Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4859–4872
Marseille, 20-25 June 2022

CorefUD 1.0: Coreference Meets Universal Dependencies

Anna Nedoluzhko[1], **Michal Novák**[1], **Martin Popel**[1],
Zdeněk Žabokrtský[1], **Amir Zeldes**[2], **Daniel Zeman**[1]
[1] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{nedoluzhko,mnovak,popel,zabokrtsky,zeman}@ufal.mff.cuni.cz
[2] Georgetown University, Department of Linguistics,
1421 37th St. NW, Washington, DC, 20057, USA
amir.zeldes@georgetown.edu

Abstract

Recent advances in standardization for annotated language resources have led to successful large scale efforts, such as the Universal Dependencies (UD) project for multilingual syntactically annotated data. By comparison, the important task of coreference resolution, which clusters multiple mentions of entities in a text, has yet to be standardized in terms of data formats or annotation guidelines. In this paper we present CorefUD, a multilingual collection of corpora and a standardized format for coreference resolution, compatible with morphosyntactic annotations in the UD framework and including facilities for related tasks such as named entity recognition, which forms a first step in the direction of convergence for coreference resolution across languages.

Keywords: coreference, anaphora, corpus, treebank, multilinguality, coreference resolution

1. Introduction

Recent years have seen a tremendous growth in the amount and quality of annotated resources available in comparable formats and using compatible guidelines for a variety of languages. Following the introduction of Google's universal part-of-speech tags (Petrov et al., 2012), the Universal Dependencies (UD) project (de Marneffe et al., 2021) has made remarkable progress in standardizing both the inventory of morphosyntactic labels as well as the guidelines for assigning them across a wide range of datasets and languages, with the result that practitioners in a variety of fields relying on syntactic analysis can have a reasonable expectation of the format and nature of automatic and manual syntactic analyses based on the framework.

The same cannot be said for coreference resolution, the task of clustering together multiple mentions of the same entity in a text (e.g. 'Joe Biden', 'the U.S. President' and 'he'), as well as other tasks concerned with anaphoric relations (e.g. event coreference resolution, bridging resolution, mention detection). As our survey below will show, even for a single, high-resource language such as English, datasets diverge broadly in the phenomena covered, the way they are analyzed and the formats and information made available by datasets. As a result, state-of-the-art methods, e.g. (Joshi et al., 2020, Kirstain et al., 2021, Dobrovolskii, 2021), are usually benchmarked only on one type of English data focusing on a limited scope of phenomena (most often OntoNotes (Weischedel et al., 2011)).

Our motivation for changing the current situation is the following. First, testing the methods on other languages is crucial, as the properties of how anaphoric relations are expressed may vary considerably (Kunz and Lapshinova-Koltunski, 2015, Novák and Nedoluzhko, 2015). For instance, languages exhibit various extents of pronoun dropping (which is very rare in English), different rules of agreement in grammatical categories hold between the anaphor and its antecedent, or definiteness of noun phrases is expressed in different ways or not at all.

Second, making other anaphoric phenomena available under the same annotation scheme and technical format may attract more attention to their computational modelling.

Last but not least, the lack of multilingual data annotated in a common scheme also hinders theoretical cross-lingual comparative studies of anaphoric phenomena. This all motivates our work in creating harmonized, multilingual and consistent resources for this task.

Our choice of using the UD scheme as the basis for our harmonization efforts in the field of coreference has not only pragmatic reasons (such as the popularity of UD and the fact that numerous technical issues, such as tokenization, are already **standardized** in some way across languages in UD), but is also grounded theoretically. We believe that it will be beneficial to intersect the world of coreference with the world of syntax as much as possible, since entity mentions often correspond to syntactically relevant notions (e.g., noun phrase, subject), some coreference relations are manifested primarily by syntactic means (such as bound reflexive and relative constructions, apposition, predication), zero expressions (such as pro-drop) are needed for coreference and syntax is useful for their identification, and specific syntactic constructions such as coordination interfere with coreference too, to name just a few connections.

The paper is structured as follows. After an overview of previous harmonization efforts for coreference corpora in Section 2, we present the corpora selected for CorefUD (Section 3) and show how diverse they are (Section 4). The CorefUD collection and its harmonization scheme is then described in Section 5. In Section 6, we refer to potential applications of CorefUD, both existing and future ones. Finally, we conclude in Section 7.

