Sign Language. While glosses are widely used as an intermediate step in the study of sign language translation, from signs to spoken text (Müller et al., 2023; Fayyazsanavi et al., 2024), and sign language production, from spoken text to signs (Varanasi et al., 2024; Fang et al., 2024), they can also support documentation, education, and linguistic research.

GEBÄRDEN1A LEHRER5 ICH1 BIS-HEUTE2 $GEST-OFF^*
(To this day, I am still a sign language teacher.)

Table 1: An example text shows the glosses and their English translation. Glosses are written in capital letters and may include affixes or markers indicating additional information.

In this paper, we explore coreference—a linguistic phenomenon crucial for natural language understanding—in German Sign Language, primarily through gloss-based analysis. Although glosses—one-dimensional sequences of words—cannot fully capture spatial constructions or represent the meaning of signs through various cues, such as non-manual features (Yin and Read, 2020; Müller et al., 2023), we focus on coreference in glosses as an initial step towards the Signed Coreference Resolution task (Yin et al., 2021a). To our knowledge, Yin et al. (2021a) is the only recent work that addresses coreference resolution in sign language, specifically for pronominal indexing signs. Following the work, we study entity coreference resolution in German Sign Language using the Public DGS Corpus (Hanke et al., 2020). Our contributions are threefold:

- A linguistic analysis of signed coreference in glosses, including noun phrase, pronoun, ellipsis, and others.
- Entity coreference annotations on the glosses of three videos from the DGS corpus.

[1] Our annotated data are publicly available at https://github.com/orcastimulatee/Coref_GLS_GSL.git

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 77–84
Suzhou, China, November 9, 2025.

- Evaluating the ability of GPT-4o (OpenAI, 2024) and DeepSeek-v3 (DeepSeek-AI, 2025) to perform coreference resolution on glosses with coreference gold annotations using prompt engineering.

2 Related Work

2.1 Coreference Resolution in Sign Language

The coreference phenomenon in sign languages has previously been studied (Steinbach and Onea, 2015), e.g., in American Sign Language (Kegl, 1987) and in German Sign Language (Wienholz et al., 2018). More recently, Yin et al. (2021a) introduced the Signed Coreference Resolution task for Sign Language Processing, thereby presenting a new challenge for the coreference research community. The work proposes DGS-Coref, a gloss-based dataset derived from the DGS corpus (Hanke et al., 2020) with coreference annotations. Similarly, to reduce the overhead of visual processing, we conduct annotations on glosses as well. The main difference is that Yin et al. (2021a) focuses solely on pronominal coreference relations, whereas our annotations cover all types of entity coreference. The authors (Yin et al., 2021a) also propose a linguistically informed unsupervised coreference resolution model for the task, using both glosses and spatial features extracted from pose estimations.

2.2 Large Language Models

Many studies focus on transformer-based large language models (LLMs) for sign language translation and production tasks (Camgoz et al., 2020; Yin and Read, 2020; Fang et al., 2024), aiming to make them accessible to deaf and signing communities. However, the extent to which an LLM truly understands the underlying structure and meaning of signed discourse remains unclear. Gan et al. (2024) examines the coreference resolution abilities of several LLMs (e.g., the GPT series and the LLaMA2 family) in English by using prompts and conducting both manual and automatic analyses. In contrast, our work evaluates LLMs in the context of sign language. Coreference in sign language involves manual features, non-manual features, and spatial referencing, which differ significantly from the devices used in spoken languages.

3 Coreference in Sign Language

In sign language, a signer can introduce a new entity into the discourse not only through explicit signs that visually depict its shape, orientation, or movement, but also by assigning it a spatial locus within the signing space in front of the signer, which can later be used to refer back to the entity. To understand how to track an entity, we conduct a basic analysis of the Public DGS Corpus.

Noun Phrase. Iconicity is one of the prominent properties of sign language (Perniss et al., 2010). For example, `$PROD SCHWIMMEN` (*swimming*) is a gloss for a productive sign used to represent an inanimate entity—swimming. It illustrates the backstroke, a swimming style, with alternating backward circular motions of both hands. It is worth noting that *backstroke* is not annotated in glosses, but only in the mouthing unit (i.e., it is coarticulated with mouthed German words for lip reading). This means that mentions referring to the same entity can appear in different units, making coreference resolution a task in a multidimensional space. Beyond standard glosses, compound glosses, e.g., `TAUB-GEHÖRLOS` (*deaf*), consist of two or more glosses connected by hyphens to express meaning more efficiently. A sequence of glosses is not annotated as a full German sentence but rather as a chunk of a sentence that conveys a core meaning. Therefore, noun phrases may be omitted from either the subject or object position, regardless of word order.

Pronoun. `$INDEX` gloss represents an indexical sign (a pointing gesture) that refers to a spatial locus associated with a lexical sign, thus defining a referent. Subsequent pronouns refer back to the referent by pointing to the same locus. It is clear when only two referents are assigned to the left-hand and right-hand areas, respectively. However, if many referents need to be assigned within a limited signing space, the loci can become very close together, making them difficult to differentiate and potentially causing ambiguity in entity resolution. Adding to the complexity, a signer may relocate a referent to a new locus (Yin et al., 2021a). In the DGS corpus, following the gloss annotation convention (Konrad et al., 2018), a number is added as a suffix to the end of a gloss. For example, `ICH1` (*I1*) and `ICH2` (*I2*) are two contextual variants of `ICH` (*I*), though the basic meaning remains the same. In our annotation presented in §4, we treat these two glosses as referring to the same entity. Unlike in English, pronouns in sign language are not morphologically marked for gender (i.e., there are no separate signs for *he* or *she*). This means that gender must be inferred from the discourse

context. We observe that, sometimes, pronouns can be replaced with an iconic sign. For example, a signer may use both hands with palms facing each other, circling slightly from the sides towards the center to express the concept of *together*, visually representing a group of people, rather than signing *us* directly. As a result, the use of pronouns is, to some extent, reduced.

Ellipsis. Ellipsis is a common phenomenon in sign languages, including but not limited to German Sign Language. It arises from features such as spatial referencing and role shift—a linguistic device that marks different characters through body shift, eye gaze, and head orientation (Proske et al., 2020)—allowing signers to omit overt pronouns or explicit entity names. Moreover, many sign languages are topic-prominent, e.g., `WASSER1 $INDEX1 FRISCH1 IMMER4A*` (*Yes, you always feel refreshed when you are in the water.*), meaning that a topic or entity (i.e., *water*) is typically introduced first in discourse. This entity may later be omitted from subject or object positions, if it is predictable and inferable from context. In Table 2, we present another example of ellipsis, in which even more expressions are not explicitly signed. Signer A was talking about playing ninepin bowling in the earlier context. Signer B inferred that the reason for the knee pain was kneeling down too much, so only the corresponding sign, `HINKNIEN-SICH1` (*kneel-down*), was used, omitting other non-essential signs. Resolving and recovering such elliptical constituents can undoubtedly benefit sign language understanding and gloss-based studies of sign language translation. In spoken language, many linguistic theories have been studied in relation to zero pronouns and focus, including topic chain theory (Tsao, 1977; Zhang et al., 2022) and centering theory (Joshi and Weinstein, 1981; Grosz et al., 1983, 1995; Walker et al., 1998; Chai and Strube, 2022). We raise a question of whether these theories could aid in resolving implicit expressions in sign language, which we leave for future work.

A:	`JETZT1* ICH2 KNIE1A* SCHMERZ3 $GEST-OFF^* RÜCKEN-UNTEN1E SCHMERZ3` *(Now I have knee and back pain.)*
B:	`HINKNIEN-SICH1` *(Well, because you had to kneel down a lot.)*

Table 2: An example illustrating ellipsis in a dialogue between two signers.

Others. In the signing space, verb inflection—through modified movements that match the loci of the subject and object to indicate agreement—can help track entities. However, glosses are written in their base form, and verbs are not morphologically inflected in the gloss itself. To this end, cues from visual processing become especially important for discourse understanding and entity resolution.

4 Human Annotation

To examine the current ability of LLMs to resolve coreference in sign language glosses (see §5), we conducted coreference annotation on the glosses of three videos totaling 990 seconds from the DGS corpus[2][3][4]. Some gloss names are followed by numerical or alphabetical suffixes to distinguish lexical and phonological variants (Konrad et al., 2018). These glosses were annotated carefully by considering the context, the English translation, and by watching the video. Entities are annotated throughout the entire duration of each video, and singletons are excluded from the annotation. We have two annotators with backgrounds in computer science and computational linguistics, both of whom have knowledge of German Sign Language. Inter-annotator agreement was measured using Krippendorff's α (Krippendorff, 1980), resulting in a high score of 0.93. For the annotations with disagreements, the annotators discussed and reached final decisions for the gold annotations. Ambiguous cases were excluded. We release the annotated data as a JSON file.[5] Table 3 shows the statistics of our annotations.

	Video1	Video2	Video3
#mentions	166	80	90
#noun phrases	78	51	42
#pronouns	63	9	29
#$INDEX	25	20	19
#entities	36	26	27
#mentions/entities	4.6	3.0	3.3

Table 3: Statistics of the annotated data. Video1, Video2, and Video3 refer respectively to the links in the footage[2][3][4].

[2] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1429737_en.html

[3] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1183720-17021701-17054739_en.html

[4] https://www.sign-lang.uni-hamburg.de/meinedgs/html/1182135_en.html

[5] Our annotated data are publicly available at https://github.com/orcastimulatee/Coref_GLS_GSL.git

	MUC			B^3			$CEAF_e$			CoNLL		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
DS-v3_zs	67.60	64.82	66.10	64.65	59.58	61.96	67.23	53.83	59.80	66.49	59.41	62.62
DS-v3_fs	**76.20**	54.58	63.60	**72.79**	50.65	59.70	71.39	53.50	61.18	**73.46**	52.91	61.49
GPT-4o_zs	64.30	70.85	67.38	58.81	66.20	62.26	65.39	52.09	57.98	62.83	63.05	62.54
GPT-4o_fs	74.50	**70.95**	**72.70**	71.68	**67.10**	**69.30**	**73.42**	**57.54**	**64.50**	73.20	**65.20**	**68.83**

Table 4: Performance on the annotated data in §4. *zs* and *fs* denote the zero-shot and few-shot settings for the two LLMs: DeepSeek-v3 (DS) and GPT-4o. Bold numbers indicate the highest score in each column.

5 Prompt Engineering

We employ prompt engineering in zero-shot and few-shot settings (Brown et al., 2020; Liu et al., 2023) to evaluate two LLMs: GPT-4o (OpenAI, 2024) and DeepSeek-v3 (DeepSeek-AI, 2025). Figure 1 shows the prompt template, which includes an instruction, input glosses, and one example in the few-shot setting. For few-shot prompting, we select examples that are similar to the cases in the input glosses under examination to provide the LLMs with additional cues for coreference resolution. Additionally, LLMs are required to provide explanations of their resolved results to enable further manual analysis of their performance. One example prompt is presented in Appendix A.

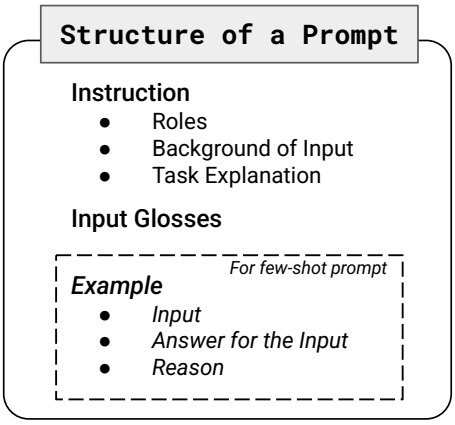

Figure 1: A prompt template shows each components of the prompt. The dashed box applies only to the few-shot setup.

6 Experiment

6.1 Setup

We conduct experiments on the annotated data (see §4). To obtain stable and reliable responses from LLMs, we divided the data into 20 prompts, grouped by topic for potential further analysis. GPT-4o and DeepSeek-v3 are prompted with a temperature of 0 and a maximum token limit of 5,025 via an AI model API platform[6]. We report macro-averaged results for the 20 prompts in both zero-shot and few-shot settings for the two LLMs. The evaluation uses the CoNLL F1 score, which averages MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAFe (Luo, 2005).

6.2 Results and Analyses

Table 4 shows that *GPT-4o_fs* achieves the best F1 score of 68.83, which is more than 7 points higher than *DS-v3_fs* and 6 points higher than *GPT-4o_zs*. This indicates that providing an example containing gold coreference annotations improves *GPT-4o*'s overall performance. We also observe that *DS-v3_fs* has high recall (73.46) but low precision (52.91), resulting in a lower CoNLL F1 score. This suggests that *DS-v3_fs* resolves many entities, but few of them are correct. Overall, the two LLMs demonstrate moderate performance on our annotated data. It is important to note that the experiments are conducted on glosses composed of words from spoken language, which are used for training the LLMs. How well they can perform directly on signs remains a worthwhile question for future research.

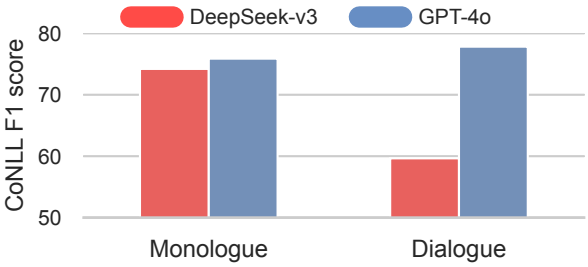

Figure 2: CoNLL F1 scores of the two LLMs in the few-shot setting across different genres.

Figure 2 presents the performance of the two LLMs on the first video, comprising 8 monologue prompts and 5 dialogue prompts. It shows that *GPT-4o_fs* performs better in dialogue than in monologue. Based on our manual analysis of the model

[6] https://aimlapi.com/

responses, we found that *GPT-4o* can understand the conversation well and can resolve *ICH* (*I*) from Signer A and *DU* (*you*) from Signer B as referring to the same entity. However, *DS-v3_fs* struggles to resolve some entities in dialogue, even when provided with an example and its explanation in the few-shot setting.

We also perform an analysis of some entities that are not resolved successfully. Specifically, for the entity in Table 5, we observe that *DS-v3_fs* can capture the hint from the selected example that is similar to the queried input and correctly resolve the entity. However, *GPT-4o_fs* ignores the example despite our various attempts and fails to resolve it. This suggests that, in this specific case, *GPT-4o_fs* relies more on its internal knowledge and reasoning, making it less influenced by the provided examples, whereas *DS-v3_fs* appears more receptive to such guidance.

A:	$LIST1:1of1d KEGELN1 $LIST1:2of2d SCHERE1* $LIST1:3of3d $NUM-EINER1A:3d BAHN-WEG1A* $GEST-OFF^* $LIST1:4of4d ASPHALT1* *(For example Bohle, Schere, three lane alleys and classic.)*
B:	$ORAL^KANN1 ALLES1A *(Can you play all four disciplines?)*

Table 5: A snippet of glosses between two signers with coreference annotations shown in red.

7 Discussions and Conclusions

While analysing the properties of coreference in sign language, our study is primarily based on linear glosses. Further research is needed in coreference resolution modeling—especially for understanding spatial relations—or in developing enhanced glosses that recover omitted elements to support downstream tasks. In this paper, as an initial step, we gain linguistic insights into coreference, annotate glosses accordingly, and evaluate the coreference resolution abilities of two LLMs, to foster future advancements.

Limitations

Our study is conducted on written glosses, which may omit some information (e.g., from mouthings or productive signs), and therefore the experiments on coreference resolution may not fully reflect a natural signing scenario. Due to limited resources, no deaf people or sign language users were involved in the annotation process.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and constructive feedback, which greatly improved the clarity and quality of this work.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation*, pages 563–566, Granada, Spain.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Haixia Chai and Michael Strube. 2022. Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2996–3002, Seattle, United States. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2024. Signllm: Sign language production large language models. *arXiv preprint arXiv:2405.10718*.

Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Kosecka. 2024. Gloss2Text: Sign language gloss translation using LLMs and semantically aware label smoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16162–16171, Miami, Florida, USA. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1645–1665, Torino, Italia. ELRA and ICCL.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Annika Herrmann and Markus Steinbach. 2011. Nonmanuals in sign languages. *Sign Language & Linguistics*, 14(1):3–8. Publisher: John Benjamins.

Aravind K. Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the IJCAI, Vancouver, CA*, pages 385–387.

Judy Kegl. 1987. *Coreference Relations in American Sign Language*, pages 135–170. Springer Netherlands, Dordrecht.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2018. Public DGS Corpus: Annotation Conventions. *Technical report, Project Note AP03–2018-01, DGS-Korpus project*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. CA: Sage Publications, Beverly Hills.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Nicholas Michael, Peng Yang, Qingshan Liu, Dimitris Metaxas, and Carol Neidle. 2011. A framework for the recognition of nonmanual markers in segmented sequences of american sign language. In *Proceedings of the British Machine Vision Conference*, pages 124.1–124.12. BMVA Press.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. ArXiv:2410.21276.

Pamela Perniss, Robin L. Thompson, and Gabriella Vigliocco. 2010. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227.

Sina Proske, Annika Herrmann, Jana Hosemann, and Markus Steinbach. 2020. *A Grammar of German Sign Language (DGS)*, 1 edition. SIGN-HUB Sign Language Grammar Series. SIGN-HUB. Accessed 31-10-2021.

Markus Steinbach and Edgar Onea. 2015. A drt analysis of discourse referents and anaphora resolution in sign language. *Journal of Semantics*, 33(3):409–448.

William C Stokoe. 1980. Sign language structure. *Annual review of anthropology*, pages 365–390.

Fengfu Tsao. 1977. *A Functional Study of Topic in Chinese: The First Step Towards Discourse Analysis*. Ph.d. dissertation, University of Southern California.

Abhishek Bharadwaj Varanasi, Manjira Sinha, Tirthankar Dasgupta, and Charudatta Jadhav. 2024. Linguistically informed transformers for text to American Sign Language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 50–56, Bangkok, Thailand. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. *Centering Theory in Discourse*. Oxford University Press.

Anne Wienholz, Derya Nuhbalaoglu, Nivedita Mani, Annika Herrmann, Edgar Onea, and Markus Steinbach. 2018. Pointing to the right side? An ERP study on anaphora resolution in German Sign Language. *PloS one*, 13(9):e0204223.

Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021a. Signed coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 4950–4961, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021b. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shulin Zhang, Jixing Li, and John Hale. 2022. Quantifying discourse support for omitted pronouns. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–12, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Prompts

A.1 An Example Prompt

Figure 3 shows a prompt we used in the experiment for *GPT-4o_zs*, along with the LLM's response. Note that the answer is summarized from the LLM's raw responses, which we manually checked.

Instruction:

Role: You are a sign language coreference resolution assistant.
Input: a list of glosses, speaker A & B are having a conversation, sentences(words separated by comma or full stop) are separated by """, each gloss with a unique gloss ID.
1. Identify which glosses refer to entities.
2. Cluster glosses that refer to the same entity.

Input Glosses:

"# Speaker A"
"""$GEST-NM-KOPFSCHÜTTELN1^#265 || KAMPF1A#266 || WILLE6#267 || ICH1*#268 || HIN-UND-HER1*#269"""

"# Speaker B"
"""KAMPF1A#270 || $ORAL^#271 || TRAINING1#272 || HIN-UND-HER1#273 || SPASS1*#274"""

"# Speaker A"
"""$GEST-NM-KOPFNICKEN1^#275 || ICH1*#276 || SPASS1#277 || WASSER1#278 || $INDEX1#279 || FRISCH1#280 || IMMER4A*#281"""

"# Speaker B"
"""ICH1#282 || ICH1#283 || LAND1A*#284 || BENUTZEN1#285 || FLACH1^#286 || ICH1#287 || $PROD*#288 || BESSER1*#289 || ALS4#290 || WASSER11#291 || $INDEX1#292 || $PROD*#293"""

"# Speaker A"
"""SPORT4B#294 || $PROD*#295 || GUT1#296 || $INDEX1#297 || SCHWIMMEN1#298 || GUT1#299"""

"# Speaker B"
"""FLEISSIG1#300 || $INDEX1#301 || FLEISSIG1#302 || ICH2#303 || FAUL1#304 || ICH2#305 || $ORAL^#306"""

"# Speaker A"
"""ICH1*#307 || GERN1#308 || ICH2#309 || HEUTE1#310 || NOCH4A*#311 || GERN1#312"""
"""$GEST-OFF^*#313 || NUR4*#314 || PRIVAT1A#315 || FREI1#316 || FREIZEIT1B*#317 || SCHWIMMEN1#318 || GERN1#319 || DORTHIN-GEHEN1#320 || $ORAL^#321 || KAMPF1A#322 || SCHLUSS1#323"""
"""$GEST-OFF^#324"""

GPT-4o's Answer:

Assistant: 1. Identify glosses that refer to entities:
- ICH1*#268, ICH1*#276, ICH1#282, ICH1#283, ICH1#287, ICH1*#307, ICH2#303, ICH2#305, ICH2#309 - KAMPF1A#266, KAMPF1A#270, KAMPF1A#322
- HIN-UND-HER1*#269, HIN-UND-HER1#273
- SPASS1*#274, SPASS1#277
- WASSER1#278, WASSER11#291
- $INDEX1#279, $INDEX1#292, $INDEX1#297, $INDEX1#301
- $PROD*#288, $PROD*#293, $PROD*#295
- FREIZEIT1B*#317
- SCHWIMMEN1#298, SCHWIMMEN1#318

2. Cluster glosses that refer to the same entity:
- [ICH1*#268, ICH1*#276, ICH1#282, ICH1#283, ICH1#287, ICH1*#307]
- [ICH2#303, ICH2#305, ICH2#309]
- [KAMPF1A#266, KAMPF1A#270, KAMPF1A#322]
- [HIN-UND-HER1*#269, HIN-UND-HER1#273]
- [SPASS1*#274, SPASS1#277]
- [WASSER1#278, WASSER11#291]
- [$INDEX1#279, $INDEX1#292, $INDEX1#297, $INDEX1#301]
- [$PROD*#288, $PROD*#293, $PROD*#295]
- [SCHWIMMEN1#298, SCHWIMMEN1#318]

Figure 3: An example prompt shows the instruction, input glosses, and the answer from *GPT-4o_zs*.

Impact of ASR Transcriptions on French Spoken Coreference Resolution

Kirill Milintsevich
Institut National de l'Audiovisuel (INA)
France
kmilintsevich@ina.fr

Abstract

This study introduces a new ASR-transcribed coreference corpus for French and explores the transferability of coreference resolution models from human-transcribed to ASR-transcribed data. Given the challenges posed by differences in text characteristics and errors introduced by ASR systems, we evaluate model performance using newly constructed parallel human-ASR silver training and gold validation datasets. Our findings show a decline in performance on ASR data for models trained on manual transcriptions. However, combining silver ASR data with gold manual data enhances model robustness. Through detailed error analysis, we observe that models emphasizing recall are more resilient to ASR-induced errors compared to those focusing on precision. The resulting ASR corpus, along with all related materials, is freely available under the CC BY-NC-SA 4.0 license at: https://github.com/ina-foss/french-asr-coreference.

1 Introduction

Coreference resolution differs between written and spoken texts and is generally more challenging for spoken data, primarily because most existing corpora are based on written texts (Amoia et al., 2012). For the French language, the large-scale coreference corpus ANCOR (Muzerelle et al., 2014) is based on interview transcripts that were produced manually (Antoine et al., 2002; Eshkol-Taravella et al., 2011). With the help of recent state-of-the-art automatic speech recognition (ASR) systems, such as Whisper (Radford et al., 2023), we can automatically transcribe large amounts of audio data. For example, the *Institut national de l'audiovisuel* stores millions of hours of recorded French TV and radio broadcasts, which are continuously and automatically transcribed and used for research in the social sciences and digital humanities.

However, unlike the manual transcripts in ANCOR, Whisper produces text that includes punctuation, capitalization, occasional rephrasing, as well as ASR errors. This might lead to poor transferability of coreference resolution models trained on the ANCOR corpus when applied to ASR data.

To date, most studies on transferability in coreference resolution have focused on cross-corpus (Xia and Van Durme, 2021; Yuan et al., 2022) and cross-lingual (Lai and Ji, 2023; Pražák et al., 2024) transferability. However, pre-trained language models are sensitive even to small text perturbations, such as punctuation (Wang et al., 2023) and casing (Moradi and Samwald, 2021). Moreover, ASR errors have negative impact on downstream tasks, such as named entity recognition (Szymański et al., 2023) or spoken language understanding (Chang and Chen, 2022). Since these models are at the heart of most recent automatic coreference resolution models, such sensitivities might hinder their performance when resolving coreference on ASR texts.

In this study, we evaluate the transferability of coreference resolution models from human-transcribed to ASR-transcribed data. We create parallel silver training and gold validation datasets and conduct a comparative study using two distinct architectures. Finally, we perform a detailed error analysis to identify the types of ASR-induced errors that most affect model performance.

2 Automatic Coreference Resolution

Most widely used end-to-end coreference resolution systems are *mention-to-link*, meaning they first predict candidate mentions—phrases referring to some entity—and then establish coreference or anaphoric links between each pair of candidates. Lee et al. (2017) developed a model that lists all overlapping spans of a certain length as possible mention candidates. However, this approach incurs high computational overhead. Subsequently, Kirstain et al. (2021) reduced the computational

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 85–94
Suzhou, China, November 9, 2025.

complexity by using only the start and end tokens to construct the mention representation. CorPipe uses a similar approach by first predicting all mentions using a sequence tagging approach and then establishing coreference links with a self-attention layer (Straka, 2024). This system has repeatedly shown top performance at the CRAC Shared Task on coreference resolution (Novák et al., 2024).

Another approach to automatic coreference resolution is the *link-to-mention* approach, where anaphoric links are first predicted between the syntactic heads, and then the mention spans are reconstructed from the coreferent heads. This approach reduces computational overhead compared to the mention-to-link approach, as constructing span representations is unnecessary. Dobrovolskii (2021) presented WL-Coref, the first model that followed the link-to-mention approach. However, D'Oosterlinck et al. (2023) found that in the case of conjunctions of multiple mentions, the same syntactic head could correspond to multiple mention spans, leading to errors in the model of Dobrovolskii (2021). Subsequently, D'Oosterlinck et al. (2023) proposed moving the syntactic head to the coordinating conjunction instead (e.g., in a mention [Tom and Mary], the head is moved from "Tom" to "and"). Finally, Liu et al. (2024) proposed another iteration of the WL-Coref model, which added a special "antecedent link" to support singletons.

3 Data

The ANCOR corpus is the largest collection of spoken French text annotated for coreference. It consists of manual transcripts from four corpora: two representing socio-linguistic interviews (Eshkol-Taravella et al., 2011) and two representing highly interactive dialogues (Antoine et al., 2002). Originally in TEI format, the corpus is now available in the CorefUD format within the CorefUD collection (Novák et al., 2025; Nedoluzhko et al., 2022). The manual transcriptions in ANCOR do not include any punctuation or casing, except for question marks and proper names, and accurately retain speech discontinuities, including repetitions and stuttering.

The coreference annotation in ANCOR has several particularities that distinguish it from other corpora. First, the deictic pronouns (e.g., *I, you, we*) are always annotated as singletons, i.e. they are never linked to any other mentions. Second, the discontinuous mentions are present in the corpus.

Statistics	ANCOR		ASR	
	Train	Val	Train	Val
#documents	365	45	54	9
#sentences	25K	2,385	16K	2,628
#words	371K	38K	193K	31K
#entities	55K	5,827	25K	4,212
#mentions	91K	9,491	40K	6,751
%singletons	80.8%	79.9%	79.9%	79.1%
%disc. mentions	0.5%	0.6%	0.2%	0.3%

Table 1: Statistics of the datasets. Here, *disc.* stands for discontinuous.

Finally, in the original corpus, each utterance is attributed to a speaker, but this information was omitted in the CorefUD format.

3.1 Re-transcribing the Corpus

To build an ASR coreference corpus, we utilized the Whisper Large multilingual model[1] (Radford et al., 2023) to transcribe the ESLO corpus (Eshkol-Taravella et al., 2011) which constitutes the largest part of ANCOR. We then performed word-level alignment of the manual transcriptions with the ASR transcriptions using the `spacy-alignments`[2] library. Since the coreference annotation in CorefUD format is also word-level, we transferred it to the ASR data (see Annex B for an example). Next, we split the ASR data into training and validation sets using the same documents as in ANCOR. Finally, we added morpho-syntactic information (lemmas, part-of-speech tags, detailed morphological features, dependency trees) using Stanza's default model for French (Qi et al., 2020) and repositioned the syntactic head of each mention with heuristics from the `udapi-python` package.[3]

Due to imperfect automatic alignment, the resulting ASR corpus often contained invalid coreference annotations. For the validation set, we manually verified and corrected these annotations. For the training set, we removed sentences containing invalid CorefUD annotations, such as unclosed mention tags or closing tags without corresponding opening tags. Table 1 shows that the resulting ASR training set is almost half the size of its ANCOR counterpart, while the ASR validation set retains nearly 80% of its original size. Furthermore, the proportion of discontinuous mentions in the ASR dataset is smaller than in the original dataset. This

[1] Using the WhisperX implementation (Bain et al., 2023).

[2] https://github.com/explosion/spacy-alignments

[3] https://github.com/udapi/udapi-python

Model	Train	MUC	B^3	$CEAF_e$	BLANC	LEA	MOR	CoNLL
Human transcription validation set								
WL-Coref	*Hum.*	**76**/77/76	59/68/63	67/58/62	55/68/59	55/65/60	**86**/85/**86**	67.26
	ASR	67/68/68	37/59/45	62/34/44	43/57/43	33/55/41	85/79/82	52.30
	H+A	**76**/**79**/**77**	**63**/**71**/**67**	**70**/**63**/**67**	**61**/**70**/**65**	**59**/**68**/**63**	**86**/**87**/**86**	**70.23**
CorPipe-24	*Hum.*	78/73/76	73/60/66	66/71/68	72/57/63	69/56/62	**86**/**84**/**85**	70.06
	ASR	64/66/65	58/54/56	55/61/58	55/53/54	52/48/50	72/82/76	59.50
	H+A	**80**/**75**/**77**	**74**/**63**/**68**	**67**/**72**/**69**	**73**/**60**/**66**	**70**/**58**/**64**	85/**84**/**85**	**71.37**
ASR transcription validation set								
WL-Coref	*Hum.*	66/**74**/70	47/**67**/55	62/51/56	43/**65**/49	43/**63**/51	80/**84**/82	60.42
	ASR	69/64/66	46/54/50	62/42/50	46/50/44	42/50/46	**85**/76/80	55.43
	H+A	**74**/72/**73**	**63**/63/**63**	**66**/**61**/**64**	**60**/61/**61**	**59**/60/**59**	84/82/**83**	**66.55**
CorPipe-24	*Hum.*	74/**69**/71	68/**59**/**63**	**64**/**68**/**66**	**66**/**56**/**60**	63/**54**/58	85/82/**83**	66.79
	ASR	73/67/70	68/56/61	62/65/64	65/52/57	63/51/56	**86**/79/82	64.79
	H+A	**75**/**69**/**72**	**69**/**59**/**63**	**64**/67/65	**66**/55/**60**	**64**/**54**/**59**	**86**/81/**83**	**66.80**

Table 2: Results on the Human (upper part) and ASR (lower part) transcription validation sets. For each validation set, the best results for each model are shown in **bold**, and the best results across the models are underlined. All metrics are reported as Recall/Precision/F1, except for the CoNLL F1 score.

reduction is due to speech discontinuities (e.g., stutterings, repetitions, talking over) being preserved in the human transcription but absent from the ASR transcription. Finally, the ASR validation set has more sentences[4] despite being smaller. This results from Whisper producing text closer to written form, while human transcriptions split the text by pauses in speech or speaker changes.

4 Experimental Setup

We trained both WL-Coref (Dobrovolskii, 2021; D'Oosterlinck et al., 2023; Liu et al., 2024) and CorPipe (Straka, 2024) models with `camembertav2-base`[5] pre-trained encoder, which currently achieves state-of-the-art results on French NLP tasks (Antoun et al., 2024) (see Appendix A for more details). For each architecture, three model variants were trained according to the training data: 1) ***Hum.*** using the original ANCOR data; 2) ***ASR*** using the automatically transcribed subset of the original data; 3) ***H+A*** using the combination of the ANCOR and ASR training datasets.

To measure the performance of the models, in addition to the ASR validation set, we created a subset of the ANCOR human transcription validation set, which includes the same documents as the ASR validation set. For evaluation metrics, we used MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998), CEAFe (Luo, 2005), BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016), MOR (Mention Overlap Ratio) (Žabokrtský et al., 2022), and the CoNLL score which is the average of the first three metrics. All metrics were calculated using the CorefUD scorer[6] with exact mention matching and excluding all singletons.

5 Results and Discussion

The upper part of Table 2 presents the results on the manually transcribed validation set. For both the WL-Coref and CorPipe models, training on human transcripts (*Hum.*) yielded better performance compared to using only automatically constructed ASR training data (*ASR*), with CoNLL score drops of -14.96 and -10.56 for the WL-Coref and CorPipe models, respectively. Utilizing a mix of manual and ASR training data (*H+A*) slightly enhanced the performance of both models on the manually transcribed data, resulting in CoNLL score increases of $+2.97$ and $+1.31$ for the WL-Coref and CorPipe models, respectively. The lower part of Table 2 illustrates similar trends for the ASR validation set. However, the WL-Coref model appears to be more sensitive to changes in the data, whereas the CorPipe model shows almost no difference between the *Hum.* and *H+A* variants.

Across both validation sets, WL-Coref achieved higher precision in all metrics except $CEAF_e$, while

[4] Defining a sentence in spoken text can be challenging. In the context of this work, a sentence is defined as a continuous sequence of words where all mentions are fully contained within it, meaning that a mention cannot span across sentence boundaries.

[5] https://huggingface.co/almanach/camembertav2-base

[6] https://github.com/ufal/corefud-scorer

Model	Train	Head Error	Span Error	Conflated Entities	Extra Mention	Extra Entity	Divided Entity	Missing Mention	Missing Entity
Human transcription validation set									
WL-Coref	*Hum.*	188	43	85	286	70	62	100	125
	ASR	163	53	58	322	48	54	133	171
	H+A	188	49	76	303	71	46	79	111
CorPipe	*Hum.*	3	64	112	277	81	97	121	73
	ASR	3	103	143	446	150	142	127	79
	H+A	3	63	105	257	79	96	93	70
ASR transcription validation set									
WL-Coref	*Hum.*	167	48	54	377	85	50	119	125
	ASR	144	51	58	257	42	50	171	197
	H+A	154	53	64	287	68	46	118	137
CorPipe	*Hum.*	0	85	91	307	87	76	130	86
	ASR	0	82	101	301	78	79	133	103
	H+A	0	86	89	285	77	77	129	102

Table 3: Error analysis on the Human (upper part) and ASR (lower part) transcription validation sets. For each validation set, the cells are color-coded in a gradient column-wise, with red representing the highest value and green representing the lowest value.

CorPipe showed higher recall. When applied to the ASR validation set, WL-Coref trained on human transcribed data exhibited a significant drop in recall and only a slight drop in precision. In contrast, CorPipe showed only a moderate decrease in recall.

We hypothesize that this discrepancy may occur because the WL-Coref model predicts links between mention heads, making it more susceptible to errors from ASR and automatic syntactic parsing, which in turn affect its performance. In contrast, the CorPipe model employs a sequence tagging approach to detect mentions, which does not rely on additional syntactic information.

