Findings of the 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], Yilun Zhu[4]

[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

[4] Georgetown University, Department of Linguistics,
Washington, DC, USA, yz565@georgetown.edu

Abstract

This paper presents an overview of the shared task on multilingual coreference resolution associated with the CRAC 2022 workshop. Shared task participants were supposed to develop trainable systems capable of identifying mentions and clustering them according to identity coreference. The public edition of CorefUD 1.0, which contains 13 datasets for 10 languages, was used as the source of training and evaluation data. The CoNLL score used in previous coreference-oriented shared tasks was used as the main evaluation metric. There were 8 coreference prediction systems submitted by 5 participating teams; in addition, there was a competitive Transformer-based baseline system provided by the organizers at the beginning of the shared task. The winner system outperformed the baseline by 12 percentage points (in terms of the CoNLL scores averaged across all datasets for individual languages).

1 Introduction

Multilingual shared tasks are an important source of momentum in various subfields of NLP research, with the CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006) being one of the most successful and influential examples. Clearly, the limiting factor for organizing such shared tasks is the availability of multilingual data whose annotations are harmonized at least to some extent, so that the experiments on individual languages can be performed and evaluated in a uniform way.

In the coreference world, one of the first multilingual shared tasks were SemEval-2010 (Recasens et al., 2010) with seven languages and CoNLL-2012 (Pradhan et al., 2012), in which OntoNotes data for three languages (English, Chinese, and Arabic) were included. With the recent advance of the CorefUD collection (Nedoluzhko et al., 2021a, 2022), harmonized coreference data for 10 languages (covered in CorefUD's publicly available edition) became available. Hence, CorefUD is the source of data for the present shared task; more information about the collection is given in Section 2. In brief, participants of this shared task are supposed to (a) identify mentions in texts and (b) predict which mentions belong to the same coreference cluster (i.e., refer to the same entity or event), using the CorefUD data both for training and evaluation of their coreference resolution systems.

A specific feature of CorefUD is that it combines coreference with dependency syntax, using the annotation scheme (and file format too) of the Universal Dependencies (UD) project (de Marneffe et al., 2021). In all datasets included in the collection, the coreference annotation is manual and the dependency annotation is either manual too, if available, or produced by a dependency parser. Empirical evidence showing advantages of such symbiosis of coreference and dependency syntax is presented in two case studies (Popel et al., 2021; Nedoluzhko et al., 2021b). Participants of this shared task can employ the dependency annotation for determining mention spans (as mentions often correspond to syntactically meaningful units) and for determining core parts of mentions (which correspond to syntactic head in CorefUD).

To the best of our knowledge, this is the first

Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution, pages 1–17
October 17, 2022

shared task on multilingual coreference resolution that accepts zeros (e.g. elided subjects) as potential members of coreference chains.[1] Zeros are an integral part of some of the CorefUD datasets, using empty nodes in enhanced UD representation to annotate them. We keep all annotated zeros, encouraging participants to resolve coreference also for this type of potential mentions.

As with other shared tasks, evaluation is crucial. Unfortunately, and unlike e.g. in dependency parsing, there is no simple and easily interpretable accuracy metric for coreference. We adhere to using the CoNLL score developed in former coreference shared tasks. More specifically, we use an average of the F_1 values of MUC, B^3 and CEAF-e scores as the main evaluation metric. More details concerning evaluation are presented in Section 3.

A Transformer-based coreference prediction system (Pražák et al., 2021) was provided as a strong baseline to the shared task participants. The baseline system as well as 8 systems submitted by the participants are briefly described in Section 4 and some of the systems are described in more detail in separate papers in this volume. Their results are summarized in Section 5. Possible directions for future editions of the shared task are outlined in Section 6.

2 Datasets

For training and evaluation purposes, the present shared task uses 13 coreference datasets for 10 languages as available in the public edition of the CorefUD 1.0 collection (Nedoluzhko et al., 2022) and follows the train/dev/test split of the collection, too.

2.1 Data Resources

Key features of the original coreference resources harmonized under the CorefUD scheme are extracted from Nedoluzhko et al. (2022) into the following paragraphs; some of their quantitative properties are summarized in Table 1.

Prague Dependency Treebank (Czech) (denoted as cs_pdt for short in this paper) is a corpus of Czech newspaper texts (∼830K tokens) with manual multi-layer annotation (Hajič et al., 2020). Coreference and bridging relations are annotated as links on the deep syntactic layer. The links lead from the node of the syntactic head of the anaphor to the node representing the syntactic head of the antecedent and the whole subtrees of these nodes are considered to be mention spans.

Prague Czech-English Dependency Treebank – the Czech part (cs_pcedt) is one side of the PCEDT parallel corpus (Nedoluzhko et al., 2016) consisting of more than 1M tokens. The annotation of coreference-like phenomena is principally similar to the Prague Dependency Treebank with some minor differences and no bridging annotation.

Georgetown University Multilayer Corpus (English) (en_gum) (Zeldes, 2017) is a growing open source corpus of 12 written and spoken English genres (∼180K tokens as of 2022). Next to UD syntax trees and discourse parses, it exhaustively annotates all mentions, including nested, named/non-named entities, singletons, and 10 entity classes and 6 information status tags. It distinguishes 8 anaphoric links: pronominal anaphora and cataphora, lexical and predicative coreference, apposition, discourse deixis, split antecedents and bridging. For licence reasons, Reddit data is excluded from both the UD_English-GUM and CorefUD 1.0 releases of GUM.

Polish Coreference Corpus (pl_pcc) (Ogrodniczuk et al., 2013, 2015) is a corpus (∼ 540K tokens) of Polish nominal coreference built upon the National Corpus of Polish (Przepiórkowski et al., 2008). Mentions are annotated as linear spans, with additionally marked semantic heads. The annotation includes identity coreference, quasi-identity relations and non-identity close-to-coreference relations.