2. Related Work

From a broad perspective, any attempt at creating a multilingual coreference corpus that follows the same annotation scheme for all languages can be considered a harmonization effort. Examples of such multilingual corpora are AnCora (Recasens and Martí, 2010, Spanish and Catalan), OntoNotes 5.0 (Weischedel et al., 2011, English, Chinese and Arabic), PCEDT 2.0 (Nedoluzhko et al., 2016, Czech and English), PAWS (Nedoluzhko et al., 2018, Czech, English, Polish and Russian), ParCor (Guillou et al., 2014, English and German), or ParCorFull (Lapshinova-Koltunski et al., 2018, English and German).

If understood in its narrow sense as merging multiple already existing corpora under the same annotation scheme, not many harmonization attempts have been undertaken to date. One of the earliest and broadest ones in terms of the number of languages was the SemEval 2010 Shared task on Coreference Resolution in Multiple Languages (Recasens et al., 2010b). The shared task took advantage of five corpora covering six languages: AnCora (Recasens and Martí, 2010), KNACK-2002 (Hoste and De Pauw, 2006), OntoNotes 2.0 (Pradhan et al., 2007), TüBa-D/Z Treebank (Hinrichs et al., 2005) and LiveMemories (Rodríguez et al., 2010). A unified format for coreference representation was devised. Inspired by CoNLL shared tasks in previous years, it combined columns with gold and automatic morpho-syntactic and semantic information. The last column was reserved for coreference information in an open-close notation with the entity number in parentheses. Identity coreference was the only anaphoric relation annotated in the scheme.

This CoNLL-like format was later adopted in the CoNLL 2011 (Pradhan et al., 2011) and CoNLL 2012 (Pradhan et al., 2012) shared tasks on modeling unrestricted coreference in OntoNotes, which set the standard for representation of identity coreference and for evaluation of coreference resolution.

In the meantime, the XML-based format of annotation produced by the MMAX (Müller and Strube, 2001) and MMAX2 (Müller and Strube, 2006) tools was established as another standard for annotation of a broad variety of linguistic phenomena, including several kinds of anaphora. It has been adopted by multiple corpora of various languages, e.g. ARRAU (Uryupina et al., 2020, English), the Polish Coreference Corpus (Ogrodniczuk et al., 2013, Polish), COREA (Hendrickx et al., 2008, Dutch), the Potsdam Commentary Corpus (Bourgonje and Stede, 2020, German), SzegedKoref (Vincze et al., 2018, Hungarian), and ParCorFull (Lapshinova-Koltunski et al., 2018, English and German). Other corpora were developed using the tabular format of the popular WebAnno tool (Yimam et al., 2013), such as German GerDraCor (Pagel and Reiter, 2020) and English GUM (Zeldes, 2017), which is edited using the GitDox interface (Zhang and Zeldes, 2017). However, when it comes to representing concrete pieces of annotated information, there are numerous variations in how these formats have been used in individual projects.

Only recently, inspired by the Universal Dependencies initiative, the community has started discussions on establishing a universal annotation scheme and using it to harmonize existing corpora. The discussions officially started at the CRAC 2020 workshop (Ogrodniczuk et al., 2020) with a plenary session[1] proposing the Universal Anaphora initiative.[2] CorefUD aims to be our contribution to realizing these goals.

3. Coreference Data Resources

There are dozens of coreference-related annotation projects which have resulted in published datasets and we are clearly unable to analyze and harmonize them all. We therefore combined the following selection criteria to decide which resources to prioritize for inclusion in the present work: license (the freer the better), size (the bigger the better), language diversity (multilingual preferred), annotation schema diversity (we did not want to limit ourselves only to a few families of "genealogically" related projects), and existence of documentation.

The selected resources are listed in Table 1 and described in the rest of this section.[3] The notation introduced in the first column of Table 1 will be used throughout the rest of the paper: we denote each dataset with a label composed of the language name and of a shortcut of the name of the original resource. In addition, there is a horizontal line in Table 1 and all remaining tables in the paper that separates resources available under free licenses from the more restricted ones.

Prague Dependency Treebank (Czech) is a corpus of Czech newspaper texts (∼830K tokens) with manual multi-layer annotation. Coreference and bridging relations are annotated as links on the deep syntactic layer. The arrows lead from the node of the syntactic

[1] `https://sites.google.com/view/crac2020`

[2] `https://github.com/UniversalAnaphora/UniversalAnaphora`

[3] An overview of 20 other coreference resources can be found in Nedoluzhko et al. (2021a); they are candidates for future extensions of our study.

