Findings of the Second Shared Task on Multilingual Coreference Resolution

Zdeněk Žabokrtský[1], **Miloslav Konopík**[2], **Anna Nedoluzhko**[1], **Michal Novák**[1],
Maciej Ogrodniczuk[3], **Martin Popel**[1], **Ondřej Pražák**[2],
Jakub Sido[2], **Daniel Zeman**[1]

[1] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czechia
{zabokrtsky,nedoluzhko,mnovak,popel,zeman}@ufal.mff.cuni.cz

[2] University of West Bohemia, Faculty of Applied Sciences,
Department of Computer Science and Engineering, Pilsen, Czechia
konopik@kiv.zcu.cz, {ondfa,sidoj}@ntis.zcu.cz

[3] Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland, maciej.ogrodniczuk@gmail.com

Abstract

This paper summarizes the second edition of the shared task on multilingual coreference resolution, held with the CRAC 2023 workshop. Just like last year, participants of the shared task were to create trainable systems that detect mentions and group them based on identity coreference; however, this year's edition uses a slightly different primary evaluation score, and is also broader in terms of covered languages: version 1.1 of the multilingual collection of harmonized coreference resources CorefUD was used as the source of training and evaluation data this time, with 17 datasets for 12 languages. 7 systems competed in this shared task.

1 Introduction

The idea of a shared task focused on resolving coreference for multiple languages goes back to SemEval-2010 (Recasens et al., 2010) with seven languages and CoNLL-2012 (Pradhan et al., 2012) with three languages included. The amount of languages has been extended to 10 languages (with multiple datasets for some of them) in the Multilingual Coreference Resolution Shared Task at CRAC 2022 (Žabokrtský et al., 2022), making use of the CorefUD 1.0 collection (Nedoluzhko et al., 2022). This paper reports on the second edition of this shared task organized in 2023,[1] associated with CRAC again.

In brief, the most important improvements in this year's edition are the following. First, the shared task employs a newer version of the CorefUD collection. CorefUD 1.1 contains updated versions of 13 datasets (for 10 languages) already included in CorefUD 1.0, one new dataset for (already included) Hungarian, and 3 new datasets for newly added languages: 2 for Norwegian and 1 for Turkish.

Second, the original morpho-syntactic features in the development and test sets were replaced by the output of UDPipe 2 (Straka, 2018) to make the evaluation scheme more realistic (with gold feature values being available, coreference prediction might be simplified to some extent, compared to real-world application scenarios).

Third, we use the head-matching approach for mentions in the primary score in this year's edition instead of partial matching. Last year, partial matching led several teams to optimize their predicted mentions by reducing them to their syntactic heads, thereby losing the information about full mention spans.

The remainder of the paper is structured as follows. Section 2 focuses on changes of this shared task's data compared to the previous edition. Section 3 explains the evaluation metrics – the primary score as well as the supplementary ones – employed in the shared task. Section 4 describes the baseline system and the 7 participating systems. Section 5 summarizes the results. Section 6 concludes.

2 Datasets

Like the previous year, the shared task draws its training and evaluation data from the public part of

[1] https://ufal.mff.cuni.cz/corefud/crac23

Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, pages 1–18
December 7, 2023

CorefUD dataset	docs	sents	words	zeros	entities	avg. len.	non-singletons
Catalan-AnCora	1298	13,613	429,313	6,377	18,030	3.5	62,417
Czech-PCEDT	2312	49,208	1,155,755	35,844	52,721	3.3	168,138
Czech-PDT	3165	49,428	834,720	22,389	78,747	2.4	154,983
English-GUM	195	10,761	187,416	99	27,757	1.9	32,323
English-ParCorFull	19	543	10,798	0	202	4.2	835
French-Democrat	126	13,057	284,883	0	39,023	2.0	46,487
German-ParCorFull	19	543	10,602	0	259	3.5	896
German-PotsdamCC	176	2,238	33,222	0	3,752	1.4	2,519
Hungarian-KorKor	94	1,351	24,568	1,988	1,134	3.6	4,103
Hungarian-SzegedKoref	400	8,820	123,968	4,857	5,182	3.0	15,165
Lithuanian-LCC	100	1,714	37,014	0	1,224	3.7	4,337
Norwegian-BokmaalNARC	346	15,742	245,515	0	53,357	1.4	26,611
Norwegian-NynorskNARC	394	12,481	206,660	0	44,847	1.4	21,847
Polish-PCC	1828	35,874	538,885	470	127,688	1.5	82,804
Russian-RuCor	181	9,035	156,636	0	3,636	4.5	16,193
Spanish-AnCora	1356	14,159	458,418	8,112	20,115	3.5	70,663
Turkish-ITCC	24	4,733	55,341	0	690	5.3	3,668

Table 1: Data sizes in terms of the total number of documents, sentences, tokens, zeros (empty words), coreference entities, average entity length (in number of mentions) and the total number of non-singleton mentions. Train/dev/test splits of these datasets roughly follow 8/1/1 ratio. See Nedoluzhko et al. (2022) for details.

the CorefUD collection (Nedoluzhko et al., 2022),[2] now in its latest release (1.1).[3] There are 17 datasets for 12 languages (3 language families). Compared to CorefUD 1.0, which was used in the previous year of the shared task, there are 4 new datasets and 2 new languages (1 new language family): Hungarian KorKor, Norwegian NARC (Bokmål and Nynorsk versions), and Turkish ITCC.

CorefUD ensures that the datasets are unified at the file format level: They use the CoNLL-U format with extra annotation in the last column.[4] The data have not been sufficiently harmonized at the level of annotation guidelines (for example, different datasets may have different rules for the extent of a mention). Table 1 gives an overview of the datasets and their sizes.

We follow the official train/dev/test splits of CorefUD 1.1.

2.1 Updated Resources

The 13 datasets that were already available in CorefUD 1.0 are introduced in Žabokrtský et al. (2022). Instead of repeating the introduction here, we focus on changes between CorefUD 1.0 and 1.1.

Catalan-AnCora (ca_ancora) and **Spanish-AnCora** (es_ancora): The 3LB section of the AnCora treebank is omitted from CorefUD 1.1 because it does not contain coreference annotation. Named entities that are not annotated for coreference are omitted also in the remaining sections (previously they appeared as singletons). There are also some corrections in the LEMMA column and in dependency relations; the `arg` and `tem` semantic attributes from the original corpus are now visible in the MISC column.