5.1 Error Analysis

To better understand the impact of ASR transcriptions on the performance of coreference resolution models, we conduct an error analysis based on the work of Kummerfeld and Klein (2013). To adapt this analysis to the exact mention matching scenario, we introduce a Move Head operation. This operation corrects a predicted mention head if the spans of the predicted and ground truth mentions match exactly, corresponding to what is termed a Head Error. The remainder of the analysis largely adheres to the methodology outlined by Kummerfeld and Klein (2013).

Table 3 presents the error analysis for the WL-Coref and CorPipe models (see Annex C for examples of errors). The WL-Coref model exhibits a high number of Head Errors but fewer Span Errors. This can be explained by the design of the CorPipe model, which is specifically tailored for the CorefUD shared task where head matching is used for evaluation. Interestingly, the WL-Coref model produces more accurate spans even when starting from incorrect heads. Lastly, WL-Coref consistently has more Missing Entity errors which is explained by the lower recall.

When evaluated on the ASR validation set, models trained on *Hum.* data demonstrate more Conflated Entities, where a predicted entity includes mentions from different ground-truth entities, and fewer Divided Entities, where different predicted entities include mentions from the same ground-truth entity. This behavior suggests that when applied to ASR transcriptions, the models group mentions into tighter clusters. A possible reason is that the lack of filler words and repetitions in ASR transcriptions reduces the distance between mentions.

The large number of Extra Mention errors mostly stems from assigning mentions, which should otherwise be singletons, to a coreference chain and linking the pronouns *ce* and *ça* (it) when they are non-referential. The increase in such errors on the ASR validation data could be explained by Whisper producing more "grammatically valid" transcriptions, adding these pronouns, e.g., by inserting *c'est* (it is) when they are absent from speech and consequently from human transcriptions.

Finally, we found that both human and ASR validation sets contain annotation errors. However, their exact impact on the evaluation is beyond the scope of this study and requires further study.

6 Conclusions

In this study, we evaluated the performance of coreference resolution models trained on human-transcribed data when applied to ASR-transcribed data, observing a general trend of decreased performance on ASR data for models trained on manual transcriptions. We proposed an approach to automatically transfer coreference annotations from human to ASR transcriptions and discovered that training only on silver ASR data harms model performance, whereas combining silver ASR data with gold manual data enhances it. Further error analysis revealed that ASR systems, which tend to overcorrect transcriptions, introduce potential errors to coreference resolution systems. We found that models prioritizing higher recall are more robust to these errors than those focusing on precision.

Limitations

Given the scarcity of spoken coreference datasets in French, this study is confined to a single corpus, primarily comprising socio-linguistic interviews. These interviews have low interactivity and cover a limited range of topics. Furthermore, participants are sampled from a restricted geographic area, specifically Orléans and Tours, which narrows the vocabulary used in the interviews. A more topically diverse corpus would be essential for a broader evaluation.

Regarding coreference resolution models, this study evaluates only two architectures: WL-Coref and CorPipe. While a more diverse set of models would enhance the robustness of the comparison, hardware limitations and variations in coreference data formats present significant challenges. Additionally, the prevalence of English-specific or OntoNotes-specific architectures complicates the adaptation of existing models to other languages and the CorefUD format, which is beyond the scope of this study.

Finally, this study only uses Whisper as the ASR system for automatically transcribing the dataset recordings. We acknowledge that other ASR systems may produce different transcriptions, potentially leading to different effects on automatic coreference resolution performance.

Acknowledgments

This work is partially funded by the ANR Pantagruel project ANR-23-IAS1-0001-02.

References

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. Coreference in spoken vs. written texts: a corpus-based analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 158–164, Istanbul, Turkey. European Language Resources Association (ELRA).

Jean-Yves Antoine, Sabine Letellier-Zarshenas, Pascale Nicolas, Igor Schadle, and Jean Caelen. 2002. Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Posters*, pages 319–324, Nancy, France. ATALA.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A smarter french language model aged to perfection. *arXiv preprint arXiv:2411.08868*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive learning for improving ASR robustness in spoken language understanding. In *Interspeech 2022*, pages 3458–3462.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Celine Dugua, and Isabelle Tellier. 2011. Un grand corpus oral disponible : le corpus d'orléans 1968-2012 [a large available oral corpus: Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, 52(3):17–46.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 14–19, Online. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Tuan Lai and Heng Ji. 2023. Ensemble transfer learning for multilingual coreference resolution. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 24–36, Toronto, Canada. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and conjunction-aware word-level coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ondřej Pražák, Miloslav Konopík, and Pavel Král. 2024. Exploring multiple strategies to improve multilingual coreference resolution in CorefUD. *Preprint*, arXiv:2408.16893.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

M Recasens and E Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we NER yet? artifacts

of ASR errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761, Toronto, Canada. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Wenqiang Wang, Chongyang Du, Tao Wang, Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Xiaochun Cao. 2023. Punctuation-level attack: Single-shot and single punctuation can fool text models. In *Advances in Neural Information Processing Systems*, volume 36, pages 49312–49324. Curran Associates, Inc.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Implementation Details

For WL-Coref, we utilized the implementation by Stanza (Qi et al., 2020), while for CorPipe, we used the implementation from their official repository.[7] The original implementation of CorPipe-2024 is based on an older version of TensorFlow, making it challenging to run on modern systems. We updated the original code to be compatible with the most recent version of TensorFlow without altering the model's architecture. Since neither model supports discontinuous mentions, they were removed from the training data. All models were trained for 40 epochs on an NVIDIA A100 40GB GPU.

[7] https://github.com/ufal/crac2024-corpipe

B Examples of Data

Table 4 shows an example of human and ASR transcribed data.

C Examples of Errors

In this section, we demonstrate the examples of different errors. Table 5 shows an example of a Head Error, Table 6 shows an example of a Span Error, Table 7 shows an example of an Extra Mention, Table 8 shows an example of an Extra Entity, Table 9 shows an example of a Missing Mention, Table 10 shows an example of a Missing Entity, Table 11 shows an example of a Divided Entity, and Table 12 shows an example of a Conflated Entity.

Human	eh bien [**monsieur**]$_s$ [**je**]$_s$ vais commencer par [**vous**]$_s$ poser [des petites **questions** préliminaires toutes simples]$_s$ n' est -ce pas et depuis combien de temps habitez-[**vous**]$_s$ [**Orléans**]$_1$? euh [dix-neuf **ans**]$_2$ [dix-neuf **ans**]$_2$ oui et qu' est -ce qui [**vous**]$_s$ a amené à vivre à [**Orléans**]$_1$?
ASR	Eh bien, [**Monsieur**]$_s$, [**je**]$_s$ vais commencer par [**vous**]$_s$ poser [des petites **questions** préliminaires, toutes simples]$_s$, n'est-ce pas ? Et depuis combien [de **temps**]$_s$ habitez[**-vous**]$_s$ à [**Orléans**]$_1$? [19 **ans**]$_2$. [19 **ans**]$_2$, oui. Et qu'est-ce qui [**vous**]$_s$ a amené à vivre à [**Orléans**]$_1$?

Table 4: Examples of human and ASR transcribed data. Each new line represents a sentence break. Mentions are enclosed in square brackets with mention heads highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton.

Gold	Est-ce que [la **langue** française] est aussi bien enseignée, ou mieux enseignée, ou moins bien enseignée, que de le temps où vous étiez vous-même à l'école ?
Predicted	Est-ce que (la langue **française**) est aussi bien enseignée, ou mieux enseignée, ou moins bien enseignée, que de le temps où vous étiez vous-même à l'école ?

Table 5: Examples of Head Error. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold and underlined**.

Gold	Je trouve que c'est [des **différences** assez grandes].
Predicted	Je trouve que c'est (des **différences**) assez grandes.

Table 6: Examples of Span Error. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**.

Gold	Demain, je serai peut-être partie ou prête à partir et je peux rester encore [six **ans**]$_s$. Je ne sais pas. J'aimerais rester. J'aimerais rester, mais... On nous a dit, n'est-ce pas, ailleurs, que la ville d'Orléans est une ville assez froide, mais je ne sais pas si vous avez des visites là-dessus, puisque vous êtes... Il y a [quelques **années**]$_s$, oui.
Predicted	Demain, je serai peut-être partie ou prête à partir et je peux rester encore (six **ans**)$_1$. Je ne sais pas. J'aimerais rester. J'aimerais rester, mais... On nous a dit, n'est-ce pas, ailleurs, que la ville d'Orléans est une ville assez froide, mais je ne sais pas si vous avez des visites là-dessus, puisque vous êtes... Il y a (quelques **années**)$_1$, oui.

Table 7: Examples of Extra Mention. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Parce que [les **Américains**]$_s$, [les **Allemands**]$_s$, les Suisses, les Japonais, [tout **ça**]$_s$, [**ça**]$_s$ ne parle pas latin.
Predicted	Parce que (les **Américains**)$_1$, (les **Allemands**)$_1$, les Suisses, les Japonais, (tout **ça**)$_1$, (**ça**)$_1$ ne parle pas latin.

Table 8: Examples of Extra Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	C'est peut-être moi qui écris le plus. Et à [votre **famille**]$_1$? disons que ma femme écrit plutôt à sa famille et moi j'écris plutôt à [la **mienne**]$_1$, encore qu'il arrive très fréquemment que j'écrive à la sienne et qu'elle écrive à [la **mienne**]$_1$.
Predicted	C'est peut-être moi qui écris le plus. Et à (votre **famille**)$_1$? disons que ma femme écrit plutôt à sa famille et moi j'écris plutôt à la mienne, encore qu'il arrive très fréquemment que j'écrive à la sienne et qu'elle écrive à la mienne.

Table 9: Examples of Missing Mention. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Mais alors, ce qui est embêtant, c'est que vous avez des gosses qui demandent à travailler. et qui ne veulent pas être les bras coincés, ou qui ne veulent pas faire de [la **pâte** à modeler]$_1$ parce qu'on [**l'**]$_1$a déjà fait à la maison.
Predicted	Mais alors, ce qui est embêtant, c'est que vous avez des gosses qui demandent à travailler. et qui ne veulent pas être les bras coincés, ou qui ne veulent pas faire de (la **pâte** à modeler)$_s$ parce qu'on (**l'**)$_s$a déjà fait à la maison.

Table 10: Examples of Missing Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Je ne parle évidemment pas des dictionnaires de [**langue** ancienne]$_1$ que nous avons à la maison. ... En tout cas, en ce qui concerne [les **langues** anciennes]$_1$, il y a 20 ans, plus de la moitié des élèves faisaient [des **langues** anciennes]$_1$, alors que maintenant, [**ça**]$_1$ représente 1 %.
Predicted	Je ne parle évidemment pas des dictionnaires de (**langue** ancienne)$_1$ que nous avons à la maison. ... En tout cas, en ce qui concerne (les **langues** anciennes)$_2$, il y a 20 ans, plus de la moitié des élèves faisaient (des **langues** anciennes)$_2$, alors que maintenant, (**ça**)$_2$ représente 1 %.

Table 11: Examples of Divided Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown. Three dots (...) show that there are several sentences in between.

Gold	Alors, [au **bureau**]$_1$, À [mon **bureau**]$_1$, j'ai un petit Larousse, mais j'ai chez moi un dictionnaire en dix volumes. ... Je parle pour mon foyer, je ne parle pas [du **bureau**]$_1$. ... Est-ce que vous pourriez dire combien par [**mois**]$_2$? Oui, au point de vue personnel, pas plus de deux ou trois lettres personnelles par [**mois**]$_2$.
Predicted	Alors, (au **bureau**)$_1$, À (mon **bureau**)$_1$, j'ai un petit Larousse, mais j'ai chez moi un dictionnaire en dix volumes. ... Je parle pour mon foyer, je ne parle pas (du **bureau**)$_1$. ... Est-ce que vous pourriez dire combien par (**mois**)$_1$? Oui, au point de vue personnel, pas plus de deux ou trois lettres personnelles par (**mois**)$_1$.

Table 12: Examples of Conflated Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown. Three dots (...) show that there are several sentences in between.

Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches?

Michal Novák[1], **Miloslav Konopík**[2], **Anna Nedoluzhko**[1], **Martin Popel**[1], **Ondřej Pražák**[2], **Jakub Sido**[2], **Milan Straka**[1], **Zdeněk Žabokrtský**[1], **Daniel Zeman**[1]

[1] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czechia
{mnovak,nedoluzko,popel,straka,zabokrtsky,zeman}@ufal.mff.cuni.cz

[2] University of West Bohemia, Faculty of Applied Sciences,
Department of Computer Science and Engineering, Pilsen, Czechia
{konopik,ondfa,sidoj}@kiv.zcu.cz

Abstract

The paper presents an overview of the fourth edition of the Shared Task on Multilingual Coreference Resolution, organized as part of the CODI-CRAC 2025 workshop. As in the previous editions, participants were challenged to develop systems that identify mentions and cluster them according to identity coreference.

A key innovation of this year's task was the introduction of a dedicated Large Language Model (LLM) track, featuring a simplified plaintext format designed to be more suitable for LLMs than the original CoNLL-U representation.

The task also expanded its coverage with three new datasets in two additional languages, using version 1.3 of CorefUD – a harmonized multilingual collection of 22 datasets in 17 languages.

In total, nine systems participated, including four LLM-based approaches (two fine-tuned and two using few-shot adaptation). While traditional systems still kept the lead, LLMs showed clear potential, suggesting they may soon challenge established approaches in future editions.

1 Introduction

Coreference is the phenomenon where multiple expressions in a text refer to the same real-world entity. For example: *"Beethoven was a revolutionary artist. The German composer changed the course of music, and he continues to inspire musicians today."* Here, *"Beethoven"*, *"the German composer"*, and *"he"* all point to the same individual. The computational task of coreference resolution is to automatically identify such links between mentions and group them into clusters that represent entities. In the multilingual setting, the task is the same, but complicated by the diversity of languages and their grammatical and discourse conventions.

In this article, we present the overall setup and results of the fourth edition of the shared task in multilingual coreference resolution. For descriptions of previous editions, as well as references to the roots and predecessors of the series, see Novák et al. (2024).

This year's edition uses an improved and expanded collection of coreference data, CorefUD 1.3 (Novák et al., 2025), currently spanning 17 languages from a few typologically different families. However, the most important novelty in this edition is the introduction of the Large Language Model (LLM) track. Although non-LLM models were still welcome, a dedicated LLM Track was introduced to highlight and explore the capabilities of LLM-based approaches. Hence, to accommodate different modeling strategies and study their effects, we defined two shared-task tracks:

- *LLM Track:* Focused on solutions that primarily rely on LLMs for coreference resolution. Allowed strategies include fine-tuning LLMs, using in-context learning, designing effective prompts, utilizing constrained decoding strategies, and building more complex agentic systems.

- *Unconstrained Track:* Open to all other approaches, including non-LLM models and hybrid systems. This track allows the use of

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 95–118
Suzhou, China, November 9, 2025.

Spanish: El conductor del tren vio el coche en la vía e intentó frenar.
English transl.: The driver of-the train saw the car on the track and tried to brake.

Our serialization:
```
El|[e22 conductor de el tren|[e5],e22] vio el|[e7 coche|e7]
en la|[e8 vía|e8] e intentó ##|[e22] frenar|[e23] .
```

Figure 1: Our plaintext serialization of a Spanish example sentence from es_ancora. For clarity, mention spans are highlighted by colored underlining, where two coreferential entities share the same color. A zero mention labeled on an empty node is greyed. Note that multi-word tokens are split in the plaintext format into syntactic words (e.g., the Spanish *"del"* appears as *"de el"*); this conversion error was identified after the data release.

additional pre-existing coreference systems, external tools, and extensive model modifications.

A major trend in NLP is the shift from traditional task-specific models to LLMs, which can address a wide range of tasks with little fine-tuning and are comparatively easy to deploy. This unification brings greater efficiency, flexibility, and scalability, but also raises challenges such as bias, computational cost, and privacy concerns. At the same time, LLMs have shown strong performance on tasks that require understanding of textual context and relations, including question answering, summarization, and commonsense reasoning.

A category of benchmarks that are commonly used to test these coreference-related capabilities are derivations of the Winograd Schema Challenge (Levesque et al., 2012), for instance KnowRef (Emami et al., 2019), WinoGrande (Sakaguchi et al., 2021), and recently WinoWhat (Gevers et al., 2025). However, these benchmarks represent an overly narrow view of coreference resolution. They primarily focus on commonsense reasoning through carefully crafted disambiguation scenarios, while real-world coreference resolution involves a much broader spectrum of phenomena.

Previous works on using LLMs for coreference resolution show that they struggle with this task and are not able to outperform systems specifically tailored for coreference resolution (Le and Ritter, 2023; Vadász, 2023; Hicke and Mimno, 2024; Gan et al., 2024; Saputa et al., 2024). One of the reasons may be that the data used to model and test the task is very heterogeneous due to practical difficulties in clearly and precisely defining the elements that coreference relations work with, specifically the scope of mentions, the degree of zero reconstruction, and the typology of coreference and anaphoric relations.

Still, the progress in LLMs is so rapid that it seems just a matter of time before these LLM-based systems will dominate also in this task. We see the LLM track of this shared task as an opportunity to test this hypothesis and encourage development in this field, providing a platform for researchers to explore the current boundaries and future potential of LLM-based coreference resolution.

The step towards LLMs does not represent only a technological change – it often requires rethinking how we approach a particular task. Structured (possibly pipelined) solutions are typically abandoned and replaced by processing "flat" sequences of (sub)words. In the particular case of this shared task, we replace the relatively richly structured CoNLL-U format in which the encoding of coreference relations is stored in the CorefUD collection with an encoding of coreference that could be added directly into plain text.

Naturally, there are many possible ways to insert coreference markup into text, and prior work on LLMs for coreference has each used its own prompt and format. So far, no widely accepted best practices have emerged for encoding or prompting coreference in plain text. We implement our own conversion from CorefUD into a plaintext serialization (example in Figure 1), but acknowledge that our design choices may limit applicability and that further optimization could improve LLM performance.

The remainder of the paper is structured as follows. Section 2 discusses the changes in the data compared to the previous (third) edition of the shared task. Section 3 outlines the evaluation metrics used in the task, including both the primary and supplementary scores. Section 4 details all participating systems, both in the LLM track and in the Unconstrained track. Section 5 presents a summary of the results and discusses some differences between the performance of LLM and Unconstrained systems. Section 6 provides the conclusion.

2 Datasets

As in previous years, the shared task takes training and evaluation data from the public part of the CorefUD collection (Nedoluzhko et al., 2022),[1] now in its latest release (1.3).[2] The public edition of CorefUD 1.3 consists of 24 datasets[3] covering 17 languages from five language families. Compared to CorefUD 1.2, used last year (Novák et al., 2024), the release adds three new datasets and two new languages including Korean, which represents a new language family. The new datasets are French ANCOR, Hindi HDTB, and Korean ECMT. In addition, several existing datasets from CorefUD 1.2 were updated. The data span diverse domains including news, fiction, Bible texts, and Wikipedia articles. French ANCOR notably introduces transcripts of originally spoken conversational data, which were previously only marginally represented in CorefUD. Table 1 gives an overview of the datasets and their sizes. See Appendix A for references of the individual datasets.

One of the goals of the CorefUD project is to encourage research on coreference resolution in languages other than English, particularly those with zero anaphora. Zero anaphora, or *zero mentions*, occur when a referent (like a subject or object) is implied but not explicitly stated. This is a common feature of pro-drop languages, where verb conjugation often provides enough information to infer the missing pronoun. In CorefUD, zero mentions are represented as *empty nodes* that are artificially inserted into Universal Dependencies (UD) trees. This allows them to be grouped with other mentions in a coreference chain, just like any other explicitly stated mention. Although the two newly added languages, Korean and Hindi, are considered pro-drop, the original datasets do not include zero mention annotation. Therefore, the collection of datasets with zero mentions remains the same as in the previous edition.

Our shared task focuses exclusively on identity coreference. The datasets in the CorefUD collection, however, may include annotations of other relations, such as bridging. Similarly, phenomena like event anaphora and abstract anaphora may be annotated in some datasets but not in others. Because CorefUD is not fully harmonized in terms of annotation guidelines, the precise nature of annotated anaphoric phenomena may vary slightly across corpora. In converting to the CorefUD format, we aim to isolate identity coreference[4] while largely preserving the original annotations.

2.1 New Resources

French ANCOR (fr_ancor; Muzerelle et al., 2014) is a collection of three different corpora of conversational speech (Accueil_UBS, OTG, ESLO), annotated for coreference. Cross-sentence mentions (caused e.g. by two speakers speaking simultaneously) are ignored in the conversion from TEI to CorefUD.

Hindi HDTB (hi_hdtb; Mujadia et al., 2016) is based on the HDTB corpus (Palmer et al., 2009) annotated with coreference and anaphoric relations and corresponding to the namesake treebank in UD. However, the coreference corpus does not constitute a strict subset of the UD treebank, as approximately 14% of its sentences are not included in the UD release. Still, each coreference-annotated document contains at least one sentence that appears in the treebank. Although the original annotations distinguish *PartOf* relations, these are often merged with identity coreference relations within the same cluster, complicating the separation of identity, bridging, and split-antecedent relations. As a result, we currently treat all mentions within a cluster as coreferential, without making finer distinctions. At present, we do not incorporate the manually annotated morpho-syntactic information from the UD treebank; instead, we replace it with automatic parses produced by UDPipe 2.

Korean ECMT (ko_ecmt; Nam et al., 2020) is a conversion of the dataset created for the paper "Effective Crowdsourcing of Multiple Tasks for Comprehensive Knowledge Extraction" (ECMT). The original dataset is based on Korean Wikipedia and KBox with crowdsourced annotations for four information extraction tasks: (1) entity detection, (2) entity linking, (3) coreference resolution, and (4) relation extraction. The original dataset seems to contain errors where distinct entities are incorrectly merged into a single coreference cluster. The CorefUD conversion did not attempt to fix these errors.

[1] https://ufal.mff.cuni.cz/corefud

[2] http://hdl.handle.net/11234/1-5896

[3] For the shared task, we used only 22 of them (see Section 2.3).

[4] We are aware that complete isolation is not possible due to near-identity relations; see Recasens et al. (2010).

document	total number of				entities				mentions			
					total	per 1k	length		total	per 1k	length	
	docs	sents	words	empty n.	count	words	max	avg.	count	words	max	avg.
Ancient_Greek-PROIEL	19	6,475	64,111	6,283	3,215	50	332	6.6	21,354	333	52	1.7
Ancient_Hebrew-PTNK	40	1,161	28,485	0	870	31	102	7.2	6,247	219	22	1.5
Catalan-AnCora	1,298	13,613	429,313	6,377	17,558	41	101	3.6	62,417	145	141	4.8
Czech-PCEDT	2,312	49,208	1,155,755	35,654	49,225	43	236	3.4	168,055	145	79	3.6
Czech-PDT	3,165	49,419	834,707	21,092	46,460	56	173	3.3	154,437	185	99	3.1
English-GUM	237	13,263	233,926	119	9,200	39	131	4.4	40,656	174	95	2.6
English-LitBank	100	8,560	210,530	0	2,164	10	261	10.8	23,340	111	129	1.6
English-ParCorFull	19	543	10,798	0	188	17	38	4.4	835	77	37	2.1
French-ANCOR	455	31,761	454,577	0	13,204	29	103	4.3	56,459	124	17	1.9
French-Democrat	126	13,057	284,883	0	7,162	25	895	6.5	46,487	163	71	1.7
German-ParCorFull	19	543	10,602	0	243	23	43	3.7	896	85	30	2.0
German-PotsdamCC	176	2,238	33,222	0	880	26	15	2.9	2,519	76	34	2.6
Hindi-HDTB	271	3,479	76,282	0	3,148	41	36	3.8	12,082	158	43	1.8
Hungarian-KorKor	94	1,351	24,568	1,569	1,122	46	41	3.6	4,091	167	42	2.2
Hungarian-SzegedKoref	400	8,820	123,968	4,857	4,769	38	36	3.2	15,165	122	36	1.6
Korean-ECMT	1,470	30,784	482,986	0	16,536	34	55	3.4	56,538	117	12	1.3
Lithuanian-LCC	100	1,714	37,014	0	1,087	29	23	4.0	4,337	117	19	1.5
Norwegian-BokmaalNARC	346	15,742	245,515	0	5,658	23	298	4.7	26,611	108	51	1.9
Norwegian-NynorskNARC	394	12,481	206,660	0	5,079	25	84	4.3	21,847	106	57	2.1
Old_Church_Slavonic-PROIEL	26	6,832	61,759	6,289	3,396	55	134	6.5	22,116	358	52	1.5
Polish-PCC	1,828	35,874	538,885	18,615	22,143	41	135	3.7	82,706	153	108	1.9
Russian-RuCor	181	9,035	156,636	0	3,515	22	141	4.6	16,193	103	18	1.7
Spanish-AnCora	1,356	14,159	458,418	8,112	19,445	42	110	3.6	70,663	154	101	4.8
Turkish-ITCC	24	4,732	55,358	11,584	4,019	73	369	5.4	21,569	390	31	1.1

Table 1: CorefUD 1.3 data sizes in terms of the total number of documents, sentences, words (i.e. non-empty nodes), empty nodes (empty words), coreference entities (total count, relative count per 1000 words, average and maximal length in number of mentions) and coreference mentions (total count, relative count per 1000 words, average and maximal length in number of words). All the counts are excluding singletons and for the concatenation of train+dev+test. Train/dev/test splits of these datasets roughly follow the 8/1/1 ratio. However, note that for the shared task we used reduced versions of dev and test: mini-dev and mini-test, respectively.

2.2 Updated Resources

More data The English GUM corpus (en_gum) is now in its version 11, which has approximately 10% more data. All the other datasets are the same size as before (except for a few minor changes resulting from annotation corrections).

New prediction of morphosyntax For datasets that do not come with manual morphosyntactic annotation, the UD relations, tags and features were predicted with newer models for UDPipe (based on UD release 2.15 instead of 2.12). This involves the following ten corpora: Czech PCEDT, English LitBank, English ParCorFull, German ParCorFull, German PotsdamCC, Hungarian KorKor, Hungarian SzegedKoref, Lithuanian LCC, Polish PCC, Russian RuCor.

Substantial changes Re-implementation of conversion from non-CorefUD formats and/or major revision of the annotation was applied to Czech PDT (cs_pdt) and Hungarian KorKor (hu_korkor). For Czech, the source dataset is now the PDT part of PDT-C 2.0 (previously it was 1.0), which has substantial improvements on the surface-syntactic layer. Many other changes were done in the PDT-to-UD conversion of morphology and syntax; coreference annotation is unchanged, except for a few corrections. For Hungarian, the conversion from the native format was almost completely rewritten. Empty copula nodes are now deleted as required in UD. DROP empty nodes now receive correct incoming dependency relations (`nsubj`, `obj`, or `nmod:att`), and there are several other small improvements.[5]

2.3 Data for the Shared Task

Compared to the public edition of CorefUD 1.3, the data provided for the shared task participants underwent slight adjustments.[6]

[5] More details on the changes can be found in the README files of the individual corpora.

[6] Both the shared task data and submissions are available at http://hdl.handle.net/11234/1-5987.

Data reduction Firstly, the English and German ParCorFull datasets were excluded from this year's shared task. These datasets are the smallest (their test sets contain less than 900 words, one third of the next smallest test set) and exhibited the largest variance, considerably influencing overall macro-averaged scores.[7]

Secondly, the development and test sets were reduced to *mini-dev* and *mini-test* sets, respectively. This change was introduced to lower the computational cost of evaluation while preserving high discriminative power. Each dev and test set is now capped at 25k words, achieved by randomly sampling complete documents. The 25k threshold was selected to cut the overall collection size by roughly half, while affecting only a few of the largest corpora and still ensuring reliable and representative results.[8]

Plaintext format For the LLM track, we provide a conversion to a simple plaintext format, along with both the conversion tool and the converted dataset files.

The plaintext format (see Figure 1) is a plain text file in which each line represents a document, and tokens are separated by spaces. Coreference annotations are appended to each token after the '|' character. Each mention, including singletons, is defined by its span boundaries, marked with opening and closing square brackets concatenated with the entity ID. Empty nodes are prefixed with '##'; if an empty node has a form or lemma in the original data, it is appended immediately after. Because empty nodes are defined by their syntactic position rather than linear order, each empty node is placed directly after its syntactic parent. This format does not encode the dependency relation type to the parent, which means it cannot distinguish between multiple empty nodes dependent on the same parent (see Section 3). While this limitation may slightly affect evaluation results, we consider the impact marginal and an acceptable trade-off for preserving the simplicity of the format.

The plaintext format is intentionally less expressive than CoNLL-U and lacks sufficient information for some evaluation metrics (e.g., head match requires mention heads derived from spans using syntactic trees). To bridge this gap, we provide a backwards conversion tool that restores plaintext annotations to CoNLL-U format, as well as an output cleaner.[9]

The cleaner addresses common issues caused by LLM outputs, such as broken annotation structure (e.g., unclosed mentions) or added/removed/modified words. It first ensures all mentions are properly opened and closed, then uses word-level edit distance to align output documents to the original input. Empty nodes are ignored in the edit-distance computation, as systems are expected to insert them themselves. Once the token sequences match exactly, the output annotations can be safely mapped back to the original CoNLL-U files.

Data variants and starting points In both tracks, two main variants of the data are provided: gold, and input data. In addition, participants of the Unconstrained track can choose from three starting points.

Gold data includes gold-standard annotations of coreference and empty nodes, intended for fine-tuning and evaluation. The data are consistent with the CorefUD 1.3 release, retaining manually annotated morpho-syntactic features (for datasets that originally included them), gold empty nodes, and gold coreference annotations. The only technical modification is the removal of empty nodes' forms in order to align the data with the output of the baseline empty node prediction, which does not predict these forms (see Section 4.1). While the gold train and mini-dev sets were available for download, the gold test set remained secret and were used internally in CodaLab for evaluation.

Input data was intended to be processed by participants' systems and subsequent submission. The following preprocessing was thus performed only on the mini-dev and mini-test sets. To better simulate a real-world scenario where no manual linguistic annotation is available, we removed the forms of empty nodes and replaced the original morpho-syntactic features with the outputs of UD 2.15 models across all datasets, including those with origi-

[7] Considering eight training runs of the last year's winning system differing in just random initialization, the standard deviation of the ParCorFull development results is more than 10 times larger than the standard deviation of the overall macro-averaged scores and 15 times larger than the standard deviation of the largest dataset.

[8] Again considering eight training runs of the last year's winning system differing in just random initialization, capping the large datasets to 25k words increase the standard deviation of the overall macro-averaged percentage results on the development sets by less than +0.03, from 0.296 to 0.324.

[9] The conversion tool and cleaner are available as a single application/Python library on GitHub: https://github.com/ondfa/text2text-coref

nally human-annotated features. Additionally, the gold empty nodes and coreference annotations were removed, forming the input data for the LLM track. On the other hand, in line with the setup of the last year's edition, participants of the Unconstrained track could choose from three different *starting points* for entering the shared task, with varying degrees of work required: (1) *Coreference and zeros from scratch* with no predictions of empty nodes and coreference (practically identical to the LLM-track variant), (2) *Coreference from scratch* with baseline predictions of empty nodes, and (3) *Refine the baseline* with baseline predictions of empty nodes and coreference.

3 Evaluation Metrics

The systems participating in the shared task are evaluated using the CorefUD scorer. In line with previous editions, the primary evaluation score is the CoNLL F_1 score, computed with head mention matching and excluding singletons. To align zero mentions, no longer guaranteed to match one-to-one due to the shift to a more realistic setup introduced last year, we apply a dependency-based matching method. In addition to the primary metric, we also compute several supplementary scores to support a more comprehensive comparison of the shared task submissions.

Official scorer We evaluate participant submissions using the CorefUD scorer[10], specifically the February 2025 version, which remains virtually unchanged from the version used in the previous edition. The scorer builds on the Universal Anaphora (UA) scorer 2.0 (Yu et al., 2023),[11] adopting all features relevant to the shared task, including implementations of widely used coreference evaluation metrics. In contrast to the UA scorer, the CorefUD scorer also supports head matching and a dependency-based method for aligning zero mentions.

The scorer takes two CoNLL-U files as input: the gold file and the predicted file. Since our plaintext format cannot capture all the information required for evaluation (e.g., mention heads), any LLM output produced in this format must first be restored into CoNLL-U before it can be properly evaluated.

Mention matching Due to the limitations of *exact* and *partial* mention matching methods (see Žabokrtský et al. (2023) for details), we have settled on the *head match* strategy for the primary evaluation metrics. In this approach, a gold and predicted mention are considered a match if their heads refer to the same token.[12] Full mention spans are ignored, except in cases where multiple mentions share the same head; in such instances, span information is used to disambiguate them.

However, this approach is not applicable to empty nodes, which frequently occur in zero anaphora. Predicted counterparts of gold zero mentions may be missing, spurious, or appear at different surface positions within a sentence, even if they serve the same syntactic or semantic role. To handle this, we devised a *dependency-based method* last year (Novák et al., 2024). The method aligns predicted and gold zero mentions within the same sentence by maximizing their overlap in enhanced dependency annotations. It formulates the task as a one-to-one matching in a weighted bipartite graph, where each candidate pair is scored based on how well the predicted zero replicates the gold zero's dependencies. Matches that correctly assign both the parent and the dependency type receive higher weights, though the method remains robust even when dependency types are not provided.

Primary score As is standard in coreference resolution, we use the CoNLL F_1 score (Denis and Baldridge, 2009; Pradhan et al., 2014) as the primary evaluation metric. This score is calculated as the unweighted average of the F_1 scores from three widely used coreference evaluation measures: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF-e (Luo, 2005). These metrics offer complementary perspectives: link-based, mention-based, and entity-based, respectively. As we aim to identify systems with stable performance across all datasets, the final ranking of submissions is determined by the macro-average of CoNLL F_1 scores across all mini-test sets in the shared task collection.[13]

[10] https://github.com/ufal/corefud-scorer

[11] The UA scorer 2.0 merges, reimplements, and extends several earlier tools, including previous versions of the CorefUD scorer.

[12] Gold mention heads in the CorefUD data are determined from the dependency tree using the Udapi block `corefud.MoveHead`.

[13] The evaluation protocol with macro-averaging CoNLL F_1 scores was announced before the start of the development phase and it was used also in previous versions of the shared task. We think it is the fairest aggregation method. As alternatives, one could average differences to the baseline or average

Supplementary scores Beyond the primary CoNLL F_1 score, we report its alternative variants based on different mention matching strategies: partial match[14] and exact match. We also compute the CoNLL score using head match for all mentions, including singletons.

To provide a more comprehensive evaluation, we report the individual coreference metrics comprising the CoNLL score (MUC, B^3, and CEAF) as well as other commonly used metrics such as BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016). Furthermore, we include the Mention Overlap Ratio (MOR) to assess mention detection independently of coreference clustering and the anaphor-decomposable score for zero anaphora, both introduced in Žabokrtský et al. (2022).

4 Participating Systems

4.1 Baseline

As in the previous edition, two baseline systems are provided: one for predicting empty nodes as slots for zero anaphora and another for coreference resolution. Only participants in the Unconstrained track are permitted to use or build upon these baseline systems.

Empty nodes prediction baseline Empty node prediction was introduced as an additional task in last year's shared task, and it is again part of the shared task this year. To support participants who wish to focus exclusively on coreference resolution, we again provide a baseline system for empty nodes prediction. We release the source code,[15] the trained multilingual model,[16] and the mini-dev and mini-test data with predicted empty nodes.

