CorPipe at CRAC 2026: Empty Nodes and Cross-Lingual Transfer in 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 introduce CorPipe 26, our winning submission to the CRAC 2026 Shared Task on Multilingual Coreference Resolution. The fifth edition of this shared task focuses mainly on the comparison of generative LLMs and specialized systems; additionally, 5 more datasets and 2 new languages are introduced. CorPipe 26 is an improved version of CorPipe 25, with a new variant predicting empty nodes together with mentions and coreference links in a single model. Our system outperforms all other submissions in the LLM track by 2.8 percent points and all submissions in the unconstrained track by 9.5 percent points. Furthermore, we perform a series of ablation experiments with different model sizes, empty node prediction methods, and cross-lingual zero-shot evaluation. The source code and the trained models are publicly available at https://github.com/ufal/crac2026-corpipe.

1 Introduction

Coreference resolution aims to identify and group together expressions that refer to the same real-world entity within a text. The CRAC 2026 Shared Task on Multilingual Coreference Resolution (Novák et al., 2026a) is the fifth edition of a shared task designed to advance multilingual coreference research. Built on the CorefUD 1.4 dataset collection, this year's iteration focuses primarily on the comparison of generative large language models (LLMs) and specialized systems. It also expands the task's scope by adding five new datasets and two new languages (Dutch and Latin), bringing the total to 27 datasets across 19 languages.

As in the previous two years, participating systems must predict *empty nodes*. These are omitted words that do not appear in the surface text but are necessary for accurate coreference modeling. Resolving empty nodes is especially important for pro-drop languages, such as those in the Slavic and Romance families. In these languages, pronouns are often omitted when they can be inferred from context, such as through verb morphology. This is shown in the Czech example *Řekl, že nepřijde*, which translates to *(He) said that (he) won't come*.

Our entry for the CRAC 2026 Shared Task, CorPipe 26, is an improvement of our past winning systems (Straka, 2025, 2024, 2023; Straka and Straková, 2022). Our submission is a two-step pipeline: the empty nodes are predicted first and then mention detection and coreference linking are performed together by a single model. The empty node prediction system, which has been improved compared to last year by predicting all available information about the empty nodes, not just that required for coreference evaluation, has been made available to all participants as a baseline. All our models are strictly multilingual (without indicating the language or dataset on the input) and are trained on all provided datasets simultaneously.

Our main contributions are as follows:

- We present the top-performing system for the CRAC 2026 Shared Task, outperforming all other submissions in the LLM track by 2.8 percent points and all submissions in the unconstrained track by 9.5 percent points.
- We investigate a one-stage variant of CorPipe 26 predicting empty nodes together with mentions and coreference links in a single model, surpassing the two-stage variant slightly on the development set.
- We evaluate the impact of model size, empty node methods, and cross-lingual zero-shot settings through ablations.
- We make the CorPipe 26 source code publicly available under an open-source license at https://github.com/ufal/crac2026-corpipe. Furthermore, we release several pretrained multilingual models of varying sizes under the CC BY-NC-SA license.

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 205–216
July 3, 2026

2 Related Work

Coreference Resolution Most neural coreference models have relied on span-based methods since the work of Lee et al. (2017), who introduced a model that identifies mentions and links them simultaneously. Lee et al. (2018) later updated this approach to make it more efficient and accurate. Joshi et al. (2020) further improved results by using SpanBERT (Joshi et al., 2019), a model designed specifically to represent text spans.

Other researchers have explored different designs to avoid the constraints of span-based methods. Wu et al. (2020) treated coreference as a question-answering task, Liu et al. (2022) developed an autoregressive system, and Bohnet et al. (2023) used a text-to-text approach. A common drawback of these methods, however, is the computational overhead requiring multiple model calls to process a single sentence.

Word-Level Coreference Resolution A major departure from span-based methods came when Dobrovolskii (2021) introduced word-level coreference. Instead of using whole phrases, this method focuses on the head word of a mention. D'Oosterlinck et al. (2023) built on this with CAW-coref to better handle complex dependencies. More recently, Liu et al. (2024) introduced MSCAW-coref, which supports multiple languages and handles mentions that appear only once. This word-level approach is now part of Stanza (Qi et al., 2020), a popular Python natural language processing framework.

Multilingual Coreference Resolution The CRAC shared tasks (Žabokrtský et al., 2022, 2023; Novák et al., 2024, 2025, 2026a) have played a central role in multilingual coreference resolution. They provide a standard evaluation framework, the CorefUD dataset (Novák et al., 2026b), and a multilingual baseline (Pražák et al., 2021).

Earlier versions of our CorPipe system have taken part in every CRAC shared task. The system has evolved from early multilingual models (Straka and Straková, 2022) to versions that handle wider context (Straka, 2023), detect zero mentions in raw text (Straka, 2024), and a PyTorch version capable of training larger models (Straka, 2025).

3 Architecture

CorPipe 26 is based heavily on CorPipe 25. As in the previous version, given input raw text it starts

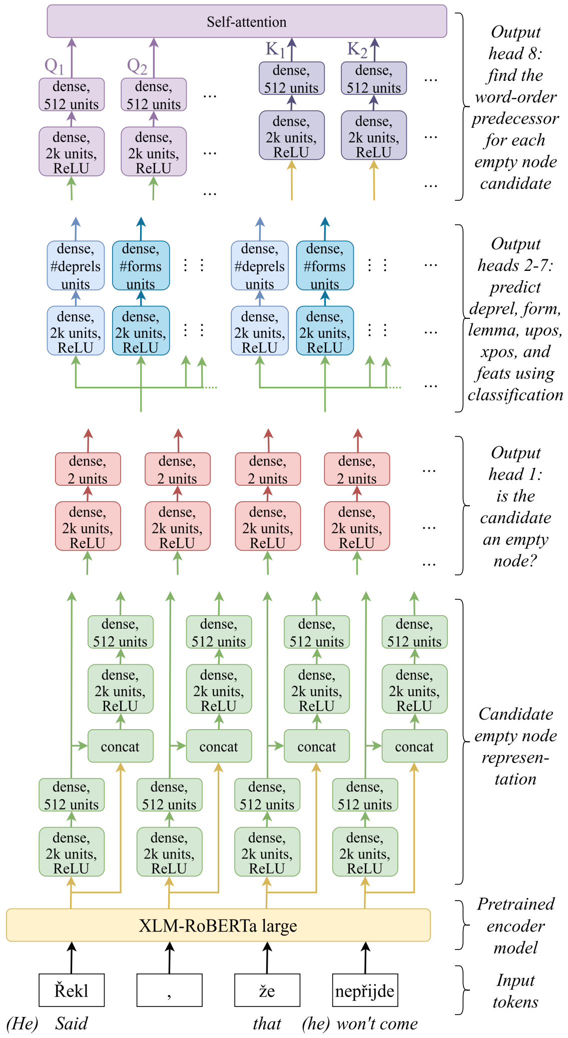

Figure 1: The system architecture of the empty node prediction baseline. Every ReLU activation is followed by a dropout layer with a dropout rate of 50%.

by predicting the empty nodes using one model and then performs mention detection and coreference linking together using another model.

Empty Nodes Baseline The empty nodes are predicted using our baseline system that was provided to all shared task participants. It is a substantial improvement over the baseline provided in previous years, which produced only the information necessary for coreference resolution evaluation (word order position, dependency parent, and dependency relation). We have extended the system to predict all available empty node information, additionally providing the form, lemma, UPOS, XPOS, and FEATS columns. The model operates non-autoregressively, predicting up to two empty nodes per input word, with each input word serving as a potential dependency head.

Figure 1 presents the architecture of this system.

Treebank	ARC	DEP	WO	DEP_WO	FORM	LEMMA	UPOS	XPOS	FEATS	ALL
ca	95.55	95.55	92.74	92.74	—	—	—	—	—	92.74
cs_pcedt	70.91	69.36	70.77	69.21	69.07	70.77	70.91	70.91	68.08	67.80
cs_pdt	79.21	78.34	78.52	78.00	77.83	78.00	79.03	79.21	77.14	76.79
cs_pdtsc	85.88	85.11	85.16	84.39	84.59	85.73	85.88	85.88	82.79	81.91
cu	80.55	79.50	79.77	78.85	—	—	—	—	—	78.85
es	95.74	95.74	93.48	93.48	—	—	—	—	—	93.48
grc	89.85	87.90	89.85	87.90	—	—	—	—	—	87.90
hu_korkor	85.44	79.61	83.50	77.67	85.44	—	—	—	—	77.67
hu_szegedkoref	92.48	89.86	91.82	89.20	—	—	—	—	—	89.20
pl	90.99	90.88	90.88	90.76	—	—	—	—	—	90.65
tr	84.82	84.82	84.72	84.72	82.26	—	—	—	—	82.17

Table 1: Empty nodes prediction baseline performance on the minidev sets of the CRAC 2026 dataset containing empty nodes. Each reported metric is an F1 score where a prediction is considered correct if both the dependency head and the given column are correct, with ARC denoting the dependency head, DEP the dependency relation, WO the word order position, and ALL the combination of all other predictions.

The input words of a single sentence are tokenized and processed through an XLM-RoBERTa-large encoder (Conneau et al., 2020), with each input word represented by its first subword embedding. For each word, we generate two empty node candidates: the first through a dense-ReLU-dropout-dense module (768→ $2k$ →512 units), and the second by concatenating the first candidate with the input word representation and applying an analogous transformation. The candidates are processed by eight heads, each first passing its input through its own 2k-unit ReLU layer and dropout: (1) a binary classification head determining whether the candidate will produce an empty node, (2-7) classification heads predicting the dependency relation, form, lemma, UPOS, XPOS, and FEATS, and (8) a self-attention word-order prediction head identifying the insertion point for the empty node. Please refer to the source code for more details.