Czech-PCEDT (cs_pcedt) and **Czech-PDT** (cs_pdt): Removed superfluous empty nodes (zeros) `#Rcp`, `#Cor` and `#QCor`. Removed empty nodes depending on the artificial root. Improved guessing of pronominal forms for empty nodes, fixed cases where conditional auxiliaries in multiword tokens are used to break mention spans. There are also some improvements in morphological and syntactic annotation. The tectogrammatical functors from the original corpus are now visible in the MISC column.

English-GUM (en_gum): new data from GUM v9 (published in Universal Dependencies 2.12), the total size increased from 164 to 187 thousand words.

English-ParCorFull (en_parcorfull) and **German-ParCorFull** (de_parcorfull): Morpho-

[2] https://ufal.mff.cuni.cz/corefud

[3] http://hdl.handle.net/11234/1-5053

[4] https://ufal.mff.cuni.cz/~popel/corefud-1.0/corefud-1.0-format.pdf

syntactic annotation updated using UD 2.10 models for UDPipe 2. In addition, the conversion of the English data was fixed so that mentions are detected even in invalid files.

French-Democrat (fr_democrat): Conversion into CorefUD reimplemented, fixing multiple bugs.

German-PotsdamCC (de_potsdam), **Hungarian-SzegedKoref** (hu_szeged), **Lithuanian-LCC** (lt_lcc), **Polish-PCC** (pl_pcc), and **Russian-RuCor** (ru_rucor): Morpho-syntactic annotation updated using UD 2.10 models.

2.2 New Resources

Hungarian-KorKor (hu_korkor) (Vadász, 2022) contains texts from two sources: articles from Hungarian Wikipedia and texts from the Hungarian website of the GlobalVoices news portal. Compared to hu_szeged, the latter contains student essays and news articles. Both corpora contain zeros in subject, object, and possessor positions, but the rules for their placement are not identical. Moreover, the tagset of coreference and anaphora relations are different as well.

Norwegian-BokmaalNARC and **Norwegian-NynorskNARC** (no_bokmaalnarc, no_nynorsknarc) (Mæhlum et al., 2022) are based on parts of the Norwegian Dependency Treebank (NDT), which contains mostly news texts, but also government reports, parliamentary transcripts, and blogs in the two varieties of written Norwegian – Bokmål and Nynorsk. Train/dev/test splits correspond to those in the UD version of the NDT treebank.

Turkish-ITCC (tr_itcc) (Pamay and Eryiğit, 2018) is based on the Marmara Turkish Coreference Corpus, which in turn contains documents from the METU Turkish Corpus. There is an overlap between ITCC and the UD Turkish IMST treebank. The gold-standard morphosyntactic annotation of sentences that occur in both datasets was taken from IMST; the remaining sentences were parsed by a model trained on IMST. Train/dev/test split in the shared task follows that of CorefUD.[5] The coreference annotation in this corpus is less advanced than in the other corpora in CorefUD: some paragraphs completely lack coreference annotation, in some other paragraphs coreference is annotated only partially. Annotation of zeros is missing in the current version.

2.3 Data pre-processing

For training and tuning purposes, we have provided the participants with the train and dev sets as they were released in CorefUD 1.1, i.e. with gold coreference annotation for all datasets and manually annotated morpho-syntactic features for the datasets that originally include them. However, in the dev and test sets intended for evaluation (and submitting), we have deleted the corefence annotation and replaced original morpho-syntax features by the outputs of UD 2.10 models for all datasets, even those in which these features were originally human-annotated. Although it makes the evaluation setup more realistic, there is still room for improvement as this has not affected zeros. Similarly to last year's edition, participants have been given the input documents with zeros already reconstructed.

3 Evaluation Metrics

Systems participating in the shared task are evaluated with the CorefUD scorer.[6] The primary evaluation score is the CoNLL F_1 score with singletons excluded and using *head* mention matching, which is a change to the last year's edition, where *partial* mention matching was used in the primary score. In addition, we calculate several other supplementary scores to compare the shared task submissions.

Official scorer We use the CorefUD scorer to evaluate participants' submissions. It is built on the Universal Anaphora (UA) scorer 1.0 (Yu et al., 2022)[7] taking advantage of the implementations of all generally used coreferential measures with no modifications. Additionally, the CorefUD scorer introduces the implementation of head match and the Mention Overlap Ratio (MOR; Žabokrtský et al., 2022). It also supports matching of potentially discontinuous mentions and anaphor-level evaluation of zeros. Naturally, it is also compatible with the CorefUD 1.0 file format.[8]

[5] The CorefUD ITCC data split is not compatible with the IMST treebank data split in Universal Dependencies 2.12 because the sentences were shuffled in IMST. An improved version of IMST is prepared for UD 2.13 to be released in November 2023: The original ordering of sentences from METU is restored, sentence identifiers refer to METU, document boundaries are marked and data split is made compatible with ITCC.

[6] https://github.com/ufal/corefud-scorer

[7] This in turn reimplements the official CoNLL-2012 scorer (Pradhan et al., 2014).

[8] After the scorer for the shared task had been frozen, the UA scorer 2.0 (Yu et al., 2023), which integrates most of the new features from the CorefUD scorer, was released as a result of the cooperation of the authors of the two scorers.

Mention matching Within the CorefUD collection, some datasets do not specify mention spans in their original annotations (e.g. cs_pdt, hu_korkor). In such datasets, a mention is primarily identified by its head and loosely associated with a dependency subtree rooted in this head. Additionally, in other datasets, it can be challenging to precisely define mention boundaries, particularly when mentions involve embedded clauses, long detailed specifications, etc. On the other hand, some of the original sources from CorefUD do not annotate mention heads at all (e.g. de_potsdam, lt_lcc). Consequently, CorefUD addresses this issue by specifying both the mention span and its head for each mention in all its datasets. While mention spans are derived using the dependency tree only if they are not present in the original source, mention heads are always determined from the tree[9] using the Udapi block `corefud.MoveHead`.[10]

The availability of both spans and heads in gold annotation allows for various possible ways of mention matching in the evaluation. Last year, the participants were asked to predict only the span boundaries in order to keep the task simple. To compensate for the drawbacks of *exact matching* (i.e., precise matching of the full span), we proposed the *partial mention matching* method and used it also in the primary score. A partial match of a predicted mention to a gold mention is found if all its words are included in the gold mention and one of them is the gold head. Nevertheless, this approach appeared to be problematic. It encouraged some participants to post-process their predictions by reducing the full mention spans to the head word only. First, since not all the participants applied this post-processing, it made the comparison of the participants' submissions slightly unfair. To rectify this imbalance, we evaluated the submissions also with a head match, deriving the mention heads automatically using the same method as for the gold spans. More importantly, forced shrinkage of predicted mention spans performed by some of the teams resulted in loss of the original mention spans produced by their systems. Consequently, such submissions failed in the evaluation with the exact match.