Democrat (French) (fr_democrat) (Landragin, 2021) is a diachronic corpus of written French texts from the 12th to the 21st century. The annotation focuses on nominal mentions (pronouns and full NPs only) and includes information of definiteness and syntactic type of mentions. Its conversion in CorefUD is based only on its automatically parsed subset of texts from 19th-21st century (Wilkens et al., 2020) (∼280K tokens).

Russian Coreference Corpus (ru_rucor) (Toldova et al., 2014) is a corpus of ∼150K tokens annotated with anaphoric and coreferential relations between noun groups. Mentions are annotated as linear spans, with additionally

[1] Recasens et al. (2010) do not state how zeros were treated for pro-drop languages such as Spanish and Catalan in SemEval-2010, and Pradhan et al. (2012) excluded all zeros from the CoNLL-2012 shared task data.

CorefUD dataset	docs	sents	words	zeros	entities	avg. len.	non-singletons
Catalan-AnCora	1550	16,678	546,665	6,377	69,239	1.6	62,416
Czech-PCEDT	2312	49,208	1,155,755	43,054	52,743	3.4	178,376
Czech-PDT	3165	49,428	834,721	32,617	78,880	2.5	169,545
English-GUM	175	9,130	164,392	92	24,801	1.9	28,054
English-ParCorFull	19	543	10,798	0	180	4.0	718
French-Democrat	126	13,054	284,823	0	40,937	2.0	47,172
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-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
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	1635	17,662	559,782	8,112	73,210	1.7	70,664

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.

distinguished syntactic heads. Only NPs which take part in coreference relations are considered and singletons are not annotated.

ParCorFull (German and English) (de_parcorfull and en_parcorfull) is a parallel corpus of ~160K tokens annotated for coreference (Lapshinova-Koltunski et al., 2018). Mentions are NPs which form part of pronoun-antecedent pairs, pronouns without antecedents or VPs if they are antecedents of anaphoric NPs (discourse deixis). The annotation includes identity coreference relations only. Due to license restrictions, CorefUD contains only its WMT News section (~20K tokens).

AnCora: Multi-level Annotated Corpora for Catalan and Spanish (ca_ancora and es_ancora) (Taulé et al., 2008; Recasens and Martí, 2010) consist of very detailed annotations of coreference (including zero anaphora, split antecedent, discourse deixis, etc.). The corpora (~1M tokens) also contain annotations of related phenomena such as argument structure, thematic roles, semantic classes of verbs, named entities, denotative types of deverbal nouns etc.

Potsdam Commentary Corpus (German) (de_potsdam) is a relatively small (~35K tokens) corpus of newspaper articles (Bourgonje and Stede, 2020) annotated for nominal and pronominal identity coreference. Mentions are further classified into primary (e.g. pronouns, definite NPs, proper names), secondary (indefinite NPs, clauses), and non-referring mentions. The corpus also contains gold constituent syntax, information structure (including topic and focus, see Lüdeling et al. (2016)), and discourse parses.

Lithuanian Coreference Corpus (lt_lcc) (Žitkus and Butkienė, 2018) is a corpus of written texts, focusing on political news (~35K tokens). Coreference annotation is link-based and additional coreference information is divided into four levels that include types of mentions, types of anaphoric relations, the direction of the relation, and annotation of split antecedents.

SzegedKoref: Hungarian Coreference Corpus (hu_szeged) (Vincze et al., 2018) is a corpus of written texts (~125K tokens) selected from the Szeged Treebank. The treebank has manual annotations at several linguistic layers such as deep phrase-structured syntactic analysis, dependency syntax and morphology. Mentions are linear spans without specially marked heads, the relations are classified into anaphoric classes such as repetitions, synonyms, hypernyms, hyponyms etc.

2.2 Annotation Details

CorefUD collection is fully compliant with the CoNLL-U format,[2] using the MISC column for annotation of coreference. Besides coreference,

[2] https://universaldependencies.org/format.html

also other anaphoric relations (e.g. bridging, split antecedents) are labeled in some CorefUD datasets. Nevertheless, the shared task focuses only on coreference. Therefore, the participants are asked to predict only the Entity attribute in the MISC column, namely the bracketing of mention spans (including possible discontinuities) and entity/cluster IDs assigning the mentions to entities. They do not need to identify mention heads or fill other coreference-related features that can be found in CorefUD data.

Reconstructed zeros are an integral part of some of the CorefUD datasets. CorefUD utilizes empty nodes in enhanced UD representation to mark them. In the shared task data, we keep all annotated zeros and ask the participants to predict coreference also for them. However, note that we decided not to strip off the empty nodes from the test data in the first edition of the shared task. Although some datasets mark also non-anaphoric zeros, presence of an empty node may indicate its anaphoricity. Its assignment to a cluster of other mentions still remains unknown, yet this makes the setup a bit unrealistic. We find it a reasonable compromise between exploring insufficiently known area of zero anaphora in coreference resolution and making the shared task simple and accessible.

Apart from annotation of coreference and anaphora, CorefUD comprises also standard UD-like annotation of parts of speech, morphological features and dependency syntax. With some exceptions, if the original resources contained manual annotation of morpho-syntax, it has been kept also in CorefUD. Otherwise, it has been obtained automatically using UDPipe 2.0 (Straka, 2018). Therefore, it must be noted that if a system takes advantage of this morpho-syntactic information, its performance on the datasets with manual morpho-syntax may be a bit overestimated, compared to real-world NLP scenarios in which manual annotations of morphology and syntax are usually not available.

3 Evaluation Metrics

Systems participating in the shared task are evaluated with the CorefUD scorer.[3] The primary evaluation score is the CoNLL F_1 score with singletons excluded and using partial mention matching. We also assess the shared task submissions by multiple supplementary scores.