We train a single multilingual model on a concatenation of all corpora containing empty nodes, with sentences sampled proportionally to the square root of their respective corpus sizes. We use the Adam optimizer (Kingma and Ba, 2015) and train for 20 epochs of 5 000 batches, each consisting of 64 sentences. The learning rate linearly increases to 1e-5 in the first epoch and then decays to zero following cosine decay (Loshchilov and Hutter, 2017). The intrinsic performance of the system is summarized in Table 1.

The source code of the system is available at `https://github.com/ufal/crac2026_empty_nodes_baseline` under the open-source MPL license. The trained model is available both on Hugging Face and via LINDAT/CLARIAH-CZ. Finally, the minidev and minitest sets of the CRAC 2026 Shared Task with predicted empty nodes are available to all participants.

As a part of the shared task submission, we also trained an improved version of the empty node prediction system by using a larger batch size of 384 sentences (the maximum batch size allowed by a single H100 GPU), compared to the 96 sentences used by the baseline. The intrinsic evaluation of this improved system is presented in Table 9 in the Appendix. Overall, the improved system provides a boost of 2.2 percent points in the DEP metric and 2.5 percent points in the ALL metric.

Coreference Resolution With the empty nodes predicted, we perform coreference resolution by predicting mentions and coreference links together using a single model. The architecture of this model is the same as the one used in CorPipe 25 (Straka, 2025) and it is presented in Figure 2 and summarized below; see the referenced work for a detailed description.

The model processes documents one sentence at a time. To include as much context as possible, we expand every input sentence with preceding tokens and at most 50 following tokens to the extent allowed by the maximum segment length (512 or 2 560 tokens). The input tokens are passed through a pretrained multilingual encoder, and coreference mentions are predicted using an enhanced BIO encoding scheme capable of representing potentially overlapping spans. The identified mentions are then represented as a concatenation of their first and last tokens, and coreference links are predicted through a self-attention layer determining the most likely antecedent for each mention, allowing self-reference to indicate the first mention of an entity.

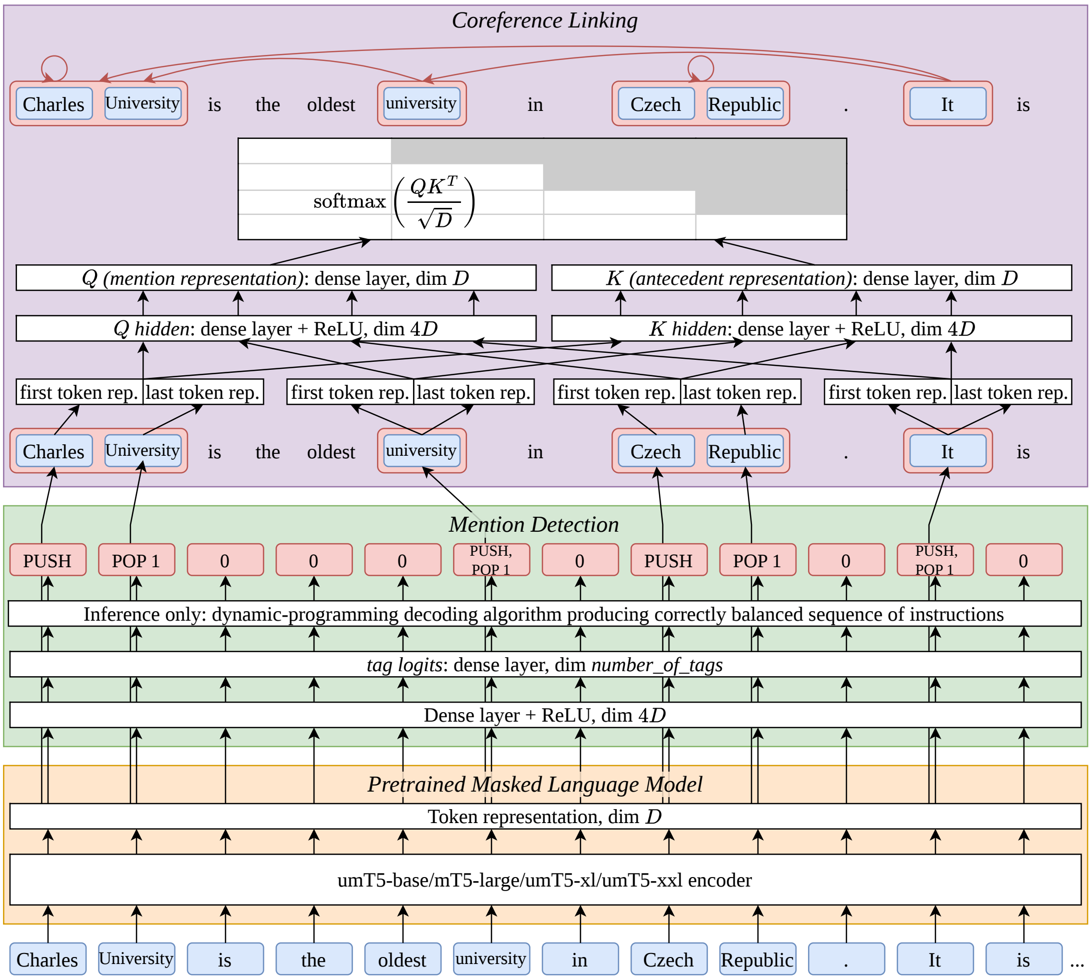

Figure 2: The CorPipe 26 model architecture.

Model	Params	Batch Size	Learning Rate	Train Time
umT5 base	269M	8	6e-4	5h
mT5 large	538M	8	6e-4	10h
umT5 xl	1605M	6	5e-4	22h
umT5 xxl	5417M	6	5e-4	36h

Table 2: Properties of 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.

We employ different segment sizes during training (512 tokens) and inference (2 560 tokens, except for two PROIEL corpora using 512 tokens) to improve modeling of long-range coreference links; this difference is allowed by the relative positional encodings used by our pretrained encoders.

We consider four sizes of the pretrained encoders: umT5-base (Chung et al., 2023), mT5-large (Xue et al., 2021), umT5-xl, and umT5-xxl. For each encoder, we train 10 models differing only in random initialization. Each model is trained for 15 epochs each consisting of 10k batches using the AdaFactor optimizer (Shazeer and Stern, 2018). The learning rate first linearly increases during the initial 10% of training and then decays to zero following cosine decay (Loshchilov and Hutter, 2017). The parameter counts of the models, as well as the learning rate, batch size, and training time, are summarized in Table 2.

One-Stage Variant In addition to the described CorPipe 26, we also evaluate a one-stage variant predicting the empty nodes together with mentions and coreference links in a single model. The architecture is inspired by the single-stage CorPipe 24 (Straka, 2024), with a few modifications improving the performance considerably.

Unlike the empty node baseline, the one-stage variant does not predict every empty node or all of its details. It only predicts empty nodes that are actually coreference mentions, providing for

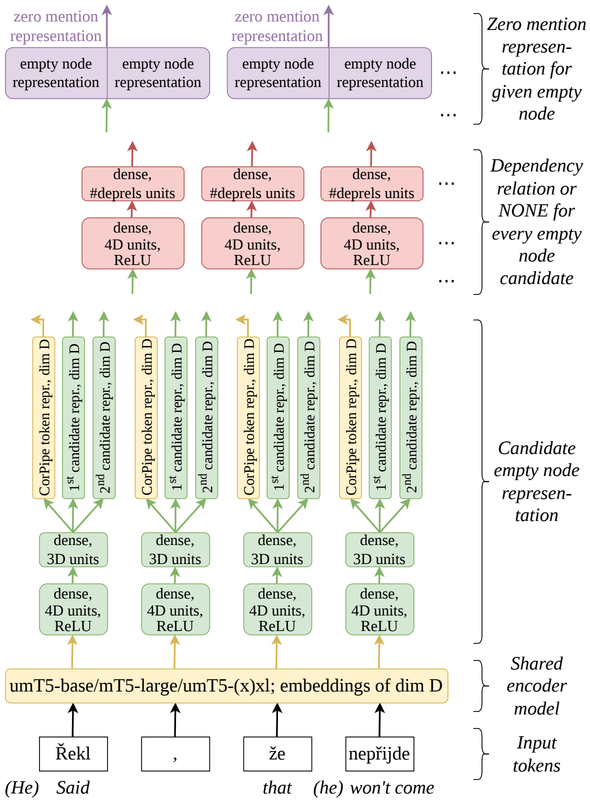

Figure 3: The modifications of the one-stage CorPipe 26 variant to the standard CorPipe 26 architecture.