For this year's edition, we decided to use *head match* in the primary metric. Two mentions are considered matching if their heads correspond to identical tokens. If there are multiple gold or predicted mentions with the same head, full spans are taken into account but only to disambiguate between multiple mentions with the same head. Otherwise, full mention spans are ignored.

Therefore, the participants were expected to predict mention heads in their submissions. However, due to the disambiguation rules we encouraged the participants to predict the mention span boundaries as well. In addition, their presence allows us to evaluate the systems with respect to exact matching as one of the supplementary scores.

Note that the participants were also free to use the Udapi block `corefud.MoveHead` in order to derive the mention head from the dependency tree, if their systems were not able to predict the heads by their own means.[11]

Singletons New additions to the CorefUD collection have not altered the dominance of the datasets without the annotation of singletons, i.e., entities comprising only a single mention. We thus keep the setup from the last year's edition and calculate the primary score excluding potential singletons in both gold and predicted coreference chains.

Primary score As is usual for coreference resolution tasks, we employed the CoNLL F_1 score (Denis and Baldridge, 2009; Pradhan et al., 2014) as the primary evaluation score. It is an unweighted average of the F_1 scores of three coreference metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAF-e (Luo, 2005), each adopting a different view on coreference relations, namely link-based, mention-based and entity-based, respectively. A single primary score providing a final ranking of participating submissions is a macro-average over all datasets in the CorefUD test collection.

Supplementary scores In addition to the primary CoNLL F_1 score, we calculate alternative versions of this metric using different ways of mention matching: partial-match and exact-match. Note

[9] Note that some datasets label a semantic head (single word) or a minimal span (multiple words possible, e.g. in ARRAU, Uryupina et al., 2020), i.e., a unit that carries the most crucial semantic information, instead. Nedoluzhko et al. (2021) have shown though that heads labeled in coreference annotation most often correspond to heads in a dependency tree.

[10] https://github.com/udapi/udapi-python/blob/master/udapi/block/corefud/movehead.py

[11] All of the participants used this Udapi block for predicting heads (or another method with identical results on the test set).

that the partial-match setup was used as the primary score in the last year's edition. Furthermore, we compute the primary metrics using the head-match for all mentions including singletons.

Besides the primary score, we also report the systems' performance in terms of the coreference measures that contribute to the CoNLL score as well as other standard measures, e.g. BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016). To evaluate the quality of mention matching while ignoring the assignment of mentions to coreferential entities, we use the MOR score. Last but not least, we also measure the performance of the systems on zeros using the anaphor-decomposable score for zeros (Žabokrtský et al., 2022), which is an application of the scoring schema proposed by Tuggener (2014).

4 Participating Systems

4.1 Baseline

Same as last year, the baseline system is the end-to-end neural coreference resolution system based on Pražák et al. (2021).[12] The model solves both tasks (mention prediction and coreference linking) at the same time. It goes through all the possible mention spans and learns to predict the antecedent of each span. In case a span is not the correct mention or it is the singleton the model learns to align it to the artificial antecedent. Therefore, the model is not able to predict singletons. During training, the marginal probability of all the correct antecedents of each mention is maximized. More details can be found in Pražák et al. (2021).

4.2 System Submissions

This year, 7 teams participated in the shared task. The descriptions below are based on the information provided by the respective participants in an online questionnaire. As the authors of the DeepBlueAI system have neither provided us with any details nor submitted their system description paper, we cannot include it among the descriptions.

Anonymous[13] The system initially drew inspiration from wl-coref (Dobrovolskii, 2021), accounting for head information. The authors found that XLM-Roberta yields the best results, leading to its selection for subsequent tests. They developed a conversion system to manage the CoNLL-U format as jsonlines. Furthermore, they efficiently incorporate new features (e.g., UPOS, DEPREL, FEATS) with Udapi assistance. Alongside the CoNLL features, a BIO-like scheme is added to the indices in mention spans. Various distance/matching features and context sizes are used to update token scores for potential antecedents. The results primarily depend on a model's ability to construct the assigned scheme, where the head (B) is the primary focus of this specific task. Future work plans include leveraging similarity- and classification properties through fine-tuning sentence embeddings to further enhance span detection and merging. The authors note that they did not conduct any ablation study, and there is still much to explore regarding the usefulness of features.

CorPipe[14] ÚFAL CorPipe is a minor evolution from the system implemented in the previous year (Straka and Straková, 2022). All models undergo training on the concatenation of all treebanks. They utilize either the mT5-large pre-trained model or the mT5-xl pre-trained model. The architecture remains the same, with a few modifications: The system employs 2560 subwords during prediction, which is possible due to the relative embeddings in mT5. Instead of using CRF to perform mention span detection (since it would be complicated to ensemble), the authors train the model using standard classification into generalized BIO encoding, allowing overlapping mentions. Subsequently, a dynamic programming algorithm performs structured prediction, whose output always presents a valid sequence of BIO tags. Ensembling takes place during both the mention span detection and the coreference linking. The ÚFAL CorPipe team submits multiple configurations – one best-performing mT5-large-sized model, one best-performing mT5-xl-sized model, a best-performing checkpoint selected for each treebank independently, and the best submission that is an ensemble of 3 checkpoints chosen for each treebank independently. See Straka (2023) in this volume for details.

DFKI-Adapt[15] The DFKI-Adapt system is based on the baseline system provided by the organizers. This system augments it by adding character embeddings for each token to the original input

[12] `https://github.com/ondfa/coref-multiling`

[13] The authors of this system asked us to anonymize this submission.

[14] The CorPipe system was submitted to CodaLab by user "straka" from team ÚFAL CorPipe.