CorefUD dataset	Original name, version	License	Reference
Catalan-AnCora	Coreferentially annotated corpora for Spanish and Catalan	CC BY 4.0	(Recasens and Martí, 2010)
Czech-PCEDT	Prague Czech-English Dependency Treebank	CC BY-NC-SA 3.0	(Nedoluzhko et al., 2016)
Czech-PDT	Prague Dependency Treebank – Consolidated 1.0	CC BY-NC-SA 4.0	(Hajič et al., 2020)
English-GUM	Georgetown University Multilayer Corpus	mixture of CC licenses (none contains ND)	(Zeldes, 2017)
English-ParCorFull	Parallel Corpus Annotated with Full Coreference	CC BY-NC 4.0 (if TED section is omitted)	(Lapshinova-Koltunski et al., 2018)
French-Democrat	Democrat	CC BY-SA 4.0	(Landragin, 2021)
German-ParCorFull	Parallel Corpus Annotated with Full Coreference	CC BY-NC 4.0 (if TED section is omitted)	(Lapshinova-Koltunski et al., 2018)
German-PotsdamCC	Potsdam Commentary Corpus	CC BY-NC-SA	(Bourgonje and Stede, 2020)
Hungarian-SzegedKoref	SzegedKored: Hungarian Coreference Corpus	CC BY 4.0	(Vincze et al., 2018)
Lithuanian-LCC	Lithuanian Coreference Corpus	CLARIN-LT End User License	(Žitkus and Butkienė, 2018)
Polish-PCC	Polish Coreference Corpus	CC BY 3.0	(Ogrodniczuk et al., 2013)
Russian-RuCor	RuCor: Russian Coreference Corpus	CC BY-SA 4.0	(Toldova et al., 2014)
Spanish-AnCora	Coreferentially annotated corpora for Spanish and Catalan	CC BY 4.0	(Recasens and Martí, 2010)
Dutch-COREA	Coreference Corpus and Resolution System for Dutch	a proprietary license	(Hendrickx et al., 2008)
English-ARRAU	The ARRAU Corpus of Anaphoric Information	a mixture of proprietary licenses	(Uryupina et al., 2020)
English-OntoNotes	OntoNotes Release 5.0	LDC	(Weischedel et al., 2011)
English-PCEDT	Prague Czech-English Dependency Treebank	LDC	(Nedoluzhko et al., 2016)

Table 1: Overview of the harmonized coreference resources. The 13 datasets in the upper part are released publicly within the CorefUD 1.0 collection. We can experiment with the 4 datasets in the bottom part only internally because of their license limitations.

head of the anaphor to the node representing the syntactic head of the antecedent and the whole sub-trees of these nodes are considered to be mention spans.

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

Georgetown University Multilayer Corpus (English). GUM 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.

Polish Coreference Corpus (Ogrodniczuk et al., 2013, Ogrodniczuk et al., 2015) is a corpus (∼ 540K tokens) of Polish nominal coreference built upon the National Corpus of Polish. 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) (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 (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 distinguished syntactic heads. Only NPs which take part in coreference relations are considered, singletons are not annotated.

ParCorFull (German and English) 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 (Taulé et al., 2008, Recasens and Martí, 2010) consists 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) 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 (Žitkus and Butkienė, 2018, Lithuanian-LCC) 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 (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.

OntoNotes (English) The English portion of the OntoNotes corpus includes 1.6M tokens in 6 written and spoken genres, annotated for identity coreference and apposition. The corpus does not include markup for singleton mentions. In addition, it contains gold constituent trees, annotations of named entities and PropBank semantic roles, which we currently do not include in CorefUD. For the time being, its Arabic and Chinese parts are not contained in CorefUD.

The ARRAU Corpus of Anaphoric Information (English) covers ∼300K tokens in 5 written and spoken genres, annotated exhaustively for all mentions, including non-referential NPs and entity types, and anaphoric relations including identify coreference, definite predication, apposition, split antecedents and bridging. For the Wall Street Journal portion of the data (∼200K tokens) gold constituent parses and discourse parses are available from other projects.

COREA: Coreference Corpus for Dutch contains more that 140K tokens of written and transcribed oral texts. Mentions are strings of text with specially distinguished heads. Pronouns and full NPs with their dependencies are annotated for coreference and bridging relations. The speciality of the corpus is distinguishing between the level of sense (identity on the type level) and the level of reference (identity on the token level).

PCEDT – the English part consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), with constituency trees automatically converted to dependencies. Coreference-like relations are annotated at the manual layer of deep syntax basically the same way as in the Czech part of the corpus. Unlike its Czech part, the texts in English-PCEDT do not have an open license.

4. Diversity of Annotation Schemes in Coreference Resources

Diversity across corpora with coreference-like annotation can be observed in multiple aspects. We will give a high level overview of most of the differences in the following sections. See also Tables 2 and 3 in Appendix A for a brief overview.