System	Head-match	Partial-match	Exact-match	With Singletons
UNCONSTRAINED				
CorPipeEnsemble	**77.11 1**	**76.30 1**	**74.07 1**	**79.11 1**
CorPipeXXL	76.18 2	75.32 2	72.90 2	78.15 2
CorPipeLarge	72.32 3	71.16 3	68.79 3	74.52 3
DAggerCoref	67.56 4	67.56 4	37.63 7	58.75 5
Stanza	67.00 5	65.87 5	63.32 4	68.53 4
BASELINE-GZ	55.39 6	55.06 6	53.91 5	48.28 6
BASELINE	54.54 7	54.16 7	52.96 6	47.45 7
AUKBC-MULCRF	35.24 8	35.21 8	20.65 8	33.10 8
LLM				
LLM-LatticeNLP	**74.32 1**	**74.32 1**	41.83 2	**76.09 1**
LLM-UWB	73.83 2	73.83 2	40.76 3	75.59 2
LLM-PortNLP	68.69 3	67.51 3	**65.30 1**	70.98 3
LLM-Landcore	46.19 4	44.79 4	40.47 4	47.78 4

Table 3: Official results of CRAC 2026 Shared Task on the minitest set with various metrics in %.

each just the dependency head and relation, the minimum needed for coreference evaluation. Additionally, each predicted empty node is treated as a full mention. While this simplification may reduce the exact-match score, it does not affect the head-match score metric.

The additions of the one-stage variant to the standard CorPipe 26 architecture are presented in Figure 3. The encoder embeddings of size D are first passed through a $4D$-unit ReLU layer and dropout followed by a $3D$ dense layer that is split into three D-dimensional embeddings for (a) the token representation used by the mention detection and coreference linking heads, and (b) two empty node candidates. All empty node candidates are processed by a shared $4D$-unit ReLU layer, dropout, and a classification layer predicting either `NONE`, indicating the candidate is not an empty node, or the dependency relation of the empty node, whose head is then the input token. The zero mentions are constructed by repeating the candidate embeddings twice; they are added to the mentions predicted from the input tokens, and coreference linking is performed as in the standard CorPipe 26.

We train the variant using the same encoders and hyperparameters as the standard CorPipe 26.

4 Shared Task Results

We submitted the maximum number of three systems to the CRAC 2026 Shared Task, all based on the standard two-stage CorPipe 26 architecture (unfortunately, we did not finish training the one-stage models before the submission deadline):

- **CorPipeLarge**, a single best-performing model according to the minidev set performance based on the mT5-large encoder;
- **CorPipeXXL**, using the best model with the umT5-xxl encoder and improved empty node prediction system;
- **CorPipeEnsemble**, an ensemble of the best 7 out of 10 umT5-xxl models, again utilizing the improved empty node prediction system.

The official results of the CRAC 2026 Shared Task are presented in Table 3 showing four minitest metrics of all the submitted systems, and in Table 4 displaying the minitest CoNLL scores of all individual treebanks. CorPipeXXL and CorPipeEnsemble outperform all other submissions, both in the LLM track and the unconstrained track, by large margins of 2.8 and 9.5 percent points for the best CorPipeEnsemble model, respectively.

System	Avg	ca	cs pced	cs pdt	cs pdts	cu	de	en fant	en gum	en litb	es	fr anco	fr demo	fr litb	grc	hbo	hi	hu kork	hu szeg	ko	la	lt	nl	no bokm	no nyno	pl	ru	tr
UNCONSTRAINED																												
CorPipeEnsemble	**77.1** **1**	**84.4** **1**	**79.3** **1**	**81.8** **1**	**76.7** **1**	**68.8** **1**	**75.0** **1**	**81.1** **1**	**77.4** **1**	**85.3** **1**	**85.3** **1**	**76.5** **1**	73.4 2	**82.5** **1**	**79.0** **1**	**74.5** **1**	**78.4** **1**	**68.7** **1**	**72.6** **1**	**70.3** **1**	**62.6** **1**	**76.1** **1**	**74.7** **1**	**78.9** **1**	**76.8** **1**	**82.1** **1**	**86.2** **1**	73.5 2
CorPipeXXL	76.2 2	83.0 2	78.8 2	81.6 2	**76.7** **1**	67.9 2	74.8 2	80.8 2	76.7 2	83.7 2	84.2 2	75.5 2	**73.8** **1**	81.5 2	77.9 2	72.0 2	77.8 2	68.2 2	71.4 2	69.7 2	58.7 2	75.2 3	73.1 2	77.4 2	76.4 2	81.5 2	84.5 2	**74.0** **1**
CorPipeLarge	72.3 3	79.7 3	73.8 4	77.9 3	72.3 2	56.3 4	69.9 5	74.9 3	73.5 3	79.4 3	82.3 3	71.5 3	71.0 3	76.5 3	69.0 3	66.4 3	76.3 3	64.6 3	67.7 3	68.6 3	57.5 3	75.7 2	69.9 3	74.4 3	74.0 3	78.0 3	82.2 3	69.5 3
DAggerCoref	67.6 4	76.8 5	70.8 5	74.6 5	68.2 5	57.2 3	71.9 3	73.0 4	68.6 5	73.6 4	77.4 5	70.2 4	69.0 4	66.9 4	65.2 4	57.9 5	72.8 5	60.5 4	61.0 5	61.1 6	56.0 4	66.5 5	64.5 4	72.9 4	72.0 4	73.7 4	79.1 5	42.9 7
Stanza	67.0 5	77.7 4	74.0 3	76.4 4	71.2 3	38.9 5	70.4 4	69.9 5	72.7 4	73.3 5	80.1 4	69.5 5	57.1 5	64.5 5	53.9 5	61.2 4	75.5 4	59.9 5	66.9 4	67.1 4	36.9 5	73.0 4	60.0 5	72.8 5	70.8 5	73.7 4	80.4 4	60.9 4
BASELINE-GZ†	55.4 6	68.8 6	68.8 6	67.9 6	69.3 4	26.9 7	52.4 6	65.1 6	62.0 6	66.2 6	70.7 6	61.8 6	55.6 6	46.1 6	31.1 7	31.7 7	66.6 6	44.6 6	55.0 6	65.0 5	6.8 7	62.4 6	40.6 6	61.4 6	61.1 6	68.9 5	68.2 6	50.8 5
BASELINE†	54.5 7	67.9 7	63.4 7	66.2 7	66.1 6	24.7 8	52.4 6	65.1 6	61.9 7	66.2 6	70.3 7	61.8 6	55.6 6	46.1 6	30.6 8	31.7 7	66.6 6	42.2 7	54.3 7	65.0 5	6.8 7	62.4 6	40.6 6	61.4 6	61.1 6	67.5 6	68.2 6	46.8 6
AUKBC-MULCRF	35.2 8	38.8 8	25.8 8	36.8 8	37.9 7	34.8 6	29.8 7	41.4 7	44.2 8	34.4 7	37.4 8	37.4 7	35.5 7	28.6 7	43.4 6	48.2 6	46.2 7	29.5 8	31.6 8	21.6 7	16.2 6	28.0 7	34.5 7	42.9 7	39.6 7	36.2 7	33.7 7	37.2 8
LLM																												
LLM-LatticeNLP	**74.3** **1**	77.0 2	72.8 2	76.3 2	70.9 2	**60.1** **1**	71.3 2	78.8 2	**76.9** **1**	**85.4** **1**	78.3 2	77.3 2	**74.5** **1**	**80.2** **1**	74.9 2	**79.8** **1**	76.8 2	**65.5** **1**	66.6 2	68.8 2	**58.7** **1**	65.3 3	**77.4** **1**	**81.2** **1**	77.6 2	78.3 2	82.6 2	**73.1** **1**
LLM-UWB	73.8 2	**82.7** **1**	**75.8** **1**	**80.0** **1**	**73.8** **1**	59.6 2	**73.3** **1**	**80.2** **1**	75.9 2	84.6 2	**83.1** **1**	**77.5** **1**	66.7 2	60.5 2	**75.3** **1**	76.8 2	**77.0** **1**	65.3 2	**69.0** **1**	66.9 3	56.2 2	**73.5** **1**	66.1 3	80.4 2	**79.5** **1**	**80.1** **1**	**84.7** **1**	69.1 2
LLM-PortNLP	68.7 3	73.7 3	71.5 3	74.1 3	69.8 3	57.9 3	67.9 3	74.8 3	70.6 3	78.4 3	75.4 3	70.5 3	53.4 3	54.9 3	71.1 3	72.7 3	75.6 3	59.1 3	62.5 3	**69.5** **1**	44.8 3	68.1 2	72.9 2	72.1 3	72.1 3	76.8 3	81.2 3	63.0 3
LLM-Landcore	46.2 4	41.3 4	36.7 4	40.3 4	40.9 4	29.4 4	49.3 4	58.1 4	55.3 4	63.0 4	47.3 4	46.4 4	20.7 4	46.7 4	43.4 4	61.4 4	60.6 4	41.6 4	37.3 4	59.5 4	32.8 4	52.8 4	39.3 4	54.5 4	53.5 4	43.2 4	52.3 4	39.8 4

Table 4: Official results of CRAC 2026 Shared Task on the minitest set (CoNLL score in %). The systems † are described in Pražák et al. (2021); the rest in Novák et al. (2026a).