[15] The DFKI-Adapt system was submitted to CodaLab by user "tatiana.anikina" from team DFKI_TR.

embeddings (based on multilingual BERT) using LSTM (300 dimensions). The training procedure starts with pre-training the joint model utilizing all languages combined into a single training set. Following this step, the team merges the datasets for the related languages (for example, all Slavic or Romance languages) and fine-tunes a separate model for each language using these combined datasets. Additionally, they train the language-specific task adapters added to the BERT model. During the training process, they sort all documents after every epoch according to their difficulty for the model, as determined by the loss function. The most challenging instances are chosen for further model fine-tuning before the next epoch begins. The DFKI-Adapt system employs no external resources for training, relying solely on the Shared Task data.

DFKI-MPrompt[16] The DFKI-MPrompt system integrates two independent modules. One module performs mention generation based on prompt learning facilitated by the OpenPrompt library. Using a prefix template and a frozen mT5-large model, the prompt model generates all possible mentions within a given sentence, including their indices. The training of this single prompt model encompasses all languages. The other module uses the baseline trained on gold mentions. Given the availability of gold mentions, the baseline's mention scorer is not utilized. The baseline also undergoes training on the combined datasets. In the final stage, the authors input the mentions generated by the prompt model to the baseline to identify coreferent pairs.

McGill[17] The McGill system is based on the Longdoc "unbounded memory" model (Toshniwal et al., 2020). It is similar to end-to-end coreference (Lee et al., 2017) adapted for BERT (Joshi et al., 2019). The primary difference is that the model has a discrete set of candidate entities. The McGill system uses the same hyperparameters that Toshniwal et al. (2021) use for the PreCo dataset, with the following exceptions: Speaker information is included at the start of each sentence if present in the dataset. A language embedding is defined for each dataset using the same configuration as the genre embedding used by Lee et al. (2017). The McGill model uses a batch size of 1, similar to most other models based on Lee et al. (2017). The authors experimented with using XLM-Roberta (Conneau et al., 2020) and mT5 (Xue et al., 2021) *Large* model sizes as the language model encoder. They found that XLM-Roberta leads to better performance, so they used XLM-Roberta Large in the final submission. The McGill team trained the model for 60k steps. In the first 50k steps, they trained their model on all datasets weighted by the number of documents in the dataset. For the last 10k steps, they trained the model on all datasets weighted equally. The model with the best performance on the development set, corresponding to 57.5k steps, was submitted. The McGill model predicts only coreferring spans. Therefore, the McGill team estimated mention heads using Udapi following the same method as the shared-task baseline. For details, see Porada and Cheung (2023) in this volume.

Morfbase[18] The Morfbase system enhances the baseline system by incorporating morphological features, drawing inspiration from Pamay Arslan and Eryiğit (2023). These linguistic features, represented as one-hot vectors, are concatenated to BERT representations. Both the mention detection and coreference linking stages utilize these hand-crafted linguistic features. The team used the provided heuristic head detection script on the model outputs to estimate the heads of the predicted mentions. The primary goal of this model is to enhance coreference performance, particularly for pro-dropped and morphologically rich languages. See Pamay Arslan et al. (2023) in this volume for details.

Ondfa The UWB system remains identical to the one submitted in the previous year, optimized for the new metric (Pražák and Konopík, 2022). It builds on the baseline system with several modifications. Initially, the team trains a joint cross-lingual model (XLMR-large) for all datasets. Subsequently, they fine-tune this model for each dataset separately. The model learns to predict the heads of the mentions from the original spans. They either use head prediction or whole span prediction with `corefud.MoveHead` (chosen for each dataset separately based on the performance on the dev dataset). Syntax trees are also incorporated as features into the model. Additionally, the UWB team modified the model to handle singletons.

[16] The DFKI-MPrompt system was submitted to CodaLab by user "natalia_s" from team DFKI_TR.

[17] The McGill system was submitted to CodaLab by user "ianpo".

[18] The Morfbase system was submitted to CodaLab by user "TugbaP" from team TrCR, originally under the name "itunlp".

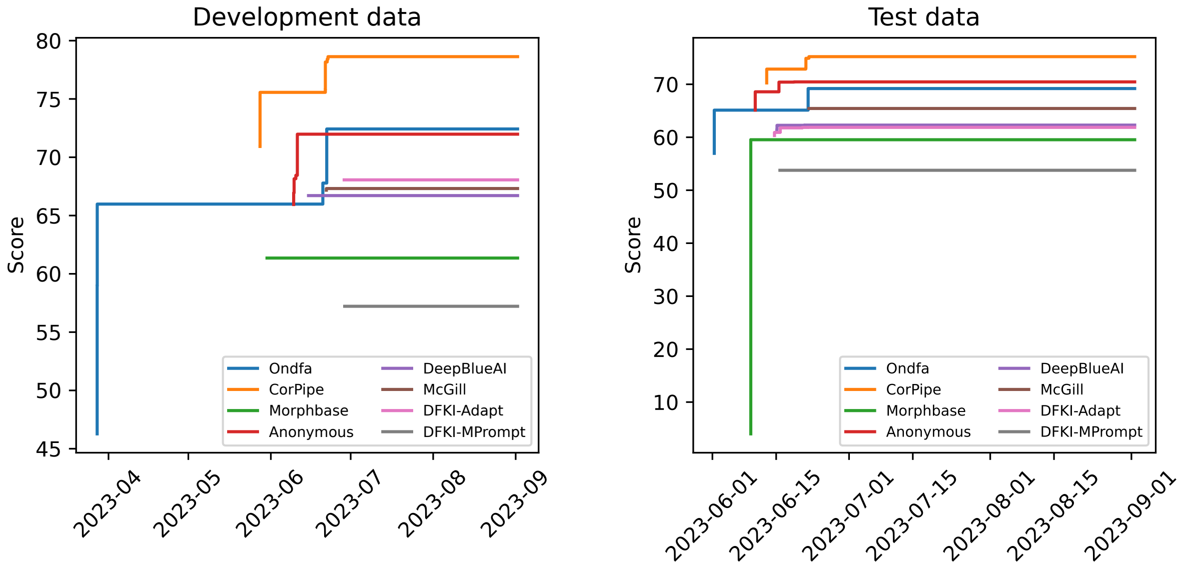

Figure 1: The evolution of the competition in the development (left) and the test phase (right).

4.3 System Comparison

Table 2 shows the basic properties of all submitted systems for evaluation. Half of the submissions based their systems on the provided baseline. The participants who used the baseline model either used it as it is, or added some modifications to it, such as soft prompt, tuning per language, or changing the sequence length.