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

Official scorer We use our modification of the coreference scorer – CorefUD scorer. It is based on the Universal Anaphora (UA) scorer (Yu et al., 2022)[4] reusing the implementations of all generally used coreferential measures without any modification. This guarantees that the measures are computed in exactly the same way. However, our scorer is capable of processing the coreference annotation files in the CorefUD 1.0 format. Among other things, it allows evaluation of coreference for zeros.[5] Moreover, it re-defines matching of key and response mentions in the way to be able to handle potentially discontinuous mentions, which are present in some CorefUD datasets. Last but not least, we proposed and implemented the MM score to measure the accuracy of mention matching (see below).

Partial matching The CorefUD collection includes datasets (e.g. cs_pdt) that do not specify mention spans in their original annotations. In these datasets, a mention is only specified by its head and loosely by a dependency subtree rooted in this head. Also in other datasets, the exact specification of mention boundaries may be difficult, for instance, if mentions comprise embedded clauses, long detailed specifications, etc. Therefore, authors of some datasets address this issue by defining a syntactic or 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 important semantic information.

CorefUD specifies a mention head only syntactically. However, as it has been shown in Nedoluzhko et al. (2021b), heads labeled within coreference annotation most often correspond to heads defined by a dependency tree.

Availability of heads/minimal spans in key (i.e. gold reference) annotation allows for *partial mention matching* during the computation of any evaluation measure. In the UA scorer, a response (i.e. predicted by a system) mention matches a key mention if the boundaries of the response span lie within the key span and surround the key minimal span at the same time. In order to support evaluation of discontinuous mentions, we modified this criterion using a set/subset relation. In the

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

[5] Nonetheless, the current implementation is not able to handle a response document whose tokens are not completely identical to ones in the key document. This holds also for empty nodes.

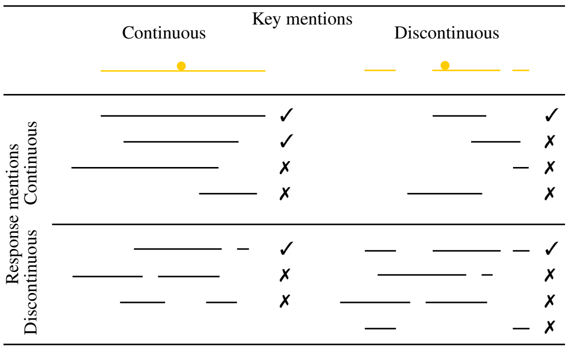

Figure 1: Examples of successful and unsuccessful partial mention matching of key mentions (the yellow ones in the top; the mention head depicted by a small circle) by various response mentions. Showing cases of both continuous and discontinuous mentions. Recall the definition of partial match: A response mention matches a key mention if all its words are included in the key mention and one of them is the key head.

CorefUD scorer, a response mention matches a key mention if all its words are included in the key mention and one of them is the key head. See Figure 1 for examples of response mentions that succeed or fail to match a key mention, depending on whether the mentions are continuous or discontinuous.

Head matching The *partial-match* approach to evaluation described above has two disadvantages. First, it suffices for the systems to predict only heads instead of full mention spans. For this reason, we report also the *exact-match* version as a secondary measure.

Second, some authors may decide to post-process predictions of their systems by reducing the span of each mention to the head word only using one of the methods described below. We can see in Table 4 that five systems (*straka*, berulasek and simple-rule-based*) applied this post-processing and improved thus their results in terms of the primary metric. However, this post-processing can be applied to any system, so we have decided to introduce it as another secondary metric called *head-match*. This way we can see what is the effect of such post-processing for systems which have not applied it. The *head-match* metric is even more benevolent than *partial-match* because it does not penalize extra words added to the span as long as the head remains the same.

The shared task did not require to predict the head in each mention. However, the head can be predicted given the span and the provided dependency tree as the "highest" node. We used Udapi block `corefud.MoveHead` for this purpose.[6]

The easiest post-processing method (chosen in all three *straka** submissions) is to reduce the span of each mention to the head.[7] However, the resulting CoNLL-U files may be invalid because two mentions may be assigned the same span.[8] One solution (chosen in the *berulasek* submission) is to merge the entities of the two mentions which got assigned the same span. In the *head-match* solution, we chose a more conservative solution: if two spans share the same head, we reduce only the smaller span and keep the larger span intact. We confirmed that differences between the three methods described in this paragraph according to the evaluation metrics are negligible because the cases of two mentions sharing the same head are rare.

Singletons The primary score is calculated excluding potential singletons, i.e., entities comprising only a single mention, in both key and response coreference chains. We selected this option as the primary metric because a majority of datasets in the CorefUD collection does not have singletons annotated.

Primary score As a primary evaluation metric, we employed the CoNLL F_1 score (Denis and Baldridge, 2009; Pradhan et al., 2014), which has been established as a standard for the evaluation of coreference resolution. It is an unweighted average of F_1 scores of three coreference measures: 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 three alternative

[6] `https://github.com/udapi/udapi-python/blob/master/udapi/block/corefud/movehead.py` This block was used also for annotating the heads in the gold data.

[7] With Udapi, it can be done using a command `udapy -s corefud.MoveHead util.Eval coref_mention='mention.words=[mention.head]' < in.conllu > out.conllu`.

[8] For example in coordinations, the mention covering the whole coordination and the mention covering the first conjunct share the same head. It should be noted we did not require the submissions to pass the official UD validation tests (`validate.py --level 2 --coref`).

versions of this metric: head-match, exact-match and with-singletons.

Besides the primary score and its three variants, we also report the systems' performance in terms of two additional scores: BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016).