System	Avg	ca	cs pced	cs pdt	cs pdts	cu	de	en fant	en gum	en litb	es	fr anco	fr demo	fr litb	grc	hbo	hi	hu kork	hu szeg	ko	la	lt	nl	no bokm	no nyno	pl	ru	tr
A) CORPIPE SINGLE MODELS																												
Single mT5-large model	72.32	79.7	73.8	77.9	72.3	56.3	69.9	74.9	73.5	79.4	82.3	71.5	71.0	76.5	69.0	66.4	76.3	64.6	67.7	68.6	57.5	75.7	69.9	74.4	74.0	78.0	82.2	69.5
Single umT5-base model	−3.13 69.19	−3.1 76.6	−0.6 73.2	−2.7 75.2	−2.1 70.2	−2.6 53.7	−3.4 66.5	−2.9 72.0	−1.5 72.0	−5.1 74.3	−3.0 79.3	−1.6 69.9	−2.0 69.0	−3.1 73.4	−5.2 63.8	−6.9 59.5	−3.4 72.9	−4.3 60.3	−2.3 65.4	−1.5 67.1	−8.1 49.4	−2.3 73.4	−3.1 66.8	−1.1 73.3	−1.7 72.3	−3.0 75.0	−4.3 77.9	−4.0 65.5
Single umT5-xl model	+2.48 74.80	+2.7 82.4	+2.5 76.3	+2.3 80.2	+2.5 74.8	+5.3 61.6	+2.1 72.0	+4.3 79.2	+2.7 76.2	+3.4 82.8	+1.8 84.1	+2.4 73.9	+1.8 72.8	+4.3 80.8	+4.8 73.8	+1.1 67.5	+1.1 77.4	+3.4 68.0	+2.5 70.2	+0.6 69.2	+0.0 57.5	**+1.4** **77.1**	+2.3 72.2	**+3.8** **78.2**	−0.5 73.5	+2.0 80.0	**+3.7** **85.9**	+2.5 72.0
Single umT5-xxl model	**+3.63** **75.95**	**+3.3** **83.0**	**+3.5** **77.3**	**+3.7** **81.6**	**+3.7** **76.0**	**+10.2** **66.5**	**+4.9** **74.8**	**+5.9** **80.8**	**+3.2** **76.7**	**+4.3** **83.7**	**+1.8** **84.1**	**+4.0** **75.5**	**+2.8** **73.8**	**+5.0** **81.5**	**+7.0** **76.0**	**+5.6** **72.0**	**+1.5** **77.8**	**+4.1** **68.7**	**+3.3** **71.0**	**+1.1** **69.7**	**+1.2** **58.7**	−0.5 75.2	**+3.2** **73.1**	+3.0 77.4	**+2.4** **76.4**	**+3.4** **81.4**	+2.3 84.5	**+3.9** **73.4**
B) CORPIPE SINGLE UMT5-XXL MODEL WITH IMPROVED EMPTY NODE PREDICTION SYSTEM																												
Baseline enode pred.	75.95	83.0	77.3	81.6	76.0	66.5	74.8	80.8	76.7	83.7	84.1	75.5	73.8	81.5	76.0	72.0	77.8	**68.7**	71.0	69.7	58.7	75.2	73.1	77.4	76.4	81.4	84.5	73.4
Improved enode pred.	**+0.23** **76.18**	**+0.0** **83.0**	**+1.5** **78.8**	+0.0 81.6	**+0.7** **76.7**	**+1.4** **67.9**	+0.0 74.8	+0.0 80.8	+0.0 76.7	+0.0 83.7	**+0.1** **84.2**	+0.0 75.5	+0.0 73.8	+0.0 81.5	**+1.9** **77.9**	+0.0 72.0	+0.0 77.8	−0.5 68.2	**+0.4** **71.4**	+0.0 69.7	+0.0 58.7	+0.0 75.2	+0.0 73.1	+0.0 77.4	+0.0 76.4	**+0.1** **81.5**	+0.0 84.5	**+0.6** **74.0**
C) CORPIPE ENSEMBLE MODELS																												
7 mT5-large models	73.44	80.8	75.1	78.5	73.1	59.1	70.8	76.2	74.1	80.3	83.5	72.7	73.3	77.8	71.7	68.6	76.9	66.6	69.2	69.8	56.5	**76.2**	71.4	75.0	74.7	79.2	82.6	68.9
7 umT5-base models	−3.38 70.06	−3.1 77.7	−1.8 73.3	−3.2 75.3	−2.1 71.0	−5.1 54.0	−4.8 66.0	−2.6 73.6	−1.8 72.3	−5.1 75.2	−3.2 80.3	−2.1 70.6	−2.8 70.5	−4.6 73.2	−7.4 64.3	−7.3 61.3	−3.2 73.7	−5.4 61.2	−4.0 65.2	−1.3 68.5	−4.2 52.3	−0.7 75.5	−2.9 68.5	−1.2 73.8	−1.8 72.9	−3.8 75.4	−1.9 80.7	−3.4 65.5
7 umT5-xl models	+1.99 75.43	+2.3 83.1	+2.2 77.3	+1.9 80.4	+2.4 75.5	+4.0 63.1	+3.4 74.2	+4.0 80.2	+2.7 76.8	+2.9 83.2	**+1.8** **85.3**	+2.4 75.1	−1.2 72.1	+3.2 81.0	+2.8 74.5	+1.8 70.4	+1.1 78.0	+0.6 67.2	+2.1 71.3	+0.2 70.0	−1.0 55.5	+0.0 76.2	+2.5 73.9	+3.0 78.0	+1.8 76.5	+1.3 80.5	+2.3 84.9	+3.4 72.3
7 umT5-xxl models	**+3.44** **76.88**	**+3.4** **84.2**	**+2.7** **77.8**	**+3.3** **81.8**	**+3.0** **76.1**	**+9.1** **68.2**	**+4.2** **75.0**	**+4.9** **81.1**	**+3.3** **77.4**	**+5.0** **85.3**	+1.8 85.3	**+3.8** **76.5**	**+0.1** **73.4**	**+4.7** **82.5**	**+5.4** **77.1**	**+5.9** **74.5**	**+1.5** **78.4**	**+1.8** **68.4**	**+3.4** **72.6**	**+0.5** **70.3**	**+6.1** **62.6**	−0.1 76.1	**+3.3** **74.7**	**+3.9** **78.9**	**+2.1** **76.8**	**+2.5** **81.7**	**+3.6** **86.2**	**+3.9** **72.8**
D) CORPIPE UMT5-XXL 7-MODEL ENSEMBLE WITH IMPROVED EMPTY NODE PREDICTION SYSTEM																												
Baseline enode pred.	76.88	84.2	77.8	81.8	76.1	68.2	75.0	81.1	77.4	85.3	85.3	76.5	73.4	82.5	77.1	74.5	78.4	68.4	72.6	70.3	62.6	76.1	74.7	78.9	76.8	81.7	86.2	72.8
Improved enode pred.	**+0.23** **77.11**	**+0.2** **84.4**	**+1.5** **79.3**	+0.0 81.8	**+0.6** **76.7**	**+0.6** **68.8**	+0.0 75.0	+0.0 81.1	+0.0 77.4	+0.0 85.3	**+0.0** **85.3**	+0.0 76.5	+0.0 73.4	+0.0 82.5	**+1.9** **79.0**	+0.0 74.5	+0.0 78.4	**+0.3** **68.7**	**+0.0** **72.6**	+0.0 70.3	+0.0 62.6	+0.0 76.1	+0.0 74.7	+0.0 78.9	+0.0 76.8	**+0.4** **82.1**	+0.0 86.2	**+0.7** **73.5**
E) CORPIPE SINGLE VS 7-MODEL ENSEMBLE COMPARISON FOR INDIVIDUAL MODEL SIZES																												
Single umT5-base model	69.19	76.6	73.2	75.2	70.2	53.7	**66.5**	72.0	72.0	74.3	79.3	69.9	69.0	**73.4**	63.8	59.5	72.9	60.3	**65.4**	67.1	49.4	73.4	66.8	73.3	72.3	75.0	77.9	**65.5**
7 umT5-base models	**+0.87** **70.06**	**+1.1** **77.7**	**+0.1** **73.3**	**+0.1** **75.3**	**+0.8** **71.0**	**+0.3** **54.0**	−0.5 66.0	**+1.6** **73.6**	**+0.3** **72.3**	**+0.9** **75.2**	**+1.0** **80.3**	**+0.7** **70.6**	**+1.5** **70.5**	−0.2 73.2	**+0.5** **64.3**	**+1.8** **61.3**	**+0.8** **73.7**	**+0.9** **61.2**	−0.2 65.2	**+1.4** **68.5**	**+2.9** **52.3**	**+2.1** **75.5**	**+1.7** **68.5**	**+0.5** **73.8**	**+0.6** **72.9**	**+0.4** **75.4**	**+2.8** **80.7**	+0.0 65.5
Single mT5-large model	72.32	79.7	73.8	77.9	72.3	56.3	69.9	74.9	73.5	79.4	82.3	71.5	71.0	76.5	69.0	66.4	76.3	64.6	67.7	68.6	**57.5**	75.7	69.9	74.4	74.0	78.0	82.2	**69.5**
7 mT5-large models	**+1.12** **73.44**	**+1.1** **80.8**	**+1.3** **75.1**	**+0.6** **78.5**	**+0.8** **73.1**	**+2.8** **59.1**	**+0.9** **70.8**	**+1.3** **76.2**	**+0.6** **74.1**	**+0.9** **80.3**	**+1.2** **83.5**	**+1.2** **72.7**	**+2.3** **73.3**	**+1.3** **77.8**	**+2.7** **71.7**	**+2.2** **68.6**	**+0.6** **76.9**	**+2.0** **66.6**	**+1.5** **69.2**	**+1.2** **69.8**	−1.0 56.5	**+0.5** **76.2**	**+1.5** **71.4**	**+0.6** **75.0**	**+0.7** **74.7**	**+1.2** **79.2**	**+0.4** **82.6**	−0.6 68.9
Single umT5-xl model	74.80	82.4	76.3	80.2	74.8	61.6	72.0	79.2	76.2	82.8	84.1	73.9	**72.8**	80.8	73.8	67.5	77.4	**68.0**	70.2	69.2	**57.5**	**77.1**	72.2	**78.2**	73.5	80.0	**85.9**	72.0
7 umT5-xl models	**+0.63** **75.43**	**+0.7** **83.1**	**+1.0** **77.3**	**+0.2** **80.4**	**+0.7** **75.5**	**+1.5** **63.1**	**+2.2** **74.2**	**+1.0** **80.2**	**+0.6** **76.8**	**+0.4** **83.2**	**+1.2** **85.3**	**+1.2** **75.1**	−0.7 72.1	**+0.2** **81.0**	**+0.7** **74.5**	**+2.9** **70.4**	**+0.6** **78.0**	−0.8 67.2	**+1.1** **71.3**	**+0.8** **70.0**	−2.0 55.5	−0.9 76.2	**+1.7** **73.9**	−0.2 78.0	**+3.0** **76.5**	**+0.5** **80.5**	−1.0 84.9	**+0.3** **72.3**
Single umT5-xxl model	75.95	83.0	77.3	81.6	76.0	66.5	74.8	80.8	76.7	83.7	84.1	75.5	**73.8**	81.5	76.0	72.0	77.8	**68.7**	71.0	69.7	58.7	75.2	73.1	77.4	76.4	81.4	84.5	**73.4**
7 umT5-xxl models	**+0.93** **76.88**	**+1.2** **84.2**	**+0.5** **77.8**	**+0.2** **81.8**	**+0.1** **76.1**	**+1.7** **68.2**	**+0.2** **75.0**	**+0.3** **81.1**	**+0.7** **77.4**	**+1.6** **85.3**	**+1.2** **85.3**	**+1.0** **76.5**	−0.4 73.4	**+1.0** **82.5**	**+1.1** **77.1**	**+2.5** **74.5**	**+0.6** **78.4**	−0.3 68.4	**+1.6** **72.6**	**+0.6** **70.3**	**+3.9** **62.6**	**+0.9** **76.1**	**+1.6** **74.7**	**+1.5** **78.9**	**+0.4** **76.8**	**+0.3** **81.7**	**+1.7** **86.2**	−0.6 72.8