The baseline model architecture is virtually unchanged from last year. Each input sentence is processed by a XLM-RoBERTa-large (Conneau et al., 2020), generating embeddings for each input word. Then, two candidate empty nodes are predicted for each word, and passed through three heads: (1) a binary classification head predicting whether the candidate is really an empty node or not, (2) a word-order prediction head implemented using self-attention selecting the word after which the empty node should be added, and (3) a dependency relation prediction head, which first concatenates the candidate representation and the representation of the word most probable according to the word-order prediction head, and then predicts the dependency relation. A single model is trained on a concatenation of all corpora with empty nodes, sampling every sentence proportionally to the square root of its corpora size. For a detailed description and a visualization of the model architecture, see Straka (2024).

We intrinsically evaluate the empty node prediction baseline using precision, recall, and the F1 score, as shown in Table 2, where a prediction is classified as correct only when all of its dependency head, dependency relation, and word order are correct. For comparison, we also include the last year's F1 score. This year's results are very consistent, with the exception of hu_korkor showing an increase of nearly 20 percent points due to improved conversion to the CorefUD format in CorefUD 1.3 (see Section 2.2).

Coreference resolution baseline The coreference resolution baseline is the same as in the past three years. It is based on the multilingual end-to-end neural coreference resolution system by Pražák et al. (2021), which adapts the original end-to-end model of Lee et al. (2017). The model considers all possible spans up to a predefined maximum length and directly predicts an antecedent for each span. Since it has no separate mention detection step, it is well suited for datasets that do not annotate singletons. The baseline uses the mBERT base model as its encoder.

Hereafter, we denote the combination of the two baseline systems as BASELINE and the coreference resolution baseline applied to gold empty nodes as BASELINE-GZ.

4.2 System Submissions

This year, nine systems were submitted to the shared task by six teams: UWB,[17] PUXAI,[18]

ranks. The former yields the same final ranking as macro-averaging, while the latter would lead to a single difference: in the LLM track, the winner would be LLM-UWB, despite this system not producing output for one dataset and not covering zero anaphora in some datasets (see Sections 4.2 and 5).

[14] Partial match was used as the primary metric in the first edition of the shared task (Žabokrtský et al., 2022).

[15] https://github.com/ufal/crac2025_empty_nodes_baseline

[16] https://www.kaggle.com/models/ufal-mff/crac2025_empty_nodes_baseline/

[17] UWB = University of West Bohemia.

[18] PUXAI refers to the system by Nguyễn Xuân Phúc.

Language	Recall	Precision	F1	*2024 F1*
ca_ancora	91.1	91.9	91.5	*91.7*
cs_pcedt	61.4	77.1	68.4	*67.8*
cs_pdt	74.9	81.0	77.8	*76.2*
cu_proiel	79.0	81.0	80.0	*80.2*
es_ancora	93.4	92.9	93.2	*92.0*
grc_proiel	86.3	89.7	88.0	*88.4*
hu_korkor	83.3	85.5	84.4	*66.7*
hu_szeged	87.8	88.9	88.3	*90.7*
pl_pcc	91.9	89.0	90.4	*89.5*
tr_itcc	94.0	79.8	86.3	*85.8*

Table 2: Empty nodes prediction baseline performance on the minidev sets of CorefUD 1.3 languages containing empty nodes. An empty node is considered correct if it has the correct dependency head, dependency relation, and word order. For comparison, we also show results from the last year on CorefUD 1.2 dev sets.

GLaRef,[19] NUST-SEECS,[20] ÚFAL CorPipe,[21] and Stanford NLP Group.[22] For clarity, we distinguish the submissions to the LLM track with the 'LLM-' prefix in the following text.

LLM-UWB (hejmanj) The UWB team fine-tunes a Llama-3.1-8B model on the official plaintext export of the CoNLL-U files. Training is done using QLoRA adaptation. The model is trained to generate the fully tagged document text, including empty nodes, by inserting them directly in the output. For some datasets, they modify the input format to use just a headword for mention representation. Two variants of the model are trained: a simple version using the provided format, but ignoring empty nodes, and an extended version with empty nodes and headword mention representation. Versions for the final submission was selected based on dev set results. The simple version is used for: cs_pcedt, cs_pdt, es_ancora, grc_proiel, hu_korkor, ko_ecmt, lt_lcc, and pl_pcc. For hbo_ptnk, the model was not properly trained due to very long sequences and inefficient tokenization, and the system failed to meet the output format. Input windows up to 4 096 tokens are used in training; at inference time, contexts of 2 048 tokens and outputs of 4 096 tokens are typical, with occasional extensions to 8 192/16 384. No additional data is used.

LLM-PUXCRAC2025 (PuxAI) This system is purely prompt-based, few-shot coreference resolver combining two closed-source LLMs (Gemini-Flash-2.0 and Grok-3). A difficulty-aware pipeline selects three hardest examples per language, reranks them by two semantic scores, and feeds them plus the test document into the model. Output chains are post-processed into CoNLL-U. No fine-tuning or extra data is used; the system runs free of charge on public tiers.

LLM-GLaRef-CRAC25 (oseminck) The authors fine-tune google/gemma-3-12b-it in two stages: a context-free end-to-end tagger, and a context-aware variant that processes chunks of sentences (8 or 10 at a time) with preceding context of 500–700 characters. The best three runs (context-free, 8sent_500char, 10sent_700char) are combined for the final submission. Training follows QLoRA + prompt tuning + quantization over plaintext inputs; no extra data are used.

LLM-NUST-FewShot (moizsajid) This system applies few-shot in-context learning with Gemini 2.5 Pro. Up to 300k tokens of input are allowed; generation limits are defined by the task. No fine-tuning or additional data are used. The system demonstrates that performance scales with the number of examples provided

GLaRef-Propp (antoine.bourgois) This work is based on a multi-stage pipeline built on google/mt5-xl. Empty nodes are detected first (pro-drop languages only), then mentions with a BiLSTM-CRF, followed by a mention-pair feed-forward coreference scorer. Windows of up to 512 subwords are used, with sliding overlaps. The three modules contain approximately 54 million trainable parameters and are all fine-tuned solely on CoNLL-U input.

CorPipeSingle (ÚFAL CorPipe) The system utilizes a PyTorch re-implementation of CorPipe24 using google/umt5-xl. Mentions and links are predicted jointly, but empty nodes are taken from the provided baseline. The model is trained multilingually for 150k gradient updates over 15 epochs;

[19] GLaRef = Group Lattice for Reference. Two systems are submitted under this name: GLaRef-CRAC25 and GLaRef-Propp.

[20] NUST-SEECS = National University of Sciences and Technology, School of Electrical Engineering and Computer Science.

[21] ÚFAL CorPipe submitted three variants: CorPipeSingle, CorPipeBestDev, and CorPipeEnsemble.

[22] Stanford NLP Group is the creator of the Stanza package.

batch sizes of 6–16 sentences with proportional sampling yield the final selected checkpoint.

CorPipeBestDev (ÚFAL CorPipe) Same architecture as CorPipeSingle, but instead of one fixed checkpoint, the best checkpoint per treebank (out of 13 models trained with different seeds and sampling) is selected on the mini-dev sets.

CorPipeEnsemble (ÚFAL CorPipe) An ensemble of the top five out of the 13 multilingual umT5-xl models from CorPipeSingle, averaging their predicted mention-pair probabilities.

Stanza (Stanford NLP Group) This work is based on a head-joining efficient word-level conference approach, built on the work of Dobrovolskii (2021); D'Oosterlinck et al. (2023); Liu et al. (2024). Mentions are first linked by head words, after which spans are resolved locally through a CNN. Embeddings for mention resolution are initialized via XLM-RoBERTa large, with a sliding window over the document 512 tokens wide.

4.3 System Comparison

Overview of tables Tables 3–5 provide a comprehensive comparison of all nine submissions. Table 3 lists each system's shared-task track, primary pretrained backbone, and key methodological components (e.g. fine-tuning, prompt tuning, few-shot prompting, pipeline modules). Table 4 details each model's maximum input context length, maximum new tokens generated at inference, and total number of trainable parameters. Finally, Table 5 outlines the training regimes: whether models were tuned per language, the batch sizes used, the total number of gradient updates, which hyperparameters were tuned, and how empty nodes were handled.

Although all nine submissions share the same official CoNLL-U training data and target format, they diverge along four main dimensions: modelling paradigm, context capacity, empty node handling, and language- or treebank-specific adaptation.

Modeling paradigms There are four contributions in the LLM track and five submissions in the unconstrainted track. The four LLM-track systems (LLM-UWB, LLM-PUXCRAC2025, LLM-GLaRef-CRAC25, LLM-NUST-FewShot) treat coreference as a text-generation or prompt-answering task. LLM-UWB and LLM-GLaRef-CRAC25 perform full fine-tuning (via QLoRA, LoRA, quantization, or prompt tuning) of large open-source models (Llama-3.1-8B, gemma-3-12b-it), teaching them to output bracketed and empty-node-annotated text. In contrast, LLM-PUXCRAC2025 and LLM-NUST-FewShot use purely few-shot or in-context prompting on closed-source models (Gemini, Grok), with no parameter updates.

Unconstrained-track submissions (GLaRef-Propp, CorPipeSingle, CorPipeBestDev, CorPipeEnsemble, Stanza) adopt a more traditional, mention detection – mention-pair scoring pipeline. These systems fine-tune XLM-RoBERTa, mT5-xl or umT5-xl in a supervised manner and build clusters via antecedent ranking and transitive closure.

Context capacity and model scale The LLM-track systems exploit the extended context windows of modern LLMs: LLM-UWB up to 8 192 input / 16 384 output tokens, LLM-PUXCRAC2025 effectively unlimited (1 048 576), and LLM-NUST-FewShot 300 000 tokens. LLM-GLaRef-CRAC25 similarly benefits from large-context inference. By contrast, the Unconstrained track systems are limited by standard transformer lengths (512–2 560 subwords), relying on sliding windows or chunking to cover long documents. Model sizes range from 54 M trainable parameters in GLaRef-Propp's BiLSTM-CRF modules to 12 B in gemma-3-12b-it, with most systems clustering around 1.7 B–8 B parameters.

Data usage All nine systems use only the official CoNLL-U data, with no additional corpora. Most train a single multilingual model rather than separate per-language models. The only exception is the CorPipeBestDev system, which picks the best checkpoint per treebank. In terms of computational cost, only LLM-NUST-FewShot reports a non-zero expense (about $234.7), while all other systems either report zero cost or rely on university computing resources.

Empty node handling Empty nodes are addressed in different ways: (1) predicted end-to-end with a fine-tuned system (LLM-UWB and LLM-GLaRef-CRAC25), (2) predicted end-to-end via in-context learning (LLM-PUXCRAC2025 and LLM-NUST-FewShot), (3) adopted from the shared task's baseline (CorPipe variants, Stanza), or (4) predicted with a custom model (GLaRef-Propp). The LLM-based systems relied on the serialized

Name	Track	Techniques
LLM-UWB	LLM	FT, LoRA, QLoRA, quant.
LLM-PUXCRAC2025	LLM	few-shot, re-rank
LLM-GLaRef-CRAC25	LLM	FT, prompt-tune, QLoRA, quant.
LLM-NUST-FewShot	LLM	few-shot in-context
GLaRef-Propp	Unconstr.	BiLSTM-CRF + feedforward
CorPipeSingle	Unconstr.	FT multistage
CorPipeBestDev	Unconstr.	FT + per-treebank select
CorPipeEnsemble	Unconstr.	FT + ensemble
Stanza	Unconstr.	FT + LoRA

Table 3: System names, task tracks, and main techniques.

Name	Model	Input ctx. len.	Output tok. len.	#Params
LLM-UWB	Llama-3.1-8B	8,192	16,384	8 B
LLM-PUXCRAC2025	Gemini-Flash-2.0 Grok-3	1,048,576	16,384	—
LLM-GLaRef-CRAC25	gemma-3-12b-it	—	—	12 B
LLM-NUST-FewShot	Gemini 2.5 Pro	300,000	—	—
GLaRef-Propp	mt5-xl	512	—	54 M
CorPipeSingle	umT5-xl	512/2,560	—	1.7 B
CorPipeBestDev	umT5-xl	512/2,560	—	1.7 B
CorPipeEnsemble	umT5-xl	512/2,560	—	8.6 B
Stanza	XLM-RoBERTa-L	512	—	31M + 560M frozen

Table 4: Models: model name, maximum input context length, maximum new tokens generated, and model sizes.

Name	Empty nodes	Batch size	Grad ups	Tuned h-params
LLM-UWB	predicted ignored	1	?	?
LLM-PUXCRAC2025	predicted	few-shot	0	—
LLM-GLaRef-CRAC25	predicted	?	?	?
LLM-NUST-FewShot	predicted	few-shot	0	—
GLaRef-Propp	predicted	16,000 mention pairs	1.26 M	batch, epochs
CorPipeSingle	baseline	6 sentences	150 k	sampling mode
CorPipeBestDev	baseline	6 sentences	150 k × 13	same as Single
CorPipeEnsemble	baseline	6 sentences	150 k × 5	same as Single
Stanza	baseline	10·512-token windows	367 k	learning rate, warmup, LoRA params, …

Table 5: Training configuration: empty-node handling, batch sizes, total gradient updates, and tuned hyperparameters. GLaRef-Propp used batch size: 16 sentences for empty nodes prediction and mention detection and 16,000 mention pairs for coreference resolution.

format, which represents empty nodes using '##' markers (see Figure 1). These varied approaches reflect different assumptions about the importance and difficulty of modeling zero-anaphora phenomena.

Language/treebank specialization and ensembling Most systems train a single multilingual model for all languages (LLM-UWB, LLM-PUXCRAC2025, GLaRef-CRAC25, NUST-FewShot, GLaRef-Propp, CorPipeSingle, Stanza). Only CorPipeBestDev and CorPipeEnsemble select or combine checkpoints: CorPipeBestDev picks the best of 195 (13 models · 15 epochs) multilingual checkpoints for each corpus, while Ensemble averages the top five multilingual models. Neither LLM-UWB nor LLM-GLaRef-CRAC25 employ per-language tuning, favoring a unified model. The few-shot systems dynamically adapt to each input

via prompt construction but do not explicitly retrain per language.

In sum, the task saw a spectrum from lightweight, prompt-only solutions on closed LLM APIs to heavyweight, quantized fine-tuned open models, and from end-to-end generation of annotations to modular neural-pipeline architectures.

5 Results and Comparison

Main results The main results are summarized in Table 6. LLM-GLaRef-CRAC25 and CorPipeEnsemble are the top-performing systems in the LLM and Unconstrained tracks, respectively, outperforming all other submissions in their respective tracks according to the primary metric. Both systems also achieve the best results within their track when evaluated with alternative mention matching strategies: partial match, exact match, and head match including singletons.

The LLM track exhibits tighter competition, with performance differences between systems significantly smaller than in the Unconstrained track. Excluding the baseline system, the standard deviation of the head match score in the Unconstrained track is 5.53, compared to just 1.27 in the LLM track. This higher level of competition is also reflected in the progression of scores over time, as shown in Figure 3 in Appendix D, which tracks the evolution of primary scores for individual submissions during the evaluation phase of the shared task.

Comparing across the tracks, all LLMs could beat the non-LLM baseline system. However, we have to admit that in this shared task the best LLM solution fell behind the best non-LLM system by a large margin of almost 13 points. For simplicity, we will be comparing the submissions from both tracks jointly in the remainder of this section.

Secondary metrics The secondary metrics in Table 7 reveal a similar trend as the primary metric: the ÚFAL CorPipe system consistently outperforms all other submissions. The most striking pattern is the pronounced contrast between the CorPipe systems and the remaining entries, particularly the LLM-based ones, in terms of the precision–recall balance across individual coreference metrics. While CorPipe systems maintain relatively small gaps between precision and recall, the other systems consistently show much higher precision than recall. This indicates that CorPipe systems are substantially more effective at capturing and following the coreference annotation guidelines reflected in the data.

Comparison across datasets Both Table 8 and Figure 2 present CoNLL F_1 scores of all systems across the datasets. To make patterns more visible, the datasets in Figure 2 are ordered from left to right by the decreasing performance of the top system, CorPipeEnsemble. For roughly the lower-performing half of the datasets, the performance gap between CorPipe and the other systems tends to be larger, and their scores are more varied, suggesting that these datasets pose greater challenges for coreference resolution.

Interestingly, CorPipeEnsemble was outperformed on two datasets: en_litbank by LLM-UWB, and hbo_ptnk by LLM-NUST-FewShot. The latter is particularly striking: on Ancient Hebrew, LLM-NUST-FewShot surpassed CorPipeEnsemble by 10 points, despite ranking among the weakest systems on many other datasets. While the exact cause of this anomaly remains unclear, a closer analysis shows that LLM-NUST-FewShot produced almost exactly the same number of non-singleton mentions as in the gold data (2,327 vs. 2,312), whereas all other system produced less mentions.

The zero score of LLM-UWB on hbo_ptnk is in line with their fine-tuning failure described in Section 4.2.

Performance on zero mentions Table 9 shows system performance on datasets containing zero mentions, evaluated using the anaphor-decomposable score for zero anaphora. Two observations stand out.

First, LLM-UWB fails to predict any zero mentions for all but two of these datasets. This is likely because several of these datasets substantially overlap with those for which the authors used an LLM variant fine-tuned on data where empty nodes had been excluded.

Second, on hu_korkor, both the winning system and the baseline outperform their counterparts from last year's edition by 8 and 10 percentage points, respectively. The winning system's score is now closer to its performance on the other Hungarian dataset, hu_szeged. These gains are consistent with the improved intrinsic performance of the empty-node prediction baseline for this dataset (see Section 4.1), resulting from fixes to its conversion pipeline described in Section 2.2.

Comparison over years Having organized this shared task for the fourth consecutive year, it is particularly interesting to examine how it has contributed to advancing the state of the art in multilingual coreference resolution. While the datasets and certain aspects of the task have evolved each year, one constant has been the coreference baseline system, which is simply retrained annually on the updated data. This stability allows us to track progress by comparing the best-performing system each year against the baseline.

The relative improvement over the baseline showed a promising upward trend in previous editions: +21% in 2022, +31% in 2023, and +39% in 2024 (Novák et al., 2024). This year, however, the improvement stands at +35%, marking a slight break in the upward trajectory. This drop is caused by the exclusion of two very small datasets from the test set, where the improvement over baseline has been exceptionally high last year (+47% in de_parcorfull and +108% in en_parcorful) perhaps by chance. Still, the results show that systems continue to deliver strong performance even as the task grows more diverse and challenging.

Further analysis Similarly to previous years, we provide several additional tables in the appendices to shed more light on the differences between the submitted systems.

Tables 10–11 show results factorized according to the different universal part of speech tags (UPOS) in the mention heads.

Tables 12–15 show various statistics on the entities and mentions in a concatenation of all the test sets. Note that such statistics are mostly influenced by larger datasets.

Differences between LLM and Unconstrained The main novelty in this year's shared task setup was the support for LLM approaches to coreference resolution. As mentioned in the Main Results above, the performance of the LLM participating systems is worse than the best Unconstrained system (CorPipe) by a large margin (with only two datasets where an LLM system outperforms all Unconstrained systems). In addition, some LLM systems seem to be sensitive to particular datasets: there are dramatic drops in performance (see e.g. the performance declines for `grc_proiel`, `tr_itcc`, `hbo_ptnk`, and `cu_proiel` in Figure 2).

However, it would be premature to conclude that LLMs are not a promising solution for coreference resolution. First, this would contradict everyday experience with public LLMs, which seem to handle coreference-related phenomena relatively well. Second, the best-performing CorPipe system has been tuned for CorefUD over years, while LLM approaches had only a few months of testing. Third, and perhaps most importantly, we are still at the beginning of learning how to best provide LLMs with coreference-annotated data and how to elicit coreference reasoning, questions that clearly require further exploration.

6 Conclusions and Future Work

The paper summarizes the fourth edition of the shared task on multilingual coreference resolution, organized in 2025. Besides relatively conservative (though important too) updates with respect to the previous editions, such as improved quality of the data integrated in CorefUD and the increased number of languages, the major innovation in this edition was the support for LLM-based solutions. With only a few exceptions, LLM-based solutions did not outperform CorPipeEnsemble, the best Unconstrained system (from the same author as the winning submissions in the previous editions). However, we believe that the lower performance of the LLM solutions should be rather attributed to our currently limited knowledge of how coreference is handled internally in LLMs, and that studying how to deal with coreference in LLMs may – in a longer-term perspective – result in rethinking how we should represent coreference in NLP in general.

Acknowledgements

This work has been supported by Charles University Research Centre program No. 24/SSH/009, Ministry of Education, Youth, and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ and CZ.02.01.01/00/23_020/0008518, and the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR). We thank all the participants of the shared task for participating and for providing brief descriptions of their systems. We thank Kirill Milintsevich for the initial conversion of French-ANCOR into CorefUD, and Ian Porada for his assistance with the conversion of Korean-ECMT. We also thank anonymous reviewers for their useful remarks.

	excluding singletons			with singletons
system	**head-match**	**partial-match**	**exact-match**	**head-match**
LLM-GLaRef-CRAC25	**62.96**	**61.66** (-1.30)	**58.98** (-3.98)	**65.61** (+2.66)
LLM-NUST-FewShot	61.74	61.14 (-0.60)	56.34 (-5.40)	63.44 (+1.69)
LLM-PUXCRAC2025	60.09	59.68 (-0.41)	55.22 (-4.87)	54.77 (-5.32)
LLM-UWB	59.84	59.55 (-0.29)	38.81 (-21.03)	62.77 (+2.93)
CorPipeEnsemble	**75.84**	**74.90** (-0.94)	**72.76** (-3.08)	**78.33** (+2.49)
CorPipeBestDev	75.06	74.08 (-0.98)	71.97 (-3.10)	77.63 (+2.57)
CorPipeSingle	74.75	73.74 (-1.01)	71.53 (-3.23)	77.43 (+2.68)
Stanza	67.81	67.03 (-0.78)	64.68 (-3.13)	70.64 (+2.83)
GLaRef-Propp	61.57	60.72 (-0.85)	58.43 (-3.14)	65.28 (+3.70)
BASELINE-GZ	58.18	57.75 (-0.42)	56.48 (-1.69)	49.88 (-8.29)
BASELINE	56.01	55.58 (-0.43)	54.24 (-1.77)	47.88 (-8.13)
WINNER-2023	74.90	73.33 (-1.57)	71.46 (-3.44)	76.82 (+1.91)
WINNER-2024	73.90	72.19 (-1.71)	69.86 (-4.04)	75.65 (+1.75)
BASELINE-2023	56.96	56.28 (-0.68)	54.75 (-2.21)	49.32 (-7.64)
BASELINE-2024	53.16	52.48 (-0.68)	51.26 (-1.90)	46.45 (-6.71)

Table 6: Main results: the CoNLL F_1 score macro-averaged over all datasets. The table shows the primary metric (head-match excluding singletons) and three alternative metrics: partial-match excluding singletons, exact-match excluding singletons and head-match with singletons. A difference relative to the primary metric is reported in parenthesis. The top section shows the LLM track, below is the Unconstrained track. The best score in each column and each of these two sections is in bold. The systems are ordered by the primary metric. The last four rows showing the winner and baseline results from CRAC 2023 and 2024 are copied from the last year Findings (Novák et al., 2024), and thus are not directly comparable with the rest of the table because both the test and training data have been changed (CorefUD 1.1 vs. 1.2 vs. 1.3). Similar notes apply to the following tables.

system	**MUC**	**B^3**	**CEAF-e**	**BLANC**	**LEA**	**MOR**
CorPipeEnsemble	**81 / 82 / 82**	**73 / 75 / 74**	**74 / 70 / 72**	72 / **75 / 73**	**70 / 73 / 71**	81 / 82 / **81**
CorPipeBestDev	81 / 81 / 81	72 / 74 / 73	73 / 70 / 71	**72** / 74 / 73	70 / 71 / 70	**81** / 81 / 81
CorPipeSingle	81 / 81 / 81	72 / 73 / 72	72 / 70 / 71	72 / 73 / 72	69 / 71 / 70	80 / 81 / 80
Stanza	72 / 80 / 76	62 / 70 / 65	62 / 64 / 63	61 / 70 / 64	59 / 67 / 62	70 / 83 / 75
LLM-GLaRef-CRAC25	67 / 76 / 71	55 / 67 / 60	55 / 61 / 58	54 / 67 / 59	51 / 64 / 56	64 / 79 / 71
LLM-NUST-FewShot	66 / 73 / 69	58 / 65 / 60	52 / 65 / 56	57 / 65 / 58	56 / 62 / 57	59 / 79 / 66
GLaRef-Propp	69 / 76 / 72	56 / 62 / 58	49 / 62 / 55	56 / 62 / 57	52 / 58 / 55	57 / 78 / 65
LLM-PUXCRAC2025	64 / 72 / 68	54 / 63 / 57	52 / 61 / 55	53 / 62 / 56	51 / 59 / 54	56 / 80 / 65
LLM-UWB	60 / 74 / 65	53 / 67 / 57	53 / 64 / 57	48 / 67 / 53	50 / 64 / 55	42 / 81 / 53
BASELINE-GZ	61 / 76 / 68	48 / 63 / 54	49 / 58 / 52	48 / 64 / 54	45 / 59 / 50	55 / **87** / 66
BASELINE	58 / 75 / 65	45 / 62 / 52	47 / 57 / 51	44 / 63 / 50	42 / 58 / 48	53 / 86 / 65

Table 7: Recall / Precision / F1 for individual secondary metrics. All scores macro-averaged over all datasets.

system	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_potsdam	en_gum	en_litbank	es_ancora	fr_ancor	fr_democrat	grc_proiel	hbo_ptnk	hi_hdtb	hu_korkor	hu_szeged	ko_ecmt	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipeEnsemble	**82.9**	**77.1**	**80.7**	**65.5**	**73.0**	**76.1**	81.8	**84.5**	**76.3**	**71.8**	**74.5**	69.8	**77.7**	**68.6**	**71.0**	**69.9**	**77.2**	**78.2**	**76.3**	**80.2**	84.2	71.2
CorPipeBestDev	82.0	76.3	80.4	62.8	72.6	75.9	81.3	83.8	75.9	69.9	74.3	68.3	77.5	68.3	70.5	69.3	76.0	77.1	74.0	79.9	**84.8**	70.4
CorPipeSingle	82.5	76.2	80.1	63.0	72.8	75.2	80.8	84.1	75.8	70.3	74.4	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	**71.6**
Stanza	79.5	72.7	75.1	40.8	67.3	69.0	74.8	80.4	67.5	62.5	54.9	62.1	74.2	60.0	64.6	67.7	72.8	72.4	71.7	73.0	80.8	47.8
LLM-GLaRef-CRAC25	73.5	65.1	71.3	58.2	59.6	58.7	69.0	74.4	66.7	60.4	65.8	44.0	56.4	52.5	59.8	63.0	62.5	64.7	61.6	72.5	68.8	56.2
LLM-NUST-FewShot	60.9	51.4	54.3	58.5	48.7	69.8	70.4	61.8	71.9	57.6	57.9	**80.2**	71.3	43.5	52.3	66.0	59.2	72.8	68.9	70.8	71.4	39.0
GLaRef-Propp	68.1	61.7	66.6	39.1	61.2	61.9	70.0	69.1	65.1	66.1	51.3	58.8	69.5	50.9	60.1	60.6	57.6	67.1	66.3	68.0	71.5	44.3
LLM-PUXCRAC2025	68.0	56.9	63.0	43.7	57.4	61.7	69.1	70.5	63.8	61.5	47.9	45.3	66.8	50.6	61.6	50.3	65.3	65.2	63.0	66.5	67.6	56.1
LLM-UWB	79.2	61.0	68.2	25.3	67.6	73.6	**84.0**	73.6	58.6	49.1	47.6	0.0	75.8	38.9	67.3	68.3	63.4	73.8	72.0	64.5	80.1	24.3
BASELINE-GZ	68.8	69.5	67.9	29.5	55.7	61.6	66.0	71.0	63.8	55.0	29.4	31.0	66.8	47.1	54.3	64.3	65.3	62.5	63.0	68.1	67.6	51.7
BASELINE	68.0	56.9	63.0	26.3	55.7	61.7	66.0	70.5	63.8	55.0	28.5	31.0	66.8	43.2	54.5	50.3	65.3	62.5	63.0	66.5	67.6	45.9

Table 8: Results for individual languages in the primary metric (CoNLL F_1).

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipeEnsemble	**91** / **87** / **89**	82 / 86 / **84**	61 / 79 / 69	**77** / 80 / **79**	93 / **92** / **92**	87 / 87 / 87	65 / 81 / 72	**85** / **73** / **78**	93 / 84 / **89**	**84** / 83 / **84**
CorPipeBestDev	90 / 87 / 88	82 / 85 / 84	60 / 77 / 68	76 / 79 / 78	**93** / 91 / 92	**87** / 88 / **88**	66 / 82 / 73	83 / 70 / 76	93 / 84 / 88	84 / 82 / 83
CorPipeSingle	90 / 86 / 88	81 / 85 / 83	61 / 78 / 68	77 / 79 / 78	93 / 92 / 92	**87** / 88 / 88	63 / **83** / 72	83 / 70 / 76	**94** / 83 / 88	84 / 82 / 83
Stanza	87 / 86 / 86	77 / **88** / 82	52 / 84 / 65	63 / 69 / 66	91 / 91 / 91	80 / 84 / 82	59 / 83 / 69	74 / 70 / 72	91 / 81 / 86	57 / 83 / 67
LLM-GLaRef-CRAC25	81 / 84 / 82	75 / 81 / 78	56 / 67 / 61	77 / 79 / 78	83 / 89 / 86	85 / 87 / 86	52 / 68 / 59	66 / 65 / 65	84 / 83 / 84	75 / 75 / 75
LLM-NUST-FewShot	53 / 82 / 64	55 / 79 / 65	35 / 81 / 48	74 / **82** / 78	56 / 91 / 69	59 / **89** / 71	23 / 83 / 36	25 / 63 / 36	72 / 86 / 79	29 / 63 / 40
GLaRef-Propp	80 / 80 / 80	74 / 83 / 78	48 / 63 / 54	49 / 56 / 53	84 / 87 / 86	70 / 74 / 72	51 / 70 / 59	66 / 66 / 66	84 / 82 / 83	60 / 83 / 70
LLM-PUXCRAC2025	79 / 75 / 77	34 / 82 / 48	9 / **93** / 17	39 / 53 / 45	88 / 87 / 87	82 / 60 / 69	50 / 48 / 49	73 / 49 / 59	86 / 78 / 82	50 / **93** / 65
LLM-UWB	83 / 82 / 82	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	71 / 73 / 72	0 / 0 / 0	0 / 0 / 0
BASELINE-GZ	84 / 83 / 84	**83** / 85 / 84	**76** / 81 / **79**	61 / 71 / 66	89 / 90 / 90	64 / 67 / 66	**73** / 76 / **74**	54 / 59 / 56	89 / **87** / 88	79 / 81 / 80
BASELINE	79 / 75 / 77	34 / 82 / 48	9 / **93** / 17	52 / 62 / 57	88 / 87 / 87	62 / 67 / 64	56 / 63 / 59	54 / 57 / 55	86 / 78 / 82	71 / 73 / 72
WINNER-2023	93 / 92 / 92	91 / 92 / 92	87 / 88 / 87	–	94 / 95 / 95	–	82 / 89 / 85	88 / 70 / 78	75 / 69 / 72	–
WINNER-2024	88 / 85 / 86	77 / 82 / 80	59 / 74 / 66	75 / 78 / 76	90 / 92 / 91	84 / 88 / 86	56 / 75 / 64	83 / 68 / 75	90 / 84 / 87	83 / 80 / 82
BASELINE-2023	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	–	87 / 88 / 87	–	60 / 68 / 64	61 / 57 / 59	50 / 80 / 62	–
BASELINE-2024	79 / 76 / 77	70 / 74 / 72	55 / 69 / 61	52 / 62 / 56	83 / 83 / 83	63 / 70 / 66	41 / 61 / 49	49 / 57 / 53	85 / 78 / 82	68 / 71 / 70

Table 9: Recall / Precision / F1 for anaphor-decomposable score of coreference resolution on zero anaphors across individual languages. Only datasets containing anaphoric zeros are listed (en_gum excluded as all zeros in its test set are non-anaphoric). Note that these scores are directly comparable to neither the CoNLL score nor the supplementary scores calculated with respect to whole entities in Table 7.

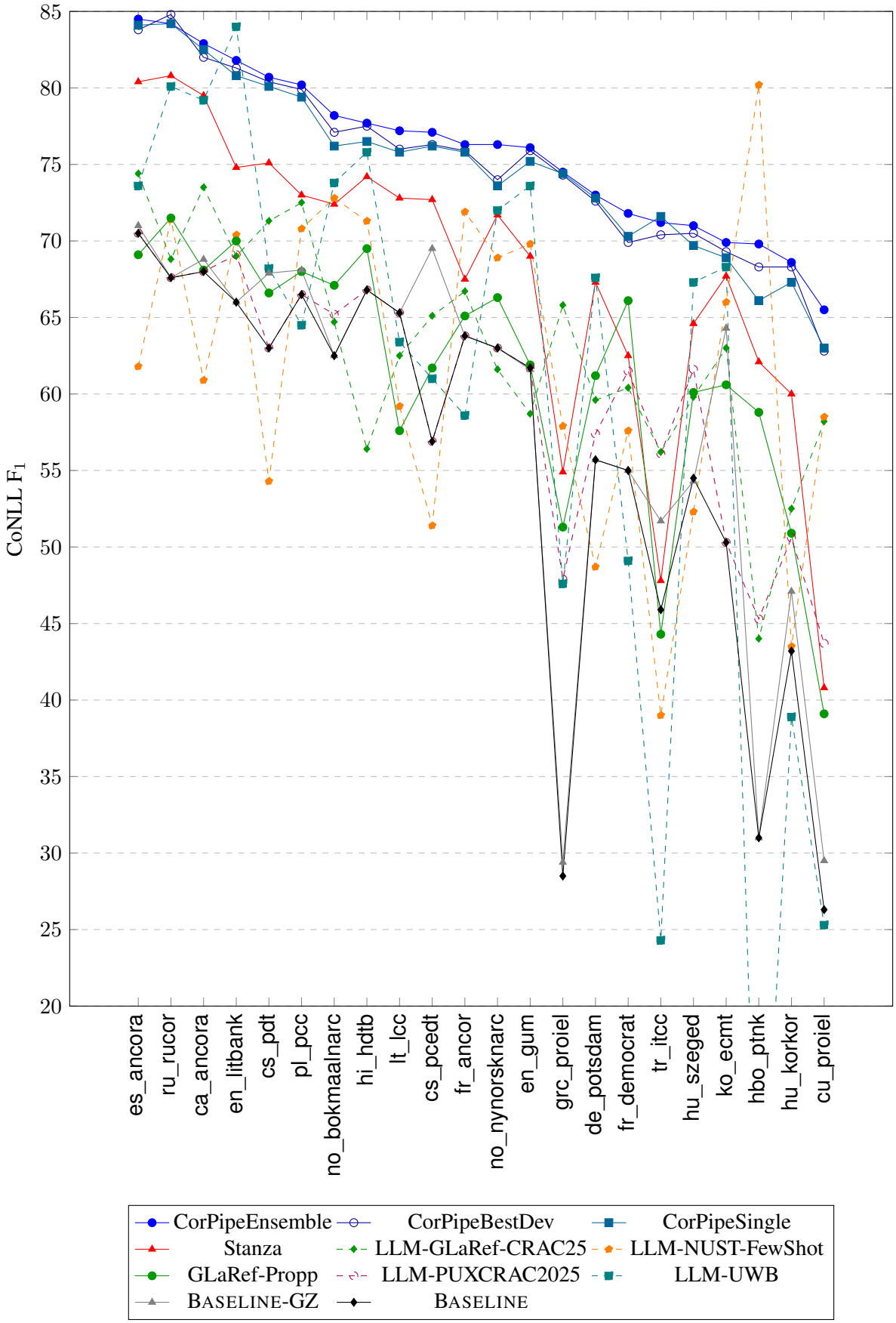

Figure 2: Plot with results for individual languages in the primary metric (CoNLL F_1). This plot shows the same information as Table 8, but languages are sorted according to the performance of the best system and LLM-based systems are shown with dashed lines.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *Proceedings of The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42.