In addition, we implement the MOR[9] score measuring to what extent key and response mentions match, no matter to which coreference entity they belong. First, we find such one-to-one alignment $A(\mathcal{K}, \mathcal{R})$ between the sets of all key mentions $\mathcal{K}$ and all response mentions $\mathcal{R}$ that maximizes the overall number of overlapping words within aligned mentions. We then calculate the recall of mention overlap as a ratio of the total number of overlapping words in mentions and the overall size of all key mentions (sum of its lengths):

$$MOR_{rec} = \frac{\sum_{(K,R) \in A(\mathcal{K},\mathcal{R})} |K \cap R|}{\sum_{K \in \mathcal{K}} |K|}$$

Precision is calculated analogously using the set of all response mentions $\mathcal{R}$ in the denominator. Note that position of the head in mentions does not play a role in MOR score.

In order to show performance of the systems on zeros, we use an anaphor-decomposable score which is an application of the scoring schema introduced by Tuggener (2014). For each zero mention other than the first one in the entity, we indicate a *true positive (tp)* case if an overlap in at least one preceding mention is found between respective key and response entities. *Wrong linkage (wl)* is indicated if no such mention is found and *False positive/negative (fp/fn)* case if the anaphoric response/key mention is not anaphoric (or it is the first mention of the entity) in the key/response document, respectively. Having these counts aggregated, recall is calculated as $\frac{tp}{tp+wl+fn}$ and precision as $\frac{tp}{tp+wl+fp}$.

4 Participating Systems

4.1 Baseline

The baseline system (BASELINE[10]) is based on the multilingual coreference resolution system presented by Pražák et al. (2021). The model uses multilingual BERT (Devlin et al., 2018) in the end-to-end setting. In high-level terms, the model goes through all potential spans and maximizes the probability of gold antecedents for each span. The same system is used for all the languages in the training dataset.

The simplified system adapted to CorefUD 1.0 is publicly available on GitHub[11] along with tagged dev data and its dev data results.

4.2 System Comparison

Table 2 shows the basic properties of all submitted systems for evaluation. The table is organized by individual teams. Some teams submitted more than one system. Roughly half of the systems exploited the provided baseline and the majority of the systems relied on machine learning.

Further details of the machine learning systems are described in Table 3. The table indicates that all machine-learning systems rely on large pretrained models consisting of hundreds of millions of parameters. The ÚFAL CorPipe team and the UWB team employ multilingual models. Karol Saputa utilizes a Polish model as he submitted results for Polish only. All teams who developed their deep-learning solution use the maximum sequence length of 512 sub-word tokens which equals the maximum allowed length of the employed models. Clearly, all the teams are aware of the necessity to model long dependencies in the coreference resolution task. The ÚFAL CorPipe trains on sentences and they put 8 samples in a batch. The UWB team works with documents and they put 1 document in a batch. Karol Saputa uses a dynamic batch to fill the buffer of 4 000 subwords. The number of gradient updates is similar to the teams that train on all languages. Karol Saputa trains with a much smaller number of updates since he trains only on one corpus.

4.3 Teams

The descriptions below are based on the information provided by the respective participants in an online questionnaire.

ÚFAL CorPipe submitted three systems (for details see (Straka and Straková, 2022) in this volume). All are based on pre-trained masked language models, either the RemBERT (Chung et al., 2020) or the XLM-RoBERTa (Conneau et al., 2019) large models. Each sentence is processed as an individual example. Additionally, the neighboring sentences from the document are included

[9]It stands for Mention Overlap Ratio.

[10]The baseline system was submitted to CodaLab under the name *sidoj*, but we rename it here to BASELINE for clarity.

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

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

Table 2: System comparison. The baseline solution, if involved, was either modified internally, or only its predictions were used and modified subsequently ("files only"). "DL" stands for a deep learning solution.

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	$\sim$10

Table 3: Machine Learning Parameters. SL means sequence length, Size is the number of trainable parameters in the models, Updates is the number of gradient updates during training and HParames shows the number of tuned hyper-parameters.

as context – the right context is limited to 50 subwords, and the size of the left context is chosen so that the whole input has 512 subwords. The model is trained jointly to perform two tasks – mention span detection and coreference linking. The mention detection is trained using a CRF sequence tagging scheme based on a generalization of BIO encoding allowing overlapping mentions. Then, for each mention, it is decided which of the preceding mentions is its antecedent (selecting the original mention if there is no antecedent). To obtain a distribution over the previous mentions, a query and a key are computed using a nonlinear transformation, and then masked dot-product attention is utilized. Some experiments include *corpus id* – a special token at the beginning of a sample indicating the source corpus of the sample.

The *straka* system is trained jointly on all training data in all languages. This strategy exhibited a considerably better performance than training on individual corpora separately. For each corpus, the optimal model and epoch is chosen according to its development score. The *straka-single-multilingual-model* system employs a single checkpoint of a single model, thus corresponding to a real deployment scenario. The chosen model is based on Rembert, samples training data according to the logarithm of the respective corpus size, and does not utilize the corpus id. The *straka-only-single-treebank-data* system uses an independent model for each corpus with corresponding training data only. The model is based on Rembert, and for each corpus the submitted predictions are from the epoch with the best development performance. All three submissions were post-processed by reducing mentions spans to the head (cf. Head matching in Section 3).

UWB submitted one system *ondfa* which extends the baseline system (for details see Pražák and Konopik (2022) in this volume). The system relies on combined datasets to employ cross-lingual training. The authors did not know the exact procedure to generate heads for mentions. Therefore, they attempted to learn the heads from the data. The system relies on XLM-Roberta large, which is a substantially bigger model than in the baseline.

Barbora Dohnalová submitted two systems, *berulasek* and *simple-rule-based*, implemented as rule-based blocks in Udapi (Popel et al., 2017).[12]

berulasek post-processes the baseline predictions by first reducing mention spans to the head (cf. Head matching in Section 3) and then adding all proper nouns (upos=PROPN) with the same lemma into the same entity cluster (potentially adding new mentions to existing entities). The second step is applied only to cs, de, es, fr, and hu because it improved the results on the dev set only for these languages.

simple-rule-based starts by linking each pronoun to the nearest previous noun of the same gender (as annotated in the provided CoNLL-U files) and then applies the "*berulasek*" post-processing.