Table 5: Additional experiments on the CorefUD 1.4 minitest set (CoNLL score in %).

System	Avg	ca	cs pced	cs pdt	cs pdts	cu	de	en fant	en gum	en litb	es	fr anco	fr demo	fr litb	grc	hbo	hi	hu kork	hu szeg	ko	la	lt	nl	no bokm	no nyno	pl	ru	tr
A) SUBMITTED CORPIPE SINGLE MODELS																												
Single mT5-large model	72.23	81.3	74.4	76.9	72.0	58.3	74.9	76.2	74.1	76.1	82.0	74.5	71.8	71.9	69.6	65.8	78.7	65.8	68.7	67.3	52.7	79.8	67.3	78.4	77.7	76.4	75.3	62.2
Single umT5-base model	−3.34	−2.8	−2.0	−2.4	−2.8	−6.6	−4.3	−4.1	−1.1	−4.5	−2.0	−1.3	−2.9	−3.6	−7.1	−11	−1.5	−3.2	−2.6	−1.8	−6.5	−3.1	−2.1	−2.4	−1.5	−2.1	−1.9	−3.0
	68.89	78.5	72.4	74.5	69.2	51.7	70.6	72.1	73.0	71.5	80.0	73.2	68.9	68.3	62.5	55.1	77.2	62.6	66.1	65.5	46.2	76.7	65.2	76.0	76.2	74.3	73.5	59.1
Single umT5-xl model	+2.76	+2.5	+3.0	+3.9	+2.3	+3.7	−0.6	+4.4	+3.4	+3.7	+2.0	+4.4	+2.1	+4.9	+4.1	−0.7	**+2.0**	**+1.7**	**+4.7**	+1.1	+3.6	**+0.2**	+4.5	+1.4	**+2.5**	+2.9	+3.9	+2.6
	74.99	83.8	77.4	80.8	74.3	62.0	74.3	80.6	77.5	79.8	84.0	78.9	73.9	76.8	73.7	65.1	**80.7**	**67.5**	**73.4**	68.4	56.3	**80.0**	71.8	79.8	**80.2**	79.3	79.3	64.8
Single umT5-xxl model	**+4.26**	**+2.7**	**+4.1**	**+4.5**	**+2.9**	**+7.3**	**+3.1**	**+6.4**	**+5.1**	**+6.5**	**+2.6**	**+5.4**	**+3.9**	**+5.9**	**+8.3**	**+6.4**	+1.9	+0.7	+3.1	**+1.5**	**+8.4**	−0.1	**+6.8**	**+2.7**	+2.1	**+4.4**	**+5.1**	**+3.4**
	76.49	**84.0**	**78.5**	**81.3**	**74.9**	**65.6**	**78.0**	**82.5**	**79.2**	**82.6**	**84.6**	**79.9**	**75.7**	**77.8**	**77.9**	**72.2**	80.7	66.5	71.8	**68.8**	**61.1**	79.7	**74.1**	**81.1**	79.8	**80.8**	**80.5**	**65.7**
B) CORPIPE ONE-STAGE SINGLE MODELS																												
One-stage mT5-large model	72.56	81.9	75.7	77.9	72.5	59.3	70.8	76.2	74.8	76.7	82.6	74.2	69.7	69.8	72.5	66.2	80.8	65.7	68.1	67.7	57.8	79.5	69.0	77.6	77.4	76.4	76.0	62.4
One-stage umT5-base model	−3.66	−3.7	−3.3	−3.5	−3.2	−6.3	−1.6	−3.3	−1.9	−4.9	−2.8	−1.8	−0.1	−0.1	−9.1	−8.5	−4.7	−5.3	−3.0	−2.1	−12	−4.1	−1.2	−1.5	−0.9	−2.6	−3.5	−4.3
	68.90	78.2	72.4	74.4	69.4	53.0	69.2	72.9	72.9	71.8	79.8	72.5	69.5	69.7	63.4	57.7	76.1	60.4	65.1	65.6	46.2	75.4	67.8	76.1	76.5	73.8	72.5	58.1
One-stage umT5-xl model	+2.84	+1.9	+2.7	+2.3	+1.6	+5.8	+5.9	+3.2	+2.8	+2.5	+1.4	+3.5	+4.4	+7.2	+2.7	−1.2	−0.2	**+3.8**	**+4.1**	+1.9	−0.4	**+1.0**	+3.6	+1.6	**+3.2**	+3.2	+3.5	+4.7
	75.40	83.8	78.3	80.2	74.2	65.1	76.7	79.4	77.6	79.2	84.0	77.7	74.1	77.0	75.2	65.0	80.6	**69.5**	**72.2**	69.6	57.4	**80.5**	72.6	79.2	**80.6**	79.6	79.5	67.1
One-stage umT5-xxl model	**+4.20**	**+2.4**	**+3.5**	**+3.6**	**+3.9**	**+10.8**	**+7.4**	**+5.6**	**+4.4**	**+7.9**	**+2.2**	**+5.4**	**+5.1**	**+8.3**	**+7.0**	**+4.8**	**+0.6**	+0.3	+2.3	**+2.4**	**+3.5**	−0.8	**+4.6**	**+2.7**	+2.4	**+4.0**	**+4.0**	**+4.8**
	76.76	**84.3**	**79.2**	**81.5**	**76.5**	**70.1**	**78.2**	**81.8**	**79.2**	**84.6**	**84.8**	**79.6**	**74.8**	**78.1**	**79.5**	**71.0**	**81.3**	66.0	70.3	**70.1**	**61.3**	78.7	**73.7**	**80.3**	79.8	**80.4**	**80.0**	**67.2**
C) CORPIPE SINGLE VS ONE-STAGE MODELS FOR INDIVIDUAL MODEL SIZES																												
Single umT5-base model	68.89	**78.5**	**72.4**	**74.5**	69.2	51.7	**70.6**	72.1	**73.0**	71.5	**80.0**	**73.2**	68.9	68.3	62.5	55.1	**77.2**	**62.6**	**66.1**	65.5	**46.2**	**76.7**	65.2	76.0	76.2	**74.3**	**73.5**	**59.1**
One-stage umT5-base model	**+0.01**	−0.3	+0.0	−0.1	**+0.2**	**+1.3**	−1.4	**+0.8**	−0.1	**+0.2**	−0.2	−0.8	**+0.7**	**+1.4**	**+0.9**	**+2.6**	−1.1	−2.2	−1.0	**+0.1**	+0.0	−1.3	**+2.6**	**+0.1**	**+0.3**	−0.5	−1.0	−1.1
	68.90	78.2	72.4	74.4	**69.4**	**53.0**	69.2	**72.9**	72.9	**71.8**	79.8	72.5	**69.5**	**69.7**	**63.4**	**57.7**	76.1	60.4	65.1	**65.6**	46.2	75.4	**67.8**	**76.1**	**76.5**	73.8	72.5	58.1
Single mT5-large model	72.23	81.3	74.4	76.9	72.0	58.3	**74.9**	**76.2**	74.1	76.1	82.0	**74.5**	**71.8**	**71.9**	69.6	65.8	78.7	**65.8**	**68.7**	67.3	52.7	**79.8**	67.3	**78.4**	**77.7**	**76.4**	75.3	62.2
One-stage mT5-large model	**+0.33**	**+0.6**	**+1.3**	**+1.0**	**+0.6**	**+1.0**	−4.1	+0.0	**+0.7**	**+0.6**	**+0.6**	−0.3	−2.1	−2.1	**+2.9**	**+0.4**	**+2.1**	−0.1	−0.6	**+0.4**	**+5.1**	−0.3	**+1.7**	−0.8	−0.3	+0.0	**+0.6**	**+0.2**
	72.56	**81.9**	**75.7**	**77.9**	**72.5**	**59.3**	70.8	76.2	**74.8**	**76.7**	**82.6**	74.2	69.7	69.8	**72.5**	**66.2**	**80.8**	65.7	68.1	**67.7**	**57.8**	79.5	**69.0**	77.6	77.4	76.4	**76.0**	**62.4**
Single umT5-xl model	74.99	**83.8**	77.4	**80.8**	**74.3**	62.0	74.3	**80.6**	77.5	**79.8**	**84.0**	**78.9**	73.9	76.8	73.7	**65.1**	**80.7**	67.5	**73.4**	68.4	56.3	80.0	71.8	**79.8**	80.2	79.3	79.3	64.8