Vladimir Dobrovolskii. 2021. Word-Level Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-Aware Word-level Coreference Resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the Capabilities of Large Language Models in Coreference: An Evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Ine Gevers, Victor De Marez, Luna De Bruyne, and Walter Daelemans. 2025. WinoWhat: A Parallel Corpus of Paraphrased WinoGrande Sentences with Common Sense Categorization. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 68–80, Vienna, Austria. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France. European Language Resources Association.

Dag Trygve Truslew Haug and Marius L. Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*.

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Frédéric Landragin. 2021. Le corpus Democrat et son exploitation. Présentation. *Langages*, 224:11–24.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a Parallel Corpus Annotated with Full Coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Nghia T. Le and Alan Ritter. 2023. Are Large Language Models Robust Coreference Resolvers? ArXiv:2305.14489 [cs].

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561, Rome, Italy. AAAI Press.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, Singleton and Conjunction-Aware Word-Level Coreference Resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT 2005, pages 25–32. Association for Computational Linguistics.

Petter Mæhlum, Dag Haug, Tollef Jørgensen, Andre Kåsen, Anders Nøklestad, Egil Rønningstad, Per Erik Solberg, Erik Velldal, and Lilja Øvrelid. 2022. NARC–Norwegian Anaphora Resolution Corpus. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 48–60, Gyeongju, Korea. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Vandan Mujadia, Palash Gupta, and Dipti Misra Sharma. 2016. Coreference Annotation Scheme and Relation Types for Hindi. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 161–168, Portorož, Slovenia. European Language Resources Association (ELRA).

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sangha Nam, Minho Lee, Donghwan Kim, Kijong Han, Kuntae Kim, Sooji Yoon, Eun-kyung Kim, and Key-Sun Choi. 2020. Effective Crowdsourcing of Multiple Tasks for Comprehensive Knowledge Extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 212–219, Marseille, France. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Portorož, Slovenia. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference Meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the Third Shared Task on Multilingual Coreference Resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Kirill Milintsevich, Vandan Mujadia, Judith Muzerelle, Sangha Nam, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Ian Porada, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Daniel Swanson, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2025. Coreference in Universal Dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Maciej Ogrodniczuk, Katarzyna Glowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2013. Polish Coreference Corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics — 6th Language and Technology Conference (LTC 2013), Revised Selected Papers*, volume 9561 of *Lecture Notes in Computer Science*, pages 215–226. Springer.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Tuğba Pamay and Gülşen Eryiğit. 2018. Turkish Coreference Resolution. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual Coreference Resolution with Harmonized Annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2010. A Typology of Near-Identity Relations for Coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association.

Marta Recasens and Eduard H. Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun. ACM*, 64(9):99–106.

Karol Saputa, Angelika Peljak-Łapińska, and Maciej Ogrodniczuk. 2024. Polish Coreference Corpus as an LLM Testbed: Evaluating Coreference Resolution within Instruction-Following Language Models by Instruction–Answer Alignment. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–32, Miami. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting Zero Mentions from Raw Text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Daniel G. Swanson, Bryce D. Bussert, and Francis Tyers. 2024. Towards Named-Entity and Coreference Annotation of the Hebrew Bible. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 36–40, Torino, Italia. ELRA and ICCL.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association.

Svetlana Toldova, Anna Roytberg, Alina Ladygina, Maria Vasilyeva, Ilya Azerkovich, Matvei Kurzukov, G. Sim, D.V. Gorshkov, A. Ivanova, Anna Nedoluzhko, and Yulia Grishina. 2014. Evaluating Anaphora and Coreference Resolution for Russian. In *Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii Dialog*, pages 681–695.

Noémi Vadász. 2022. Building a Manually Annotated Hungarian Coreference Corpus: Workflow and Tools. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 38–47, Gyeongju, Korea. Association for Computational Linguistics.

Noémi Vadász. 2023. Resolving Hungarian Anaphora with ChatGPT. In *Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*, page 45–57, Berlin, Heidelberg. Springer-Verlag.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. SzegedKoref: A Hungarian Coreference Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2023. The Universal Anaphora Scorer 2.0. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 183–194, Nancy, France. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, and Daniel Zeman. 2023. Findings of the Second Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Korea. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Voldemaras Žitkus and Rita Butkienė. 2018. Coreference Annotation Scheme and Corpus for Lithuanian Language. In *Fifth International Conference on Social Networks Analysis, Management and Security, SNAMS 2018, Valencia, Spain, October 15-18, 2018*, pages 243–250. IEEE.

A CorefUD 1.3 Details

Ancient Greek	PROIEL	grc_proiel	(Haug and Jøhndal, 2008)
Ancient Hebrew	PTNK	hbo_ptnk	(Swanson et al., 2024)
Catalan	AnCora	ca_ancora	(Taulé et al., 2008; Recasens and Martí, 2010)
Czech	PCEDT	cs_pcedt	(Nedoluzhko et al., 2016)
Czech	PDT	cs_pdt	(Hajič et al., 2020)
English	GUM	en_gum	(Zeldes, 2017)
English	LitBank	en_litbank	(Bamman et al., 2020)
English	ParCorFull	en_parcorfull	(Lapshinova-Koltunski et al., 2018)
French	ANCOR	fr_ancor	(Muzerelle et al., 2014)
French	Democrat	fr_democrat	(Landragin, 2021)
German	ParCorFull	de_parcorfull	(Lapshinova-Koltunski et al., 2018)
German	PotsdamCC	de_potsdam	(Bourgonje and Stede, 2020)
Hindi	HDTB	hi_hdtb	(Mujadia et al., 2016)
Hungarian	KorKor	hu_korkor	(Vadász, 2022)
Hungarian	SzegedKoref	hu_szeged	(Vincze et al., 2018)
Korean	ECMT	ko_ecmt	(Nam et al., 2020)
Lithuanian	LCC	lt_lcc	(Žitkus and Butkienė, 2018)
Norwegian	Bokmål NARC	no_bokmaalnarc	(Mæhlum et al., 2022)
Norwegian	Nynorsk NARC	no_nynorsknarc	(Mæhlum et al., 2022)
Old Church Slavonic	PROIEL	cu_proiel	(Haug and Jøhndal, 2008)
Polish	PCC	pl_pcc	(Ogrodniczuk et al., 2013, 2015)
Russian	RuCor	ru_rucor	(Toldova et al., 2014)
Spanish	AnCora	es_ancora	(Taulé et al., 2008; Recasens and Martí, 2010)
Turkish	ITCC	tr_itcc	(Pamay and Eryiğit, 2018)

B CoNLL results by head UPOS

system	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM
CorPipeEnsemble	**71.78**	**71.67**	**78.11**	52.58	47.92	**37.36**	**32.03**	37.40
CorPipeBestDev	71.07	71.13	77.69	49.22	**48.35**	36.62	27.62	38.22
CorPipeSingle	70.96	70.47	77.28	**53.01**	44.69	35.45	31.96	**38.76**
Stanza	62.55	64.24	70.94	41.78	32.77	21.73	21.89	29.58
LLM-GLaRef-CRAC25	58.81	61.23	64.30	41.83	29.26	23.08	20.90	34.52
LLM-NUST-FewShot	58.01	59.21	69.88	32.79	34.39	14.39	20.59	26.36
GLaRef-Propp	56.44	57.99	63.20	36.10	28.43	17.88	20.26	21.56
LLM-PUXCRAC2025	54.71	56.22	64.51	36.55	27.53	15.36	17.86	25.76
LLM-UWB	57.19	55.95	64.72	36.83	29.57	22.30	23.53	26.25
BASELINE-GZ	50.74	58.46	57.21	37.24	25.85	14.15	18.15	23.11
BASELINE	48.44	52.03	54.96	36.75	24.04	13.44	16.98	22.81

Table 10: CoNLL F_1 score (head-match) evaluated only on entities with heads of a given UPOS. In both the gold and prediction files we deleted some entities before running the evaluation. We kept only entities with at least one mention with a given head UPOS (universal part of speech tag). For the purpose of this analysis, if the head node had deprel=flat children, their UPOS tags were considered as well, so for example in "Mr./NOUN Brown/PROPN" both NOUN and PROPN were taken as head UPOS, so the entity with this mention will be reported in both columns NOUN and PROPN. Otherwise, the CoNLL F_1 scores are the same as in the primary metric, i.e. an unweighted average over all datasets, head-match, without singletons. Note that when distinguishing entities into events and nominal entities, the VERB column can be considered as an approximation of the performance on events. One of the limitations of this approach is that copula is not treated as head in the Universal Dependencies, so, e.g., phrase *She is nice* is not considered for the VERB column, but for the ADJ column (head of the phrase is *nice*).

system	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM
CorPipeEnsemble	**63.91**	**61.69**	**64.74**	**53.28**	**51.12**	**50.58**	**50.81**	**50.46**
CorPipeBestDev	62.42	60.85	63.57	52.51	49.91	48.72	49.33	49.00
CorPipeSingle	62.91	60.69	64.05	52.66	49.98	49.66	49.92	49.72
Stanza	54.67	54.66	56.77	44.31	42.51	41.37	42.31	41.78
LLM-GLaRef-CRAC25	50.80	51.80	52.12	41.98	39.11	38.75	39.08	38.81
LLM-NUST-FewShot	52.16	52.84	54.26	42.09	40.05	39.47	40.28	39.96
GLaRef-Propp	47.57	48.85	49.46	36.41	33.83	33.37	34.09	33.58
LLM-PUXCRAC2025	47.37	46.07	49.09	34.88	33.11	31.91	32.71	32.48
LLM-UWB	51.82	47.99	53.14	40.23	37.45	36.91	37.44	36.99
BASELINE-GZ	42.44	49.49	45.96	33.76	31.16	30.43	31.05	30.61
BASELINE	40.99	42.45	44.50	31.94	29.42	28.58	29.17	28.80

Table 11: CoNLL F_1 score (head-match) evaluated only on mentions with heads of a given UPOS. In both the gold and prediction files we deleted some mentions before running the evaluation. We kept only mentions with a given head UPOS (again considering also deprel=flat children). These results may be a bit misleading because e.g. the PRON column does not consider all pronominal coreference, but only pronoun-to-pronoun coreference. An entity with one pronoun and one noun mention is excluded from this table (because it becomes a singleton after deleting noun or pronoun mentions and singletons are excluded from the evaluation in this table).

C Statistics of the submitted systems on concatenation of all test sets

The systems are sorted alphabetically in tables in this section.

	entities				distribution of lengths				
system	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
gold	39,576	108	509	2.1	67.4	17.3	5.9	2.8	6.6
BASELINE	10,591	29	347	4.2	0.0	55.8	17.6	7.8	18.9
BASELINE-GZ	10,977	30	354	4.2	0.0	55.5	17.6	7.8	19.2
CorPipeBestDev	40,392	111	248	2.1	66.6	17.7	6.2	2.8	6.6
CorPipeEnsemble	40,615	111	461	2.0	66.5	17.8	6.3	2.9	6.5
CorPipeSingle	40,377	111	362	2.1	66.6	17.7	6.2	3.0	6.6
GLaRef-Propp	40,481	111	563	1.9	75.0	12.4	4.6	2.3	5.7
LLM-GLaRef-CRAC25	39,664	109	280	1.9	70.6	15.1	5.6	2.7	6.0
LLM-NUST-FewShot	35,703	98	393	2.0	71.1	13.5	5.5	2.8	7.1
LLM-PUXCRAC2025	19,896	55	545	2.9	44.3	29.4	10.1	4.8	11.5
LLM-UWB	35,542	97	317	1.9	70.0	15.6	5.6	2.8	6.0
Stanza	38,464	105	523	2.0	67.8	17.4	5.9	2.8	6.2

Table 12: Statistics on coreference entities. The total number of entities and the average number of entities per 1000 tokens in the running text. The maximum and average entity "length", i.e., the number of mentions in the entity. Distribution of entity lengths (singletons have length = 1). The two baselines and LLM-PUXCRAC2025 heavily undergenerate (i.e. predict less entities than in the gold data) and the baselines also predict on average longer entities (i.e. with more mentions) than in the gold data. The remaining systems have the statistics similar to the gold data, (although the CorPipe* systems and GLaRef-Propp slightly overgenerate, while LLM-NUST-FewShot and LLM-UWB undergenerate).

system	non-singleton mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	55,333	152	100	2.5	9.8	50.1	19.1	7.0	3.3	10.8
BASELINE	44,110	121	27	1.9	10.0	54.9	18.8	6.3	2.6	7.3
BASELINE-GZ	45,989	126	27	1.9	11.4	54.2	18.5	6.2	2.6	7.1
CorPipeBestDev	56,020	154	149	2.4	9.6	51.0	19.1	6.9	3.1	10.3
CorPipeEnsemble	55,668	153	149	2.4	9.6	51.0	19.0	6.9	3.1	10.2
CorPipeSingle	56,026	154	140	2.5	9.6	50.9	19.1	6.9	3.1	10.4
GLaRef-Propp	48,362	133	51	1.9	9.9	55.3	19.2	6.4	2.6	6.6
LLM-GLaRef-CRAC25	49,311	135	96	2.3	10.7	52.1	18.6	6.4	3.0	9.2
LLM-NUST-FewShot	47,681	131	104	2.0	6.9	58.0	19.1	6.2	2.6	7.2
LLM-PUXCRAC2025	48,593	133	27	1.8	8.4	57.8	18.4	5.9	2.5	6.9
LLM-UWB	42,852	117	58	1.8	1.2	80.6	8.3	2.9	1.4	5.6
Stanza	50,811	139	100	2.3	9.3	52.8	18.9	6.6	2.9	9.6

Table 13: Statistics on non-singleton mentions. The total number of mentions and the average number of mentions per 1000 words of running text. The maximum and average mention length, i.e., the number of nonempty nodes (words) in the mention. Distribution of mention lengths (zeros have length = 0). Only the CorPipe* systems generate a similar number of non-singleton mentions as in the gold data, all other systems generate less mentions. The average length of mentions predicted by LLM-UWB is notably lower than in the gold data because LLM-UWB predicted single-word mentions only in most datasets. All other systems have the distribution of mention lengths similar to the gold data, although no system predicts long mentions (4 and 5+ words) more frequently than in the gold data, (but CorPipe* systems are near to the gold distribution).

system	singleton mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	26,661	73	81	3.0	0.7	39.4	24.0	12.2	6.3	17.3
BASELINE	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BASELINE-GZ	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CorPipeBestDev	26,919	74	112	3.1	0.7	38.1	24.8	12.7	6.4	17.3
CorPipeEnsemble	27,014	74	112	3.0	0.7	38.5	25.0	12.5	6.2	17.0
CorPipeSingle	26,885	74	85	3.1	0.7	38.5	24.9	12.6	6.3	17.1
GLaRef-Propp	30,343	83	33	2.3	2.4	40.4	27.7	13.0	6.1	10.5
LLM-GLaRef-CRAC25	28,021	77	80	2.9	0.9	40.5	25.1	12.2	5.8	15.5
LLM-NUST-FewShot	25,379	70	63	2.8	0.2	41.8	24.9	12.0	5.9	15.3
LLM-PUXCRAC2025	8,807	24	17	2.0	0.4	52.5	23.8	11.4	4.1	7.8
LLM-UWB	24,889	68	86	1.7	0.0	78.2	10.0	4.3	2.1	5.4
Stanza	26,060	71	100	2.9	1.4	40.2	24.5	11.8	6.1	16.0

Table 14: Statistics on singleton mentions. See the caption of Table 13 for details. The two baseline systems do not attempt to predict singletons at all. LLM-PUXCRAC2025 heavily undergenerates singletons. GLaRef-Propp overgenerates singletons (including zeros), but note that singletons are not annotated in all the (gold) datasets.

	mention type [%]			distribution of head UPOS [%]									
system	w/empty	w/gap	non-tree	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	_	other
gold	11.0	0.7	1.4	38.6	31.5	17.7	4.2	1.3	1.9	1.4	0.5	2.1	0.8
BASELINE	10.5	0.0	1.4	35.4	26.9	18.7	4.8	1.1	0.9	1.2	0.4	10.0	0.6
BASELINE-GZ	12.0	0.0	1.5	35.1	34.7	18.5	4.7	1.1	0.9	1.5	0.4	2.5	0.8
CorPipeBestDev	10.6	0.0	1.9	39.0	23.7	17.6	4.3	1.2	1.8	1.4	0.5	9.6	0.8
CorPipeEnsemble	10.6	0.0	1.8	39.0	23.8	17.7	4.3	1.2	1.7	1.4	0.5	9.6	0.8
CorPipeSingle	10.5	0.0	1.9	39.1	23.7	17.6	4.3	1.2	1.7	1.4	0.5	9.6	0.8
GLaRef-Propp	9.9	0.0	1.4	35.5	26.9	18.4	4.7	1.1	0.8	1.4	0.4	9.9	0.9
LLM-GLaRef-CRAC25	11.4	0.0	1.8	37.5	24.7	17.0	4.7	1.3	1.4	1.4	0.5	10.7	1.0
LLM-NUST-FewShot	7.1	0.0	1.3	39.4	25.9	18.6	3.5	1.2	1.5	1.5	0.5	6.9	1.1
LLM-PUXCRAC2025	8.9	0.0	1.4	37.2	25.7	18.7	4.1	1.3	2.0	1.3	0.5	8.4	0.8
LLM-UWB	1.2	0.0	0.8	42.9	24.6	20.9	4.8	1.3	1.1	1.7	0.5	1.2	1.0
Stanza	10.0	0.0	1.4	39.0	24.0	18.8	4.1	1.1	1.1	1.4	0.4	9.3	0.8

Table 15: Detailed statistics on non-singleton mentions. The left part of the table shows the percentage of: mentions with at least one empty node (w/empty); mentions with at least one gap, i.e. discontinuous mentions (w/gap); and non-treelet mentions, i.e. mentions not forming a connected subgraph (catena) in the dependency tree (non-tree). Note that these three types of mentions may be overlapping. We can see that none of the systems attempts to predict discontinuous mentions. LLM-UWB has a notably lower percentage (0.8%) of non-treelet mention spans, but this is simply explained by its higher percentage (80%) of single-word mentions. The right part of the table shows the distribution of mentions based on the universal part-of-speech tag (UPOS) of the head word. Note that this distribution has to be interpreted with the total number of non-singleton mentions predicted (as reported in Table 13) in mind. For example, 34.7% of non-singleton mentions predicted by BASELINE-GZ are pronominal (head=PRON), while there are only 31.5% of pronominal non-singleton mentions in the gold data. However, BASELINE-GZ predicts actually less pronominal non-singleton mentions ($45{,}989 \times 34.7\% \approx 15{,}958$) than in the gold data ($55{,}333 \times 31.5\% \approx 17{,}430$). Note that the same word may be assigned a different UPOS tag in the predicted and gold data (in case of empty nodes or if the gold data includes manual annotation). The empty UPOS tag (_) is present only in the empty nodes and none of the systems attempts to predict the actual UPOS tag of empty nodes (they all keep the empty tag from the baseline predictor of empty nodes, although about 78% of the empty nodes in the gold devset are pronouns).

D Evolution of CodaLab Submissions

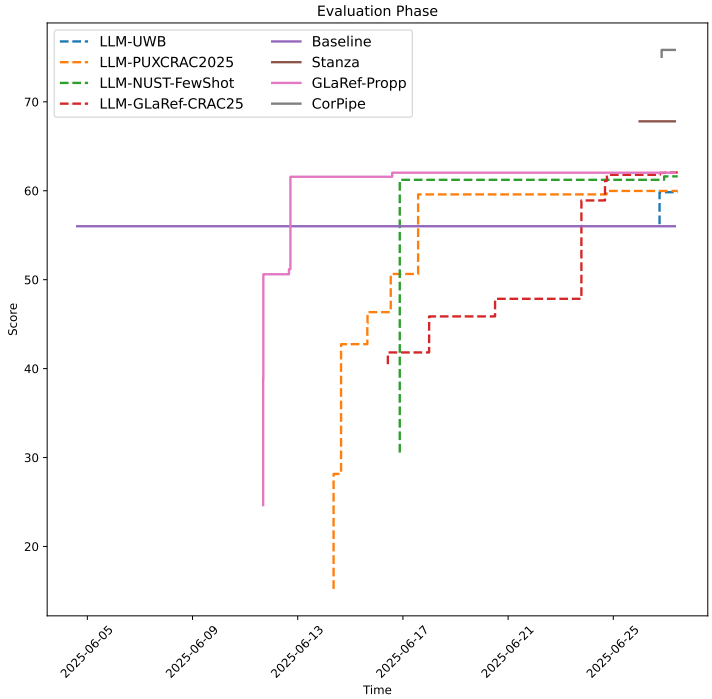

Figure 3: Evolution of CodaLab Submissions in the evaluation phase. The submissions to the LLM and Unconstrained track are shown by using the dashed and solid lines, respectively. For clarity, all submissions of the ÚFAL CorPipe team are represented by the scores of CorPipeEnsemble.

GLaRef@CRAC2025: Should We Transform Coreference Resolution into a Text Generation Task?

Olga Seminck* **and Antoine Bourgois*** **and Yoann Dupont***
and Mathieu Dehouck* **and Marine Delaborde†**

* Lattice (CNRS UMR 8094 & ENS-PSL & Université Sorbonne Nouvelle), Montrouge, France
† LT2D (EA 7518, CY Cergy Paris Université), Cergy-Pontoise, France

olga.seminck@cnrs.fr, antoine.bourgois@protonmail.com,
yoann.dupont@sorbonne-nouvelle.fr, mathieu.dehouck@cnrs.fr,
marine.delaborde@cyu.fr

Your scientists were so preoccupied with whether they could, they didn't stop to think if they should.[1]

Abstract

We present the submissions of our team to the Unconstrained and LLM tracks of the Computational Models of Reference, Anaphora and Coreference (CRAC2025) shared task, where we ended respectively in the fifth and the first place, but nevertheless with similar scores: average CoNLL-F1 scores of 61.57 and 62.96 on the test set, but with very large differences in computational cost. Indeed, the classical pairwise resolution system submitted to the Unconstrained track obtained similar performance but with less than 10% of the computational cost. Reflecting on this fact, we point out problems that we ran into using generative AI to perform coreference resolution. We explain how the framework of text generation stands in the way of a reliable text-global coreference representation. Nonetheless, we realize there are many potential improvements of our LLM-system; we discuss them at the end of this article.

1 Coreference Resolution

Coreference resolution, the task of identifying and grouping textual linguistic expressions (mentions) that refer to the same entity, has been studied since the 1970s, beginning with rule-based systems for pronouns (Winograd, 1972; Hirst, 1981). The Message Understanding Conference (MUC) initiated a standardised framework for a coreference resolution shared task with the MUC-6 challenge (Grishman and Sundheim, 1995). Data-driven machine learning methods appeared with the availability of annotated corpora, initially in English. Subsequently, detection systems using statistical classifiers and pairs of mentions were developed (Soon et al., 2001), then mention-ranking systems like Denis and Baldridge (2008), usually in two stages: mention detection then coreference resolution. End-to-end global models later emerged and were evaluated in the CoNLL shared tasks (Pradhan et al., 2011, 2012). The arrival of deep neural models marked a turning point for the coreference resolution with models often inspired by Lee et al. (2017) later being replaced by BERT-based models (Joshi et al., 2019) and encoder-decoder architectures (Raffel et al., 2020), all contributing to improvements on benchmark datasets (Porada et al., 2024). In recent years, solutions based on seq2seq models (Zhang et al., 2023) and generative LLMs (Zhu et al., 2025) have also been proposed. These have been praised for their performance, while also revealing limitations (Gan et al., 2024); prompting reflection on the relevance of using such approaches for coreference resolution.

2 CRAC: Task Description and Corpora

The CRAC shared task 2025 is part of a series of annual challenges since 2016[2].

In 2024, the detection of zero mentions was added to the task[3] as were 4 new datasets (ancient Greek, Old Church Slavonic, Ancient Hebrew and English litBank) (Novák et al., 2024). CorPipe 24 (Straka, 2024), the winning entry in 2024, used a pretrained language encoder model with two variants: a two stages model (mentions detection then coreference resolution) and a single stage model.

Since 2025, the task corpus is based on CorefUD 1.3. (Novák et al., 2025) and contains 22 datasets for 17 languages, including for the first time ANCOR (Muzerelle et al., 2011), a French spoken language corpus. In addition to the Unconstrained track, a new LLM track was introduced this year, which

[1] Dr. Ian Malcolm - *Jurassic Park* (dir. S. Spielberg, 1993).

[2] https://corbon.nlp.ipipan.waw.pl/.

[3] With three possible starting points: coreference and zeros from scratch, coreference from scratch, refine the baseline.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 119–129
Suzhou, China, November 9, 2025.

focuses on using only large language models to resolve coreference, via prompting, fine-tuning, or in-context learning.

The Universal Anaphora corpus (which is the source corpus for the CRAC task) brings together independently created corpora in different languages. The different annotation schemes (when available) indicate that the concept of coreference can include various phenomena depending on each corpus. Indeed, some corpora contain annotations for all the referring expressions, while some others include selected expressions only, such as the English-LitBank corpus, which is annotated in coreferences only for a subset of entity types (Bamman, 2020). Despite efforts to standardise the format, some phenomena are represented differently in several languages. For example, zero mentions are generally represented by adding empty nodes to the UD trees, such as for the Spanish Ancora (Taulé et al., 2008). Yet, in the French Democrat corpus (Landragin, 2016), zero subjects are annotated on the verb[4].

3 System Descriptions and Results

Our team participated in both the Unconstrained and the LLM tracks submitting results for two entirely different systems. In this section, we describe the two approaches.

3.1 Unconstrained: Mention-Pair System

3.1.1 Architecture

The baseline system used in the Unconstrained track is a mention-pair based multi-stage coreference resolution system adapted from the existing *Propp* processing pipeline.[5]

As a first step, it extracts contextualized token embeddings using a frozen multilingual pretrained transformer encoder (`mt5-xl`[6]), applying overlapping sliding windows to capture maximum context and averaging embeddings across overlaps.

Mention spans are identified using stacked BiLSTM-CRF models trained to predict nested BIOES tags (Ratinov and Roth, 2009) at the sentence level. A separate BiLSTM model is used to identify head tokens for zero mentions.

Mentions are encoded using either the head token (for zero mentions) or the average of the first and last token embeddings (for multi-token spans). Mention-pair representations are the concatenation of the embeddings of two mentions with a rich set of linguistic and positional features, and are scored using a feedforward neural network.

To reduce complexity, the number of antecedent candidates is limited to 80 per mention. Clusters are formed using a highest-ranked-antecedent strategy and refined via transitive closure. Global decisions are improved through leveraging local high-confidence non-coreference links to avoid erroneous later merges.

3.1.2 Training and Computational Resources

Training our unconstrained system involves three main modules: mention detection, zero mention head detection, and coreference resolution. All components rely on word-level embeddings generated by the frozen encoder.[7]

- **Embedding Stage.** We use the `mt5-xl` model to extract contextualized embeddings for all tokens in the training and development datasets. The embedding model alone requires approximately 7.6 GiB of GPU memory. Processing all 12,187 documents (training + minidev) takes 55 minutes[8].

- **Mention Detection Stage.** The mention detection module is trained separately for each nesting level using the precomputed embeddings. The best models were obtained at epoch 23 (~4h46) for nested level 0 and epoch 21 (~4h32) for nested level 1, with a peak memory usage of 3.8 GiB.

- **Zero Mention Head Detection.** Trained similarly to the mention detection module, the best model was obtained after 24 epochs (~2h36), with a peak memory usage of 1.7 GiB.

- **Coreference Resolution.** The coreference resolution module is trained on all mention pairs using a batch size of 16,000 pairs per batch. The best model was obtained after 25 epochs (~3h57), with a peak memory usage of 1.8 GiB.

In the best-case scenario, the different modules are trained in parallel, so the total training time for the entire pipeline corresponds to the embedding

[4] A choice partly motivated by the annotation tool.

[5] https://github.com/lattice-8094/propp

[6] https://huggingface.co/google/mt5-xl

[7] More details about hyperparameters used for training each components can be found in the Appendix A.

[8] All experiments for the unconstrained track are performed on a single 48 GiB Nvidia RTX 6000 Ada Generation GPU.

time plus the duration of the longest individual module, resulting in a total of under 6 hours. Due to the size of the pretrained model used, the embedding step remains the most memory-intensive part of the pipeline and ultimately determines the minimum required GPU size (~8 GiB in our case).

Inference on the test set takes approximately 16 minutes, with peak GPU memory usage of 7.5 GiB. As with training, the embedding remains the main bottleneck, meaning that coreference resolution with this pipeline can be performed on any GPU capable of holding the embedding model.

3.1.3 Unconstrained Track Results

Despite its relatively simple design, our system achieves substantial improvements over the CRAC-2025 provided baseline (Table 1). On average, it yields a 5.56-point absolute gain in CoNLL F1-score across the test corpora. These gains are consistent across most languages, with particularly strong improvements observed on lower-resource corpora such as *grc_proiel* (+22.8), *hbo_ptnk* (+27.8), and *cu_proiel* (+12.8). This demonstrates the system's robustness and its capacity to generalize effectively across diverse linguistic settings.

Corpus	CRAC-coref	GLaRef
ca_ancora	68.01	**68.06**
cs_pcedt	56.94	**61.68**
cs_pdt	62.96	**66.59**
de_potsdamcc	55.70	**61.18**
en_gum	61.71	**61.86**
es_ancora	**70.52**	69.09
fr_democrat	54.99	**66.13**
hu_szegedkoref	54.54	**60.08**
lt_lcc	**65.35**	57.60
pl_pcc	66.55	**67.98**
ru_rucor	67.59	**71.45**
hu_korkor	43.17	**50.87**
no_bokmaalnarc	62.45	**67.09**
no_nynorsknarc	63.00	**66.28**
tr_itcc	**45.92**	44.28
cu_proiel	26.33	**39.10**
en_litbank	65.96	**69.96**
grc_proiel	28.54	**51.34**
hbo_ptnk	31.04	**58.80**
fr_ancor	63.77	**65.11**
hi_hdt	66.85	**69.51**
ko_ecmt	50.32	**60.57**
Average	**56.01**	**61.57**

Table 1: Test results for the Unconstrained track compared to the provided baseline (CRAC-coref).

Our system, adapted from the `Propp` architecture, follows a modular pipeline in which each stage depends on the previous one. This design introduces a key limitation: error propagation. The mention detection module plays a critical role, as errors at this stage directly affect downstream components such as mention pairing and clustering.

A notable challenge arises in datasets where singleton mentions (i.e., mentions not involved in any coreference chain) are not annotated. In such cases, the mention detector is trained only on spans that are part of coreference chains, resulting in an incomplete learning signal. This weakens its ability to identify valid mentions in general, particularly when the coreference resolution component depends entirely on the output of this detector.

This problem is further compounded by inconsistent annotation guidelines across datasets. As mentioned in Section 2, some corpora provide exhaustive mention annotations, while others are more selective. Such inconsistencies make it difficult for the system to generalize across languages and domains, and can lead to performance drops on datasets with different annotation guidelines.

3.2 LLM Track: Fine-tuning Gemma 3

For the LLM track, we developed two models based on fine-tuning of the `Gemma-3-12B-it` model using quantization and one single LoRA (Low-Rank Adaptation) (Biderman et al., 2024) adapter for all corpora. We proceeded to *peft* (parameter-efficient fine-tuning) with 4-bit NormalFloat quantization using QLora (Dettmers et al., 2024). The choice for the Gemma model was motivated by participation of members of our team in the shared task for Multilingual Grammatical Error Correction (MultiGEC-2025) (Masciolini et al., 2025), where they experienced particular problems with Llama 3 for under-resourced languages, in particular Icelandic and Slovene (Seminck et al., 2025). The task was won by a system build on Gemma 2 (Staruch, 2025), which is known to be a reliable multilingual model. Therefore, we decided to work with Gemma models for the current shared task.

We used the text2text-coref tool[9] provided by the CRAC organizers to transform the CoNLL data into a plain text format with in-text annotations and also to transform the system's output in plain-text back to CoNLL format. We proceeded to two distinct fine-tunings: a context-

[9] https://github.com/ondfa/text2text-coref

free model and a context-aware model. Our systems can be found on https://github.com/lattice-8094/GLaRef-CRAC25-LLM-Track.

3.2.1 Context-free Model

This model has the simplest design imaginable for coreference resolution using LLMs. We model the problem as just an annotation of coreference of the text: we give the whole text unannotated as an input, and the gold annotated text in the plain-text format as an output. The text is treated as a whole and there is no modelling of context. The prompt is given in (1). We experimented with different prompts, also leveraging `ChatGPT-4o` to enhance the prompt and give detailed instructions of the annotation schema. But in preliminary experiments, it turned out that a shorter prompt led to better performances and that the annotation schema can be learned implicitly during the fine-tuning of the model. Therefore, we kept a small prompt that is language agnostic.

(1) You are a linguist, expert in anaphora and coreference resolution. You have to annotate in the text which nouns, pronouns and other linguistic expressions refer to the same entity. Do only insert annotations. Do not insert extra linguistic material, nor punctuation markers and do not delete elements from the input texts.

Gemma 3 models can take up to 128K input tokens, so there is theoretically no problem of input length.

Our model was trained for 10 epochs, using batch size of one, for bigger batch sizes, the code threw an out of memory error. The training lasted about 3 days on two Nvidia RTX 6000 Ada Generation GPUs, featuring each 48 GB of memory capacity.

In Table 2, we can see that the results differ substantially across corpora. Whereas for some languages we observe scores above 70 points, for others the system's performance is poor. The main reason for this is the length of the texts per corpus. Despite the promise of handling up to 128K tokens of input, we soon realized that Gemma 3 was not capable of handling long texts properly, at least for this task, but it has been demonstrated for other tasks as well that output tends to degrade for longer texts, even if the maximum input length is respected (Levy et al., 2024; Liu et al., 2024). The system diverges from the original text when it is too long, for example by producing repetitive text (cycles), a well known problem of generative models (Fu et al., 2021; Ildiz et al., 2024). When the original text is not present anymore, it is impossible to gain points on in-text coreference resolution annotation. But what exactly a long text is depends on the language and the model's knowledge of the language. That has to do with the system's tokenizer. Tokens of under-resourced languages tend to be smaller than the ones of well-represented ones. This problem led us to the development of a second model.

3.2.2 Context-aware Model

The second fine-tuning splits the data into chunks of 8 sentences at a time. In the prompt, the most recent context (500 characters) that the model has already annotated is given, in order to preserve the coreference chains that were found earlier in the text.

If the chunk of sentences is the beginning of the text, the previous context is empty. In Example (2), we can see that the prompt is almost the same as the one of the Context-free Model.

(2) You are a linguist, expert in anaphora and coreference resolution. Based on the previous context, you have to annotate in the new sentence which nouns, pronouns and other linguistic expressions refer to the same entity.

Previous context: {gold_previous_context}

Do only insert annotations. Do not insert extra linguistic material, nor punctuation markers and do not delete elements from the input texts.