The purpose of these two submissions was to show what results can be achieved with just a few lines of code and without using the training data.

Matouš Moravec submitted one system *moravec*. The system is based on postprocessing existing coreference prediction using named entity information. Specifically, the submission starts with baseline predictions, runs the NameTag web service[13] (Straková et al., 2019) on the underlying texts and applies the following three postprocessing rules using Udapi (Popel et al., 2017): (1) changing coreference spans to spans of named entities, (2) removing coreference links between different named entity types, and (3) adding coreference links between named entities of the same type that have a high string similarity. The author was not able to obtain any results that were better than the baseline for a whole dataset, although in some individual documents within these datasets coreference prediction was improved.

Karol Saputa submitted one system *k-sap* (for details see (Saputa, 2022) in this volume). It employs BERT-based antecedent scoring for possible spans based on representation of span start and end tokens. The submission employs the approach described by Kirstain et al. (2021).

5 Results and Comparison

The *straka* system by the ÚFAL CorPipe team is clearly the winner of the shared task. It surpasses other systems not only in terms of the primary score (see the *primary* column in Table 4) but consistently also in almost all coreference metrics, both in precision and recall (see Table 5).

Table 6 shows that systems submitted by the ÚFAL CorPipe team are dominant on the great majority of datasets. They are outperformed only by the *ondfa* system, namely on de_parcorfull and hu_szeged datasets. Per-dataset evaluation also reveals that the last place of the *k-sap* system in the overall ranking is unequivocally caused by ignoring all but the pl_pcc dataset where it ranks 3rd.

In comparison to the baseline system, most systems ouperformed it by a relatively large margin. The winning *straka* system achieves over 12 points in the primary score, which is more than 20% improvement over the baseline performance. This is an extremely beneficial effect of the shared task, which may drive further development in multilingual coreference resolution.

Results unsurprisingly also confirm a doubtless dominance of machine learning approaches. Although rule-based postprocessing has been employed by some teams (also encouraged by availability of the baseline predictions), its incorporation is either motivated by the nature of the primary score (*straka** systems) or it improves upon the baseline only marginally (the *berulasek* system) or not at all (the *moravec* system).

We observe almost the same picture in evaluation with singletons (see Table 4) – the *straka** systems outperforming all the other systems. Moreover, these submissions are the only ones that are positively affected by the inclusion of singletons. It suggests that unlike other teams, ÚFAL CorPipe have optimized for singletons as well (confirmed by statistics on mentions and entities in Table 9).

Interestingly, no system outperformed the baseline in the exact-match evaluation (see the *exact-match* column in Table 4). Considerably low scores compared to the partial matching setting are apparently caused by the choice of partial matching as part of the primary score, which most of the teams optimized for. Two teams (ÚFAL CorPipe and Barbora Dohnalová) even utilize the present dependency structure to reduce their mentions to heads only in post-processing (cf. Head matching in Section 3).[14] The preference of most systems in

[12]The *simple-rule-based* system was originally called *simple_baseline* in CodaLab, but we renamed it here to prevent confusing it with the official baseline (described in Section 4.1 and named *sidoj* in CodaLab).

[13]http://lindat.mff.cuni.cz/services/nametag/api-reference.php

[14]It would be interesting to evaluate the ÚFAL CorPipe (*straka**) systems before this post-processing, which slightly improves the primary metric (partial-match), but substantially worsens the exact-match.

underspecified mentions is confirmed by the head-match scores (Table 4), which are almost identical to the primary scores, and by MOR scores (see Table 5), reaching high precision but failing in recall.

5.1 Automatic analysis

To the best of our knowledge, this is the first shared task on multilingual coreference resolution that includes zeros. Therefore, Table 7 focuses more on the performance with respect to zero anaphora (cf. Table 1 for proportion of all zeros in the data). It shows anaphor-decomposable scores achieved by the systems on zeros across the datasets that comprise anaphoric zeros. The best-performing systems surpass 90 F1 points for most of the languages. Nevertheless, recall that the setup for zeros is slightly unrealistic as participants have been given the input documents with zeros (both anaphoric and non-anaphoric) already reconstructed.

We provide several additional tables in the appendices to shed more light on the differences between the submitted systems. Table 8 shows results factorized according the different part of speech tags in the mention heads. Tables 9–11 show various statistics on the entities and mentions in a concatenation of all the test sets. Tables 12–14 show the same statistics for cs_pcedt, which is the largest dataset in CorefUD 1.0.

5.2 Manual analysis

In addition to numerical scores, we also want to gain some insight into the types of errors that individual systems do. Such error analysis is inevitably incomplete, as we cannot manually check over 50,000 non-singleton mentions from all the test datasets, times eight system submissions. Nevertheless, here are some observations for illustration:

BASELINE, cs_pdt

It often does not recognize a mention. For example, adjectives derived from locations (*ostravské* "Ostrava-based") tend to be mentions in CorefUD, often nested ones (*ostravské firmy* "Ostrava-based companies") but the system does not recognize them. It also fails to recognize many mentions that are regular noun phrases.

Once the system detects a mention, it often has the correct mention span, although there are some odd failures, too.

In case of a newspaper interview, first and second person pronouns are recognized as mentions, coreference between mentions of the same person is found correctly, but their link to a person's name is easily misinterpreted.

***straka*, cs_pdt**

It detects some mentions that BASELINE does not see (e.g. *ostravské*).

Linking names to first and second person pronouns is also a problem, although the system got right one instance where the baseline failed.

BASELINE, es_ancora

There is an even more dramatic disproportion between number of mentions found and those in the gold data. This is probably caused by the large number of singletons in AnCora.