One-stage umT5-xl model	**+0.41**	+0.0	**+1.0**	−0.6	−0.1	**+3.1**	**+2.4**	−1.2	**+0.1**	−0.6	**+0.0**	−1.2	**+0.2**	**+0.2**	**+1.5**	−0.1	−0.1	**+2.0**	−1.2	**+1.2**	**+1.1**	**+0.5**	**+0.8**	−0.6	**+0.4**	**+0.3**	**+0.2**	**+2.3**
	75.40	83.8	**78.3**	80.2	74.2	**65.1**	**76.7**	79.4	**77.6**	79.2	**84.0**	77.7	**74.1**	**77.0**	**75.2**	65.0	80.6	**69.5**	72.2	**69.6**	**57.4**	**80.5**	**72.6**	79.2	**80.6**	**79.6**	**79.5**	**67.1**
Single umT5-xxl model	76.49	84.0	78.5	81.3	74.9	65.6	78.0	**82.5**	**79.2**	82.6	84.6	**79.9**	**75.7**	77.8	77.9	**72.2**	80.7	**66.5**	**71.8**	68.8	61.1	**79.7**	**74.1**	**81.1**	79.8	**80.8**	**80.5**	65.7
One-stage umT5-xxl model	**+0.27**	**+0.3**	**+0.7**	**+0.1**	**+1.6**	**+4.5**	**+0.2**	−0.8	+0.0	**+2.0**	**+0.2**	−0.3	−0.9	**+0.3**	**+1.6**	−1.2	**+0.8**	−0.5	−1.4	**+1.3**	**+0.2**	−1.0	−0.5	−0.8	**+0.0**	−0.4	−0.5	**+1.6**
	76.76	**84.3**	**79.2**	**81.5**	**76.5**	**70.1**	**78.2**	81.8	79.2	**84.6**	**84.8**	79.6	74.8	**78.1**	**79.5**	71.0	**81.3**	66.0	70.3	**70.1**	**61.3**	78.7	73.7	80.3	**79.8**	80.4	80.0	**67.2**
D) ZERO-SHOT RESULTS FOR INDIVIDUAL MODEL SIZES																												
Single umT5-base model	**68.89**	**78.5**	**72.4**	**74.5**	**69.2**	**51.7**	**70.6**	**72.1**	**73.0**	**71.5**	**80.0**	**73.2**	**68.9**	**68.3**	**62.5**	**55.1**	**77.2**	**62.6**	**66.1**	**65.5**	**46.2**	**76.7**	**65.2**	**76.0**	**76.2**	**74.3**	**73.5**	**59.1**
Zshot one-stage umT5-base	−18.5	−6.7	−17	−14	−19	−39	−11	−8.2	−13	−14	−3.4	−29	−6.7	−20	−22	−43	−10	−13	−25	−22	−32	−26	−7.0	−18	−21	−14	−17	−25
	50.44	71.8	54.8	59.6	49.6	12.8	59.8	63.9	60.0	58.1	76.6	43.8	62.2	48.1	40.1	11.7	67.1	49.6	41.4	43.1	14.4	51.0	58.2	58.1	55.5	60.4	55.8	34.4
Zshot enodes+umT5-base	−18.7	−4.7	−18	−14	−18	−28	−13	−13	−14	−16	−2.7	−32	−9.2	−21	−14	−47	−14	−11	−25	−29	−37	−26	−8.5	−17	−20	−13	−18	−18
	50.24	73.8	54.5	60.2	50.8	23.8	58.0	59.0	58.5	55.8	77.3	40.9	59.7	47.3	47.9	7.6	63.0	52.5	41.2	36.0	9.4	50.8	56.7	59.0	56.1	61.1	54.6	40.7
Enodes + zshot umT5-base	−15.5	−4.7	−15	−12	−15	−24	−12	−7.1	−13	−13	−2.6	−31	−7.3	−19	−12	−46	−11	−3.9	−14	−21	−30	−24	−5.4	−15	−18	−9.7	−15	−14
	53.44	73.8	56.8	61.8	54.1	27.7	58.9	65.0	59.8	58.6	77.4	42.4	61.6	49.4	49.9	9.2	66.2	58.7	52.3	44.0	15.9	52.9	59.8	60.7	58.1	64.6	57.6	45.2
Single mT5-large model	**72.23**	**81.3**	**74.4**	**76.9**	**72.0**	**58.3**	**74.9**	**76.2**	**74.1**	**76.1**	**82.0**	**74.5**	**71.8**	**71.9**	**69.6**	**65.8**	**78.7**	**65.8**	**68.7**	**67.3**	**52.7**	**79.8**	**67.3**	**78.4**	**77.7**	**76.4**	**75.3**	**62.2**
Zshot one-stage mT5-large	−16.9	−4.9	−19	−17	−18	−26	−12	−6.8	−15	−14	−3.3	−29	−10	−19	−8.9	−25	−13	−14	−26	−21	−31	−29	−9.0	−13	−18	−12	−16	−26
	55.25	76.4	55.1	60.1	54.2	31.9	62.6	69.5	59.1	62.0	78.7	44.9	61.7	52.5	60.7	41.1	65.7	51.8	42.6	45.7	21.7	51.5	58.3	64.9	59.7	64.2	59.2	36.0
Zshot enodes+mT5-large	−18.2	−4.5	−20	−17	−19	−24	−14	−13	−16	−16	−1.5	−31	−12	−24	−7.6	−25	−17	−9.0	−25	−28	−36	−30	−15	−14	−18	−13	−17	−24
	54.04	76.8	54.4	60.0	53.3	33.6	61.1	63.0	58.0	60.2	80.5	42.9	60.0	48.1	62.0	41.0	62.0	56.8	44.5	39.3	16.9	50.1	52.4	64.0	59.5	62.9	57.9	37.8
Enodes + zshot mT5-large	−14.7	−4.4	−17	−15	−14	−19	−13	−6.8	−15	−13	−1.4	−30	−9.7	−22	−5.8	−17	−15	−1.0	−13	−19	−24	−28	−12	−12	−16	−9.9	−14	−27
	57.51	76.9	56.6	62.1	57.8	38.6	61.7	69.4	59.3	62.7	80.6	44.2	62.1	49.7	63.8	48.8	64.3	64.8	56.2	48.1	29.0	51.7	54.9	66.0	61.7	66.5	60.8	34.5
Single umT5-xl model	**74.99**	**83.8**	**77.4**	**80.8**	**74.3**	**62.0**	**74.3**	**80.6**	**77.5**	**79.8**	**84.0**	**78.9**	**73.9**	**76.8**	**73.7**	**65.1**	**80.7**	**67.5**	**73.4**	**68.4**	**56.3**	**80.0**	**71.8**	**79.8**	**80.2**	**79.3**	**79.3**	**64.8**
Zshot one-stage umT5-xl	−18.2	−5.5	−20	−18	−19	−33	−9.2	−7.6	−16	−18	−2.2	−33	−7.7	−25	−5.6	−29	−11	−9.9	−28	−24	−35	−26	−13	−17	−19	−15	−18	−25
	56.81	78.3	57.2	62.8	55.0	28.9	65.1	73.0	61.3	61.9	81.8	45.9	66.2	52.5	68.1	36.2	70.2	57.6	45.2	43.8	20.8	54.0	58.8	63.2	60.9	64.4	61.2	39.6
Zshot enodes+umT5-xl	−19.0	−5.0	−20	−18	−19	−27	−10	−14	−20	−19	−2.1	−34	−8.5	−21	−5.4	−28	−18	−12	−29	−28	−37	−31	−19	−13	−19	−12	−22	−20
	56.04	78.8	57.2	62.7	55.4	35.0	63.7	67.2	57.3	60.8	81.9	44.8	65.4	56.2	68.3	36.9	62.6	55.9	43.9	40.2	19.0	49.4	53.2	66.5	61.3	66.7	57.5	45.3
Enodes + zshot umT5-xl	−14.9	−5.1	−17	−17	−13	−21	−9.6	−7.5	−19	−16	−2.0	−32	−6.3	−17	−3.7	−22	−16	−4.4	−16	−19	−30	−29	−13	−11	−17	−8.9	−18	−10
	60.07	78.7	59.8	64.5	60.6	40.9	64.7	73.1	58.5	63.7	82.0	46.8	67.6	59.7	70.0	42.9	65.4	63.1	57.4	49.0	26.1	51.2	58.9	68.6	63.1	70.4	60.5	54.6

Table 6: Ablation experiments on the CorefUD 1.4 minidev set (CoNLL score in %). The results of submitted models are averages of 7 runs, other results are averages of 2 or more runs.