Before deciding to train a model with this context size, we experimented by giving it the entire context annotated thus far. It led to a disastrously bad result. Inspecting manually the output, it seems that the LLM does not 'understand' prompts that are too long. If there is already a long context that has been annotated, the LLM can no more make sense even of the task. We thus strictly restrained the given context to 500 characters (we choose characters in order to keep a similar context length across different languages in the corpora as token length is highly variable).

This model was trained on the same hardware as the Context-free model, but only on 3 epochs (mostly motivated by limited time and an increased number of training examples dues to cutting up long texts into chunks of 8 sentences). Training lasted

about two days.

First, we tested the context-aware model by preprocessing the development and test datasets the same way as the training data (chunks of 8 sentences and a context of 500 characters). Again, the results can be found in Table 2.

What we first observe is that there are some 'FAIL' results. There are two types of FAIL:

(a) The system cannot predict the corpus due to 'Torch Dynamo Hit Recompile Limit' Errors.

(b) The system has produced output that is incompatible with the text2text-coref toolkit, which prevents it from producing a CoNLL file from a plain-text output of the model.

The first problem is caused by the on the fly construction of data to predict, which leads to recompilation of the NN graph. As every chunk is accompanied by the most recent annotated context, the model has to base each prompt on its previous output. This leads to prediction data that is unstable and incompatible with the Torch library (or at least disfavored by it). Even though we found after the deadline of this shared task that there is a parameter that can be changed to enlarge the capacity of the prompt cache (which would increase the tolerance of the system to changing the prompt), it would have slowed down the system even more, meaning that prediction times would even be higher than the 1,5 days it takes the system already to predict the test set. Another option to solve this problem in the future could be to create fixed-sized prompts at the subword level, using pruning or padding when necessary, to avoid recompilation.

The second problem can undoubtedly be solved by working on the transformation scripts. We solved a small part by searching for and deleting invalid hash-tag sequences. For example, in *en_gum*, the model often generated sequences of "##", which causes errors when executing the text2text-coref tool. Unfortunately, we did not have enough time to address all the text2text-coref related issues and hence, there are some corpora that we did not manage to predict. But in the end, our context-aware approach seems to solve the problem of long texts. The performance increases significantly for the majority corpora that we managed to process.

For some corpora on which the context-free model obtained good results, the context-aware model did not manage to improve the scores (for example ca_ancora or es_ancora). We noticed that these corpora feature rather short texts and our conclusion was that the 500 characters context given in the prompt might be too short. We therefore wanted to develop a new model that had a larger context. We also wanted to address the problem of the torch dynamo recompilation limit by making less requests by enlarging the chunks.

As time fell short, we decided to use the context-aware model trained on contexts of 500 characters and chunks of 8 sentences but with different prediction parameters without retraining. We predicted chunks of 10 sentences giving 700 characters of context. The results can be found in Table 2.

We see that for most corpora, this run yielded the best results and we got a number of FAILs that is much lower. However, there are corpora performing best in the 8sent_500char setting and even two corpora where the context-free model is the best. This indicates that the trade-off between smaller texts to predict (thanks to chunking) and having only access to the most recent context is different for each corpus, depending on the LLMs knowledge of the language, and the size of the texts. It seems that each corpus would have its own optimal parameters.

Our final submission, combined best scores of all the LLM-predictions, leading to an average score of 62.96.

4 Discussion

Even though our LLM-approach yielded the highest scores in the LLM track (with only one point ahead of the second best submission), performances of systems in the unconstrained track cannot be ignored. Indeed, when we just compare our two submissions (mention-pair and Gemma 3 fine-tuning), we have to conclude that performance is very similar. And that is without taking into account the fact that the winner of the unconstrained track, the corpipe-ensemble system, largely encompasses our endeavours with an average score of 75.84. So, an important question that needs to be asked is: is it worth the trouble to use LLMs for coreference resolution? After all, their use is very costly in computation resources. For example, the training time for our two submissions differs significantly: only 6 hours for the classic model versus 2 or 3 days for the LLM-based system. The gap is even more striking at inference time, where the unconstrained system requires approximately 16 minutes to process the test-set, compared to about a day and half for the

Corpus	C-F	8s_500c	10s_700c
ca_ancora	71.83	70.44	**73.45**
cs_pcedt	53.39	64.47	**65.12**
cs_pdt	70.13	FAIL-a	**71.33**
cu_proiel	8.92	57.22	**58.25**
de_potsdamcc	58.75	FAIL-b	**59.60**
en_gum	44.34	FAIL-a	**58.73**
en_litbank	44.00	64.70	**69.01**
es_ancora	**74.43**	71.72	72.61
fr_ancor	14.40	64.73	**66.74**
fr_democrat	16.85	**60.43**	FAIL-a
grc_proiel	13.68	**65.75**	65.16
hi_hdtb	56.36	51.64	**52.74**
hbo_ptnk	1.00	FAIL-b	**43.96**
hu_korkor	46.39	**52.53**	52.46
hu_szegedkoref	56.42	56.41	**59.82**
ko_ecmt	60.52	61.09	**63.04**
lt_lcc	56.38	**62.55**	62.28
no_bokmaalnarc	57.40	64.14	**64.74**
no_nynorsknarc	**61.63**	61.60	FAIL-a
pl_pcc	70.81	72.21	**72.55**
ru_rucor	65.40	68.26	**68.79**
tr_itcc	6.08	51.92	**56.23**
Average	**47.85**	**58.91**	**62.67**

Table 2: CoNLL F1-scores of the LLM track on the test set. C-F: Context-free. Xs_Yc: X sentences, Y characters. FAIL-a: Torch Dynamo Recompilation Limit Error. FAIL-b: Text2text-coref Tool Error.

LLM-based approach. This is a substantial difference for a performance that remains comparable to that of a traditional mention-pair system.

There is a lot of room for improvement in the design of our context-aware model. In the first place by optimizing the context size, the length of the chunks, pre-treatment of prompts to avoid recompilation problems, and the machine learning parameters —which would undoubtedly allow us to gain a number of extra points in performance— and in the second place by design modifications which we will discuss broadly in Section 5. But according to us, one of the core problems of using LLMs for coreference resolution is that it asks to transform coreference resolution into a text generation task. In the remainder of this section we will explain what are the fundamental problems of doing so.

When used for coreference resolution in the plain-text format, LLMs are optimized to perform annotation. So in fact, our context-aware model handles coreference as an annotation problem, that should be handled as a text generation problem. Although using LLMs for annotation tasks is commonly done (Tan et al., 2024), conceptually it has important consequences when dealing with coreference.

Firstly, it defines coreference resolution necessarily as an incremental task: chunks are annotated in the order of the text and this leads inevitably in making only local decisions. Even if, from a cognitive point of view of coreference resolution, it seems reasonable to treat coreference as incremental (Seminck, 2018), many coreference systems are in fact not incremental, for example our pair-wise system performs resolution based on highest scoring mention-pair clustering, instead of incremental clustering in the order of the text.

As a result, it takes away the abstract representation of coreference chains, by providing only local annotations on word levels in text. The text-global modeling of coreference is at best only implicitly present, but in the setting of our context-aware model, more likely, absent. This led to annoying mistakes in long text. What can happen is that when a context is presented with entities numbered for example ‘51’, ‘67’ and ‘98’, the system will use lower numbers, starting again from ‘1’ to annotate new coreference chains. Although we could imagine simple ways to prevent this behaviour (for example by explicitly stating in the prompt that it is forbidden to restart numbering from ‘1’), it would be interesting to think about a way to make the system aware of the coreference annotation of the entire text, without giving the entire annotated preceding context.

Lastly, we would like to point out the problem of transforming coreference resolution into a text generation problem. The objects we have to deal with are necessarily string variables and only string variables. Of course, this could be seen as a general problem for using generative AI for any scientific problem. Coreference resolution is particularly impacted by the previous problem: how to represent a global and abstract presentation of the coreference chains using only a single string variable?

Even though the learning power of LLMs is impressive and one can try to insert abstract representations into the prompt to be handled, the way the LLM treats this information is a black box. For the LLM this information is part of the string, just as both the original text and the text annotations: there is no actual distinction between these things. There is no guarantee that from the output, the original text, readable annotations and abstract global coreference chain annotations can be recovered. Of

course, performance could be increased by enhancing the post-hoc scripts that parse and align the original text with the LLM-output by foreseeing more unwanted scenarios and creating patch-work solutions for them. But it does not change the problem fundamentally, we still have no guarantee of stability of our research objects.

Moreover, the larger the amount of additional information we may want to inject, treat with the LLM and then recover from the output, the lower the chances we actually succeed, as the probability of mixing up information increases. The LLM framework puts us out of control of the objects we want to calculate and manipulate. This is true for many uses of LLMs, stretching far beyond the problem of coreference. But we have to reflect on the question whether we can and want to accept it.

5 Future Work

Despite our conclusion that generative large language models are not easy to use to model coreference, participating in the shared task has given us a lot of ideas about how we could enhance our contribution next year. Even though we are not convinced that putting into practice these solutions would take away our reservations about the unsuitedness of text generation for coreference resolution, we are confident that they will enable us to increase significantly our scores. We will discuss these ideas and hope that we (or other teams and researchers) could benefit from them when developing new systems.

5.1 Improving Modelling of Coreference

Currently, in the context-aware model, as texts are split into chunks, the model never has access to the entire representation of coreference, as it is only implicitly present as the previously annotated most recent context. We could try to enhance the model by making it explicitly state all the clusters constructed so far and feed it as additional information into the prompt. Then, after annotation, extract the newly formed clusters and re-build the global coreference annotation. We expect this to help against restarting numbering coreference clusters from '1', but foresee the possibility that this representation might be unstable across the text, as it could be corrupted during text generation.

A second idea to improve the global representation of coreference resolution is to model a text in the memory of a chat conversation where each chunk is user-turn followed by a model's response. Although correctly memorizing very long conversations is still a challenge for LLMs (Maharana et al., 2024), we would like to test their abilities to keep track of global coreference chains using the memory of the chat conversation.

5.2 Task-Specific Loss Function

The fine-tuning we performed for the LLM track currently relies on the standard cross-entropy loss used in language modeling, as implemented in the `gemma-3-12b-it` model. However, this loss function is not well aligned with the specific needs of coreference resolution; while maintaining overall textual fidelity is important, assigning correct coreference identifiers is absolutely critical.

In standard text generation, two outputs such as `[e111]` and `[e112]` are nearly indistinguishable in terms of loss. The model is only minimally penalized for generating a slightly incorrect entity ID, even though such mistakes can drastically impact the coreference resolution.

One direction for future work would be to implement a task-specific loss function. After generating a batch of annotated text, we could compute a batch-level coreference evaluation metric (e.g. CoNLL F1-score). Though technically challenging, it could make LLM fine-tuning more sensitive to the actual goals of coreference resolution.

5.3 Improving the Input Format

The current plain-text format provided by the CRAC shared task uses a custom inline annotation style to mark entity spans and coreference chains. For example:

> Down the**|[e1** Rabbit-Hole**|e1]** Alice**|[e2]** was beginning to get very tired of sitting by her**|[e2],[e3** sister**|e3]** on the**|[e4** bank**|e4]** , and of having nothing to do : once or twice she**|[e2]** had peeped into the book her**|[e2],[e3** sister**|e3]** was reading

We propose exploring alternative tagging schemes better suited to LLMs, such as formats inspired by markup languages like HTML or XML. These clearly mark span boundaries with readable, nested tags, explicitly marking start and end of each span (`<entity_start> </entity_end>`):

> Down **<e1>**the Rabbit-Hole**</e1>** **<e2>**Alice**</e2>** was beginning to get very tired of sitting by **<e3><e2>**her**</e2>** sister**</e3>** on **<e4>**the bank**</e4>** , and of having nothing to do : once or twice **<e2>**she**</e2>** had peeped into the book **<e3><e2>**her**</e2>** sister**</e3>** was reading

Such a structure might be easier to tokenize and interpret by LLMs and may lead to better generalization and consistency in generation-based settings. Adopting this alternative would require adapting the conversion scripts from CoNLL-U to plain text, and from LLM outputs back to CoNLL-U. We believe this modification could help bridge the gap between coreference annotation conventions and LLM-friendly input formats, potentially improving model performance.

We could also try, together with the newly developed task-specific loss function, to fine-tune directly on the CoNLL-U format. This would limit error propagation caused by the transformation scripts.

5.4 LLM-based Pair-Wise Resolver

To limit the undesirable effects of text generation (loss of control on our study objects), we could split the coreference resolution task into sub-problems and come back to a pair-wise resolution system using LLMs. We would first use an LLM for mention detection, and then another for pair-wise classification, where pairs of mentions are classified as coreferent or not, fine-tuning the LLM to produce a binary response.

While this system would undoubtedly be computationally extremely heavy, as it asks for tens of thousands of calls to the LLM in order to perform pair-wise resolution, it would be an interesting experiment to see whether performance on the mention-detection and the pair-wise resolution increases with respect to classical systems, such as our mention-pair system. According to the results, we could also consider to replace a given module by an LLM-based system. If the LLM results are high but very costly computationally, we could also use it only for the more difficult cases of resolution. The current pair-wise system outputs confidence scores for its calculations, we could use the LLM-based system only for low confidence scores.

5.5 Student Training of LLM with Oracles

We only have access to gold data in order to fine-tune the coreference resolution systems. But, the incremental setting imposed by the LLM puts us in a situation where error propagation can be an issue. Therefore, we could want to teach the LLM to resolve coreference based on its previous predictions even if they contain errors. However, learning to predict the gold annotation given what has already been predicted (the context) can actually be detrimental. For example, if due to early errors, two chains have seen their indices swapped in the context, trying to predict the original gold indices is actually incoherent. To remedy this, we would need to relabel the current chunk to replace gold tags, given what has already been predicted in the context. This is computationally very expensive, likely NP-hard, given that the coreference metrics consider the annotation of the whole text. We consider to train oracles to predict good relabeling of the gold data at a reasonable cost, inspired by works done in syntactic parsing where oracles are trained to predict sequences of transitions of a system that reconstruct a parse tree (Coavoux and Crabbé, 2016; Shen et al., 2021).

6 Conclusion

We fine-tuned the `Gemma-3-12B-it` model to perform coreference resolution in the LLM track of the CRAC shared task and ended first. We found that our approach was adaptable to all the languages of the shared task, but that the systems were computationally very costly, especially compared to our pair-wise coreference resolution system submitted in the Unconstrained track of the CRAC shared task. Analyzing our results, we come to the conclusion that it is not obvious use generative LLMs for coreference resolution. Coreference resolution being a global discourse phenomenon, it is difficult to model it as a text generation task. Notwithstanding this fundamental problem, our work can be seen as one of the first attempts to fit the problem resolution task in the framework of LLMs and provides a rich ground for reflection on multiple areas of improvement for future work.

Acknowledgements

This research was funded in part by PRAIRIE-PSAI (Paris Artificial intelligence Research institute–Paris School of Artificial Intelligence, reference ANR-22-CMAS-0007).

References

David Bamman. 2020. Litbank: Born-literary natural language processing. *Computational Humanites, Debates in Digital Humanities (2020, preprint)*.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. Lora learns less and forgets less. *Preprint*, arXiv:2405.09673.

Maximin Coavoux and Benoît Crabbé. 2016. Neural greedy constituent parsing with dynamic oracles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 172–182, Berlin, Germany. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 660–669.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856. Association for the Advancement of Artificial Intelligence (AAAI).

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Ralph Grishman and Beth Sundheim. 1995. Design of the muc-6 evaluation. In *Proceedings of the 6th conference on Message understanding*, pages 1–11. Association for Computational Linguistics.

Graeme Hirst. 1981. *Anaphora in natural language understanding: a survey*. Springer.

M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. 2024. From self-attention to markov models: unveiling the dynamics of generative transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20955–20982.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Frédéric Landragin. 2016. Description, modélisation et détection automatique des chaînes de référence (democrat). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92):11–15.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia. University of Tartu Library.

Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol. 2011. Ancor, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, pages 555–563.

Michal Novák, Barbora Dohnalová, Miloslav Konopík, Anna Nedoluzhko, Martin Popel, Ondřej Pražák, Jakub Sido, Milan Straka, Zdeněk Žabokrtskỳ, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. *arXiv preprint arXiv:2410.15949*.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ian Porada, Xiyuan Zou, and Jackie Chi Kit Cheung. 2024. A controlled reevaluation of coreference resolution models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 256–263, Torino, Italia. ELRA and ICCL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012

shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.

Sameer Pradhan, Lance Ramshaw, Mitch Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the fifteenth conference on computational natural language learning: shared task*, pages 1–27.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Olga Seminck. 2018. *Cognitive computational models of pronoun resolution*. Ph.D. thesis, Université Sorbonne Paris Cité.

Olga Seminck, Yoann Dupont, Mathieu Dehouck, Qi Wang, Noé Durandard, and Margo Novikov. 2025. Lattice @MultiGEC-2025: A spitful multilingual language error correction system using LLaMA. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 34–41, Tallinn, Estonia. University of Tartu Library.

Yikang Shen, Shawn Tan, Alessandro Sordoni, Siva Reddy, and Aaron Courville. 2021. Explicitly modeling syntax in language models with incremental parsing and a dynamic oracle. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1660–1672, Online. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Ryszard Staruch. 2025. UAM-CSI at MultiGEC-2025: Parameter-efficient LLM fine-tuning for multilingual grammatical error correction. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 42–49, Tallinn, Estonia. University of Tartu Library.

Milan Straka. 2024. Corpipe at crac 2024: Predicting zero mentions from raw text. *arXiv preprint arXiv:2410.02756*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.

Lixing Zhu, Jun Wang, and Yulan He. 2025. Llm-Link: Dual LLMs for dynamic entity linking on long narratives with collaborative memorisation and prompt optimisation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11334–11347, Abu Dhabi, UAE. Association for Computational Linguistics.

A Unconstrained Track Model Architecture and Hyperparameters

A.1 Mention Detection Model

A.1.1 Architecture

- **Locked Dropout** (0.5) applied to embeddings for regularization.
- **Projection Layer**: Highway network mapping $1024 \rightarrow 2048$ dimensions.
- **BiLSTM Layer**: Single bidirectional LSTM (256 hidden units per direction).
- **Linear Layer**: Maps 512-dimensional BiLSTM outputs to BIOES label scores.
- **CRF Layer**: Enforces structured consistency in predictions.

A.1.2 Training Parameters

- **Data Splitting**: 85%/15% train-validation split.
- **Batch Size**: 16 sentences per batch.
- **Optimization**: Adam optimizer (lr $= 1.4 \times 10^{-4}$, weight decay = 10^{-5}).
- **Learning Rate Scheduling**: ReduceLROnPlateau (factor = 0.5, patience = 2).
- **Average Training Epochs**: 22.
- **Hardware**: Trained on a single 48 GiB Nvidia RTX 6000 Ada Generation GPU.

A.2 Coreference Resolution Model

A.2.1 Architecture

- **Model Input**: 2,063-dimensional vector, composed of concatenated:
  - **CamemBERT embeddings**: Maximum context embeddings for both mentions (2 × 1,024 = 2,048 dimensions).
  - **Mention Features** (15 dimensions):
    - Mention length.
    - Position of the mention's start token in the sentence.
    - Dependency relation of the mention's head (one-hot encoded).
  - **Mention Pair Features** (8 dimensions):
    - Distance between mention IDs.
    - Distance between start and end tokens of mentions.
    - Sentence and paragraph distance.
    - Difference in nesting levels.
    - Ratio of shared tokens between mentions.
    - Exact text match (binary).
    - Exact match of mention heads (binary).
    - Match of syntactic heads (binary).
- **Hidden Layers**:
  - Three fully connected layers.
  - 1,900 hidden units per layer with ReLU activation.
  - Dropout rate of 0.6 for regularization.
- **Final Layer**:
  - Linear layer mapping from 1,900 dimensions to a single scalar score.
  - Output: Continuous value between 0 (not coreferent) and 1 (coreferent).

A.3 Model Training

- **Data Splitting**: 85%/15% train-validation split.
- **Batch Size**: 16,000 mention-pairs per batch.
- **Optimization**: Adam optimizer (**lr** $= 4 \times 10^{-4}$, weight decay = 10^{-5}).
- **Antecedent Candidates**: 80 maximum.
- **Antecedent Candidates**:
- **Hardware**: Trained on a single 48 GiB Nvidia RTX 6000 Ada Generation GPU.

CorPipe at CRAC 2025: Evaluating Multilingual Encoders for Multilingual Coreference Resolution

Milan Straka
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague, Czech Republic
straka@ufal.mff.cuni.cz

Abstract

We present CorPipe 25, the winning entry to the CRAC 2025 Shared Task on Multilingual Coreference Resolution. This fourth iteration of the shared task introduces a new LLM track alongside the original unconstrained track, features reduced development and test sets to lower computational requirements, and includes additional datasets. CorPipe 25 represents a complete reimplementation of our previous systems, migrating from TensorFlow to PyTorch. Our system significantly outperforms all other submissions in both the LLM and unconstrained tracks by a substantial margin of 8 percentage points. The source code and trained models are publicly available at https://github.com/ufal/crac2025-corpipe.

1 Introduction

Coreference resolution seeks to identify and cluster multiple references to the same entity within text. The CRAC 2025 Shared Task on Multilingual Coreference Resolution (Novák et al., 2025a) represents the fourth iteration of this shared task, designed to advance research in multilingual coreference resolution across diverse languages and domains. Building upon the CorefUD 1.3 collection, this year's task introduces several notable changes: a new LLM track that relies on large language models (LLMs) for coreference resolution, reduced development and test sets (minidev and minitest) to lower computational demands, and the inclusion of additional datasets expanding language coverage.

As in the previous year, the submitted systems must also predict the *empty nodes*, which represent elided elements that are not explicitly present in the surface text but are necessary for coreference analysis. Empty nodes are especially important in pro-drop languages (like Slavic and Romance languages), where pronouns can be dropped from a sentence when they can be inferred, for example according to verb morphology, as in the Czech example *"Řekl, že nepřijde"*, translated as *"(He) said that (he) won't come"*.

CorPipe 25, our submission to the CRAC 2025 Shared Task, represents a complete reimplementation of our previous winning systems (Straka, 2024, 2023; Straka and Straková, 2022), transitioning from TensorFlow to PyTorch while preserving the architecture that has proven successful. Our system employs a three-stage pipeline approach: first predicting empty nodes,[1] then detecting mentions, and finally performing coreference linking through antecedent maximization on the identified spans. As in previous CorPipe versions, mention detection and coreference linking are trained jointly using a shared pretrained encoder model, and all models are fully multilingual, trained across all available corpora.

Our contributions are as follows:

- We present the winning entry to the CRAC 2025 Shared Task, surpassing other participants in both tracks by a substantial margin of 8 percentage points.
- We provide a complete reimplementation of CorPipe in PyTorch. The reimplementation enables us to leverage more pretrained multilingual models, allowing us to perform an evaluation of various models and providing insights into their relative performance for coreference resolution across diverse languages.
- We present performance comparisons between TensorFlow and PyTorch implementations, demonstrating the practical benefits of the migration.
- The CorPipe 25 source code is released at https://github.com/ufal/crac2025-corpipe under an open-source license. Three pretrained multilingual models of different sizes are also released, under the CC BY-NC-SA licence.

[1] Our empty node prediction system was provided to all participants as a baseline implementation.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 130–139
Suzhou, China, November 9, 2025.

2 Related Work

Neural Coreference Resolution Neural coreference resolution has been dominated by span-based approaches since the seminal work of Lee et al. (2017), who introduced an end-to-end neural model that jointly performs mention detection and coreference resolution. This approach was further refined by Lee et al. (2018) with coarse-to-fine inference, significantly improving both efficiency and accuracy. Joshi et al. (2020) demonstrated substantial improvements by incorporating SpanBERT (Joshi et al., 2019), a pretrained model specifically designed for span prediction tasks.

Alternative paradigms have emerged to address the limitations of span-based methods. Wu et al. (2020) formulated coreference as a question-answering task, while Liu et al. (2022) introduced a specialized autoregressive system and Bohnet et al. (2023) employed a text-to-text paradigm. However, all these architectures must evaluate the trained model repeatedly during processing of a single sentence.

Word-Level Coreference Resolution A significant departure from span-based approaches came with Dobrovolskii (2021), who proposed word-level coreference resolution, which represents mentions by their head-words only. The approach has been extended by D'Oosterlinck et al. (2023) with CAW-coref, which introduces conjunction-aware handling to better manage complex mention structures. More recently, Liu et al. (2024) proposed MSCAW-coref that aims to work in a multilingual setting and accounts for singleton mentions. This approach has been adopted by Stanza (Qi et al., 2020), a widely-used Python natural language processing toolkit.

Multilingual Coreference Resolution The CRAC shared tasks on multilingual coreference resolution (Žabokrtský et al., 2022, 2023; Novák et al., 2024, 2025a) have been instrumental in advancing the field, providing standardized evaluation framework, the CorefUD dataset (Novák et al., 2025b), and a multilingual baseline (Pražák et al., 2021).

Previous versions of CorPipe have participated in all CRAC shared tasks, evolving from basic multilingual models (Straka and Straková, 2022) to incorporating larger contexts (Straka, 2023) and performing zero mention prediction from raw text (Straka, 2024).

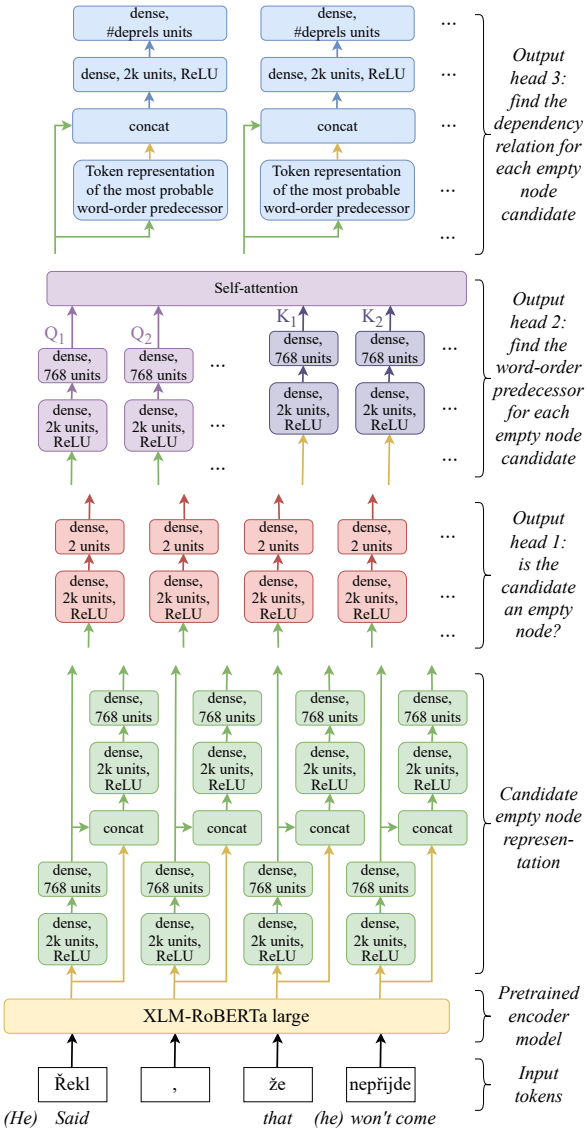

Figure 1: The system architecture of the empty node prediction baseline. Every ReLU activation is followed by a dropout layer with a dropout rate of 50%.

3 Architecture

Our system is essentially a PyTorch reimplementation of CorPipe 24 (Straka, 2024).

Empty Nodes Baseline First, empty nodes are predicted using a baseline system that was available to all shared task participants. The architecture of this system is illustrated in Figure 1.

Our approach for empty node prediction focuses on generating the essential information required for coreference evaluation: the word order position (determined by which input word the empty node follows), along with the dependency head and dependency relation. We do not predict forms or lemmas, even when available in training data. The model operates non-autoregressively, predicting up

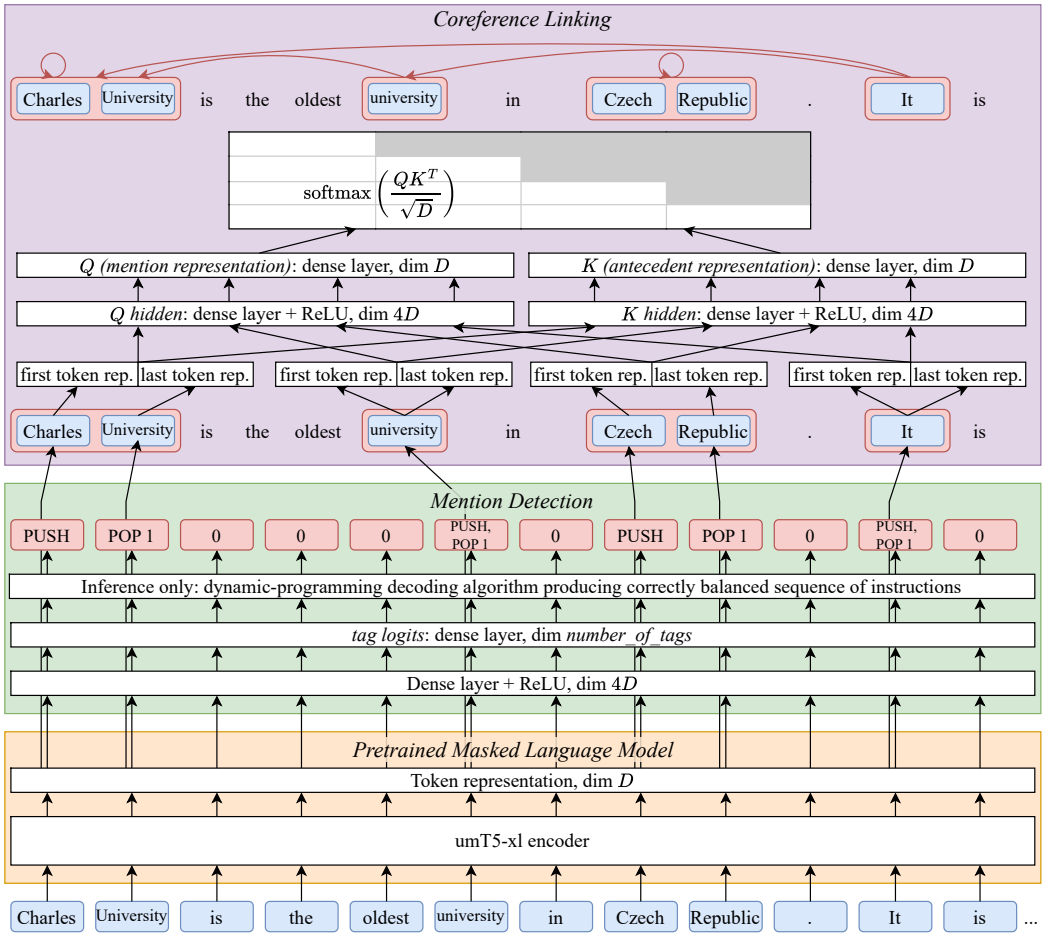

Figure 2: The CorPipe 25 model architecture.

to two empty nodes per input word, with each input word serving as the potential dependency head.

The architecture processes tokenized input through a XLM-RoBERTa-large (Conneau et al., 2020), representing each word by its first subword embedding. For each word, we generate two empty node candidates: the first through a dense-ReLU-dropout-dense module (768→2k→768 units), and the second by concatenating the first candidate with the input word representation and applying an analogous transformation. The candidates are processed by three heads, each following its own 2k-unit ReLU layer and dropout: (1) binary classification for empty node existence, (2) self-attention for word order position selection, and (3) dependency relation classification using the candidate representation concatenated with the embedding of the most likely word preceding it.

Training employs a single multilingual model with Adam optimizer (Kingma and Ba, 2015) for 20 epochs of 5 000 batches (64 sentences each). The learning rate linearly increases to 1e-5 in the first epoch and then decays to zero in the rest of the training following cosine decay (Loshchilov and Hutter, 2017). Sentences are sampled from all empty node corpora, proportionally to the square root of corpus size. Training required 19 hours on a single L40 GPU with 48GB RAM.

The source code is released under the MPL license at `https://github.com/ufal/crac2025_empty_nodes_baseline`, together with the full set of hyperparameters used. The trained model is available under the CC BY-SA-NC license at `https://www.kaggle.com/models/ufal-mff/crac2025_empty_nodes_baseline/`. Finally, the minidev and minitest sets of the CRAC 2025 Shared Task with predicted empty nodes are available to all participants.

Coreference Resolution Once the empty nodes have been predicted, we employ coreference resolution system based on CorPipe 23 from Straka (2023). The architectural overview is shown in Figure 2 and summarized below; detailed implementation specifics are available in the referenced work.

Our model processes documents sentence-by-sentence. To maximize available context for each sentence, we expand it with preceding tokens and

Model	Params	Batch Size	Learning Rate	Train Time
mT5 base	264M	8	6e-4	4h
umT5 base	269M	8	6e-4	4h
mT5 large	538M	8	6e-4	9.5h
mT5 xl	1593M	6	5e-4	22.5h
umT5 xl	1605M	6	5e-4	22.5h
mT5 xxl	5393M	6	5e-4	33h
umT5 xxl	5417M	6	5e-4	33h

Table 1: Properties of mT5 encoder models used. The training time is measured for 15 epochs 10k updates each using a single A100 GPU, with the exception of the xxl models, which are trained using a single H100 GPU.

at most 50 subsequent tokens, constrained by the maximum segment length (512 or 2 560 tokens). Input tokens first pass through a pretrained multilingual encoder. Subsequently, we predict coreference mentions using an enhanced BIO encoding scheme that handles potentially overlapping span sets. Each identified mention is then encoded as a concatenation of its boundary tokens (first and last), and coreference links are established through a self-attention mechanism that determines the most probable antecedent for each mention (including self-reference utilized by first entity mentions).

We employ different segment sizes during training versus inference: training always uses 512-token segments, while inference leverages extended 2 560-token segments (with the exception of two PROIEL corpora always using 512 tokens), exploiting relative positional encoding capabilities for improved long-range context modeling.

Training For the shared task submission, we train 13 multilingual models based on umT5-xl (Chung et al., 2023), differing only in random initialization and whether we express corpus size during sampling using sentences or words. The sentences are sampled proportionally to the square root of the corpus size; for ablations, we consider also values of this *sampling ratio* different from 0.5.

Every model is trained for 15 epochs with 10k batches each, with every batch consisting of 6 sentences. The model is trained using the AdaFactor optimizer (Shazeer and Stern, 2018). The learning rate follows a warmup schedule: linear increase to 5e-4 during the initial 10% of training, followed by a cosine decay (Loshchilov and Hutter, 2017) to 0. The model trains for 22.5 hours on a single A100 GPU with 40GB RAM. For ablation experiments, we also consider other umT5 and mT5 (Xue et al., 2021) models, whose properties and corresponding hyperparameters are summarized in Table 1.

For each model, we save checkpoints after every epoch, obtaining a pool of $13 \cdot 15$ checkpoints.

System	Head-match	Partial-match	Exact-match	With Singletons
UNCONSTRAINED				
CorPipeEnsemble	**75.84 1**	**74.90 1**	**72.76 1**	**78.33 1**
CorPipeBestDev	75.06 2	74.08 2	71.97 2	77.63 2
CorPipeSingle	74.75 3	73.74 3	71.53 3	77.43 3
Stanza	67.81 4	67.03 4	64.68 4	70.64 4
GLaRef-Propp	61.57 5	60.72 5	58.43 5	65.28 5
BASELINE-GZ	58.18 6	57.75 6	56.48 6	49.88 6
BASELINE	56.01 7	55.58 7	54.24 7	47.88 7
LLM				
GLaRef-CRAC25	62.96 1	61.66 1	58.98 1	65.61 1
NUST-FewShot	61.74 2	61.14 2	56.34 2	63.44 2
PUXCRAC2025	60.09 3	59.68 3	55.22 3	54.77 4
UWB	59.84 4	59.55 4	38.81 4	62.77 3

Table 2: Official results of CRAC 2025 Shared Task on the minitest set with various metrics in %.