On the other hand, it correctly identified mentions (including coreference) that were not annotated in the gold data: M_1 = *tanto China como Perú* "China as well as Peru", M_2 = *estas dos naciones* "these two nations".

Elsewhere, the coreference resolver got misled by similar titles of two different people: *el canciller peruano* "the Peruvian secretary" was linked to *el canciller chino* "the Chinese secretary".

***straka*, es_ancora**

Much more successful in identifying mentions; unlike the baseline, it seems to be able to identify singletons.

Unlike the baseline, *straka* did not recognize *tanto China como Perú* as a mention. It also did not link the word *China* from this expression to a previous (singleton) instance of *China*; but since the same surprising annotation appears in the gold data, the system scored here.

6 Conclusions and Future Work

This paper summarizes the outcomes of the Multilingual Coreference Resolution Shared Task held with the CRAC 2022 workshop. We hope that the presented shared task establishes a new state of the art in multilingual coreference resolution.

Possible future editions of the shared task could be improved or extended along the following directions:

- We will fix minor errors in CorefUD's harmonization procedure that have been identified during the shared task.
- We would like to include additional datasets, especially for languages that have not been covered in CorefUD yet; about 20 resources

system	**primary**	**head-match**	**exact-match**	**with-singletons**
straka	**70.72**	**70.72** (+0.00)	33.18 (-37.54)	**72.98** (+2.26)
straka-single...	69.56	69.56 (+0.00)	33.06 (-36.51)	71.81 (+2.25)
ondfa	67.64	68.51 (+0.87)	54.73 (-12.91)	58.06 (-9.58)
straka-only...	64.30	64.30 (+0.00)	32.28 (-32.02)	67.93 (+3.63)
berulasek	59.72	59.72 (+0.00)	31.50 (-28.22)	50.84 (-8.88)
BASELINE	58.53	59.67 (+1.13)	**56.72** (-1.82)	49.69 (-8.84)
moravec	55.05	56.35 (+1.29)	52.68 (-2.37)	46.79 (-8.27)
simple-rule-based	18.14	18.14 (+0.00)	12.60 (-5.54)	17.13 (-1.00)
k-sap	5.90	5.93 (+0.03)	5.84 (-0.05)	3.83 (-2.07)

Table 4: Main results: the CoNLL metric macro-averaged over all datasets. The table shows the primary metric (partial-match, excluding singletons) and its three versions: head-match, exact-match and with-singletons. The best score in each column is in bold.

system	**MUC**	**B^3**	**CEAF-e**	**BLANC**	**LEA**	**MOR**
straka	**74** / 76 / **74**	**67** / **72** / **68**	**71** / **70** / **70**	**63** / 70 / **65**	**63** / **69** / **65**	32 / 83 / 45
straka-single...	72 / 76 / 73	65 / 72 / 67	67 / 70 / 68	61 / **71** / 64	62 / 68 / 64	32 / 84 / 45
ondfa	69 / **76** / 72	61 / 71 / 65	62 / 69 / 65	59 / 69 / 63	58 / 67 / 62	**52** / 84 / **62**
straka-only...	65 / 71 / 68	58 / 68 / 62	61 / 67 / 63	55 / 66 / 59	54 / 63 / 58	30 / 83 / 43
berulasek	58 / 76 / 64	50 / 70 / 57	52 / 67 / 58	46 / 70 / 53	45 / 66 / 53	27 / **88** / 40
BASELINE	56 / 74 / 63	48 / 69 / 56	51 / 66 / 57	45 / 68 / 51	44 / 64 / 51	49 / 86 / 61
moravec	53 / 70 / 60	45 / 65 / 52	50 / 59 / 53	41 / 59 / 46	41 / 60 / 48	49 / 81 / 60
simple-rule-based	14 / 22 / 16	14 / 26 / 17	23 / 27 / 22	10 / 20 / 11	7 / 17 / 9	16 / 55 / 23
k-sap	6 / 7 / 6	5 / 7 / 6	5 / 6 / 6	5 / 7 / 6	5 / 6 / 6	5 / 7 / 6

Table 5: Recall / Precision / F1 for individual secondary metrics. All scores macro-averaged over all datasets. Note that the high recall and F1 MOR scores of ONDFA (relative to STRAKA* systems) is caused by the fact that ONDFA does not use any post-processing restricting mention spans to the head.

system	ca_ancora	cs_pcedt	cs_pdt	de_parcorfull	de_potsdam	en_gum	en_parcorfull	es_ancora	fr_democrat	hu_szeged	lt_lcc	pl_pcc	ru_rucor
straka	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
straka-single...	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
ondfa	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
straka-only...	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
berulasek	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
BASELINE	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.99	55.55	52.35	64.81	65.34	67.66
moravec	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
simple-rule-based	15.58	5.51	9.48	29.81	19.41	21.99	11.37	16.64	21.74	17.00	27.53	15.69	24.06
k-sap	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.67	0.00

Table 6: Results for individual languages in the primary metric (CoNLL).

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_szeged	pl_pcc
straka	**91** / 91 / **91**	**91** / **92** / **92**	87 / **90** / **89**	**94** / **95** / **95**	79 / 71 / 75	62 / 60 / 61
straka-single...	91 / **92** / 91	91 / 92 / 92	**88** / 90 / 89	94 / 95 / 95	76 / **76** / 76	**79** / 83 / **81**
ondfa	88 / 88 / 88	88 / 92 / 90	85 / 89 / 87	92 / 94 / 93	**81** / 74 / **77**	62 / 60 / 61
straka-only...	89 / 88 / 88	90 / 92 / 91	87 / 89 / 88	92 / 92 / 92	74 / 70 / 72	71 / 63 / 67
berulasek	82 / 83 / 82	84 / 86 / 85	80 / 84 / 82	87 / 89 / 88	55 / 54 / 54	42 / 50 / 45
BASELINE	82 / 82 / 82	84 / 86 / 85	80 / 83 / 82	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
moravec	81 / 82 / 82	84 / 85 / 84	80 / 83 / 81	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
simple-rule-based	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
k-sap	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	4 / **100** / 8

Table 7: 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 neither to the CoNLL score nor to the supplementary scores calculated with respect to whole entities in Table 5.

that have not been harmonized yet due to various reasons are listed in Nedoluzhko et al. (2021a) (or have been harmonized, but cannot be distributed publicly because of license limitations).