Table 5 presents the performance of all four model sizes as well as model ensembles of all sizes on the minitest set. The model performance increases reliably with the model size, both for single models and ensembles, resulting in increases of circa +3, +2.5, and +3.5 percent points for gradually increasing model sizes. Furthermore, the ensembles improve the performance by circa 0.9 percent points on average. Lastly, the improved empty node prediction system provides a modest boost of 0.23 percent points for the xxl-sized models.

Ablation Experiments Table 6 presents our ablation experiments on the minidev sets, starting with the mean performance of the 7 CorPipe 26 models of all sizes used for official ensemble submission shown in Section A. The one-stage variants for all four model sizes are evaluated in Section B and compared to the standard two-stage variants in Section C. The one-stage variants outperform the two-stage variants for all model sizes (even if the best 7 out of 10 two-stage models are reported), by an average of 0.25 percent points, demonstrating

System	Avg	ca	cs pced	cs pdt	cs pdts	cu	de	en fant	en gum	en litb	es	fr anco	fr democ	fr litb	grc	hbo	hi	hu korkor	hu szeg	ko	la	lt	nl	no bokm	no nyno	pl	ru	tr
A) CorPipe One-stage Single Models																												
One-stage mT5-large model	72.51	80.2	76.0	77.4	73.0	61.4	72.6	75.1	73.0	79.2	81.8	70.1	69.9	75.5	70.7	67.0	75.9	65.9	68.1	69.1	55.3	74.8	69.9	74.7	73.3	77.6	80.0	70.2
One-stage umT5-base model	−3.42	−3.8	−2.3	−3.1	−2.5	−7.9	−4.5	−3.8	−2.5	−5.0	−3.7	−1.0	−0.5	−2.0	−7.6	−8.9	−3.0	−3.9	−5.2	−1.4	−3.1	−1.0	−1.1	−1.0	−1.8	−2.1	−2.8	−6.7
	69.09	76.4	73.7	74.3	70.5	53.5	68.1	71.3	70.5	74.2	78.1	69.1	69.4	73.5	63.1	58.1	72.9	62.0	62.9	67.7	52.2	73.8	68.8	73.7	71.5	75.5	77.2	63.5
One-stage umT5-xl model	+2.23	+1.8	+1.8	+2.9	+1.7	+3.4	**+0.9**	+4.2	+3.4	+3.2	+2.3	+3.7	+2.5	+2.3	+2.4	+0.0	+1.2	+1.1	+1.9	−0.1	+2.1	**+1.8**	+3.5	+2.4	+2.0	+2.6	+3.9	+1.4
	74.74	82.0	77.8	80.3	74.7	64.8	**73.5**	79.3	76.4	82.4	84.1	73.8	72.4	77.8	73.1	67.0	77.1	67.0	70.0	69.0	57.4	**76.6**	73.4	77.1	75.3	80.2	83.9	71.6
One-stage umT5-xxl model	**+3.76**	**+3.5**	**+2.2**	**+3.9**	**+3.4**	**+7.7**	−0.5	**+5.5**	**+4.2**	**+4.7**	**+3.0**	**+5.0**	**+3.8**	**+5.8**	**+7.8**	**+4.1**	**+1.7**	**+1.4**	**+3.2**	**+0.7**	**+5.1**	+1.0	**+5.6**	**+4.1**	**+2.1**	**+3.6**	**+5.3**	**+4.1**
	76.27	**83.7**	**78.2**	**81.3**	**76.4**	**69.1**	72.1	**80.6**	**77.2**	**83.9**	**84.8**	**75.1**	**73.7**	**81.3**	**78.5**	**71.1**	**77.6**	**67.3**	**71.3**	**69.8**	**60.4**	75.8	**75.5**	**78.8**	**75.4**	**81.2**	**85.3**	**74.3**
B) CorPipe Single vs One-stage Comparison for Individual Model Sizes																												
Single umT5-base model	**69.19**	**76.6**	73.2	**75.2**	70.2	**53.7**	66.5	**72.0**	**72.0**	**74.3**	**79.3**	**69.9**	69.0	73.4	**63.8**	**59.5**	**72.9**	60.3	**65.4**	67.1	49.4	73.4	66.8	73.3	**72.3**	75.0	**77.9**	**65.5**
One-stage umT5-base model	−0.10	−0.2	**+0.5**	−0.9	**+0.3**	−0.2	**+1.6**	−0.7	−1.5	−0.1	−1.2	−0.8	**+0.4**	**+0.1**	−0.7	−1.4	+0.0	**+1.7**	−2.5	**+0.6**	**+2.8**	**+0.4**	**+2.0**	**+0.4**	−0.8	**+0.5**	−0.7	−2.0
	69.09	76.4	**73.7**	74.3	**70.5**	53.5	**68.1**	71.3	70.5	74.2	78.1	69.1	**69.4**	**73.5**	63.1	58.1	72.9	**62.0**	62.9	**67.7**	**52.2**	**73.8**	**68.8**	**73.7**	71.5	**75.5**	77.2	63.5
Single mT5-large model	72.32	79.7	73.8	**77.9**	72.3	56.3	69.9	74.9	**73.5**	**79.4**	**82.3**	**71.5**	**71.0**	**76.5**	69.0	66.4	**76.3**	64.6	67.7	68.6	**57.5**	**75.7**	**69.9**	74.4	**74.0**	**78.0**	**82.2**	69.5
One-stage mT5-large model	**+0.19**	**+0.5**	**+2.2**	−0.5	**+0.7**	**+5.1**	**+2.7**	**+0.2**	−0.5	−0.2	−0.5	−1.4	−1.1	−1.0	**+1.7**	**+0.6**	−0.4	**+1.3**	**+0.4**	**+0.5**	−2.2	−0.9	**+0.0**	**+0.3**	−0.7	−0.4	−2.2	**+0.7**
	72.51	**80.2**	**76.0**	77.4	**73.0**	**61.4**	**72.6**	**75.1**	73.0	79.2	81.8	70.1	69.9	75.5	**70.7**	**67.0**	75.9	**65.9**	**68.1**	**69.1**	55.3	74.8	**69.9**	**74.7**	73.3	77.6	80.0	**70.2**
Single umT5-xl model	**74.80**	**82.4**	76.3	80.2	**74.8**	61.6	72.0	79.2	76.2	**82.8**	84.1	**73.9**	**72.8**	**80.8**	**73.8**	**67.5**	**77.4**	**68.0**	**70.2**	**69.2**	**57.5**	**77.1**	72.2	**78.2**	73.5	80.0	**85.9**	**72.0**
One-stage umT5-xl model	−0.06	−0.4	**+1.5**	**+0.1**	−0.1	**+3.2**	**+1.5**	**+0.1**	**+0.2**	−0.4	**+0.0**	−0.1	−0.4	−3.0	−0.7	−0.5	−0.3	−1.0	−0.2	−0.2	−0.1	−0.5	**+1.2**	−1.1	**+1.8**	**+0.2**	−2.0	−0.4
	74.74	82.0	**77.8**	**80.3**	74.7	**64.8**	**73.5**	**79.3**	**76.4**	82.4	**84.1**	73.8	72.4	77.8	73.1	67.0	77.1	67.0	70.0	69.0	57.4	76.6	**73.4**	77.1	**75.3**	**80.2**	83.9	71.6
Single umT5-xxl model	75.95	83.0	77.3	**81.6**	76.0	66.5	**74.8**	**80.8**	76.7	83.7	84.1	**75.5**	**73.8**	**81.5**	76.0	**72.0**	**77.8**	**68.7**	71.0	69.7	58.7	75.2	73.1	77.4	**76.4**	**81.4**	84.5	73.4
One-stage umT5-xxl model	**+0.32**	**+0.7**	**+0.9**	−0.3	**+0.4**	**+2.6**	−2.7	−0.2	**+0.5**	**+0.2**	**+0.7**	−0.4	−0.1	−0.2	**+2.5**	−0.9	−0.2	−1.4	**+0.3**	**+0.1**	**+1.7**	**+0.6**	**+2.4**	**+1.4**	−1.0	−0.2	**+0.8**	**+0.9**
	76.27	**83.7**	**78.2**	81.3	**76.4**	**69.1**	72.1	80.6	**77.2**	**83.9**	**84.8**	75.1	73.7	81.3	**78.5**	71.1	77.6	67.3	**71.3**	**69.8**	**60.4**	**75.8**	**75.5**	**78.8**	75.4	81.2	**85.3**	**74.3**

Table 7: Two- and one-stage post-competition comparison on the CorefUD 1.4 minitest set (CoNLL score in %).

the effectiveness of the proposed one-stage variant architecture. Note that in CorPipe 24, the one-stage variant was outperformed by the two-stage variant by more than 1 percent point.

While we did not finish training the one-stage models before the shared task deadline, the organizers enabled us to perform a post-competition evaluation of our one-stage models on the minitest sets. The resulting comparison with the two-stage models is presented in Table 7. While the relative improvements of the one-stage models are less consistent on the minitest sets, the xxl-sized one-stage model outperforms the two-stage model both with the original and the improved empty node prediction system. Finally, note that the evaluated one-stage models are the best out of 2 models trained, compared to the best out of 10 models for the two-stage variant.

Zero-Shot Cross-Lingual Transfer Section D of Table 6 quantifies the zero-shot cross-lingual transfer of the models. When showcasing the performance on a dataset in a given language, we train a model on all datasets except those in the same language. Considering the one-stage variants, the average drop in performance across all datasets and model sizes is 17.9 percent points. While the performance drop is large, both the large-sized and the xl-sized one-stage zero-shot models still outperform the baseline system provided by the organizers.