4.1. Mentions

Coreferential relations hold between mentions, which are linguistic expressions, i.e. fragments of texts. There are three ways of how mentions can be represented. As shown in Table 2, they are most frequently defined by a linear span of tokens, usually specified by its start and end tokens, or by offsets (e.g. Russian-RuCor, Lithuanian-LCC). Some corpora also allow discontinuous mentions (e.g. Polish-PCC, English-ARRAU) and may possibly specify a mention head (e.g. Polish-PCC, Dutch-COREA) or a minimal span (English-ARRAU, English-GUM) for fuzzy matching. If a mention is represented by a node in a dependency tree (Czech-PDT, English-Czech-PCEDT), which is actually the mention head, the mention span is understood only implicitly as its subtree and thus requires heuristics to be transformed into the linear span representation. The third way is representing the mention by a node in a constituency tree (e.g. Spanish-Catalan-AnCora).

Prototypically, a mention is a nominal (in UD terminology), meaning a full NP, PP, or a pronoun.

Some projects also allow verbal mentions as antecedents, especially in case of discourse deixis or event anaphora (e.g. Spanish-Catalan-AnCora, Czech-PDT, English-ARRAU), or arbitrary spans (English-GUM). As regards NPs, some corpora limit themselves only to definite NPs (e.g. English-German-ParCorFull or German-PotsdamCC), referring expressions (e.g. English-GUM) or only co-referring ones (e.g. English-OntoNotes). Generic and abstract NPs are often ignored in order to increase inter-annotator agreement (see Zeldes (2022) for discussion and criticism). However, the corpora for most Slavic languages (e.g. Czech-PDT, Polish-PCC, partly Russian-RuCor) have to deal with all types of NPs as these languages do not possess grammaticalized definiteness, which is most often used to distinguish NP types.

CorefUD languages vary in their level of pro-dropping from almost none (e.g. English, German, Dutch, French), through rare subject zeros (Russian), frequent subject zeros (e.g. Czech, Catalan, Spanish, Polish) to zeros also in non-subject positions (e.g. Hungarian), which is also reflected in the coreference corpora for these languages (see Table 2). Beside the subject zeros, corpora built upon the theory of Functional Generative Description (Sgall et al., 1986, e.g. Czech-PDT, English-Czech-PCEDT) introduce syntactic zeros, for example in control constructions and participles. Another type of ellipsis, where syntactic heads of NPs and VPs are omitted, may also take part in anaphoric relations, thus being specially marked by some annotation schemes (e.g. Czech-PDT, Polish-PCC, Spanish-Catalan-AnCora, Lithuanian-LCC). While most schemes technically treat zeros as special nodes/tokens, Polish-PCC marks them directly on a governing verb or its inflection suffix.

4.2. Coreference

There are two styles of grouping mentions with identical reference: *cluster-based* and *link-based* (see Table 3). In the cluster-based style, the basic building block is affiliation of the mention to a named coreference cluster, which is subsequently formed as an equivalence class of all mentions with the same cluster name. The link-based style uses coreference links, each connecting two mentions: an anaphor (or cataphor) and its antecedent (postcedent). Every mention thus has to be labeled by a unique identifier. Representing the link as an edge in a directed graph, clusters then correspond to weakly connected components; that is, unlike in the cluster-based style, data need to be post-processed in order to access the whole cluster. On the other hand, it is harder for the cluster-based style to represent clusters' inner structure or non-equivalence relations, e.g. near-identity or bridging.

Singletons are coreference clusters consisting of only a single mention. Their presence has been shown to affect the performance of coreference resolution (Kübler and Zhekova, 2011), while their absence limits the possible range of linguistic studies, and the approaches that resolution systems trained on the data can take. Nevertheless, as shown in Table 3, they are ignored in many corpora.

Split antecedents (e.g. in the sentence '**My father**$_i$ met **my mother**$_j$ twenty years ago, but **they**$_{i+j}$ got married after I was born.') may be treated in different ways: (1) not annotated at all (e.g. Polish-PCC, Dutch-COREA), (2) as a specific category (e.g. English-ARRAU, French-Democrat, Spanish-Catalan-AnCora), or (3) as a subset type of bridging (e.g. Czech-PDT).

Bridging relations are anaphoric relations between non-coreferential nominal phrases (e.g. the relation 'part – whole' between **apple** and **stub** in the sentence 'I finished my **apple** and threw the **stub** out the window.'). Bridging relations are annotated within some annotation projects (e.g. Czech-PDT, Dutch-COREA, Polish-PCC, English-ARRAU), however the types of the annotated relations differ across annotation schemes substantially.