- We will try to find ways to include also coreference data from the OntoNotes project, which would be extremely valuable because of their size, quality, and popularity.
- We will make the setup more realistic. Firstly, we will delete empty nodes from the test data to be processed by participants' systems. It also requires adjusting the scorer so that it can evaluate pairs of documents with different sets of empty nodes. Secondly, we will replace the manual morpho-syntax annotation with the automatic one for the shared task.
- We will consider introducing subtasks focused on other anaphoric relations than just identity coreference (see Yu et al. (2022) for a description of Universal Anaphora Scorer that is capable of evaluating also non-identity coreference relations); for instance, some CorefUD datasets contain hand-annotated bridging relations already now.

Acknowledgements

This work was supported by the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR); LM2018101 (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 two anonymous reviewers and Jana Straková for very insightful and 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.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing.

In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Proces. del Leng. Natural*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

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 (ELRA).

Anke Lüdeling, Julia Ritz, Manfred Stede, and Amir Zeldes. 2016. Corpus Linguistics and Information Structure Research. In Caroline Féry and Shinichiro Ichihara, editors, *The Oxford Handbook of Information Structure*, pages 599–617. Oxford University Press.

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 '05, pages 25–32. 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'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021a. Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. Technical Report 66, ÚFAL MFF UK, Praha, Czechia.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of LREC*.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021b. 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, Poznań, Poland, December 7-9, 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.

Martin Popel, Zdeněk Žabokrtský, Anna Nedoluzhko, Michal Novák, and Daniel Zeman. 2021. Do UD trees match mention spans in coreference annotations? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3570–3576, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *NoDaLiDa 2017 Workshop on Universal Dependencies*, pages 96–101, Göteborg, Sweden. Göteborgs universitet.

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 Konopik. 2022. End-to-end Multilingual Coreference Resolution with Mention Head Prediction. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marta Recasens and Eduard H. Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510. Tex.bibsource= dblp computer science bibliography, https://dblp.org tex.biburl= https://dblp.org/rec/bib/journals/nle/RecasensH11.

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. *Lang. Resour. Eval.*, 44(4):315–345.

Karol Saputa. 2022. Coreference Resolution for Polish and Beyond: Description of the Herferencer System for the CRAC 2022 Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics.

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 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*. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. 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'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

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.

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 (ELRA).

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu. 2020. French coreference for spoken and written language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 80–89, Marseille, France. European Language Resources Association.

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.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Voldemaras Žitkus and Rita Butkienė. 2018. Coreference Annotation Scheme and Corpus for Lithuanian Language. In *Fifth International Conference on Social Networks Analysis, Management and Security, SNAMS 2018, Valencia, Spain, October 15-18, 2018*, pages 243–250. IEEE.

A Partial CoNLL results by head UPOS

system	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM
straka	**68.71**	**73.72**	**72.29**	66.58	47.71	**38.44**	**49.85**	**48.30**
straka-single…	67.17	73.25	70.35	62.65	**49.84**	36.91	45.77	44.97
ondfa	66.04	71.43	70.72	**69.01**	39.67	25.47	38.51	33.52
straka-only…	61.46	67.08	63.89	60.60	41.38	30.71	35.70	39.55
berulasek	56.43	61.55	59.47	48.91	32.74	18.37	23.67	31.02
BASELINE	55.24	60.44	58.23	48.65	30.43	18.29	23.44	29.87
moravec	52.91	58.82	52.43	46.80	27.49	18.19	23.41	29.22
simple-rule-based	10.22	18.27	17.78	6.32	2.96	3.31	1.58	4.97
k-sap	5.74	5.80	5.99	5.84	4.72	5.77	4.08	5.98

Table 8: 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*).

B 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	41,001	104	509	2.2	54.9	23.5	9.2	4.3	8.0
BASELINE	4,541	11	217	11.2	0.0	33.1	8.6	6.0	52.3
berulasek	4,583	12	242	11.1	0.4	32.8	8.9	6.1	51.8
k-sap	1,744	4	41	4.0	0.1	50.1	18.8	8.6	22.4
moravec	5,469	14	210	10.8	1.8	28.2	9.6	4.6	55.8
ondfa	4,628	12	174	11.7	0.0	31.6	9.5	5.4	53.5
simple-rule-based	1,729	4	149	16.3	0.0	4.5	1.3	7.8	86.5
straka	12,669	32	200	7.1	27.1	4.5	3.6	6.8	58.0
straka-only…	12,552	32	338	7.2	25.5	4.4	4.1	7.3	58.7
straka-single…	12,669	32	243	7.1	26.2	4.4	4.0	6.9	58.5

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., number of mentions in the entity. Distribution of entity lengths (singletons have length = 1).