Comparing Tables 2 and 3 reveals that results generally improve with larger model sizes, apart from some exceptions. This is expected, as larger models have more parameters and can capture more information and nuances from the data. However, larger models also require more computational resources and time to train and run, which could be a challenge for some participants.

5 Results and Comparison

5.1 Evolution of CodaLab Submissions

Across the two phases of the competition, participants had access to the official evaluation scripts, enabling them to track and evaluate the metrics dynamically. We also encouraged them to send continuous results into the CodaLab system.[19] After the competition, we collected all continuously received results from all contributors. The competition evolved as participants refined their models and strategies. We can see non-negligible progress in Figure 1 in terms of observed metrics during both phases, which was caused most probably by competition among participants, who could check the results of others during all phases.

5.2 Main Results

The main results are summarized in Table 3. The CorPipe system is the best one according to the official primary metric (head-match excluding singletons) as well as according to three alternative metrics: partial-match excluding singletons (which was the primary metric last year), exact-match excluding singletons and head-match including singletons. The Anonymous system is the second best according to all four metrics. All metrics result in the same ordering of systems with a single exception of the Ondfa system, which is the second worst according to exact-match, but the third best according to other metrics. This is caused by the fact that for some datasets (cf. description of Ondfa in Section 4.2), Ondfa predicted only the head word and the span was always just this single word.

Table 4 shows recall, precision, and F1 for six metrics. The F1 scores of the first five metrics (MUC. B^3, BLANC, and LEA) result in exactly the same ordering of systems (same as the primary metric). Most of the systems have higher precision than recall for all the metrics, but the highest disbalance is in the BASELINE system. CorPipe is the only system that has higher recall than precision for at least some metrics (MUC and CEAF-e), but other metrics have similar precision and recall.

The MOR metric (mention overlap ratio) mea-

[19] https://codalab.lisn.upsaclay.fr/competitions/11800

Name	Baseline?	Pretrained model	Model size	Seq. length
Anonymous	No	xlm-roberta-base	1-20M (various)	512
BASELINE	Yes	bert-base	220M	512
CorPipe	No	google/mt5-large, google/mt5-xl	567M, 1.7G (two sizes)	512, 2560
DFKI-Adapt	Yes	bert-base	259M	512
DFKI-MPrompt	Yes	bert-base + soft prompt	221M	512
McGill	No	xlm-roberta-large	596M	512
Morfbase	Yes	bert-base	219M	512
Ondfa	Yes	xlm-roberta-large	600M	512

Name	Tuned per lang.?	Batch size	Tuned hyperparameters
Anonymous	Some (l. families)	16	2 – Input size, learning rate
BASELINE	No	1 doc	0
CorPipe	No	8, 12, 16, 32	4 – Model size, batch size, learning rate, epochs
DFKI-Adapt	Yes	1 doc	3 – Dropout, mention loss coef, task LR
DFKI-MPrompt	No	1 sent + 1 doc	0
McGill	No	1	1 – Number of training steps
Morfbase	No	256	0
Ondfa	Yes	1 doc	4 – Specific for the model

Table 2: The table compares properties of systems participating in the task (except for the DeepBlueAI system, as there are no details available) . The systems are ordered alphabetically. The shortcuts in headings are defined as follows: **Name** is the name of the submission, **Baseline?** indicates whether they used a baseline model or not, **Tuned per lang.?** indicates whether they tuned their model for each language or not. **various** in Anonymous means various settings depending on features and architecture.

sures only the mention matching quality, while ignoring the coreference, but even then the ordering of systems is similar to the primary metric (Ondfa is the third worst according to MOR, again because it does not predict full spans for some datasets).

Table 5 shows that the CorPipe system consistently outperforms the other submissions across all datasets and languages. Furthermore, the low results on tr_itcc confirm that the annotation of coreference is unfinished in this dataset. Similarly, we experienced an unexpectedly low performance of submissions on en_parcorfull in the 2022 edition of the shared task. This was a consequence of the small size of the dataset and an error in the CorefUD conversion pipeline, making one of the two documents in the test set completely missing all coreference annotation. The error was fixed this year, but the English and German ParCorFull datasets remain the smallest ones in CorefUD, so there is a high risk of overfitting. We admit such outliers may have a negative impact on the overall score, especially if macro-averaging is used in the primary score to weigh performance on individual datasets. However, we still believe that due to differences in languages and annotation standards, each dataset should contribute equally. The impact of potential errors in some datasets is then mitigated by the number of contributing datasets.

5.3 Evaluation of Zeros

Table 6 focuses on the evaluation of zero anaphors for individual languages where anaphoric zeros are annotated.[20] The F1 scores are again highly correlated with the primary score, with the exception of pl_pcc, where CorPipe was outperformed by Ondfa (4 points better) and DeepBleuAI (1 point better). However, according to Table 1, pl_pcc has a very small number of zeros annotated, so these results are not reliable.

5.4 Further analysis

Similarly to last year, we provide several additional tables in the appendices to shed more light on the differences between the submitted systems.

Tables 7–8 show results factorized according to the different universal part of speech tags (UPOS) in the mention heads. Table 7 contains results on datasets where all entities without any mention with a given UPOS as head were deleted. Table 8 contains results on datasets where all mentions without a given UPOS as head were deleted, so these results may be a bit misleading because e.g. the PRON

[20] Recall that the setup for zeros is slightly unrealistic (see Section 2.3).

	excluding singletons			with singletons
system	**head-match**	**partial-match**	**exact-match**	**head-match**
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**76.82** (+1.91)
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	73.20 (+2.79)
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	68.37 (-0.82)
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	68.23 (+2.80)
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	54.51 (-7.78)
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	53.94 (-7.92)
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	52.07 (-7.47)
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49.32 (-7.64)
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	46.83 (-6.93)

Table 3: Main results: the CoNLL metric 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 best score in each column is in bold. The systems are ordered by the primary metric.