To evaluate the zero-shot performance of the two-stage model, we train both the empty node baseline and the coreference resolution model in a zero-shot manner. The average drop in this zero-shot two-stage setting is marginally larger, 18.6 percent points on average. For comparison, we also evaluate the zero-shot performance of the two-stage model when using the provided empty node prediction baseline, which is applied only on the datasets containing empty nodes. The resulting performance is 3.6 percent points higher than the zero-shot two-stage model, still significantly worse than the non-zero-shot model.

Comparing the zero-shot performances across languages, when a closely related language (maybe with a similar annotation scheme) is still present in the training data, like for Catalan when Spanish is still present and vice versa, the performance drop is much smaller, around 3-5 percent points. Furthermore, we hypothesize that the performance drop depends considerably on the annotation scheme, as suggested by the zero-shot performance on French, where the three datasets demonstrate performance decrease of -10, -20, and -30 percent points, respectively, indicating that it is not just the language that affects the zero-shot performance.

Effect of Empty Node Quality Finally, Table 8 shows the effect of empty node prediction quality on the overall coreference resolution performance for every dataset with empty nodes. Compared to

System	Avg	ca	cs pced	cs pdt	cs pdts	cu	es	grc	hu kork	hu szeg	pl	tr
A) Empty Node Prediction Ablations for mT5-large Model												
Empty node baseline, single mT5-large model	71.59	81.3	74.4	76.9	72.0	58.3	82.0	69.6	65.8	68.7	76.4	62.2
No empty nodes, single mT5-large model	−11.8 59.84	−6.1 75.2	−8.3 66.1	−4.8 72.1	−16 55.9	−12 46.2	−5.8 76.2	−16 54.4	−13 53.2	−6.7 62.0	−11 64.7	−30 32.3
One-stage mT5-large model	+0.67 72.26	+0.6 81.9	+1.3 75.7	+1.0 77.9	+0.6 72.5	+1.0 59.3	+0.6 82.6	+2.9 72.5	−0.1 65.7	−0.6 68.1	+0.0 76.4	+0.2 62.4
Empty node improved baseline, single mT5-large model	+0.66 72.25	+0.2 81.5	+0.7 75.1	+0.0 76.9	+0.7 72.7	+1.9 60.2	+0.2 82.2	+1.4 71.0	+0.9 66.7	+0.1 68.8	+0.1 76.5	+1.1 63.3
Gold empty nodes, single mT5-large model	**+3.55 75.14**	**+0.8 82.1**	**+4.5 78.9**	**+1.5 78.4**	**+4.3 76.3**	**+5.5 63.8**	**+0.7 82.7**	**+5.0 74.6**	**+3.1 68.9**	**+0.6 69.3**	**+1.7 78.1**	**+11.1 73.3**
B) Empty Node Prediction Ablations for umT5-xl Model												
Empty node baseline, single umT5-xl model	74.65	83.8	77.4	80.8	74.3	62.0	84.0	73.7	67.5	73.4	79.3	64.8
No empty nodes, single umT5-xl model	−12.6 62.10	−6.4 77.4	−8.7 68.7	−4.7 76.1	−16 57.7	−13 49.4	−6.2 77.8	−17 57.0	−15 53.4	−8.2 65.2	−12 67.0	−31 33.6
One-stage umT5-xl model	+0.74 75.39	+0.0 83.8	+1.0 78.3	−0.6 80.2	−0.1 74.2	+3.1 65.1	+0.0 84.0	+1.5 75.2	+2.0 69.5	−1.2 72.2	+0.3 79.6	+2.3 67.1
Empty node improved baseline, single umT5-xl model	+0.51 75.16	+0.0 83.8	+0.3 77.7	+0.0 80.8	+0.5 74.8	+0.9 62.9	+0.1 84.1	+1.7 75.4	+1.4 68.9	−0.1 73.3	+0.1 79.4	+0.8 65.6
Gold empty nodes, single umT5-xl model	**+3.87 78.52**	**+0.9 84.7**	**+5.0 82.4**	**+1.9 82.7**	**+4.4 78.7**	**+5.9 67.9**	**+0.8 84.8**	**+3.7 77.4**	**+4.9 72.3**	**+0.8 74.2**	**+1.8 81.1**	**+12.7 77.5**
C) Empty Node Prediction Ablations for umT5-xxl Model												
Empty node baseline, single umT5-xxl model	75.61	84.0	78.5	81.3	74.9	65.6	84.6	77.9	66.5	71.8	80.8	65.7
No empty nodes, single umT5-xxl model	−12.6 63.05	−6.7 77.3	−8.9 69.6	−5.0 76.4	−17 58.2	−13 53.3	−6.6 78.0	−18 60.1	−15 52.2	−7.2 64.6	−12 69.0	−31 34.5
One-stage umT5-xxl model	+0.74 76.35	+0.3 84.3	+0.7 79.2	+0.1 81.5	+1.6 76.5	+4.5 70.1	+0.2 84.8	+1.6 79.5	−0.5 66.0	−1.4 70.3	−0.4 80.4	+1.6 67.2
Empty node improved baseline, single umT5-xxl model	+0.53 76.14	+0.1 84.1	+0.9 79.4	+0.0 81.3	+0.6 75.5	+0.7 66.3	+0.0 84.6	+0.9 78.8	+1.3 67.8	+0.0 71.8	−0.1 80.7	+1.6 67.2
Gold empty nodes, single umT5-xxl model	**+4.00 79.61**	**+0.6 84.6**	**+5.8 84.3**	**+2.0 83.4**	**+4.5 79.4**	**+6.2 71.8**	**+0.8 85.4**	**+2.0 79.9**	**+4.7 71.2**	**+1.3 73.1**	**+1.7 82.5**	**+14.7 80.3**

Table 8: Ablation experiments on the CorefUD 1.4 minidev set for datasets that contain empty nodes (CoNLL score in %). The results of submitted models are averages of 7 runs, other results are averages of 2 or more runs.

the submitted models, not using any empty node predictions results in a drop of 12.3 percent points on average, while using gold empty nodes provides an average boost of 3.8 percent points. While using the improved empty node prediction system improves the results by 0.55 percent points on average, the best results are achieved by the one-stage variant delivering an average increase of 0.7 percent points compared to the standard two-stage variant.

5 Conclusions

We introduced CorPipe 26, the winning submission to the CRAC 2026 Shared Task on Multilingual Coreference Resolution (Novák et al., 2026a). Our approach encompasses two variants. The first is a two-stage pipeline architecture that first predicts empty nodes using a dedicated pretrained encoder model, and then performs mention detection and coreference linking through a jointly trained system utilizing another pretrained encoder. The second variant predicts empty nodes together with mentions and coreference links in a single model. Our system significantly outperforms all other submissions in both the LLM track and the unconstrained track by 2.8 and 9.5 percent points, respectively. The source code and trained models are publicly available at `https://github.com/ufal/crac2026-corpipe`.

Future Work While CorPipe has been the top-performing system in the CRAC Shared Tasks for the past years, there is always room for improvement. First, CorPipe does not handle coreference links longer than the maximum segment length (by default 2 560 tokens). Our initial attempts to address this issue by linking mentions across segments using the current models were unsuccessful, but modifying the training procedure to include cross-segment links could help.

Second, CorPipe currently relies on pretrained *encoder* models. Given the abundance of new pretrained decoder-only models compared to new multilingual encoders, exploring decoder-only architectures is a possible future direction.

Last, making CorPipe available as a user-friendly tool or a web service could facilitate its adoption by the research community and practitioners.

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.

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 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 Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 95–118, Suzhou, China. Association for Computational Linguistics.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondřej Pražák, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2026a. Findings of the Fifth Shared Task on Multilingual Coreference Resolution: Expanding Datasets for Long-Range Entities. In *Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026)*, San Diego, California, USA. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David

Bamman, Antoine Bourgois, Peter Bourgonje, Silvie Cinková, Eleonora Delfino, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Sooyoun Han, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, and 35 others. 2026b. Coreference in universal dependencies 1.4 (CorefUD 1.4). 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. 2025. CorPipe at CRAC 2025: Evaluating multilingual encoders for multilingual coreference resolution. In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 130–139, Suzhou, China. 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.

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.

A Improved Empty Node Prediction Baseline

Treebank	ARC	DEP	WO	DEP_WO	FORM	LEMMA	UPOS	XPOS	FEATS	ALL
ca	96.73	96.73	93.93	93.93	—	—	—	—	—	93.93
cs_pcedt	76.86	74.84	76.74	74.72	74.21	76.11	76.74	76.49	74.08	73.58
cs_pdt	78.06	77.44	77.75	77.13	75.88	76.50	78.06	77.60	76.03	75.41
cs_pdtsc	88.63	87.54	87.89	86.85	86.45	87.79	88.53	88.08	86.11	85.27
cu	83.37	82.97	82.58	82.32	—	—	—	—	—	82.32
es	96.09	96.09	93.82	93.82	—	—	—	—	—	93.82
grc	91.54	89.88	91.54	89.88	—	—	—	—	—	89.88
hu_korkor	91.45	88.16	91.45	88.16	91.45	—	—	—	—	88.16
hu_szegedkoref	92.46	89.80	92.02	89.36	—	—	—	—	—	89.36
pl	90.67	90.55	90.67	90.55	—	—	—	—	—	90.44
tr	86.84	86.84	86.84	86.84	84.40	—	—	—	—	84.40

Table 9: Performance of the improved empty nodes prediction system on the minidev sets of the CRAC 2026 dataset containing empty nodes. Each reported metric is an F1 score where a prediction is considered correct if both the dependency head and the given column are correct.