system	mentions total count	per 1k words	length max	avg.	distribution of lengths 0 [%]	1 [%]	2 [%]	3 [%]	4 [%]	5+ [%]
gold	69,406	175	104	3.3	10.2	39.4	19.6	8.5	4.4	17.9
BASELINE	50,783	128	26	2.2	13.3	46.3	19.1	7.3	3.4	10.7
berulasek	50,935	129	1	0.9	13.4	86.6	0.0	0.0	0.0	0.0
k-sap	6,941	18	29	1.6	0.0	75.1	14.1	4.1	2.0	4.7
moravec	58,883	149	26	2.1	11.5	50.2	18.5	7.2	3.2	9.5
ondfa	54,018	137	30	1.7	12.5	65.8	9.6	3.8	1.9	6.4
simple-rule-based	28,130	71	1	1.0	0.0	100.0	0.0	0.0	0.0	0.0
straka	86,412	218	1	0.9	8.4	91.6	0.0	0.0	0.0	0.0
straka-only…	87,059	220	1	0.9	8.4	91.6	0.0	0.0	0.0	0.0
straka-single…	86,689	219	1	0.9	8.4	91.6	0.0	0.0	0.0	0.0

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., number of nonempty nodes in the mention. Distribution of mention lengths (zeros have length = 0).

system	mention type [%] w/empty	w/gap	non-tree	distribution of head UPOS [%] NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
gold	9.9	0.7	2.6	52.6	17.9	13.9	5.4	2.5	3.5	1.0	1.1	2.0
BASELINE	15.0	0.0	2.1	38.7	28.6	14.0	8.4	2.6	3.9	1.1	0.3	2.3
berulasek	13.4	0.0	0.0	38.2	28.5	14.7	8.4	2.6	3.8	1.1	0.3	2.2
k-sap	0.2	0.0	1.5	39.9	14.1	13.3	3.0	1.2	19.5	0.5	0.1	8.4
moravec	12.9	0.0	2.4	35.0	24.6	21.7	7.7	2.3	3.5	1.0	0.4	3.9
ondfa	13.3	0.0	1.4	40.7	27.6	13.6	8.1	2.6	3.6	1.2	0.4	2.3
simple-rule-based	0.0	0.0	0.0	15.6	62.6	21.8	0.0	0.0	0.0	0.0	0.0	0.0
straka	8.1	0.0	0.0	52.4	18.4	13.9	5.6	2.2	3.5	0.9	1.1	2.1
straka-only…	8.1	0.0	0.0	52.0	18.3	14.0	5.5	2.3	3.8	0.9	1.0	2.2
straka-single…	8.1	0.0	0.0	52.4	18.3	14.1	5.6	2.2	3.5	0.8	1.0	2.1

Table 11: Detailed statistics on mentions. The left part of the table shows 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. The right part of the table shows distribution of mentions based on the universal part-of-speech tag (UPOS) of the head word. Note that the participants were not required to predict the head, so we used Udapi block `corefud.MoveHead` on all submissions for the purpose of these statistics.

C Statistics of the submitted systems on cs_pcedt

system	entities				distribution of lengths				
	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
gold	2,533	45	89	3.3	1.8	63.7	14.8	6.4	13.3
BASELINE	2,048	37	78	3.5	0.0	62.1	16.5	6.1	15.4
berulasek	2,062	37	80	3.5	0.7	62.2	15.8	6.0	15.3
moravec	2,284	41	77	3.6	2.1	55.8	18.3	6.8	16.9
ondfa	2,136	38	74	3.5	0.0	61.9	16.1	6.3	15.7
simple-rule-based	271	5	57	6.1	0.0	46.1	14.4	11.1	28.4
straka	2,770	49	81	3.0	16.4	50.1	15.2	6.4	11.9
straka-only…	2,741	49	80	3.0	16.9	48.9	15.0	6.8	12.4
straka-single…	2,773	49	82	3.0	18.1	48.6	15.3	6.1	11.9

Table 12: Statistics on coreference entities in cs_pcedt.

system	mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
gold	8,365	149	61	3.6	22.6	26.9	17.4	8.6	3.9	20.6
BASELINE	7,258	129	22	2.5	24.6	28.2	18.7	9.0	4.1	15.4
berulasek	7,262	130	1	0.8	24.9	75.1	0.0	0.0	0.0	0.0
moravec	8,228	147	22	2.4	21.7	31.7	19.2	9.2	4.1	14.1
ondfa	7,527	134	21	2.7	23.4	27.4	18.3	9.0	4.5	17.3
simple-rule-based	1,640	29	1	1.0	0.0	100.0	0.0	0.0	0.0	0.0
straka	7,890	141	1	0.8	24.0	76.0	0.0	0.0	0.0	0.0
straka-only…	7,888	141	1	0.8	24.1	75.9	0.0	0.0	0.0	0.0
straka-single…	7,831	140	1	0.8	24.1	75.9	0.0	0.0	0.0	0.0

Table 13: Statistics on non-singleton mentions in cs_pcedt.

system	mention type [%]			distribution of head UPOS [%]								
	w/empty	w/gap	non-tree	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
gold	29.2	1.2	4.5	44.9	28.0	6.4	12.4	0.9	2.7	1.5	0.7	2.6
BASELINE	28.3	0.0	3.8	45.1	30.2	6.4	12.2	0.6	1.5	1.3	0.7	2.0
berulasek	24.9	0.0	0.0	44.6	30.1	7.2	12.2	0.5	1.5	1.3	0.7	1.9
moravec	24.8	0.0	3.8	41.5	26.5	12.4	10.7	0.6	1.3	1.1	0.7	5.2
ondfa	27.5	0.0	3.5	45.3	29.0	6.1	12.7	0.7	2.0	1.4	0.6	2.3
simple-rule-based	0.0	0.0	0.0	3.4	78.2	18.4	0.0	0.0	0.0	0.0	0.0	0.0
straka	23.3	0.0	0.0	45.0	28.1	5.9	12.7	0.8	2.7	1.3	0.7	2.8
straka-only…	23.2	0.0	0.0	44.9	28.2	6.1	12.5	1.0	2.8	1.3	0.6	2.7
straka-single…	23.3	0.0	0.0	45.0	28.2	6.0	12.7	0.8	2.6	1.3	0.6	2.8

Table 14: Detailed statistics on mentions in cs_pcedt.