system	**MUC**	**B^3**	**CEAF-e**	**BLANC**	**LEA**	**MOR**
CorPipe	**80 / 79 / 80**	**73 / 73 / 73**	**73 / 71 / 72**	**72 / 73 / 72**	**70 / 71 / 70**	**79** / 80 / **79**
Anonymous	74 / 78 / 76	65 / 72 / 68	67 / 68 / 67	63 / 71 / 66	62 / 69 / 65	74 / 78 / 76
Ondfa	74 / 78 / 75	64 / 71 / 67	64 / 67 / 66	62 / 70 / 65	61 / 68 / 64	52 / 83 / 63
McGill	69 / 76 / 71	60 / 69 / 63	58 / 68 / 62	58 / 68 / 61	57 / 66 / 60	59 / 82 / 67
DeepBlueAI	67 / 74 / 70	56 / 65 / 59	55 / 63 / 58	53 / 64 / 56	53 / 61 / 56	61 / 81 / 67
DFKI-Adapt	66 / 73 / 69	56 / 65 / 59	56 / 62 / 58	53 / 63 / 56	52 / 61 / 55	58 / 80 / 66
Morfbase	63 / 71 / 66	51 / 65 / 56	56 / 58 / 56	47 / 62 / 52	47 / 61 / 52	59 / 78 / 66
BASELINE	56 / 76 / 63	46 / 69 / 54	48 / 62 / 54	44 / 67 / 51	42 / 64 / 49	49 / **87** / 61
DFKI-MPrompt	57 / 67 / 61	45 / 60 / 50	49 / 56 / 51	41 / 57 / 45	40 / 55 / 45	57 / 71 / 62

Table 4: Recall / Precision / F1 for individual secondary metrics. All scores macro-averaged over all datasets.

system	ca_ancora	cs_pcedt	cs_pdt	de_parcorfull	de_potsdam	en_gum	en_parcorfull	es_ancora	fr_democrat	hu_korkor	hu_szeged	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipe	**82.59**	**79.33**	**79.20**	**72.12**	**71.09**	**76.57**	**69.86**	**83.39**	**69.82**	**68.92**	**69.47**	**75.87**	**78.74**	**78.77**	**79.54**	**82.46**	**55.63**
Anonymous	79.51	75.88	76.39	64.37	68.24	72.29	59.02	80.52	66.13	64.65	66.25	70.09	75.32	73.33	77.58	80.19	47.22
Ondfa	76.02	74.82	74.67	71.86	69.37	71.56	61.62	77.18	60.32	66.38	65.75	68.52	72.39	70.91	76.90	76.50	41.52
McGill	71.75	67.67	70.88	41.58	70.20	66.72	47.27	73.78	65.17	60.74	65.93	65.77	73.73	72.43	76.14	77.28	45.28
DeepBlueAI	67.55	70.38	69.93	48.81	63.90	63.58	43.33	69.52	55.69	54.38	63.14	66.75	69.86	68.53	73.11	74.41	36.14
DFKI-Adapt	68.21	68.72	67.34	52.52	69.28	65.11	36.87	69.19	58.96	51.53	58.56	66.01	70.05	68.21	67.98	72.48	40.67
Morfbase	68.23	64.89	64.74	39.96	64.87	62.80	40.81	69.01	53.18	52.91	56.41	64.08	68.17	66.35	67.88	68.53	39.22
BASELINE	65.26	67.72	65.22	44.11	57.13	63.08	35.19	66.93	55.31	40.71	55.32	63.57	65.10	65.78	66.08	69.03	22.75
DFKI-MPrompt	55.45	60.39	56.13	40.34	59.75	57.83	34.32	58.31	52.96	44.53	48.79	56.52	65.12	62.99	61.15	61.96	37.44

Table 5: Results for individual languages in the primary metric (CoNLL).

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

Table 6: Recall / Precision / F1 for anaphor-decomposable score of coreference resolution on zero anaphors across individual languages. Only the datasets that contain 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 to the supplementary scores calculated with respect to whole entities in Table 4.

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 these tables).

Tables 9–12 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. Tables 13–16 show the same statistics for cs_pcedt, which is the largest dataset in CorefUD 1.1 (as for the number of words and non-singleton mentions).

6 Conclusions and Future Work

Both editions of the shared task attracted a substantial number of participants and led to an increase in the state of the art. Hence, the success of the two completed shared tasks supports us in the idea of continuing this initiative in the future.

However, there are challenges, too. For instance, the underlying data collection is still somewhat limited from the typological perspective, and thus our ambition is to add more languages with substantially different typological structures, experiment with other writing systems, or add a historical perspective with data from classical languages.

There are also more technical questions that would deserve a discussion in the future, such as whether weightless macro-averaging is the best approach for data collections with order-of-magnitude differences in training and testing data sizes. Similarly, substantial differences in internal annotation consistency in individual resources is also an issue from the evaluation viewpoint, since, for example, optimizing performance for a low-quality resource might lead to substantial performance gains, which, however, may correspond to systematic deficiencies present in the data rather than objective quality.

Finally, we aim to progress to a fully realistic evaluation setup which starts from raw or pretokenized text. Participants would be then expected to reconstruct zeros.

Acknowledgements

This work was supported by the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR); LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic; FW03010656 of the Technology Agency of the Czech Republic; the European Regional Development Fund as a part of 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; the project co-financed by the Polish Ministry of Education and Science under the agreement 2022/WK/09; and SGS-2022-016 Advanced methods of data processing and analysis. We thank all the participants of the shared task for participating and for providing brief descriptions of their systems. We also thank anonymous reviewers for very useful remarks.

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.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In *Proceedings of the 12th 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):87–96.

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.

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.

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.

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.

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.

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.

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.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Tuğba Pamay and Gülşen Eryiğit. 2018. Turkish Coreference Resolution. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7.

Tuğba Pamay Arslan, Kutay Acar, and Gülşen Eryiğit. 2023. Neural End-to-End Coreference Resolution using Morphological Information. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 34–40.

Tuğba Pamay Arslan and Gülşen Eryiğit. 2023. Incorporating Dropped Pronouns into Coreference Resolution: The case for Turkish. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 14–25.

Ian Porada and Jackie Chi Kit Cheung. 2023. McGill at CRAC 2023: Multilingual Generalization of Entity-Ranking Coreference Resolution Models. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 52–57.

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.

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.

Ondřej Pražák and Miloslav Konopík. 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. Association for Computational Linguistics.

Marta Recasens and Eduard H. Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

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. 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.

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. Association for Computational Linguistics.

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.

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.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 231–235, Gothenburg, Sweden. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering*, 26(1):95–128.

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.

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.

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.

Juntao Yu, Sopan Khosla, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022. The Universal Anaphora Scorer. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4873–4883, Marseille, France. 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 (IWCS)*, Nancy, France. 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.

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.