4 Shared Task Results

In the shared task, teams were permitted to submit up to three systems. We selected the following configurations based on our checkpoint selection strategy:

- **CorPipeSingle**, a single best-performing checkpoint selected based on overall minidev performance across all corpora;
- **CorPipeBestDev**, employing corpus-specific optimal checkpoints selected individually based on minidev performance for each corpus from the pool of $13 \cdot 15$ checkpoints;
- **CorPipeEnsemble**, an ensemble of 5 best-performing checkpoints based on overall minidev performance across all corpora.

The first configuration CorPipeSingle corresponds to practical deployment, where a single model handles all corpora, while the others aim at maximizing performance.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
UNCONSTRAINED																							
CorPipeEnsemble	**75.81**	**82.91**	**77.11**	**80.71**	**65.51**	**73.01**	**76.11**	81.81	**84.51**	**76.31**	**71.81**	**74.51**	69.81	**77.71**	**68.61**	**71.01**	**69.91**	**77.21**	**78.21**	**76.31**	**80.21**	84.23	71.22
CorPipeBestDev	75.12	82.03	76.32	80.42	62.83	72.63	75.92	81.32	83.83	75.92	69.93	74.33	68.32	77.52	68.32	70.52	69.32	76.02	77.12	74.02	79.92	**84.81**	70.43
CorPipeSingle	74.83	82.52	76.23	80.13	63.02	72.82	75.23	80.83	84.12	75.83	70.32	74.42	66.13	76.53	67.33	69.73	68.93	75.83	76.23	73.63	79.43	84.22	**71.61**
Stanza	67.84	79.54	72.74	75.14	40.84	67.34	69.04	74.84	80.44	67.54	62.55	54.94	62.14	74.24	60.04	64.64	67.74	72.84	72.44	71.74	73.04	80.84	47.85
GLaRef-Propp	61.65	68.16	61.76	66.66	39.15	61.25	61.95	70.05	69.17	65.15	66.14	51.35	58.85	69.55	50.95	60.15	60.66	57.67	67.15	66.35	68.06	71.55	44.37
BASELINE-GZ†	58.26	68.85	69.55	67.95	29.56	55.76	61.67	66.06	71.05	63.86	55.06	29.46	31.06	66.86	47.16	54.37	64.35	65.35	62.56	63.06	68.15	67.66	51.74
BASELINE†	56.07	68.07	56.97	63.07	26.37	55.76	61.76	66.06	70.56	63.86	55.06	28.57	31.06	66.86	43.27	54.56	50.37	65.35	62.56	63.06	66.57	67.66	45.96
LLM																							
GLaRef-CRAC25	63.01	73.52	65.11	71.31	58.22	59.62	58.74	69.04	74.41	66.72	60.42	65.81	44.03	56.44	52.51	59.83	63.03	62.53	64.74	61.64	72.51	68.83	56.22
NUST-FewShot	61.72	60.94	51.44	54.34	58.51	48.74	69.82	70.42	61.84	71.91	57.63	57.92	**80.21**	71.32	43.53	52.34	66.02	59.24	72.82	68.92	70.82	71.42	39.03
PUXCRAC2025	60.13	68.03	56.93	63.03	43.73	57.43	61.73	69.13	70.53	63.83	61.51	47.93	45.32	66.83	50.62	61.62	50.34	65.31	65.23	63.03	66.53	67.64	56.11
UWB	59.84	79.21	61.02	68.22	25.34	67.61	73.61	**84.01**	73.62	58.64	49.14	47.64	0.04	75.81	38.94	67.31	68.31	63.42	73.81	72.01	64.54	80.11	24.34

Table 3: Official results of CRAC 2025 Shared Task on the minitest set (CoNLL score in %). The systems † are described in Pražák et al. (2021); the rest in Novák et al. (2025a).

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) CORPIPE SINGLE MODELS																							
Single mT5-large model	72.84	80.1	74.6	78.0	58.5	67.2	73.3	77.4	82.0	72.1	68.5	71.2	67.9	76.3	67.3	68.0	69.8	74.4	75.2	74.0	77.5	81.2	67.7
Single umT5-base model	–3.54	–2.7	–1.1	–2.9	–5.0	–5.2	–2.3	–4.6	–3.5	–0.9	–1.8	–6.3	–8.9	–3.6	–5.8	–4.3	–2.0	–1.5	–2.0	–3.5	–3.0	–3.5	–3.8
	69.27	77.4	73.5	75.1	53.5	62.0	71.0	72.8	78.6	71.2	66.7	64.9	59.0	72.7	61.5	63.7	67.8	72.9	73.2	70.4	74.5	77.8	63.9
Single umT5-xl model	+1.96	+2.4	+1.6	+2.1	+4.5	+5.6	+1.9	+3.4	+2.0	+3.7	+1.8	+3.2	–1.8	+0.2	–0.0	+1.7	–0.9	+1.4	+1.0	–0.4	+1.9	+3.0	+3.9
	74.75	82.5	76.2	80.1	63.0	72.8	75.2	80.8	84.1	75.8	70.3	74.4	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	71.6
Single mT5-xxl model	+3.16	**+3.0**	**+2.9**	**+3.3**	**+10.8**	**+6.1**	+3.6	+5.0	+2.4	+3.6	**+4.0**	+6.2	**+6.5**	+0.4	–5.1	**+3.9**	–0.5	**+3.2**	+1.7	+1.5	+2.7	+3.7	+1.9
	76.04	**83.1**	**77.5**	**81.3**	**69.3**	**73.3**	76.9	82.4	84.4	75.7	**72.4**	77.5	**74.4**	76.7	62.2	**71.9**	69.3	**77.6**	76.9	75.4	80.2	84.9	69.6
Single umT5-xxl model	**+3.46**	+3.0	+2.7	+3.2	+9.1	+4.5	**+3.8**	**+6.8**	+2.4	**+5.4**	+3.0	**+6.8**	+4.8	+1.4	–1.2	+3.2	–0.6	+0.6	**+3.1**	**+2.3**	**+2.8**	**+3.8**	**+4.7**
	76.26	83.1	77.3	81.2	67.6	71.7	**77.1**	**84.2**	84.4	**77.5**	71.5	**78.1**	72.7	77.7	66.1	71.2	69.2	75.0	**78.3**	**76.3**	**80.3**	**85.0**	**72.4**
B) CORPIPE ENSEMBLE MODELS																							
Single umT5-xl model	74.75	82.5	76.2	80.1	63.0	72.8	75.2	80.8	84.1	75.8	70.3	74.4	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	71.6
5 umT5-xl models	+1.05	+0.4	+0.9	+0.6	+2.5	+0.2	+0.8	+1.0	+0.4	+0.5	+1.5	+0.1	+3.7	+1.2	**+1.3**	+1.3	+1.0	+1.4	**+2.0**	**+2.7**	+0.8	–0.0	–0.4
	75.84	82.9	77.1	80.7	65.5	73.0	76.1	81.8	84.5	76.3	71.8	74.5	69.8	77.7	**68.6**	71.0	69.9	77.2	**78.2**	**76.3**	80.2	84.2	71.2
3 mT5-xxl models	+2.15	+1.4	+1.4	+1.0	+7.0	**+2.1**	+2.0	+2.9	+1.0	+1.0	+2.4	+6.1	**+8.9**	+1.3	–0.1	+2.8	+1.1	**+1.6**	+0.5	+2.3	+1.4	+0.6	–0.6
	76.93	83.9	77.6	81.1	70.0	**74.9**	77.3	83.7	85.1	76.7	72.7	80.5	**75.0**	77.8	67.2	72.5	70.0	**77.4**	76.7	75.9	80.8	84.8	71.0
3 umT5-xxl models	+2.05	+1.1	+1.5	**+1.7**	+5.8	+1.0	+2.0	+2.9	+0.9	**+2.2**	+2.5	+3.9	+7.9	+1.3	–0.5	**+3.5**	+1.2	–1.0	+1.1	+2.3	+1.9	**+1.8**	**+0.2**
	76.80	83.6	77.7	**81.8**	68.8	73.8	77.2	83.7	85.0	**78.0**	72.8	78.3	74.0	77.8	66.8	**73.2**	70.1	74.8	77.3	75.9	81.3	**86.0**	**71.8**
3 mT5-xxl models +	**+2.45**	**+1.7**	**+1.8**	+1.5	**+7.2**	+0.6	**+2.5**	**+3.3**	**+1.3**	+1.8	**+2.6**	**+6.2**	+8.9	**+1.6**	+0.5	+2.7	**+1.6**	+0.1	+1.3	+2.5	**+2.1**	+1.6	+0.2
+3 umT5-xxl models	**77.20**	**84.2**	**78.0**	81.6	**70.2**	73.4	**77.8**	**84.1**	**85.4**	77.6	**72.9**	**80.6**	75.0	**78.1**	67.8	72.4	**70.5**	75.9	77.5	76.1	**81.5**	85.8	71.8
C) CORPIPE PER-CORPUS BEST MODELS																							
Single umT5-xl model	74.75	**82.5**	76.2	80.1	**63.0**	**72.8**	75.2	80.8	**84.1**	75.8	**70.3**	**74.4**	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	**71.6**
Per-corpus best umT5-xl model	**+0.35**	–0.5	**+0.1**	**+0.3**	–0.2	–0.2	**+0.7**	**+0.5**	–0.3	**+0.2**	–0.4	–0.1	**+2.2**	**+1.0**	**+1.0**	**+0.8**	**+0.4**	**+0.2**	**+0.9**	**+0.4**	**+0.5**	**+0.6**	–1.2
	75.06	82.0	**76.3**	**80.4**	62.8	72.6	**75.9**	**81.3**	83.8	**75.9**	69.9	74.3	**68.3**	**77.5**	**68.3**	**70.5**	**69.3**	**76.0**	**77.1**	**74.0**	**79.9**	**84.8**	70.4

Table 4: Additional experiments on the CorefUD 1.3 minitest set (CoNLL score in %). The models in italics are post-competition submissions (i.e., submitted after the shared task deadline).

The official results of the CRAC 2025 Shared Task are summarized in Table 3 showing the CoNLL score and individual corpora performance, and in Table 2 showing four metrics across all corpora. All CorPipe 25 configurations substantially surpass all other participants, by 7 percent points for CorPipeSingle and 8 for CorPipeEnsemble. The CorPipeBestDev configuration only marginally outperforms CorPipeSingle, which we attribute to the exclusion of the two smallest corpora this year.

We evaluate additional mT5 and umT5 models on the minitest in Table 4. The xxl-sized models provide a boost of more than 1 percent point over the xl size; the ensemble of 3 mT5-xxl and umT5-xxl models provide an additional 1 percent point gain, achieving the best performance of 77.2%, a 1.4 percent point increase compared to the best competition submission.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) SUBMITTED CRAC25 SYSTEMS																							
CorPipeEnsemble	**76.51**	**84.1**	**76.9**	**81.1**	**64.2**	77.9	**77.5**	80.0	**85.1**	**79.6**	**72.5**	76.1	**66.8**	**82.0**	**69.7**	**73.1**	**69.4**	**81.6**	**80.1**	79.7	**80.3**	**80.0**	65.5
CorPipeSingle	75.69	83.2	75.9	80.2	62.7	76.9	76.5	**80.1**	84.2	79.0	71.9	**76.2**	66.0	80.6	68.1	71.9	67.6	80.2	79.2	**80.3**	79.2	78.3	**66.7**
Stanza	69.37	80.3	72.8	74.5	38.0	**78.0**	70.7	73.0	79.5	69.8	63.2	54.1	63.6	78.9	65.3	68.6	64.9	78.8	74.9	75.3	74.1	78.4	49.5
GLaRef-Propp	62.96	68.9	61.9	62.9	40.0	64.0	65.2	72.1	68.8	69.0	65.0	54.1	57.8	72.1	52.5	60.4	60.6	74.2	69.8	70.3	65.0	67.8	42.8
BASELINE-GZ	58.64	70.5	68.0	67.4	27.7	57.9	65.0	66.6	71.7	65.4	56.3	29.8	23.8	69.9	49.9	59.0	63.0	69.3	66.1	66.8	65.6	63.4	47.1
BASELINE	56.39	69.9	57.3	63.2	24.1	57.9	65.0	66.6	71.3	65.4	56.3	27.0	23.8	69.9	46.6	58.3	48.3	69.3	66.1	66.8	64.1	63.4	40.1
B) CORPIPE SINGLE MODELS																							
mT5-large	73.26	81.3	73.8	77.0	57.7	75.3	74.1	75.9	81.7	74.9	69.7	72.1	65.2	79.7	66.4	68.7	67.7	80.0	77.2	77.5	76.8	76.2	62.8
mT5-base	−4.43 68.83	−3.7 77.6	−3.4 70.4	−4.3 72.7	−5.0 52.7	−7.8 67.5	−2.6 71.5	−5.3 70.6	−3.1 78.6	−3.7 71.2	−4.2 65.5	−9.1 63.0	−8.1 57.1	−3.6 76.1	−5.7 60.7	−3.7 65.0	−2.2 65.5	−2.8 77.2	−1.9 75.3	−3.1 74.4	−3.8 73.0	−5.4 70.8	−4.8 58.0
umT5-base	−3.38 69.88	−2.6 78.7	−1.3 72.5	−2.2 74.8	−4.5 53.2	−6.4 68.9	−2.9 71.2	−3.6 72.3	−1.5 80.2	−1.8 73.1	−3.2 66.5	−7.8 64.3	−9.3 55.9	−2.9 76.8	−2.5 63.9	−3.4 65.3	−2.2 65.5	−5.3 74.7	−1.4 75.8	−1.6 75.9	−2.6 74.2	−3.0 73.2	−2.5 60.3
XLM-RoBERTa-base	−5.23 68.03	−6.2 75.1	−5.0 68.8	−5.3 71.7	−6.5 51.2	−4.4 70.9	−3.9 70.2	−7.4 68.5	−4.5 77.2	−4.4 70.5	−4.2 65.5	−8.3 63.8	−14.7 50.5	−2.5 77.2	−1.3 65.1	−3.1 65.6	−6.0 61.7	−2.7 77.3	−4.2 73.0	−4.0 73.5	−5.0 71.8	−5.6 70.6	−5.9 56.9
XLM-RoBERTa-large	−1.36 71.90	−1.6 79.7	−2.3 71.5	−0.7 76.3	+0.6 58.3	+0.1 75.4	+0.1 74.2	−2.4 73.5	−1.0 80.7	−1.4 73.5	−1.1 68.6	−0.5 71.6	−6.6 58.6	+0.2 79.9	+1.2 67.6	+1.2 69.9	−3.1 64.6	−2.0 78.0	−2.4 74.8	−1.5 76.0	−1.8 75.0	−2.2 74.0	−2.9 59.9
RemBERT	−1.84 71.42	−2.5 78.8	−2.2 71.6	−0.8 76.2	+0.3 58.0	+1.0 76.3	−0.4 73.7	−1.4 74.5	−1.1 80.6	−0.4 74.5	−1.4 68.3	−1.6 70.5	−4.7 60.5	+0.2 79.9	−1.0 65.4	−2.1 66.6	−4.4 63.3	−5.2 74.8	−2.3 74.9	−2.5 75.0	−1.5 75.3	−2.7 73.5	−3.5 59.3
InfoXLM-large	−1.44 71.82	−1.9 79.4	−1.9 71.9	−0.4 76.6	+1.3 59.0	+0.0 75.3	−0.3 73.8	−3.0 72.9	−1.0 80.7	−0.8 74.1	−1.1 68.6	−1.4 70.7	−7.3 57.9	+1.0 80.7	+0.5 66.9	+0.4 69.1	−3.3 64.4	−1.5 78.5	−1.6 75.6	−1.9 75.6	−1.7 75.1	−3.3 72.9	−2.4 60.4
T5Gemma-large-ul2	−3.13 70.13	−2.5 78.8	−1.2 72.7	−3.2 73.8	−6.1 51.6	−0.9 74.4	+0.7 74.8	+1.3 77.2	−0.6 81.1	−1.2 73.7	+0.5 70.2	−9.5 62.6	−6.0 59.2	−4.4 75.3	−5.5 60.9	−5.2 63.5	−1.7 66.0	−7.2 72.8	−1.8 75.4	−2.9 74.6	−2.9 73.9	−2.2 74.0	−6.6 56.2
T5Gemma-xl-ul2	−0.55 72.71	+0.0 81.3	+0.4 74.2	−0.9 76.1	−1.0 56.7	+1.1 76.4	+2.9 77.0	+5.5 81.4	+1.1 82.8	+1.5 76.4	+1.3 71.0	−2.2 69.9	−0.1 65.1	−2.2 77.5	−5.6 60.9	−3.0 65.7	−0.5 67.2	−3.3 76.7	+0.0 77.2	−1.3 76.2	−1.5 75.3	+0.5 76.7	−4.8 58.0
T5Gemma-xl-ul2-it	−0.07 73.19	+0.1 81.4	+0.8 74.7	−0.7 76.3	−0.2 57.5	+2.8 78.1	+2.9 77.0	+5.4 81.3	+0.9 82.6	+1.5 76.4	+2.2 71.9	−1.9 70.2	−0.7 64.5	−1.5 78.2	−3.4 63.0	−1.0 67.7	−0.4 67.3	−2.8 77.2	+0.2 77.4	−0.6 76.9	−0.9 75.9	+0.2 76.5	−4.6 58.2
T5Gemma-xl-prefixlm	−0.50 72.76	−1.1 80.2	+1.0 74.8	−0.2 76.8	−1.9 55.8	+1.4 76.7	+2.3 76.4	+5.4 81.3	+0.9 82.6	+0.5 75.4	+2.5 72.2	−1.7 70.4	+0.0 65.2	−1.3 78.4	−5.1 61.3	−2.8 65.9	+0.5 68.2	−5.3 74.7	+0.6 77.8	−1.3 76.2	−0.8 76.0	+0.9 77.1	−5.2 57.5
T5Gemma-xl-prefixlm-it	−1.89 71.37	−1.6 79.7	+0.3 74.1	−2.0 75.0	−5.4 52.3	+0.5 75.8	+1.9 76.0	+3.3 79.2	−1.2 80.5	+0.1 75.0	+0.8 70.5	−4.8 67.3	−3.4 61.9	−3.1 76.6	−5.4 61.0	−2.4 66.3	−0.7 67.0	−5.8 74.2	−1.7 75.5	−2.0 75.5	−1.8 75.0	−1.4 74.8	−5.7 57.1
T5Gemma-2B-ul2	+1.16 74.42	+1.3 82.5	+2.0 75.8	+1.4 78.4	+2.1 59.8	+1.7 77.0	+3.1 77.2	**+7.6** **83.5**	+1.5 83.2	+3.1 78.0	+1.8 71.5	+1.3 73.4	+2.9 68.1	+0.2 79.9	−4.1 62.3	−0.2 68.5	−0.7 67.0	−3.0 77.0	+1.8 79.0	+0.6 78.1	+1.0 77.8	+2.0 78.2	−1.9 60.9
mT5-xl	+0.16 73.42	−0.2 81.1	+0.2 74.0	+0.3 77.3	+1.4 59.1	−1.0 74.3	+0.1 74.2	+1.3 77.2	+0.4 82.1	+0.5 75.4	−0.6 69.1	+0.0 72.1	+0.9 66.1	−0.3 79.4	+1.6 68.0	+0.2 68.9	+0.1 67.8	−1.0 79.0	+0.0 77.2	−0.6 76.9	+0.4 77.2	+0.6 76.8	−0.8 62.0
umT5-xl	+2.40 75.66	+2.1 83.4	+2.5 76.3	+3.2 80.2	+5.0 62.7	+1.9 77.2	+2.8 76.9	+3.2 79.1	+2.3 84.0	+3.9 78.8	+2.1 71.8	+3.5 75.6	+0.3 65.5	+0.9 80.6	+1.7 68.1	+3.1 71.8	+0.6 68.3	**+0.3** **80.3**	+2.3 79.5	+1.9 79.4	+2.7 79.5	+2.4 78.6	**+4.0** **66.8**
mT5-xxl	+3.54 76.80	+2.4 83.7	+2.8 76.6	**+4.1** **81.1**	**+10.2** **67.9**	+2.0 77.3	+3.0 77.1	+5.8 81.7	+2.4 84.1	+3.8 78.7	+2.8 72.5	**+8.0** **80.1**	**+8.1** **73.3**	+1.6 81.3	**+2.4** **68.8**	+2.6 71.3	+0.8 68.5	+0.0 80.0	+2.0 79.2	**+2.9** **80.4**	+3.4 80.2	+3.3 79.5	+3.6 66.4
umT5-xxl	**+3.77** **77.03**	**+2.5** **83.8**	**+3.1** **76.9**	+3.9 80.9	+8.7 66.4	**+3.6** **78.9**	**+3.6** **77.7**	+6.5 82.4	**+2.8** **84.5**	**+4.9** **79.8**	**+3.7** **73.4**	+7.2 79.3	+6.0 71.2	**+1.7** **81.4**	+1.9 68.3	**+3.2** **71.9**	**+1.8** **69.5**	+0.3 80.3	**+3.2** **80.4**	+2.4 79.9	**+3.7** **80.5**	**+4.4** **80.6**	+3.8 66.6

Table 5: Ablations experiments on the CorefUD 1.3 minidev set (CoNLL score in %). The results are averages of 3 or more runs and for every run the epoch with best average score over the whole CorefUD is used.

5 Ablations Experiments

We perform a series of ablation experiments on the CorefUD 1.3 minidev set (to avoid overfitting on the minitest set). The presented results are averages of 3 or more runs, and for every run the epoch with the best average score across all corpora is used.

For reference, the minidev scores of the systems submitted to the CRAC 2025 Shared Task are summarized in Table 5.A.

The first set of experiments evaluates the impact of different models beyond the mT5 and umT5 families. Notably, we also evaluate the XLM-RoBERTa-base and XLM-RoBERTa-large models (Conneau et al., 2020), the RemBERT model (Chung et al., 2021), InfoXLM-large (Chi et al., 2021), and several variants of the recently introduced T5Gemma model (Zhang et al., 2025).

The results are summarized in Table 5.B. The umT5 models consistently outperform the mT5 ones, which is why we used them in the official submission.[2] The mT5 and umT5 models outperform the other evaluated models, particularly because they support longer contexts (Table 6.C and Straka, 2023, Table 4). When restricting the context to 512 tokens, XLM-RoBERTa-large model achieves the best performance, surpassing both InfoXLM-large and RemBERT. Finally, the recently introduced T5Gemma encoder-decoder model adapted from the Gemma decoder-only model seems to lag behind the umT5 models of corresponding sizes, despite supporting longer contexts too.

[2] In this context, it is unfortunate that the umT5-large model has not been released as it would likely outperform the mT5-large model, which is a size very suitable for deployment.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) Cross-Lingual Zero-Shot Evaluation of mT5-large Model																							
Single mT5-large Model	**73.26**	**81.3**	**73.8**	**77.0**	**57.7**	**75.3**	**74.1**	**75.9**	**81.7**	**74.9**	**69.7**	**72.1**	**65.2**	**79.7**	**66.4**	**68.7**	**67.7**	**80.0**	**77.2**	**77.5**	**76.8**	**76.2**	**62.8**
Zero-Shot Multilin. Models	−14.21	−4.2	−16.8	−15.1	−18.5	−13.4	−16.2	−13.0	−7.4	−30.5	−9.3	−8.0	−16.3	−13.5	−2.5	−12.2	−20.5	−24.2	−14.4	−18.4	−8.3	−14.8	−14.9
	59.05	77.1	57.0	61.9	39.2	61.9	57.9	62.9	74.3	44.4	60.4	64.1	48.9	66.2	63.9	56.5	47.2	55.8	62.8	59.1	68.5	61.4	47.9
B) Cross-Lingual Zero-Shot Evaluation of umT5-xl Model																							
Single umT5-xl Model	**75.66**	**83.4**	**76.3**	**80.2**	**62.7**	**77.2**	**76.9**	**79.1**	**84.0**	**78.8**	**71.8**	**75.6**	**65.5**	**80.6**	**68.1**	**71.8**	**68.3**	**80.3**	**79.5**	**79.4**	**79.5**	**78.6**	**66.8**
Zero-Shot Multilin. Models	−14.39	−4.6	−17.4	−15.5	−19.3	−13.5	−18.2	−14.9	−2.0	−34.3	−6.5	−6.3	−21.2	−12.6	−1.9	−14.9	−19.3	−24.0	−15.7	−18.4	−8.5	−16.9	−10.7
	61.27	78.8	58.9	64.7	43.4	63.7	58.7	64.2	82.0	44.5	65.3	69.3	44.3	68.0	66.2	56.9	49.0	56.3	63.8	61.0	71.0	61.7	56.1
C) Various Segment Sizes of mT5-large Model																							
Segment 2560	**73.26**	**81.3**	**73.8**	77.0	**57.7**	**75.3**	**74.1**	**75.9**	**81.7**	**74.9**	**69.7**	**72.1**	65.2	79.7	66.4	68.7	**67.7**	**80.0**	**77.2**	**77.5**	**76.8**	**76.2**	**62.8**
Segment 1024	−0.31	−0.1	−0.7	**+0.0**	**+0.0**	**+0.0**	−0.2	−0.1	−0.3	−0.7	−0.7	**+0.0**	**+1.7**	**+0.1**	**+0.2**	**+0.0**	−1.0	+0.0	−0.3	−0.7	−1.2	−0.9	−2.0
	72.95	81.2	73.1	**77.0**	**57.7**	**75.3**	73.9	75.8	81.4	74.2	69.0	**72.1**	**66.9**	**79.8**	**66.6**	**68.7**	66.7	80.0	76.9	76.8	75.6	75.3	60.8
Segment 512	−2.54	−4.2	−2.8	−2.2	**+0.0**	−0.2	−2.0	−4.8	−2.5	−2.2	−3.3	**+0.0**	−3.5	−1.0	−1.1	−2.1	−4.0	−1.5	−3.5	−3.2	−3.2	−4.6	−4.0
	70.72	77.1	71.0	74.8	**57.7**	75.1	72.1	71.1	79.2	72.7	66.4	**72.1**	61.7	78.7	65.3	66.6	63.7	78.5	73.7	74.3	73.6	71.5	58.8
D) Various Segment Sizes of umT5-xl Model																							
Segment 2560	**75.66**	**83.4**	**76.3**	80.2	**62.7**	**77.2**	**76.9**	**79.1**	**84.0**	**78.8**	**71.8**	**75.6**	65.5	**80.6**	68.1	71.8	**68.3**	**80.3**	79.5	**79.4**	**79.5**	**78.6**	**66.8**
Segment 1024	−0.45	−0.2	−0.9	**+0.0**	**+0.0**	−0.1	−0.4	−0.2	−0.5	−1.8	−0.4	**+0.0**	**+0.2**	+0.0	**+0.6**	**+0.1**	−1.1	−0.1	**+0.0**	+0.0	−1.1	−0.9	−3.1
	75.21	83.2	75.4	**80.2**	**62.7**	77.1	76.5	78.9	83.5	77.0	71.4	**75.6**	**65.7**	80.6	**68.7**	**71.9**	67.2	80.2	**79.5**	79.4	78.4	77.7	63.7
Segment 512	−2.30	−3.2	−2.7	−1.6	**+0.0**	−0.1	−1.9	−3.8	−2.1	−3.6	−1.7	**+0.0**	−5.1	−0.5	+0.2	−1.7	−4.4	−1.8	−3.1	−2.0	−2.3	−3.7	−5.3
	73.36	80.2	73.6	78.6	**62.7**	77.1	75.0	75.3	81.9	75.2	70.1	**75.6**	60.4	80.1	68.3	70.1	63.9	78.5	76.4	77.4	77.2	74.9	61.5
E) Various Sampling Ratios of mT5-large Model																							
Ratio 4/8	73.26	81.3	73.8	77.0	57.7	**75.3**	74.1	75.9	81.7	74.9	**69.7**	72.1	65.2	79.7	66.4	68.7	67.7	80.0	77.2	77.5	76.8	76.2	62.8
Ratio 0/8	−0.23	+0.1	−1.5	−0.5	+0.7	−0.4	−0.1	−0.4	+0.1	−0.7	−0.5	**+0.0**	−0.1	+0.5	−0.6	+0.2	−0.1	−1.0	+0.8	+0.5	+0.1	−0.7	−1.6
	73.03	81.4	72.3	76.5	58.4	74.9	74.0	75.5	81.8	74.2	69.2	**72.1**	65.1	80.2	65.8	68.9	67.6	79.0	78.0	78.0	76.9	75.5	61.2
Ratio 1/8	−0.18	−0.3	−1.0	−0.4	+0.9	−1.9	+0.0	−0.1	−0.1	−0.3	−0.5	−0.7	+0.8	+0.3	+1.0	+0.2	−0.2	−1.8	+0.2	+0.5	−0.2	+0.3	−0.6
	73.08	81.0	72.8	76.6	58.6	73.4	74.1	75.8	81.7	74.6	69.2	71.4	66.0	80.0	67.4	68.9	67.5	78.2	77.4	78.0	76.6	76.5	62.2
Ratio 2/8	−0.36	**+0.4**	−0.5	−0.2	+0.1	−0.9	−0.6	+0.1	+0.2	−0.6	−1.3	−1.8	+0.1	−0.2	−0.7	+0.1	−0.5	−1.7	+0.0	+0.0	+0.2	+0.4	−0.6
	72.90	**81.7**	73.3	76.8	57.8	74.4	73.5	76.0	81.9	74.3	68.4	70.3	65.3	79.5	65.7	68.8	67.2	78.3	77.2	77.5	77.0	76.6	62.2
Ratio 3/8	−0.24	+0.1	−0.2	−0.4	**+1.3**	−1.8	−0.3	−0.3	+0.6	−0.1	−0.5	−1.6	−0.2	−0.1	−0.4	−0.8	**+0.3**	−0.9	+0.0	−0.5	−0.4	+0.6	+0.1
	73.02	81.4	73.6	76.6	**59.0**	73.5	73.8	75.6	82.3	74.8	69.2	70.5	65.0	79.6	66.0	67.9	**68.0**	79.1	77.2	77.0	76.4	76.8	62.9
Ratio 5/8	**+0.09**	+0.3	−0.6	+0.4	−2.6	−0.7	**+0.5**	**+0.7**	**+0.7**	+0.2	−0.6	−3.1	+1.2	**+0.6**	**+2.1**	**+0.8**	−0.3	−0.2	**+1.0**	+0.4	**+0.5**	+0.4	**+0.2**
	73.35	81.6	73.2	77.4	55.1	74.6	**74.6**	**76.7**	**82.4**	75.1	69.1	69.0	66.4	**80.3**	**68.5**	**69.5**	67.4	79.8	**78.2**	77.9	**77.3**	76.6	**63.0**
Ratio 6/8	−0.31	+0.0	+0.3	+0.4	+0.2	−2.5	−0.6	−0.6	+0.2	**+0.8**	−1.0	−1.9	−0.4	−0.2	−0.4	−0.1	−0.1	**+0.5**	−0.1	−0.1	−0.3	−0.9	+0.0
	72.95	81.3	74.1	77.4	57.9	72.8	73.5	75.3	81.9	**75.7**	68.7	70.2	64.8	79.5	66.0	68.6	67.6	**80.5**	77.1	77.4	76.5	75.3	62.8
Ratio 7/8	−0.32	+0.0	+0.0	+0.4	−2.4	−1.3	+0.1	−0.6	−0.4	−0.6	−1.1	−3.9	+0.0	−0.1	+0.9	+0.6	−0.3	−0.6	+0.2	**+0.7**	+0.1	**+1.2**	−0.1
	72.94	81.3	73.8	77.3	55.3	74.0	74.2	75.3	81.3	74.3	68.6	68.2	65.2	79.6	67.3	69.3	67.4	79.4	77.4	**78.2**	76.9	**77.4**	62.7
Ratio 8/8	−0.13	−0.3	**+0.3**	**+0.5**	−0.4	−0.9	−0.7	+0.0	+0.2	+0.0	−0.5	−1.7	**+1.4**	−0.1	+0.5	−0.4	−0.9	−0.6	+0.5	+0.0	+0.2	−0.3	+0.1
	73.13	81.0	**74.1**	**77.5**	57.3	74.4	73.4	75.9	81.9	74.9	69.2	70.4	**66.6**	79.6	66.9	68.3	66.8	79.4	77.7	77.5	77.0	75.9	62.9
F) Various Sampling Ratios of umT5-xl Model																							
Ratio 4/8	75.66	83.4	76.3	80.2	62.7	77.2	76.9	79.1	84.0	78.8	71.8	75.6	65.5	80.6	68.1	71.8	68.3	80.3	79.5	79.4	**79.5**	78.6	66.8
Ratio 0/8	−0.15	**+0.4**	−0.4	−0.9	**+0.9**	−0.4	−0.7	+0.7	−0.2	−0.6	−0.2	**+1.6**	+0.3	+0.3	+0.0	−0.8	+0.2	−0.5	−0.4	−0.1	−0.7	+0.1	−1.6
	75.51	**83.8**	75.9	79.3	**63.6**	76.8	76.2	79.8	83.8	78.2	71.6	**77.2**	65.8	80.9	68.1	71.0	68.5	79.8	79.1	79.3	78.8	78.7	65.2
Ratio 1/8	−0.11	+0.3	−0.7	−0.4	+0.4	−0.8	−0.2	+0.9	−0.2	−0.8	+0.2	+0.3	−0.2	+0.3	+0.4	+0.2	−0.1	−0.5	−0.2	−0.3	−0.5	**+0.7**	−1.3
	75.55	83.7	75.6	79.8	63.1	76.4	76.7	80.0	83.8	78.0	72.0	75.9	65.3	80.9	68.5	72.0	68.2	79.8	79.3	79.1	79.0	**79.3**	65.5
Ratio 2/8	**+0.06**	+0.4	+0.1	+0.2	+0.8	−0.3	−0.4	**+1.1**	+0.0	−0.4	+0.5	+0.6	+0.1	+0.6	+0.9	−0.5	−0.2	−1.0	+0.0	+0.4	−0.3	+0.6	−1.6
	75.72	83.8	76.4	80.4	63.5	76.9	76.5	**80.2**	84.0	78.4	72.3	76.2	65.6	81.2	69.0	71.3	68.1	79.3	79.5	79.8	79.2	79.2	65.2
Ratio 3/8	−0.04	+0.2	−0.1	−0.1	+0.4	−0.3	−0.3	+0.5	+0.1	−0.3	+0.1	+0.2	+0.4	+0.4	**+1.2**	−0.5	**+0.4**	−0.9	−0.3	−0.2	−0.2	−0.1	−1.2
	75.62	83.6	76.2	80.1	63.1	76.9	76.6	79.6	84.1	78.5	71.9	75.8	65.9	81.0	**69.3**	71.3	**68.7**	79.4	79.2	79.2	79.3	78.5	65.6
Ratio 5/8	+0.00	+0.3	+0.2	**+0.4**	−0.3	−0.2	−0.3	+0.5	+0.0	+0.2	−0.5	+0.3	−0.1	+0.2	−0.2	+0.1	+0.1	−0.4	+0.4	+0.1	−0.1	+0.1	−0.7
	75.66	83.7	76.5	**80.6**	62.4	77.0	76.6	79.6	84.0	79.0	71.3	75.9	65.4	80.8	67.9	71.9	68.5	79.9	79.9	79.5	79.4	78.7	66.1
Ratio 6/8	−0.05	+0.1	+0.5	+0.3	−0.9	**+0.2**	−0.2	+0.5	−0.2	−0.1	+0.4	−1.7	+0.8	+0.6	+0.0	−0.1	−0.3	+0.1	+0.4	+0.3	−0.6	+0.2	−1.2
	75.61	83.5	76.8	80.5	61.8	**77.4**	76.7	79.6	83.8	78.7	72.2	73.9	66.3	81.2	68.1	71.7	68.0	80.4	79.9	79.7	78.9	78.8	65.6
Ratio 7/8	−0.12	+0.1	**+0.6**	+0.3	−0.9	−2.1	**+0.1**	+0.4	**+0.3**	−0.1	+0.6	−1.4	**+1.3**	+0.3	+0.0	+0.5	+0.1	−0.4	−0.2	−0.4	−0.2	−0.2	−1.0
	75.54	83.5	**76.9**	80.5	61.8	75.1	**77.0**	79.5	**84.3**	78.7	72.4	74.2	**66.8**	80.9	68.1	72.3	68.4	79.9	79.3	79.0	79.3	78.4	65.8
Ratio 8/8	−0.07	−0.1	+0.4	+0.3	−1.8	−3.2	−0.1	+0.0	+0.0	**+0.4**	**+0.8**	−3.1	+0.8	**+0.9**	−0.1	**+0.7**	+0.3	**+1.4**	**+0.5**	**+0.4**	−0.1	+0.0	**+0.2**
	75.59	83.3	76.7	80.5	60.9	74.0	76.8	79.1	84.0	**79.2**	**72.6**	72.5	66.3	**81.5**	68.0	**72.5**	68.6	**81.7**	**80.0**	**79.8**	79.4	78.6	**67.0**

Table 6: Ablations experiments on the CorefUD 1.3 minidev set (CoNLL score in %). The results are averages of 3 or more runs and for every run the epoch with best average score over the whole CorefUD is used.