A Data References

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	ParCorFull	en_parcorfull	(Lapshinova-Koltunski et al., 2018)
French	Democrat	fr_democrat	(Landragin, 2021)
German	ParCorFull	de_parcorfull	(Lapshinova-Koltunski et al., 2018)
German	PotsdamCC	de_potsdam	(Bourgonje and Stede, 2020)
Hungarian	KorKor	hu_korkor	(Vadász, 2022)
Hungarian	SzegedKoref	hu_szeged	(Vincze et al., 2018)
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)
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 Partial CoNLL results by head UPOS

system	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM
CorPipe	**72.21**	**77.05**	**76.33**	**51.58**	**44.38**	**40.13**	33.88	37.44
Anonymous	68.25	72.70	70.84	50.98	38.42	34.15	**35.91**	**41.86**
Ondfa	66.98	71.27	70.16	48.52	33.78	24.98	33.76	40.82
McGill	62.67	68.07	63.76	51.03	39.00	23.68	32.87	28.60
DeepBlueAI	59.54	65.05	60.08	40.34	36.57	17.57	28.26	31.68
DFKI-Adapt	57.80	64.02	61.82	39.53	26.72	14.71	21.29	33.03
Morfbase	55.39	61.74	58.45	44.61	28.58	20.74	30.26	29.17
BASELINE	51.82	57.79	56.32	33.89	25.80	14.12	19.43	27.51
DFKI-MPrompt	50.07	57.37	54.84	42.28	21.37	12.30	25.36	17.81

Table 7: CoNLL F1 score 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 F1 scores are the same as in the primary metric, i.e. an unweighted average over all datasets, partial-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
CorPipe	**63.51**	**65.25**	**63.85**	**52.93**	**49.85**	**50.48**	**51.24**	**50.30**
Anonymous	57.32	59.16	57.80	49.09	46.65	46.39	46.02	46.08
Ondfa	56.39	58.32	57.08	45.55	42.93	42.79	42.64	42.48
McGill	53.13	55.73	52.97	42.50	39.46	39.50	38.79	38.94
DeepBlueAI	50.43	51.93	49.63	40.39	37.60	38.02	37.36	37.14
DFKI-Adapt	48.56	50.95	50.60	34.66	32.05	32.32	31.76	31.59
Morfbase	47.08	48.93	49.23	36.41	33.90	33.92	33.36	33.19
BASELINE	40.50	43.28	45.60	30.62	27.74	28.48	27.74	27.65
DFKI-MPrompt	39.56	43.31	42.67	29.20	26.53	26.64	26.22	26.33

Table 8: CoNLL F1 score 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).

C Statistics of the submitted systems on concatenation of all test sets

	entities				distribution of lengths				
system	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
gold	44,806	107	509	2.0	61.7	22.0	6.7	3.2	6.5
Anonymous	46,367	110	232	2.0	64.0	20.3	6.7	3.0	6.0
BASELINE	14,059	33	237	3.8	0.0	57.7	17.3	7.6	17.4
CorPipe	47,054	112	540	2.0	62.6	21.0	6.8	3.2	6.3
DFKI-Adapt	14,808	35	230	3.8	0.0	56.6	17.7	8.0	17.7
DFKI-MPrompt	12,884	31	85	3.7	0.0	55.5	18.2	8.6	17.7
DeepBlueAI	14,635	35	165	3.9	0.0	54.1	18.4	8.4	19.1
McGill	44,059	105	425	1.9	67.8	17.7	5.8	2.7	6.0
Morfbase	15,118	36	92	3.6	0.0	56.9	18.2	8.2	16.8
Ondfa	55,232	131	135	1.8	70.8	16.3	5.2	2.4	5.3

Table 9: 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 systems are sorted alphabetically. We can see that the Ondfa system notably overgenerates, i.e. predicts more entities than in the gold data. On the contrary, DeepBlueAI, DFKI-Adapt, BASELINE, DFKI-MPrompt, and Morfbase undergenerate and predict on average longer entities (i.e. with more mentions) than in the gold data. The best two systems, CorPipe and Anonymous, have the statistics similar to the gold data.

	mentions				distribution of lengths					
system	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	66,520	158	100	3.1	8.3	45.1	18.7	8.0	3.9	15.9
Anonymous	87,664	209	101	3.3	6.7	41.5	20.5	9.3	4.7	17.3
Baseline	53,063	126	29	2.2	9.9	50.0	19.0	7.2	3.3	10.6
CorPipe	91,081	217	163	3.2	6.5	41.4	20.8	9.5	4.8	16.9
DFKI-Adapt	56,749	135	29	2.3	9.4	49.0	19.2	7.4	3.5	11.5
DFKI-MPrompt	47,796	114	71	2.9	10.7	50.2	17.2	5.8	2.7	13.2
DeepBlueAI	57,329	136	26	2.3	9.2	48.3	19.5	7.7	3.7	11.7
McGill	81,989	195	20	2.3	7.1	43.8	21.8	9.8	5.0	12.5
Morfbase	54,668	130	29	2.3	9.6	48.8	19.0	7.4	3.5	11.6
Ondfa	97,081	231	29	2.6	6.0	49.7	17.6	7.8	4.3	14.5

Table 10: 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). We can see that Ondfa, CorPipe, and Anonymous notably overgenerate mentions, i.e. predict more mentions than in the gold data, but these are the three best systems, so it seems a reasonable strategy. Note that CorPipe is the only system that has higher Recall than Precision in MUC and CEAF-e, according to Table 4. The average length of mentions predicted by Ondfa is lower than in the gold data (and it is caused by the single-word mentions in some datasets). CorPipe and Anonymous are the only two systems that predict long mentions (5+ words) more frequently than in the gold data.

	mentions				distribution of lengths					
system	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	24,961	59	81	3.5	1.3	30.7	25.1	13.6	7.4	21.9
Anonymous	3,088	7	57	3.9	0.0	31.2	25.3	12.3	7.8	23.4
Baseline	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CorPipe	2,674	6	78	3.7	0.1	31.5	25.7	12.2	8.2	22.4
DFKI-Adapt	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DFKI-MPrompt	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepBlueAI	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
McGill	3,160	8	15	2.9	0.0	33.7	27.3	12.7	7.6	18.7
Morfbase	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ondfa	3,226	8	21	3.3	0.1	32.5	26.1	12.2	7.2	21.9