Cross-Lingual Zero-Shot Evaluation Given that our model is multilingual, it can be used to perform coreference resolution in languages not exposed to during training. In order to evaluate the performance of our model in such a setting, we train several multilingual models on corpora from all but one language, and then evaluate their performance on the excluded corpora. The results are summarized in Table 6.A for the mT5-large model and in Table 6.B for the umT5-xl model. While

Model	TensorFlow			PyTorch				
	Compile time	Training throughput	Max batch	Cold-start compile	Warm-start compile	Eager throughput	Compiled throughput	Max batch
mT5 base	50s	7.1batch/s	39	55s	27s	8.0batch/s	10.3batch/s	58
mT5 large	91s	3.1batch/s	13	92s	50s	3.3batch/s	4.4batch/s	21
mT5 xl	95s	2.5batch/s	5	97s	51s	2.3batch/s	2.9batch/s	9

Table 7: Comparison of compilation and training times of CorPipe using the latest TensorFlow 2.19 and PyTorch 2.7 with the latest transformers 4.52.4 on a single A100 40GB GPU. The training throughput is measured using batch size of 4 for the xl model and 8 otherwise.

the cross-lingual zero-shot performance is substantially lower by roughly 14 percentage points, it is still higher than the baseline system of Pražák et al. (2021) and on par with the best LLM-track submission. Interestingly, the performance of umT5-xl is higher by more than 2 points, an increase consistent with the results in the supervised setting.

Segment Size The effect of context larger than the usual 512 tokens is quantified in Table 6.C for the mT5-large model and in Table 6.D for the umT5-xl model. The results show that the increase from 512 to 1 024 tokens leads to a significant performance increase of more than 2 percentage points, and the further increase to 2 560 tokens brings a smaller increase by less than 0.5 points.

Sampling Ratio During training, we sample sentences from the training corpora proportionally to the square root of their size, following for example van der Goot et al. (2021); Straka (2024); Straka et al. (2024). We quantify the impact of using different exponents (sampling ratios) in Table 6.E for the mT5-large model and in Table 6.F for the umT5-xl model. The results show that while the choice of 0.5 is reasonable, the sampling ratio has very little impact on the average performance. However, we can see a minor effect of the sampling ratio on the performance of the two largest corpora (the Czech ones), with the decrease of 0.5 to 1.5 percentage points for uniform sampling (sampling ratio 0) to the increase of 0.3 to 0.5 percentage points for proportional sampling (sampling ratio 1).

6 PyTorch vs TensorFlow

Having both PyTorch and TensorFlow implementations of CorPipe, we can compare the two variants in terms of training throughput and memory usage. To this end, we compare the CorPipe 23 using the latest TensorFlow 2.19 and CorPipe 25 utilizing the latest PyTorch 2.7, both with the latest transformers library 4.52.4, on a single A100 40GB GPU.

The results are presented in Table 7. For all the base, large, and xl sizes, the PyTorch implementation outperforms the TensorFlow implementation:

- The training throughput is higher by 16% for the xxl model up to 45% for the base model, when comparing compiled PyTorch models to compiled TensorFlow models.
- The PyTorch model cold-start compilation time is quite similar to TensorFlow; however, the warm-start compilation (reusing cached compilation files from preceding executions; happens automatically) is significantly shorter, being circa half of the TensorFlow time.
- The eager PyTorch model has comparable or slightly better performance than the compiled TensorFlow model.
- The PyTorch implementation has lower memory requirements, allowing batches larger by at least 50% to fit into the GPU memory.

Note that the difference might stem just from different mT5 implementations (FlashAttention, etc.), not necessarily from the frameworks themselves.

7 Conclusions

We introduced CorPipe 25, the winning submission to the CRAC 2025 Shared Task on Multilingual Coreference Resolution (Novák et al., 2025a). Our approach employs a three-stage pipeline architecture that first predicts empty nodes using a dedicated pretrained encoder model, then performs mention detection and coreference linking through a jointly trained system utilizing another pretrained encoder. This complete PyTorch reimplementation significantly outperforms all other submissions by substantial margins of 7 and 8 percentage points for our single model and ensemble variants, respectively. The source code and trained models are publicly available at `https://github.com/ufal/crac2025-corpipe`.

Acknowledgements

Our research has been supported by the OP JAK project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic and uses data provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and conjunction-aware word-level coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Miloslav Konopík, Anna Nedoluzhko, Martin Popel, Ondřej Pražák, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025a. Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches? In *Proceedings of The Sixth Workshop on Computational Approaches to Discourse and The Eight Workshop on Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2025)*, Suzhou, China. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, and 29 others. 2025b. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *Preprint*, arXiv:2504.06225.

Fine-Tuned Llama for Multilingual Text-to-Text Coreference Resolution

Jakub Hejman and Ondřej Pražák and Miloslav Konopík

{hejmanj,ondfa,konopik}@kiv.zcu.cz

Department of Computer Science and Engineering,
NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Technická 8, 306 14 Plzeň
Czech Republic

Abstract

This paper describes our approach to the CRAC 2025 Shared Task on Multilingual Coreference Resolution. We compete in the LLM track, where the systems are limited to generative text-to-text approaches. Our system is based on Llama 3.1-8B, fine-tuned to tag the document with coreference annotations. We have made one significant modification to the text format provided by the organizers: The model relies on the syntactic head for mention span representation. Additionally, we use joint pre-training, and we train the model to generate empty nodes. We provide an in-depth analysis of the performance of our models, which reveals several implementation problems. Although our system ended up in last place, we achieved the best performance on 10 datasets out of 22 within the track. By fixing the discovered problems in the post-evaluation phase, we improved our results substantially, outperforming all the systems in the LLM track and even some unconstrained track systems.

1 Introduction

Coreference resolution is the task of identifying mentions of entities and grouping the mentions of the same real-world entity together. It is a fundamental NLP task that is increasingly left to the implicit understanding of LLMs rather than being explicitly computed as an intermediate step of an NLP pipeline. As such, investigating the models' ability to accurately identify entities in real-world scenarios is a direct way of ensuring that their understanding of the material is robust. Additionally, coreference resolution is an unsolved task, and findings from it may well contribute to progress in related NLP problems. This task can be very challenging, especially in cases where coreferences span the whole document.

CorefUD (Nedoluzhko et al., 2022) is an extension of Universal Dependencies (Nivre et al., 2020) to include coreference harmonized across multiple languages. The recent version of CorefUD 1.3 (Novák et al., 2025b) contains 24 datasets in 17 languages. All data is stored in the CoNLL-U format, which stores the pretokenized text, dependency trees, and entity annotations within the miscellaneous column in a unified format. Basic statistics of individual datasets are shown in Table 1. CRAC shared task on multilingual coreference resolution is built upon this dataset, and 2025 is the fourth year this task has been running.

For generative LLMs, the coreference resolution task is still challenging, and standard benchmarks like SuperGLUE are mostly limited to the Winograd Schema Challenge (WSC) (Levesque et al., 2012). WSC was originally proposed as an alternative to the Turing test. It is a pronoun resolution problem that cannot easily be solved based on statistical patterns. General coreference resolution is typically not present in standard multi-task LLM benchmarks, yet there are many papers focusing on coreference resolution with LLMs. However, the experiments are often limited to a single dataset (Zhang et al., 2023; Stano and Horák, 2025).

As suggested last year (Novák et al., 2024), the CRAC 2025 coreference resolution shared task includes the LLM track, where the participants are asked to use a pure text-to-text approach to solve the task. The organizers also provide a recommended plaintext format of the CorefUD dataset together with the conversion tool. There are several other differences from previous years. As every year, several new datasets were added into CorefUD. The smallest datasets (en_parcorfull and de_parcorfull) were discarded due to very unstable results of all the systems across previous years.

This paper describes how we fine-tune Llama 3.1-8B in a text-to-text manner to participate in this track. Our approach relies on mention head prediction, joint pre-training, and empty node generation.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 140–148
Suzhou, China, November 9, 2025.

dataset	total number of				entities				mentions			
					total	per 1k	length		total	per 1k	length	
	docs	sents	words	empty n.	count	words	max	avg.	count	words	max	avg.
ca_ancora	1,298	13,613	429,313	6,377	17,558	41	101	3.6	62,417	145	141	4.8
cs_pcedt	2,312	49,208	1,155,755	35,654	49,225	43	236	3.4	168,055	145	79	3.6
cs_pdt	3,165	49,419	834,707	21,092	46,460	56	173	3.3	154,437	185	99	3.1
cu_proiel	26	6,832	61,759	6,289	3,396	55	134	6.5	22,116	358	52	1.5
de_potsdam	176	2,238	33,222	0	880	26	15	2.9	2,519	76	34	2.6
en_gum	237	13,263	233,926	119	9,200	39	131	4.4	40,656	174	95	2.6
en_litbank	100	8,560	210,530	0	2,164	10	261	10.8	23,340	111	129	1.6
es_ancora	1,356	14,159	458,418	8,112	19,445	42	110	3.6	70,663	154	101	4.8
fr_ancor	455	31,761	454,577	0	13,204	29	103	4.3	56,459	124	17	1.9
fr_democrat	126	13,057	284,883	0	7,162	25	895	6.5	46,487	163	71	1.7
grc_proiel	19	6,475	64,111	6,283	3,215	50	332	6.6	21,354	333	52	1.7
hbo_ptnk	40	1,161	28,485	0	870	31	102	7.2	6,247	219	22	1.5
hi_hdtb	271	3,479	76,282	0	3,148	41	36	3.8	12,082	158	43	1.8
hu_korkor	94	1,351	24,568	1,569	1,122	46	41	3.6	4,091	167	42	2.2
hu_szegedkoref	400	8,820	123,968	4,857	4,769	38	36	3.2	15,165	122	36	1.6
ko_ecmt	1,470	30,784	482,986	0	16,536	34	55	3.4	56,538	117	12	1.3
lt_lcc	100	1,714	37,014	0	1,087	29	23	4.0	4,337	117	19	1.5
no_bokmaal	346	15,742	245,515	0	5,658	23	298	4.7	26,611	108	51	1.9
no_nynorsk	394	12,481	206,660	0	5,079	25	84	4.3	21,847	106	57	2.1
pl_pcc	1,828	35,874	538,885	18,615	22,143	41	135	3.7	82,706	153	108	1.9
ru_rucor	181	9,035	156,636	0	3,515	22	141	4.6	16,193	103	18	1.7
tr_itcc	24	4,732	55,358	11,584	4,019	73	369	5.4	21,569	390	31	1.1

Table 1: CorefUD 1.3 data sizes in terms of the total number of documents, sentences, words (i.e. non-empty nodes), empty nodes (empty words), coreference entities (total count, relative count per 1000 words, average and maximal length in number of mentions) and coreference mentions (total count, relative count per 1000 words, average and maximal length in number of words). All the counts are excluding singletons and for the concatenation of train+dev+test. Train/dev/test splits of these datasets roughly follow the 8/1/1 ratio. Taken from Novák et al. (2025a)

2 Related Work

Neural coreference resolution has traditionally been approached using encoder-only models (Joshi et al., 2020; Straka, 2023; Pražák et al., 2021; Pražák and Konopik, 2022) and Higher-Order Inference (HOI) (Xu and Choi, 2020). Recently, text-to-text models have gained popularity for this task (Zhang et al., 2023).

The most commonly used model for multilingual coreference resolution is mT5 (Raffel et al., 2020), which has been applied in both end-to-end (Straka, 2023) and text-to-text approaches (Bohnet et al., 2023; Stano and Horák, 2025; Skachkova, 2024). It was also utilized by the top system at CRAC 2024 (Novák et al., 2024).

A notable text-to-text approach is the Link-Append method proposed by Bohnet et al. (2023), which avoids an intermediate mention detection step by training a seq2seq model to predict actions that incrementally build coreference clusters.

Skachkova (2024) introduced a direct annotation scheme where the model generates document text along with brackets and cluster identifiers. Their system employs prompt tuning and incremental generation to label entities progressively, along with data augmentations to address common failure modes such as unchanged inputs, repeated outputs, and duplicate mentions.

Zhang et al. (2023) propose an output scheme which combines tag generation with a second operator that copies tokens from the input to avoid repetition.

An alternative direction to fine-tuning is prompting. Stano and Horák (2025) demonstrate this approach on the simpler anaphora resolution task. This result suggests that some LLMs possess in-context learning capabilities powerful enough to tackle coreference resolution without any specialized training.

Dobrovolskii (2021) suggested reducing the mention space by selecting a single word to represent each mention, using the syntactic head as

the representative word. Their experiments were conducted on the English OntoNotes corpus. In the next step, after antecedent prediction, they employ a CNN-based span predictor to reconstruct the original mentions.

3 Model

We use the provided CoNLL-U-to-Text converter and train the model to generate document texts with entity tags inserted. Our model benefits from joint cross-lingual training, headword mention representation, and zero-mentions modeling.

Inspired by word-level coreference resolution and by previous CorefUD experiments (Pražák et al., 2024; Prazak and Konopík, 2024), we also evaluate the model with headword mention representation. Here, we represent mentions only by their syntactic heads (highest nodes in a dependency tree). The plaintext format suggested by the organizers does not include any syntactic information, so we modified the converter to extract syntactic heads of mentions from CoNLL-U. Considering that the official evaluation metric uses head-match, we do not need to reconstruct the original spans for evaluation. But this step would be fairly straightforward and can be done similarly to Dobrovolskii (2021).

We implement an optional document splitting pre-processing step to deal with datasets dominated by documents that are too long to train on in our setup. The documents are split hierarchically first by paragraphs, then by sentences, and then by words to fit into a limit of 250 words. We chose this limit empirically to fit all the datasets into our training context length. We manually enable this step for datasets that are problematic otherwise.

We train a joint model on a concatenation of all the datasets in the CorefUD 1.3 collection in the first step. In the second step, we fine-tune the joint model on each dataset separately.

Our model also predicts empty nodes and zero mentions. We fine-tune the model to insert empty nodes into the text, directly following its syntactic parent, as suggested by the provided CoNLL-U-to-Text converter.

4 Training & Inference

We fine-tune pre-trained Llama 3.1 8B (Grattafiori et al., 2024) using QLoRA (Dettmers et al., 2023) on a single NVIDIA A40 GPU. The frozen foundational model is quantized to 8 bits, and a LoRA adapter with a rank of 64 is optimized. We use completion-only training, which means that gradients are computed only on completion tokens and not on prompt tokens. This ensures that the model focuses on filling in the entity annotations instead of predicting the original document text.

Our models are trained with a maximum sequence length of 4096 tokens. Sequences that surpass the sequence length limit are filtered from the dataset before training starts. For some datasets, this leads to the removal of all documents from either the evaluation or training split. In these cases, we split the samples so that we effectively utilize the dataset as described in Section 3.

When generating the model's predictions, we use an increased sequence length. For most experiments and datasets, we allow up to 2048 tokens in the prompt and 4096 generated tokens because some datasets contain documents that are, on average, about 2 times longer with labels than without them (more in Section 5.3). For certain datasets, we increase the limits up to 8,192 for the prompt and 16,384 for generation. We do not observe issues with these implicit sequence length extensions between training and inference; scores continue to improve as inference context increases up to the maximum document length.

5 Results & Discussion

Table 2 shows the results of our system on development sets. It is split into two parts: submitted predictions and post-evaluation experiments. Since we did not have enough time to search a complete hyperparameter grid during the evaluation period, we evaluated just two variants of the model:

1. **standard model** – Full-span mention representation, zero mentions are ignored.

2. **heads_zeros model** – Headword mention representation, empty nodes generated, zero mention coreference predicted.

5.1 Submission-time Problems

We performed post-evaluation experiments to address the system's main shortcomings, since we could not resolve all the dataset-specific issues before the deadline. Our original submission exhibited the following problems:

1. **Improper training continuation for joint pre-training** – Due to a bug, joint pre-training

dataset	submitted		post-evaluation experiments		
	standard	heads_zeros	from joint	+ heads_zeros	+ long
ca_ancora	73.27	79.91	74.49	**82.19**	**82.19**
cs_pcedt	57.27	0	59	67.38	**68.89**
cs_pdt	68.75	0	71.24	**76.37**	**76.37**
cu_proiel	14	29	34.5	34.36	**42.95**
de_potsdam	74.4	77	78.95	80.14	**82.83**
en_gum	73.7	76.05	76.57	76.96	**77.16**
en_litbank	81.5	83	82.1	82.1	**84.75**
es_ancora	74.57	0	75.47	80.45	**81.68**
fr_ancor	25.5	26.06	30.7	35.5	**59.95**
fr_democrat	33.89	37.58	49.64	47.78	**57.65**
grc_proiel	50.33	0	54.26	51.61	**65.48**
hbo_ptnk	0	0	46.7	38.04	**69.45**
hi_hdtb	75.7	78.83	75.9	79.92	**80.95**
hu_korkor	40.94	0	46.91	64.72	**65.14**
hu_szegedkoref	62.88	68.52	62.92	67.83	**69.58**
ko_ecmt	66.46	62.02	**65.7**	63.75	**65.7**
lt_lcc	78.26	74.93	**79.33**	76.84	**79.33**
no_bokmaal	77.05	79.12	80.27	80.11	**80.69**
no_nynorsk	74.61	77.63	78.43	79.72	**82.06**
pl_pcc	61.85	0	60.27	**72.3**	**72.3**
ru_rucor	53.96	55.28	59.22	62.53	**63.71**
tr_itcc	24.72	30.76	-	-	**59.4**
avg	56.53	41.13	63.93	66.70	71.28
median	64.67	46.43	65.7	72.3	70.94

Table 2: Results on development splits. Best results are bold. The results on which the best submission is based are underlined. Results marked as '-' could not be evaluated due to massive overfitting and degradation of the output format.

did not improve performance and was therefore omitted from all dataset submissions.

2. **Conversion to CoNLL-U fails if there are more than nine subsequent empty nodes** – this is why there are many 0 scores for the *heads_zeros* model at evaluation time.

3. **Insufficient sequence length** – Causes 0 results for *hbo_ptnk* dataset and very low results for *tr_itcc*.

We solved all the above-mentioned problems later,[1] and the improvement achieved is shown in the second part of Table 2.

[1]Note that test data evaluation is still available only through CodaLab submission, so the post-evaluation entries have exactly the same conditions as the regular ones, except for the extended deadline. We made only 4 test submissions overall, when the limit is 10.

5.2 General Discussion

Table 2 shows that our baseline system achieves satisfactory performance (over 60%) on half of the evaluated datasets. For most of the remaining datasets, the main problem was insufficient maximum sequence length (for details, see Section 5.3).

Joint pre-training helps, but the improvements are somewhat modest (mostly 1-4%). This is a very different result compared to the participating systems from previous years. One factor is the difference in datasets. The two smallest datasets in CorefUD: `en_parcorfull` and `de_parcorfull` were removed from this year's CRAC competition. Such small datasets typically see the largest gain from joint pre-training, because the models tend to overfit more easily without it. The second factor is the difference in model architecture. Previous results make use of Transformers with task-specific heads, but our system trains only an adapter. The

difference here comes from the ability to leverage the pre-trained models' representations. A randomly initialized head has no connection to the knowledge from pre-training, while the adapted transformer can quickly adjust by reusing its latent knowledge.

After fixing all the evaluation issues, we achieve reasonable performance (over 60%) for almost all the datasets with a few exceptions. For both French datasets, our performance is relatively low. We believe the main reason is still in long sequences and long-distance coreferences. The last problematic dataset is Turkish, where we achieve significantly better results on the test set than on the development set. We believe there is an issue with a document in the development set, which contains just two documents.

5.3 Sequence Lengths and Non-Latin Scripts

In our original submission, we had issues with documents or entire datasets surpassing our training context length limit. This limit was originally set to 4096 to compromise between the practical feasibility of the training and processing enough documents to efficiently train the models. More extensive analysis of the actual dataset sequence lengths and tokenization, whose main results are shown in Table 3, shows that this proves problematic for certain datasets.

The average sample length in a majority of datasets within CorefUD fits well into our original context length limit. In all cases except for `fr_democrat`, the median samples happen to fit exactly when the average sample length does too, guaranteeing a suitable amount of data to sufficiently train our models. In the case of fr_democrat, the average is swayed heavily by exceedingly long samples, and the dataset is, in principle, trainable under these conditions as well.

The datasets with training issues due to sequence length issues are `cu_proiel`, `en_litbank`, `grc_proiel`, `hbo_ptnk`, and `tr_itcc`. In the case of `en_litbank` and `tr_itcc`, this can be resolved either by increasing the training sequence length up to 8,192 or by splitting the documents for training.

For `cu_proiel`, `grc_proiel`, and `hbo_ptnk`, the excessive sequence lengths can be attributed to using non-Latin scripts and vocabulary that was not prevalent in the training data of the tokenizer. All three datasets suffer from high number of subword tokens per word, with Hebrew in `hbo_ptnk` reaching 7.7 tokens per word. This comes from the fact that some of the scripts' code points do not have a dedicated token and fall back to byte encoding.

Context length limitations cause issues during inference as well. Having some documents that are truncated by a small amount for inference does not lower model performance as drastically as having a large amount of unused training documents. Truncated documents during inference will decrease the maximum achievable score proportionally to the truncated length, but missing training documents may lead to drastic over-fitting and near-zero scores. In addition, increasing the inference sequence length is less memory-intensive than increasing the training sequence length, and we manage to run inference at up to 8,192 input tokens and 16,384 output tokens while still recovering additional score points. Because long-context inference is much more practical than long-context training, we settled on running inference for entire documents and invested our time in other optimizations.

5.4 Effective Context Length

To determine how much context is truly necessary for coreference resolution in the CorefUD datasets, we investigate the distances between entity mentions within documents. We compute the distance between all consecutive pairs of mentions of each entity within each document. To match our results with the application, we use the outer bounds, from the beginning of the first mention to the end of the second mention.[2] The distribution of these distances across all datasets is heavily right-skewed. The median distance is 16 words, with partial medians spanning between 6 (`tr_itcc`) and 25 (`es_ancora`). The 90%, 95%, and 99% quantiles are 118, 220, and 728 words, respectively. The longest distance in any dataset is 12,398 words in `fr_democrat`.

These values suggest that most mentions of an entity are close together, but there are some long-distance dependencies that require large context windows. Generally, a sliding context window of 4096 tokens should be sufficient for 95-99% of most datasets if implemented carefully. This way, just about all mentions would have at least one other mention within their context window. However, the remaining 1-5% of mentions would still need a larger context window. Without a method

[2] Our processing of discontinuous mentions is simplified. Zero mentions are counted as full words. Each part of a discontinuous mention counts as a separate mention.

dataset name	toks/word	word length	max text	max label	mean text	mean label
ca_ancora	1.60	5.18	5,404	8,152	528.4	782.5
cs_pcedt	1.78	5.90	7,255	9,831	888.5	1,230.8
cs_pdt	1.84	5.85	5,231	8,415	473.7	761.4
cu_proiel	3.56	5.55	**42,169**	**56,978**	**15,507.5**	**21,134.8**
de_potsdamcc	1.70	6.24	420	746	319.0	503.6
en_gum	1.10	5.02	2,152	5,403	1,103.4	2,629.0
en_litbank	1.09	4.86	3,624	5,958	2,301.0	3,747.6
es_ancora	1.43	5.35	2,471	3,765	485.7	755.4
fr_ancor	1.34	4.90	**20,768**	**41,700**	1,362.7	2,679.0
fr_democrat	1.45	4.98	**23,161**	**51,495**	6,619.8	**14,166.1**
grc_proiel	3.53	5.87	**53,486**	**71,886**	**22,042.8**	**29,976.9**
hbo_ptnk	7.70	5.55	**10,317**	**11,876**	5,951.6	6,918.5
hi_hdtb	2.53	4.83	1,682	2,286	742.2	1,004.9
hu_korkor	2.67	6.55	1,493	1,844	683.2	861.9
hu_szegedkoref	2.28	5.77	4,152	4,836	715.6	905.0
ko_ecmt	2.49	3.98	4,433	6,433	817.9	1,230.9
lt_lcc	2.70	6.37	2,217	2,773	1,016.5	1,244.5
no_bokmaalnarc	1.71	5.46	**10,989**	**21,353**	1,221.1	2,356.5
no_nynorsknarc	1.81	5.50	4,812	8,846	932.5	1,743.1
pl_pcc	2.13	5.85	5,784	11,327	629.2	1,126.3
ru_rucor	1.83	5.93	6,449	8,514	1,562.2	1,987.0
tr_itcc	1.86	6.38	4,920	7,181	4,411.8	6,634.8

Table 3: All statistics are computed on the train split of the dataset, using the meta-llama/Llama-3.1-8B tokenizer. Token counts above 10,000 tokens are highlighted in bold red, samples that fit into our initial training context length are colored blue and underlined. The "toks/word" column contains the average number of tokens per word in the data. Because of how the text is pre-tokenized, punctuation such as periods and commas count as words as well. The "word length" column contains the mean word length in Unicode code points. The last four columns contain the maximum number of tokens in a sample and the average sample length in tokens for both the model input and completion. Note that the training sequences actually consist of the concatenation of both sequences along with additional overhead for the completion marker.

to recover broken chains in long documents, these long-distance mentions could account for a disproportionately large portion of the final score.

The main factor in long context mentions and document length appears to be the type and source of the data. The longest distance comes from the short story "Sarrasine" by Honoré de Balzac, which is present in `fr_democrat`. The entity in question refers to the Lanty family and has many mentions throughout the story.

This analysis suggests that while most coreference relations occur within manageable context windows, a certain portion of datasets contain long-distance dependencies that prove challenging to our approach. These long-distance coreferences are especially prevalent in both French datasets. In contrast, other datasets with shorter average document length tend to have their mentions closer together. This raises the question of whether modeling long-distance mentions separately would improve efficiency and possibly performance, or whether simply scaling the context window is more practical.

5.5 Dataset Discrimination Capabilities

Our experiments included joint models trained on a mixture of all datasets without dataset-specific fine-tuning. We never invested the resources to fully evaluate these models. Partial results suggest that this general version of the model is typically weaker than the specialized models trained on each dataset individually. Investigating the joint approach gives insights into how a single model is able to generalize between datasets.

The originally employed prompt template does not explicitly contain information about which

system	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_potsdam	en_gum	en_litbank	es_ancora	fr_democrat	fr_ancor	grc_proiel	hbo_ptnk	hi_hdtb	hu_korkor	hu_szeged	ko_ecmt	lt_lcc	no_bokmaal	no_nynorsk	pl_pcc	ru_rucor	tr_itcc	average
corpipe-best	**84.20**	**76.94**	80.64	62.63	**78.71**	77.38	80.91	84.64	**80.03**	**73.26**	**76.80**	67.27	81.90	**70.24**	**73.16**	69.21	**82.61**	**80.10**	**80.74**	**80.31**	79.71	67.51	76.77
corpipe-ens	84.09	76.92	**81.08**	**64.20**	77.89	**77.48**	80.04	**85.07**	79.64	72.51	76.14	66.75	**81.99**	69.72	73.09	**69.44**	81.62	80.09	79.66	80.29	80.05	65.50	76.51
corpipe-1	83.25	75.94	80.22	62.66	76.90	76.54	80.09	84.23	78.97	71.93	76.18	66.02	80.64	68.11	71.86	67.59	80.22	79.20	80.33	79.23	78.30	66.69	75.69
ours_post*	81.35	72.12	74.97	56.69	69.78	75.76	82.67	82.01	58.56	49.14	60.53	48.20	77.42	65.76	69.81	67.83	69.17	76.78	72.06	76.56	**84.41**	**69.09**	70.03
stanza	80.30	72.83	74.49	37.95	77.97	70.74	72.96	79.53	69.75	63.22	54.07	63.57	78.87	65.32	68.61	64.86	78.81	74.93	75.32	74.10	78.42	49.48	69.37
antoine.b	68.92	61.85	62.88	39.95	63.95	65.20	72.12	68.82	69.00	65.02	54.08	57.83	72.11	52.52	60.39	60.59	74.21	69.80	70.30	65.00	67.80	42.80	62.96
oseminck	73.45	65.12	71.33	58.25	59.60	58.73	69.01	74.43	66.74	60.43	65.75	43.96	56.36	52.53	59.82	63.04	62.55	64.74	61.63	72.55	68.79	56.23	62.96
moizsajid	60.87	51.36	54.30	58.48	48.74	69.78	70.38	61.75	71.94	57.59	57.85	**80.15**	71.32	43.49	52.27	66.05	59.16	72.76	68.86	70.83	71.40	39.00	61.74
PuxAI	68.01	56.94	62.96	43.74	57.41	61.71	69.12	70.52	63.77	61.54	47.86	45.31	66.85	50.58	61.61	50.32	65.35	65.18	63.00	66.55	67.59	56.06	60.09
ours	79.17	61.02	68.17	25.34	67.63	73.64	**84.05**	73.63	58.56	49.14	47.64	0.00	75.84	38.91	67.32	68.30	63.44	73.77	71.96	64.49	**80.12**	24.31	59.84
baseline-gz	70.53	68.00	67.43	27.69	57.90	64.97	66.59	71.71	65.37	56.27	29.78	23.77	69.86	49.86	59.05	63.04	69.32	66.11	66.76	65.63	63.39	47.14	58.64
baseline	69.94	57.32	63.20	24.10	57.90	64.96	66.59	71.32	65.37	56.27	26.98	23.77	69.86	46.61	58.34	48.34	69.32	66.11	66.76	64.08	63.39	40.06	56.39

Table 4: Results of all competing models in both tracks on the test set. Best overall scores are bold. Best scores within the LLM track are underlined (if they are not already bold). Row marked with * shows post-evaluation experiments. Post-evaluation results are also highlighted in the same manner, in addition to the official results. LLM track systems have names in bold.

dataset the current sample comes from. There are differences between how the individual datasets are annotated, and using a model trained on one while evaluating on another usually degrades model performance significantly. If the model did not know which annotation rule set to apply to each sample, it would be at a disadvantage. There are two options: either the fine-tuned LLMs already implicitly model the distinction between the datasets, or their performance can be further improved by giving them this information.

We hypothesize that it is possible that the different datasets are easily distinguishable due to factors like the length or domain of the document, or the tokenization used. Of the 22 datasets, only 5 pairs share language: `cs_pcedt` and `cs_pdt`, `en_gum` and `en_litbank`, `fr_ancor` and `fr_democrat`, `hu_korkor` and `hu_szegedkoref`, `no_bokmaalnarc` and `no_nynorsknarc`. We train a model to predict the dataset name before completing the annotations and find that it achieves 100% accuracy in classifying all datasets' evaluation splits. This result confirms that it is possible to distinguish all datasets based on the input text alone and that, when necessary, the LLM will implicitly utilize this information.

5.6 Final Results

Table 4 shows the final results on test sets. The column *ours_post** shows scores from our post-evaluation experiments, which are not a part of the official competition. From the results, we can see that though our system ended up in the last place, we achieved the best results within the LLM track for 10 datasets out of 22, which is the highest number of datasets won by a single system in this track. The reason for our low average score was in dataset-specific problems, which led to very low performance on these datasets. After fixing all issues in the post-evaluation phase, our system outperformed all other systems within the LLM track by a large margin. It would take the fourth place overall and become the second unique system (the first three *Corpipe* entries are variants of the same system by a single team).

6 Conclusion

We proposed a Llama-based text-to-text multilingual coreference resolution system with headword mention representation and joint pre-training for the CRAC 2025 shared task. We provide an extended analysis of different model configurations.

We found that generative tagging approaches struggle with large documents due to limited sequence length when running an open-weight model on a single machine. Languages with non-Latin scripts often tokenize inefficiently, leading to very long sequences.

Our system ended up in last place. However, we achieved the best results on 10 datasets out of 22. The main problem of our submission was the very low performance for a small subset of datasets, which was caused by some mistakes we did not manage to fix on time. After fixing all identified issues in the post-evaluation phase, we achieved the best results in the LLM track by a large margin, and we even outperformed some systems in the unconstrained track.

Considering the relatively small size of our model, we believe LLMs can achieve state-of-the-art results on CorefUD in the near future.

Acknowledgments

This work has been supported by the Grant no. SGS-2025-022 - New Data Processing Methods in Current Areas of Computer Science.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561, Rome, Italy. AAAI Press.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of LREC*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025a. Findings of the fourth shared task on multilingual coreference resolution. In *Proceedings of the eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, Suzhou. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025b. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ondřej Pražák and Miloslav Konopik. 2022. End-to-end multilingual coreference resolution with mention head prediction. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 23–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ondrej Prazak and Miloslav Konopík. 2024. End-to-end multilingual coreference resolution with head-word mention representation. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 107–113, Miami. Association for Computational Linguistics.

Ondřej Pražák, Miloslav Konopík, and Pavel Král. 2024. Exploring multiple strategies to improve multilingual coreference resolution in corefud. *arXiv preprint arXiv:2408.16893*.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Natalia Skachkova. 2024. Multilingual coreference resolution as text generation. In *Proceedings of the*

Seventh Workshop on Computational Models of Reference, Anaphora and Coreference, pages 114–122, Miami. Association for Computational Linguistics.

Patrik Stano and Aleš Horák. 2025. Evaluating prompt-based and fine-tuned approaches to czech anaphora resolution. *Preprint*, arXiv:2506.18091.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.

Few-Shot Coreference Resolution with Semantic Difficulty Metrics and In-Context Learning

Nguyen Xuan Phuc, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23521213@gm.uit.edu.vn
thindv@uit.edu.vn

Abstract

This paper presents our submission to the CRAC 2025 Shared Task on Multilingual Coreference Resolution in the LLM track. We propose a prompt-based few-shot coreference resolution system where the final inference is performed by Grok-3 using in-context learning. The core of our methodology is a difficulty-aware sample selection pipeline that leverages Gemini Flash 2.0 to compute semantic difficulty metrics, including mention dissimilarity and pronoun ambiguity. By identifying and selecting the most challenging training samples for each language, we construct highly informative prompts to guide Grok-3 in predicting coreference chains and reconstructing zero anaphora. Our approach secured 3rd place in the CRAC 2025 shared task.