Table 11: Statistics on singleton mentions. See the caption of Table 10 for details. Only four systems (Anonymous, CorPipe, McGill, and Ondfa) attempt to predict singletons and none of them as frequently as in the gold data. 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	10.5	0.6	2.0	44.1	23.3	14.7	7.1	2.7	4.2	1.2	0.5	2.2
Anonymous	8.5	0.0	3.4	51.9	19.1	13.6	5.8	2.5	3.6	1.0	0.6	1.8
Baseline	11.2	0.0	1.8	39.0	26.6	16.1	8.4	2.5	3.8	1.2	0.3	2.1
CorPipe	8.1	0.0	2.6	52.9	18.6	13.8	5.7	2.6	3.2	0.9	0.6	1.7
DFKI-Adapt	10.8	0.0	1.8	40.3	25.8	15.9	8.0	2.5	3.9	1.2	0.4	2.0
DFKI-MPrompt	12.6	0.0	2.0	37.7	29.0	14.7	9.1	1.7	4.0	1.1	0.2	2.5
DeepBlueAI	10.6	0.0	1.9	41.4	25.2	15.3	7.9	2.7	3.8	1.3	0.4	2.0
McGill	8.0	0.0	2.1	51.5	20.4	13.6	6.3	2.4	2.6	1.0	0.6	1.6
Morfbase	11.0	0.0	1.8	40.1	26.1	16.1	8.1	2.4	3.8	1.1	0.4	1.9
Ondfa	7.3	0.1	2.0	54.1	17.6	14.2	5.4	2.5	2.8	1.0	0.9	1.5

Table 12: 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 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 (the 0.1% of such mentions in Ondfa seems to be rather a technical error). 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 10) in mind. For example, only 18.6% of mentions predicted by CorPipe are pronominal (head=PRON), while there are 23.3% of pronominal mentions in the gold data. However, UDPipe predicts actually more pronominal mentions (16941) than in the gold data (15500).

D Statistics of the submitted systems on cs_pcedt

	entities				distribution of lengths				
system	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
gold	2,533	45	84	3.2	7.2	60.1	13.8	6.2	12.8
Anonymous	2,804	50	74	2.9	21.0	47.5	14.2	5.4	11.8
Baseline	1,963	35	77	3.5	0.0	61.7	16.4	6.9	15.0
CorPipe	2,918	52	81	3.0	20.5	47.5	13.4	5.8	12.7
DFKI-Adapt	2,034	36	73	3.6	0.0	60.4	16.2	7.5	15.9
DFKI-MPrompt	1,767	32	36	3.4	0.0	58.7	18.8	8.1	14.3
DeepBlueAI	2,069	37	71	3.6	0.0	60.7	15.9	7.2	16.3
McGill	2,627	47	83	2.8	33.4	39.4	11.2	4.5	11.5
Morfbase	2,038	36	37	3.4	0.0	60.7	17.0	8.0	14.3
Ondfa	2,844	51	74	3.0	23.9	45.4	13.0	5.3	12.3

Table 13: Statistics on coreference entities in cs_pcedt.

	mentions				distribution of lengths					
system	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	7,905	141	61	3.7	19.8	27.9	18.0	8.8	3.9	21.5
Anonymous	7,594	135	60	3.7	20.5	28.7	17.8	8.1	3.8	21.1
Baseline	6,931	124	23	2.6	21.1	29.6	19.5	9.1	4.0	16.7
CorPipe	8,083	144	59	3.7	19.0	28.5	18.3	9.0	4.3	21.0
DFKI-Adapt	7,292	130	23	2.7	20.3	29.2	19.7	9.4	4.2	17.2
DFKI-MPrompt	6,050	108	61	3.5	23.7	31.1	16.7	6.0	2.9	19.4
DeepBlueAI	7,420	132	21	2.8	20.3	28.5	19.4	9.3	4.5	18.0
McGill	6,448	115	16	2.2	22.8	29.2	19.8	10.0	4.8	13.5
Morfbase	6,843	122	26	2.7	21.4	29.5	18.9	9.2	4.2	16.8
Ondfa	7,745	138	22	3.0	19.6	28.7	19.1	9.4	4.5	18.7

Table 14: Statistics on non-singleton mentions in cs_pcedt.

	mentions				distribution of lengths					
system	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	182	3	34	3.3	20.9	21.4	18.1	11.0	8.2	20.3
Anonymous	590	11	47	4.5	9.0	18.3	24.9	15.6	7.3	24.9
Baseline	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CorPipe	598	11	30	4.0	12.4	13.4	26.1	14.0	9.2	24.9
DFKI-Adapt	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DFKI-MPrompt	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepBlueAI	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
McGill	877	16	15	2.0	15.5	40.7	19.4	8.4	5.4	10.6
Morfbase	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ondfa	679	12	22	3.7	12.8	21.8	19.0	12.7	7.7	26.1

Table 15: Statistics on singleton mentions in cs_pcedt.

	mention type [%]			distribution of head UPOS [%]								
system	w/empty	w/gap	non-tree	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
gold	25.9	0.7	4.5	46.2	25.5	6.7	13.2	0.9	2.7	1.6	0.7	2.5
Anonymous	26.3	0.0	2.4	46.6	24.2	7.0	15.9	1.2	2.4	1.5	0.5	0.7
Baseline	24.7	0.0	1.9	47.1	24.8	7.6	15.4	1.1	1.5	1.6	0.6	0.2
CorPipe	24.6	0.0	1.7	48.9	22.5	7.2	15.3	1.3	2.1	1.6	0.6	0.5
DFKI-Adapt	24.0	0.0	1.8	48.1	24.2	7.4	15.0	1.1	1.8	1.7	0.7	0.2
DFKI-MPrompt	29.0	0.0	2.1	42.8	28.3	6.7	17.7	1.1	1.4	1.3	0.3	0.3
DeepBlueAI	24.2	0.0	1.6	48.2	23.9	7.3	15.2	1.1	2.1	1.6	0.5	0.2
McGill	25.9	0.0	1.3	48.1	26.7	6.7	15.0	0.9	0.8	1.2	0.6	0.1
Morfbase	25.1	0.0	1.8	46.9	25.3	7.2	15.4	1.3	1.7	1.6	0.5	0.1
Ondfa	23.9	0.0	1.7	49.2	23.2	7.4	15.0	1.2	1.4	1.7	0.6	0.3

Table 16: Detailed statistics on non-singleton mentions in cs_pcedt.