1 Introduction

Coreference resolution is the task of identifying and grouping linguistic expressions in a text that refer to the same real-world entity or event, which may occur within the same sentence or be separated across multiple sentences, sometimes requiring analysis of the entire document for accurate identification. This task involves two subtasks: identifying entity mentions (including zero mentions such as pro-drops) and clustering them into groups corresponding to actual entities or events.

This paper describes our approach to the CRAC 2025 Shared Task on Multilingual Coreference Resolution, which is the fourth iteration of this ongoing challenge organized in conjunction with the CODI-CRAC 2025 Workshop at EMNLP 2025. Building on the successes of previous editions in 2022 (Žabokrtský and Ogrodniczuk, 2022), 2023 (Žabokrtský and Ogrodniczuk, 2023), and 2024 (Novák et al., 2024), the 2025 shared task emphasizes multilingual capabilities and introduces a dedicated LLM Track to explore the potential of large language models (LLMs) in handling coreference across typologically diverse languages. Participants are tasked with developing systems that not only detect mentions, including the reconstruction of zero mentions but also accurately cluster them, while accommodating linguistic variations such as different annotation styles and the presence of pro-drops.

The data for the shared task is based on the public edition of CorefUD 1.3 (Novák et al., 2025), comprises 22 different datasets across 17 languages in a harmonized scheme. Compared to CRAC 2024, two additional languages, Korean and Hindi, with a new French dataset, while excluding the English-ParCorFull and German-ParCorFull datasets. The data is provided in CoNLL-U format, with coreference annotations in the MISC column, and a plaintext variant is available for LLM-based approaches to facilitate prompt engineering and in-context learning. To promote realism, development and test sets are reduced to mini-dev and mini-test splits (approximately 25,000 words each), and morpho-syntactic features in input data are generated using UDPipe 2 (Straka, 2018), simulating scenarios without gold annotations.

A key innovation in CRAC 2025 is the bifurcation into two tracks: the LLM Track, which restricts systems to primarily LLM-driven methods such as fine-tuning, in-context learning, prompt tuning, and constrained decoding, and the Unconstrained Track, which allows hybrid or non-LLM approaches. Our participation in the LLM Track leverages to address the challenges of multilingual coreference, including zero mention reconstruction and cross-lingual transfer. The evaluation employs the CorefUD scorer[1], with the primary metric being the macro-averaged CoNLL F_1 score across all datasets, using head-matching for mentions and excluding singletons. This setup encourages the development of robust, multilingual systems capable

[1] https://github.com/ufal/corefud-scorer

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 149–153
Suzhou, China, November 9, 2025.

of handling diverse linguistic phenomena.

2 Related Work

The field of coreference resolution has evolved from early rule-based and statistical models (Snyder et al., 2009) to end-to-end neural architectures that framed the task as a span-ranking problem (Wang et al., 2017). The advent of pre-trained language models like BERT (Devlin et al., 2019) further advanced the state-of-the-art, with models like SpanBERT (Joshi et al., 2020) achieving new performance benchmarks by capturing richer contextual information. However, these models still largely rely on a fine-tuning paradigm, which requires substantial annotated data, a resource scarce for most languages.

More recently, the landscape has shifted with the emergence of LLMs such as GPT-3 (Brown et al., 2020), which excel at zero-shot and few-shot learning. These models can perform complex NLP tasks through in-context learning (ICL) (Dong et al., 2024) without parameter updates, often guided by a few examples in a prompt (Chen et al., 2023). The effectiveness of ICL, however, is highly sensitive to the quality and relevance of the selected exemplars (Nie et al., 2022). While most work has relied on random or heuristic-based sample selection, our approach focuses on a difficulty-aware strategy to curate the most informative examples.

3 Method

Our approach for the LLM Track employs in-context learning through few-shot prompting and carefully designed instructions to enable LLMs to perform multilingual coreference resolution. To construct effective few-shot demonstrations, we curate samples from the mini-dev sets, which contain both raw text inputs and corresponding gold-standard annotations, making them ideal for instructional purposes.

The selection of exemplars is guided by a custom difficulty metric designed to identify challenging instances that reflect diverse linguistic phenomena, such as nominal ambiguity, pronominal interference, and zero mentions (e.g., pro-drops). This approach ensures that few-shot examples expose the LLM to complex scenarios, enhancing its robustness across the 17 languages in the dataset and potential unseen languages in the mini-test set. The difficulty score is computed as a weighted linear combination of three components: the Nominal Dissimilarity Score, the Pronoun Ambiguity Score, and the Zero Mention Score. Each component is detailed below, followed by its integration methodology.

3.1 Nominal Dissimilarity Score

The Nominal Dissimilarity Score quantifies semantic dissimilarity within coreference clusters, reflecting the resolution challenge posed by diverse or partial matches between mentions. For each text sample (in plaintext format with annotations such as [eX mentionleX]), we employ the Gemini Flash 2.0 API to execute a multi-step analysis as follows:

- **Mention Extraction**: A regular expression-based parser extracts all mentions and their corresponding cluster assignments from the annotated plaintext, producing structured JSON output containing mention spans and entity identifiers.
- **Representative Phrase Selection**: Using the extracted mentions, the LLM identifies a representative phrase for each cluster (e.g., the most descriptive or head noun phrase).
- **Semantic Similarity Computation**: For each cluster, we use Gemini Flash 2.0 with a tailored instruction prompt to compute a semantic similarity score (on a 0-100 scale) between each mention and its representative phrase, using LLM embeddings. The overall Nominal Dissimilarity Score is the average of inverted similarity scores (100 - similarity) across all mentions in the sample, reflecting resolution difficulty.

This score identifies instances where mentions within a cluster exhibit low semantic similarity, increasing resolution difficulty. In cases of processing errors (e.g., invalid LLM output), the score defaults to 0.0 in case of processing errors.

3.2 Pronoun Ambiguity Score

The Pronoun Ambiguity Score evaluates pronominal ambiguity by measuring the level of interference from distracting antecedents for pronouns within the text. We utilize the Gemini Flash 2.0 API for this analysis:

Pronoun Extraction The LLM extracts all pronouns, recording their positions and contexts, and outputs them in structured JSON format.

Relationship Analysis For each pronoun, the LLM identifies potential antecedents within a ±150 character window, categorizing them as either supporting" (those that align with the correct coreferential entity) or distracting" (plausible but incorrect alternatives based on gender, number, or semantic agreement). The ambiguity score for each pronoun is calculated as: distracting count minus supporting count (higher values indicate greater ambiguity).

Aggregation The overall Pronoun Ambiguity Score is the average of positive per-pronoun scores (where distractors outnumber supporters), normalized to a 0–100 scale by dividing by the maximum observed score in the mini-dev set and multiplying by 100. This focuses on genuinely ambiguous cases. Summary statistics, including total pronoun count and score distribution, are computed for validation purposes. In cases of processing errors (e.g., invalid LLM output), the score defaults to 0.0. This metric captures discourse-level challenges where multiple candidate antecedents can mislead the resolution process.

3.3 Zero Mention Score

The Zero Mention Score quantifies the difficulty posed by implicit references (e.g., pro-drops), which require significant syntactic and semantic inference for reconstruction. These are marked by "##" in the plaintext format. Instead of using a tiered system with arbitrary boundaries, we employ a continuous function to provide a more robust and principled score. The score is calculated as a capped linear function of the number of zero-mention occurrences (N_{zero}) within the sample:

$$S_{\text{zero}} = \min(N_{\text{zero}} \times C, 100)$$

where C is a scaling factor chosen to make the score's magnitude comparable to the other two metrics. For our experiments, we set $C = 2$.

3.4 Difficulty Score Integration

The final difficulty score (S_{diff}) is computed as a weighted linear combination of the three component scores:

$$S_{\text{diff}} = 0.4 \cdot S_{\text{nom}} + 0.4 \cdot S_{\text{pron}} + 0.2 \cdot S_{\text{zero}}$$

where S_{nom}, S_{pron}, and S_{zero} represent the Nominal Dissimilarity, Pronoun Ambiguity, and Zero Mention scores, respectively. The final score is capped at 100. The weights for this combination were established to reflect the distinct nature of the challenges that each metric captures.

Nominal Dissimilarity (S_{nom}) and Pronoun Ambiguity (S_{pron}) were assigned equal, high weights of 0.4. We posit that these metrics are the primary indicators of deep inferential complexity. S_{nom} reflects semantic challenges requiring world knowledge, while S_{pron} captures structural ambiguity at the discourse level. Prioritizing these equally ensures that we select for samples rich in complex reasoning tasks, which are most beneficial for challenging the LLM in a few-shot setting.

The Zero Mention score (S_{zero}) was assigned a lower weight of 0.2. While identifying zero anaphora is crucial, this metric primarily quantifies the *frequency* of the phenomenon rather than the reasoning complexity of a single instance. Therefore, it serves as an important secondary factor that modulates the final score but is weighted less than the core semantic and discourse challenges. This weighting scheme is deliberately designed to favour exemplars that are structurally and semantically complex, aiming to maximize the learning signal provided to the LLM.

3.5 Sample Selection and Final Inference

To accommodate linguistic diversity, we compute difficulty scores and perform sample selection on a per-language basis. For each language, we rank its mini-dev samples by their difficulty scores and select the top 3 most challenging examples as few-shot demonstrations. For unseen languages, we employ a zero-shot prompt.

The prompts are tailored to each language. Finally, the constructed prompt, containing instructions and the selected few-shot examples, is fed to **Grok-3**. Grok-3 processes the input test data to generate coreference predictions, which are subsequently converted back to CoNLL-U format using the provided text2text-coref [2] tool for evaluation.

4 Results and Discussion

Our system secured the third position among four submissions in the LLM Track, with an average CoNLL F1-score of 60.09 across all datasets. This result places our system 2.87 points behind the top-performing entry's score of 62.96. Detailed per-dataset results are presented in Table 1 in the Appendix.

[2] https://github.com/ondfa/text2text-coref

Table 1: Results and rankings of our method on the datasets in the competition.

Dataset	F1-score	Dataset	F1-score	Dataset	F1-score	Dataset	F1-score
ca_ancora	68.01 (3)	pl_pcc	66.55 (3)	no_bokmaalnarc	65.18 (3)	hbo_ptnk	45.31 (2)
cs_pcedt	56.94 (3)	es_ancora	70.52 (3)	no_nynorsknarc	63.00 (3)	fr_ancor	63.77 (3)
cs_pdt	62.96 (3)	fr_democrat	61.54 (1)	tr_itcc	56.06 (2)	hi_hdtb	66.85 (3)
de_potsdamcc	57.41 (3)	hu_szegedkoref	61.61 (2)	cu_proiel	43.74 (3)	ko_ecmt	50.32 (4)
en_gum	61.71 (3)	ru_rucor	67.59 (4)	en_litbank	69.12 (3)		
lt_lcc	65.35 (1)	hu_korkor	50.58 (2)	grc_proiel	47.86 (3)		

The strongest performance of our system was observed on the `fr_democrat` and `lt_lcc` datasets, where it achieved the top rank. We attribute this success to our use of in-context learning with preselected examples, which proved particularly effective for identifying essential mentions. Similarly, the system demonstrated competitive performance on datasets characterized by a high frequency of zero mentions—such as `tr_itcc`, `ca_ancora`, and `hu_szegedkoref` securing ranks from 2 to 3. The high CoNLL F1 scores on these datasets suggest that our few-shot approach successfully captured zero-mentions, a key objective of the shared task.

Limitations and Ablation Analysis Despite these successes, our approach has several limitations that warrant discussion. A key limitation in our post-hoc analysis is the absence of a direct baseline comparing our difficulty-aware selection against a random sampling strategy using the same Grok-3 model. Such a comparison would have precisely quantified the performance gain attributable solely to our selection methodology. While time and resource constraints of the shared task prevented this ablation study, we acknowledge its importance for a more thorough evaluation. This lack of a direct baseline is a primary area for future work.

Beyond the need for a baseline, the overall performance was constrained by several factors. Firstly, the model struggled with long input contexts, particularly evident in datasets like `fr_ancor`. When faced with extensive texts, the model often failed to maintain context over long distances, leading to the fragmentation of coreference chains and an inability to resolve long-distance dependencies. This issue was compounded by occasional failures to adhere to the specified output format, which caused critical errors during the `text2conllu` conversion phase and prevented the establishment of semantic links between distant mentions.

Secondly, our strategy's reliance on only three few-shot examples per language, while computationally efficient, likely provided the LLM with an insufficient representation of linguistic diversity. Although our metrics aimed to select the *most informative* examples, this low quantity may have constrained the model's ability to generalize across the full spectrum of coreference phenomena present in the test data.

Finally, performance was impacted by domain and language mismatches. The `ko_ecmt` dataset, with its admixture of Korean, English, and Chinese text, posed a significant challenge for our clustering algorithm. Similarly, the presence of ancient languages (`hbo_ptnk`, `cu_proiel`, `grc_proiel`), which are out-of-domain relative to the LLM's pretraining data, highlighted the difficulty of adapting modern LLMs to historically distant linguistic contexts, even with targeted few-shot prompting.

5 Conclusion

In this paper, we introduced a few-shot coreference resolution system for the CRAC 2025 Shared Task using difficulty-aware in-context learning, achieving third place in the LLM track. Our approach demonstrated the viability of no-fine-tuning methods but was limited by using only three training exemplars per language. This led to coreference chain fragmentation and reduced performance on long documents and out-of-domain ancient languages. Future work will focus on two key areas: first, establishing a clear baseline against random sampling to rigorously validate the impact of our difficulty metrics; and second, exploring methods to incorporate a larger, yet still curated, set of diverse training examples to improve generalization and overall performance.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuanjing Huang. 2023. KNSE: A knowledge-aware natural language inference framework for dialogue symptom status recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10278–10286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *Preprint*, arXiv:2212.02216.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 73–81.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207. Association for Computational Linguistics.

Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2017. Affinity-preserving random walk for multi-document summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 210–220.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2022. *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2023. *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*.

Few-Shot Multilingual Coreference Resolution Using Long-Context Large Language Models

Moiz Sajid, Seemab Latif, Zuhair Zafar, Muhammad Moazam Fraz
National University of Science and Technology, Islamabad, Pakistan
msajid.phdai24seecs@seecs.edu.pk; seemab.latif@seecs.edu.pk;
zuhair.zafar@seecs.edu.pk; moazam.fraz@seecs.edu.pk

Abstract

In this work, we present our system, which ranked second in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track). For multilingual coreference resolution, our system mainly uses long-context large language models (LLMs) in a few-shot in-context learning setting. Among the various approaches we explored, few-shot prompting proved to be the most effective, particularly due to the complexity of the task and the availability of high-quality data with referential relationships provided as part of the competition. We employed Gemini 2.5 Pro, one of the best available closed-source long-context LLMs at the time of submission. Our system achieved a CoNLL F1 score of 61.74 on the mini-test set, demonstrating that performance improves significantly with the number of few-shot examples provided, thanks to the model's extended context window. While this approach comes with trade-offs in terms of inference cost and response latency, it highlights the potential of long-context LLMs for tackling multilingual coreference without task-specific fine-tuning. Although direct comparisons with traditional supervised systems are not straightforward, our findings provide valuable insights and open avenues for future work, particularly in expanding support for low-resource languages.

1 Introduction

Ever since the work of (Brown et al., 2020) showed that a general-purpose language model, trained on diverse internet-scale data, could perform well across a vast range of NLP tasks, most complex NLP tasks are now tackled first through text generation LLMs. However, there have been only a handful of works on the task of coreference resolution using text generation LLMs. This work presents the second-best approach on the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track) that utilizes the current state-of-the-art LLM model named Gemini 2.5 Pro (Comanici et al., 2025) from Google. This is the fourth edition of the shared task. Previous editions of the shared task were conducted successfully in 2022 (Žabokrtský et al., 2022), 2023 (Žabokrtský et al., 2023), and 2024 (Novák et al., 2024).

Coreference resolution remains one of the most challenging tasks in natural language processing (NLP), as it requires a comprehensive understanding of language at multiple levels, including semantics, syntax, discourse structure, and pragmatics. The complexity of this task is further amplified in multilingual settings, where variations in linguistic phenomena, grammatical structures, and referential expressions across languages introduce additional challenges. Despite its importance, research on multilingual coreference resolution remains relatively limited, leaving significant gaps in methodologies and resources for addressing this problem effectively.

Our work is, to the best of our knowledge, the first to leverage large language models (LLMs) with extended context lengths for the task of multilingual coreference resolution. Inspired by prior approaches that formulate coreference resolution as a text generation problem (Skachkova, 2024) (Le and Ritter, 2023) (Gan et al., 2024), our method processes raw text as input and directly generates text annotated with coreference clusters as output.

The main contributions of this work are as follows:

- We present the second-best performing system in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track), trailing the top submission by a margin of only 1.22.
- We model multilingual coreference resolution as an end-to-end text generation task, enabling the system to learn from few-shot examples with long context spans.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 154–162
Suzhou, China, November 9, 2025.

- We utilized the current state-of-the-art Gemini 2.5 Pro model for this challenging task, demonstrating its effectiveness in handling large-context, multilingual coreference resolution.

2 Related Work

Early approaches to coreference resolution typically adopted a two-stage framework, first identifying coreferent mentions and then using these mentions to construct coreference clusters. This paradigm was also employed by last year's shared task winner (Straka, 2024). In contrast, several subsequent studies have explored coreference resolution as an end-to-end problem, jointly performing mention detection and coreference clustering. Notable contributions in this direction include the works of (Lee et al., 2017), (Lee et al., 2018), (Joshi et al., 2019), and (Joshi et al., 2020). Most end-to-end methods build upon the foundational work of (Lee et al., 2017), which was the first to train a coreference model in a fully end-to-end manner, unlike prior approaches that relied on external systems for mention detection or clustering.

(Liu et al., 2022) explicitly models the structure of coreference resolution using language models, achieving state-of-the-art results on the OntoNotes benchmark from the CoNLL-12 English shared task dataset (Pradhan et al., 2013). Similarly, (Bohnet et al., 2023) demonstrates strong performance on the same benchmark through a text-to-text generation paradigm, where the text is processed in an autoregressive manner. Their approach employs a Link-Append transition system that encodes previously established coreference links and incrementally predicts new ones.

Our approach leverages end-to-end text generation LLMs for multilingual coreference resolution, enabling the effective application of few-shot prompting strategies on raw input texts paired with their gold annotations from the training split. Furthermore, we extend this methodology by utilizing recent long-context LLMs, which have demonstrated state-of-the-art performance across a wide range of NLP tasks.

3 Experiments

3.1 Dataset

The data utilized in this work is derived from the CRAC 2025 Shared Task, now in its fourth edition, and based on the CorefUD 1.3 collection. This year's dataset encompasses 17 languages, representing an increase of 6, 4, and 1 languages compared to the 2022, 2023, and 2024 editions, respectively. To facilitate the use of large language models (LLMs) such as GPT-4o, LLaMA, and Claude for the coreference resolution task, the organizers released a text-to-text version of the dataset in addition to the standard CoNLL-U format. This alternative representation proved advantageous for our method, as it enabled the effective application of few-shot prompting strategies. Examples of input–output pairs employed in our system for few-shot prompting are provided in Figures 1 and 2. The token counts for the 22 datasets across the training, development, and test splits are reported in Appendix A.1.

LLM	Few-Shot Examples Token Count	en gum	en litbank
Gemini 2.5 Flash	0	32.10	45.65
	100,000	46.68	63.14
	200,000	47.25	51.61
	300,000	50.84	44.16

Table 1: Results on the English-GUM and English-LitBank development sets when varying the token counts of few-shot examples are reported below. In certain cases, Gemini 2.5 Flash returned empty responses, which contributed to the observed performance degradation.

```
How to Prepare Quinoa Quinoa is known as the
little rice of Peru . The Incas treated the crop
as sacred and referred to quinoa as " chisaya
mama " or " mother of all grains . " [ 1 ] By
tradition , the Inca emperor would sow the first
seeds of the season using " golden implements .
" Quinoa is rich in protein and much lighter
than other grains .
```

Figure 1: A sample raw text instance from the English-GUM training split.

3.2 Approach

As mentioned earlier, we adopted an end-to-end text-to-text approach, where the model receives an input text and is required to return the same text with coreference annotations. Initially, we experimented with zero-shot prompting; however, this strategy yielded poor results since our prompt failed to capture all the complexities necessary for effective coreference resolution. We then moved to a 4-shot prompting setup using the same instructions, which produced more reasonable results. These four-shot examples were incorporated into

```
How to Prepare Quinoa|[e1] Quinoa|[e1] is known
as the|[e1 little rice of Peru|[e2],e1] . The|
[e3 Incas|e3] treated the|[e1 crop|e1] as sacred
and referred to quinoa|[e1] as " chisaya|[e1
mama|e1] " or " mother|[e1 of all|[e4 grains|e1],
e4] . " [ 1|[e5] ] By tradition|[e6] , the|[e7
Inca emperor|e7] would sow the|[e8 first seeds
of the|[e9 season|e8],e9] using " golden|[e10
implements|e10] . " Quinoa|[e11] is rich in
protein|[e12] and much lighter than other|[e13
grains|e13] .
```

Figure 2: A gold-annotated version of the same text shown in Figure 1, taken from the English-GUM training split.

the system prompt used for all our experiments and submissions, with the complete prompt provided in Appendix A.2. These early experiments suggested that few-shot prompting, supported by well-curated examples, could be a more effective approach to this problem.

To validate our hypothesis that incorporating a large number of well-curated few-shot examples enhances model performance, we conducted a small ablation study on the development splits of the English-GUM and English-LitBank datasets from the shared task. The token count distributions for these datasets, along with others used in the task, are provided in Appendix A.1. For this experiment, we employed Gemini 2.5 Flash, as it offers a more cost-efficient alternative to Gemini 2.5 Pro, particularly when handling long context lengths. The results, shown in Table 1, indicate that increasing the number of high-quality examples generally improves performance, although occasional instances where Gemini 2.5 Flash produced empty outputs adversely impacted the overall outcomes of the ablation study.

To build on this insight, we employed a dynamic few-shot learning strategy using Google's Gemini 2.5 Pro model for coreference resolution tasks. For each test instance, the system dynamically constructs a context window by selecting language-specific training examples and their corresponding gold-standard annotations, then shuffling them randomly to avoid ordering bias. For instance, we combined multiple training datasets of the same language before selecting them as few-shot examples. The approach leverages adaptive context management, progressively adding training examples as human-AI dialogue pairs until reaching a 300,000-token limit, ensuring optimal use of the model's context window while maintaining computational efficiency. For certain datasets, such as Czech-PCEDT, English-GUM, English-LitBank, and Hungarian-KorKor, we utilized up to 500,000 tokens for few-shot examples, while for the remaining datasets, we limited the few-shot examples to 300,000 tokens in our submission. It is important to note that we did not use a fixed number of shots across all datasets.

Each test query is processed within a structured prompt framework that includes a system prompt, a few randomized few-shot examples, and the target input, enabling the model to learn task-specific patterns in-context without parameter updates. This methodology supports language-adaptive processing by automatically selecting relevant examples for the target language and provides a scalable, multilingual framework for evaluating coreference resolution across diverse linguistic settings. We set the temperature parameter to zero across all experiments to ensure deterministic outputs and suppress stochastic or creative variations in the LLM's responses.

The complete system prompt used in our approach is provided in Appendix A.2. Our prompt design did not capture all the intricacies required for effectively solving the coreference resolution task. We included only a limited set of instructions and omitted explicit guidance for handling zero mentions. Nevertheless, through few-shot prompting strategies, our system was able to implicitly learn and annotate zero mentions successfully.

4 Shared Task Results

The official results of the shared task are summarized in Table 2. Our system, NUST-FewShot, ranked second among four participating submissions, achieving an average CoNLL F1 score of 61.74, the primary evaluation metric of the task. The CoNLL F1 score is computed as the unweighted average of the F1 scores from MUC, B-cubed, and CEAFe. Given the multilingual nature of the datasets, the final score is reported as the macro-average of the individual CoNLL F1 scores across all languages.

In addition to the primary CoNLL F1 metric, three alternative evaluation metrics are reported in Table 2: partial matching, exact matching, and head matching with singletons included. Under partial matching without singletons, our system performs nearly on par with the top-ranked system. However, the performance gap becomes slightly

System	head match	partial match	exact match	head match (with singletons)
GLaRef-CRAC25	**62.96**	**61.66**	**58.98**	**65.61**
NUST-FewShot	61.74	61.14	56.34	63.44
PUXCRAC2025	60.09	59.68	55.22	54.77
UWB	59.84	59.55	38.81	62.77

Table 2: The table presents the results of all systems participating in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track). The primary evaluation metric is the CoNLL F1 score, reported in the second column labeled head-match. Our system, NUST-FewShot, achieved the second-best overall performance among the submitted systems.

more pronounced under the exact matching without singletons and head matching with singletons metrics.

Table 3 reports our system's official test results on the shared-task language-specific datasets, whereas Table 5 (A.3) reports those of the three best overall systems. Our system achieved the best performance on 10 out of the 22 datasets. Nonetheless, a notable performance gap remains between our system and the top-performing language-specific models for several datasets, particularly Catalan-AnCora, Czech-PCEDT, and Czech-PDT. We hypothesize that this discrepancy may be due to our model's limited ability to capture fine-grained linguistic nuances unique to these languages, despite the availability of a substantial number of gold-annotated examples in their training sets. Overall, while our approach demonstrates competitive results on nearly half of the datasets, further ablation studies are necessary to better understand its strengths and weaknesses and to explore strategies for improving cross-linguistic adaptability.

5 Conclusion

This paper presented the second-best performing system in the CRAC 2025 Shared Task (LLM Track). Our approach achieved top performance on 10 out of the 22 datasets in the competition. We employed an end-to-end text generation framework leveraging few-shot learning with Gemini 2.5 Pro, a state-of-the-art long-context LLM, which processes raw text as input and produces coreference-annotated text as output. Coreference resolution with LLMs remains a nascent area of research, with only a handful of recent studies addressing this challenge. We hope that our work not only demonstrates the potential of LLM-based approaches for this task but also paves the way for future research exploring this promising direction.

Dataset	CoNLL score
ca anc	60.87
cs pce	51.36
cs pdt	54.30
cu pro	58.48
de pot	48.74
en gum	69.78
en lit	70.38
es anc	61.75
fr anc	71.94
fr dem	57.59
grc pro	57.85
hbo ptn	80.15
hi hdt	71.32
hu kor	43.49
hu sze	52.27
ko ecm	66.05
lt lcc	59.16
no bok	72.76
no nyn	68.86
pl pcc	70.83
ru ruc	71.40
tr itc	39.00

Table 3: The table shows CoNLL scores for our system across the 22 test datasets from the CRAC 2025 Shared Task.

Limitations

The primary limitations of our work are the reliance on multiple LLM calls and the associated computational cost, which can become significant. Our approach utilizes Gemini 2.5 Pro with a large context window via an API interface, leading to high costs due to the extensive token usage from few-shot examples, lengthy input texts, and generated outputs with predictions. Furthermore, we did not investigate other advanced long-context models, such as OpenAI's GPT-4.1 and Meta's LLaMA 4 Scout, which support context lengths of up to 1 million and 10 million tokens, respectively.

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *Preprint*, arXiv:2305.14489.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Natalia Skachkova. 2024. Multilingual coreference resolution as text generation. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 114–122, Miami. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk,

Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Appendix

A.1 Data Distribution

Table 4 presents the token count distribution across the various datasets included in the shared task. The token counts are computed specifically using the Gemini 2.5 Pro tokenizer[1], as our submission was based on this model. Token counts may vary slightly when calculated with alternative tokenizers due to differences in their tokenization strategies.

A.2 Prompt

We used the following system prompt in all our experiments and results:

"""

You are a coreference resolution annotator.

Your job is to read a multilingual passage and annotate all mentions that refer to the same underlying entity (could be a word or many words) using a unique identifier with a bracketed pipe-based format. Understand proper context of the text before making the annotations. Prioritize resolving pronouns based on proximity and grammatical role, but consider the semantic context to avoid incorrect annotations.

Format Rules:

1. Surround every span referring to a shared entity with the format: ‘mention text|[eID]‘ where eID is a unique entity ID (e1, e2, ...).

2. If an entity contain multiple words, start the annotation with a single pipe as ‘|[eID‘ and close it with a single pipe as ‘|eID]‘ for example ‘|[eID - entity comes here|eID]‘

3. If a span refers to **multiple entities**, use: ‘text|[eA,[eB] tail|eA],eB]‘ (notice)

4. Use the **same ID** consistently for all mentions of the same entity, even across paragraphs.

5. Do **not annotate singleton mentions** (those that appear only once in the text).

6. Annotate **all types of coreference**: full noun phrases, pronouns, nested noun mentions, and even abstract or generic references like "such outcomes", "it", etc.

7. If there is **nested structure**, use proper nesting with comma-separated closing IDs.

8. Do not resolve ’it’ if it refers to an implied or abstract concept (e.g., ’It is widely believed...’).

Example 1:

Original:
Alice went to the park . She brought her dog .

Annotated:
Alice|[e1] went to the park|[e2] . She|[e1] brought her|[e1] dog|[e3] to the park|[e2] .

Example 2:

Original:
Education and early loves Alina gained her early formal education at Aberdeen Grammar School , and in August 1799 entered the school of Dr. William Glennie , in Dulwich . [17] Placed under the care of a Dr. Bailey , she was encouraged to exercise in moderation but not restrain herself from " violent " bouts in an attempt to overcompensate for her deformed foot .

Annotated:
Education|[e1] and early|[e2 loves|e2] Alina|[e3] gained her|[e1,[e3] early formal education at Aberdeen|[e4],[e5 Grammar School|e1],e5] , and in August|[e6 1799|[e7],e6] entered the|[e8 school of Dr.|[e9 William Glennie|e9] , in Dulwich|[e10],e8] . [17|[e11]] Placed under the|[e12 care of a|[e13 Dr. Bailey|e12],e13] , she|[e3] was encouraged to

[1] https://ai.google.dev/gemini-api/docs/tokens

Language	Train	Val	Test	Train (including gold annotations)	Val (including gold annotations)
Catalan	483,179	36,371	36,363	795,896	60,446
Czech	2,947,128	92,391	91,706	4,785,013	156,544
Old Church Slavonic	150,355	24,262	18,980	272,045	41,028
German	38,405	4,841	4,623	68,321	8,613
English	375,074	49,229	49,547	864,829	113,484
Spanish	448,164	30,796	30,909	806,930	56,228
French	728,738	61,596	61,100	1,972,972	159,321
Ancient Greek	170,245	10,787	13,288	311,317	18,477
Ancient Hebrew	35,346	48,014	46,851	46,473	63,500
Hindi	48,633	12,482	28,728	88,993	21,456
Hungarian	234,845	28,304	27,543	333,437	40,497
Korean	957,191	61,049	60,973	1,557,657	99,991
Lithuanian	63,638	7,353	7,500	82,708	9,231
Norwegian	595,703	61,901	59,054	1,399,765	143,566
Polish	747,247	43,515	43,105	1,739,784	102,615
Russian	194,110	33,118	18,434	262,653	45,792
Turkish	81,379	8,757	9,523	196,850	20,726
Total	8,299,380	614,766	608,227	15,585,643	1,161,515

Table 4: The token counts for all languages in the shared task, after merging datasets for languages with multiple sources, are reported on a split-wise basis. The last two columns additionally account for tokens from the gold annotations.

exercise in moderation but not restrain herself|[e3] from "|[e14 violent " bouts|e14] in an|[e15 attempt to overcompensate for her|[e16,[e3] deformed foot|e15],e16] .

Example 3:

Original:
Los jugadores de el Espanyol aseguraron hoy que prefieren enfrentar se a el Barcelona en la final de la Copa de el Rey en lugar de en las semifinales , tras clasificar se ayer ambos equipos catalanes para esta ronda . La mayoría de los jugadores españolistas expresaron su opinión de que sería más fácil vencer a su máximo rival en un solo partido que tener que enfrentar se a el conjunto de Louis Van Gaal en las semifinales , donde tendrían que disputar una eliminatoria de ida y vuelta .

Annotated:
Los|[e1 jugadores de el Espanyol|[e2],e1] aseguraron hoy que prefieren ##|[e1] enfrentar se a el Barcelona|[e3] en la|[e4 final de la|[e5 Copa de el Rey|e4],e5] en lugar de en las|[e6 semifinales|e6] , tras clasificar se ayer ambos|[e7 equipos catalanes|e7] para esta|[e6 ronda|e6] . La|[e1 mayoría de los jugadores españolistas|e1] expresaron su|[e1] opinión de que sería más fácil vencer a su|[e2],[e3 máximo rival|e3] en un solo partido que tener que enfrentar se a el conjunto|[e3 de Louis Van Gaal|e3] en las|[e6 semifinales , donde|[e6] tendrían ##|[e7] que disputar una|[e8 eliminatoria de ida y vuelta|e6],e8] .

Example 4:

Original:
Zèbre Zèbre est un nom vernaculaire , ambigu en français , pouvant désigner plusieurs espèces différentes d’ herbivores de la famille de les équidés , et de le genre Equus , vivant en Afrique . Ils se trouvent principalement en Afrique centrale et australe . Ces animaux se caractérisent par des bandes de rayures verticales noires et blanches .

Annotated:
Zèbre|[e1] Zèbre|[e1] est un|[e2 nom vernaculaire , ambigu|e2] en français|[e3] , pouvant désigner plusieurs|[e4 espèces différentes d’ herbivores|[e5] de la|[e6 famille de les|[e7 équidés|e6],e7] , et de le|[e8 genre Equus|e8] , vivant en Afrique|[e9],e4] . Ils|[e4] se trouvent principalement en Afrique|[e10 centrale et australe|e10] .
"""

A.3 Top Systems Results

Table 5 presents the CoNLL scores for the top three overall best-performing systems across the 22 test datasets from the CRAC 2025 Shared Task.

System	ca anc	cs pce	cs pdt	cu pro	de pot	en gum	en lit	es anc	fr anc
GLaRef-CRAC25	**73.45**	**65.12**	**71.33**	58.25	**59.60**	58.73	69.01	**74.43**	66.74
NUST-FewShot	60.87	51.36	54.30	**58.48**	48.74	**69.78**	**70.38**	61.75	**71.94**
PUXCRAC2025	68.01	56.94	62.96	43.74	57.41	61.71	69.12	70.52	63.77

System	fr dem	grc pro	hbo ptn	hi hdt	hu kor	hu sze	ko ecm	lt lcc	no bok
GLaRef-CRAC25	60.43	**65.75**	43.96	56.36	**52.53**	59.82	63.04	62.55	64.74
NUST-FewShot	57.59	57.85	**80.15**	**71.32**	43.49	52.27	**66.05**	59.16	**72.76**
PUXCRAC2025	**61.54**	47.86	45.31	66.85	50.58	**61.61**	50.32	**65.35**	65.18

System	no nyn	pl pcc	ru ruc	tr itc
GLaRef-CRAC25	61.63	**72.55**	68.79	**56.23**
NUST-FewShot	**68.86**	70.83	**71.40**	39.00
PUXCRAC2025	63.00	66.55	67.59	56.06

Table 5: The table shows the CoNLL scores for the top three overall best-performing systems across the 22 test datasets from the CRAC 2025 Shared Task. Our system, NUST-FewShot, achieved the best performance on 10 of the 22 datasets, surpassing the overall top-ranked system, GLaRef-CRAC25, which led on 9 of the 22 datasets.

Author Index