Apposition and predication (e.g. 'Bob, my father' and 'Bob is my father', respectively) takes place between NPs that refer to the same entity. There is a wide range of identificational and predicative phenomena subsumed under both syntactic environments, most commonly including proper identity coreference ('Elizabeth II is the Queen of England', 'Bob, my father') and predication proper ('Elizabeth is a queen' and 'Elizabeth, a queen'), with predications generally not being bidirectional (the Queen of England is also Elizabeth II, but 'a queen' is not necessarily 'Elizabeth'), but still requiring anaphoric interpretation. As the distinction between various types of apposition, predication and coreference is not clear-cut, it is approached differently in various schemes: (1) ignoring it as a syntactic relation (e.g. Czech-PDT), (2) marking it as a special type (e.g. Russian-RuCor, Spanish-Catalan-AnCora for predication, English-OntoNotes for apposition, English-GUM for both), (3) not distinguishing them from identity coreference (e.g. Dutch-COREA for apposition), or (4) capturing both components in one mention span (e.g. Polish-PCC).

Properties of *discourse deixis* relations differ from the identity coreference between entities, in that the antecedent is often a verb phrase, clause, sentence or a passage of text (often with fuzzy boundaries) which is not normally identified as a mention, whereas the anaphor is limited to some pronouns, shell nouns with demonstratives or definite deverbatives (Webber, 1988), which may be reflected in the corpora. Some corpora annotate it as a special type (e.g. English-ARRAU, English-German-ParCorFull, Spanish-Catalan-AnCora, English-GUM), some as coreference (e.g. English-OntoNotes, Czech-PDT, German-PotsdamCC) and some do not annotate this phenomenon at all (e.g. French-Democrat or Lithuanian-LCC).

4.3. Non-coreferential relations

Some corpora contain annotation of relations beyond identity coreference, including bridging (Clark, 1977, see Table 3), near-identity relations (Recasens et al., 2010a, e.g. Polish-PCC, Dutch-COREA) and bound anaphora (Reinhart, 1983). Categories of *bridging* relations differ considerably across the schemes, with only the part–whole relation present in most corpora. Although due to their semantically-oriented definition, some bridging relations may be interpreted as relations between the entities represented by identity coreference clusters, other types of relations are valid only for particular mentions in the given context, e.g., the relation of contrast annotated in Czech-PDT and Polish-PCC.

Bound anaphora, where an anaphoric pronoun functions as a bound variable referring to non-specific antecedent with a quantifier (e.g. 'Almost **every husband** is proud of **his** wife.'), is treated in different ways: (1) as a special type (Dutch-COREA), (2) as bridging (Polish-PCC), or (3) as identity coreference (Czech-PDT, English-German-ParCorFull, English-GUM).

4.4. Additional coreference-related features

Further information on anaphoric relations is often added, which has almost no formal overlap across corpora. Such features are related to mentions, links[4] as well as entities, and include categorizations of different kinds and granularity, such as entity types (person, place, etc., e.g. in Spanish-Catalan-AnCora, English-ARRAU) or information status (discourse-new, given, etc., e.g. in English-GUM).

4.5. Other NLP annotations

The range of additional linguistic annotation in the corpora is broad, from none (e.g. Lithuanian-LCC) to multilayer annotation including morphology, syntax, named entities, semantic roles, information structure, discourse relations etc. (e.g. English-GUM, Czech-PDT, Spanish-Catalan-AnCora). Most corpora are already tokenized[5] which is important as the mention spans are usually defined on top of the given tokenization. Any change in tokenization during harmonization must thus be done with respect to mention spans, and in some cases, additional annotation layers are only available with different tokenizations which must be mapped (e.g. for corpora based on the Penn Treebank, whose tokenization has changed over time, such as English-ARRAU or English-OntoNotes).

5. CorefUD and its Harmonization Scheme

The main building blocks in the target representation are **mentions** and **clusters**. A mention in our scheme is a set of words in the sense of UD, that is, nodes in the dependency structure, including empty nodes – zeros. Mentions spanning multiple sentences are supported, too. A mention is specified by its span, i.e., the nodes it contains. Spans of two different mentions can overlap but they cannot be identical. While a typical mention is a contiguous span of the surface text, this is not a requirement and discontinuous mentions are allowed. Analogously, from the perspective of the dependency structure, a typical mention is a connected component of a dependency tree (*catena*, following (Osborne et al., 2012)), yet we do not require this to be the case, and for automatically parsed corpora we expect recurring violations of this expectation (Popel et al., 2021).

Every mention is a member of one (and only one) cluster; the cluster contains all mentions referring to the same **entity** (incl. events). Singletons are clusters that contain only one mention. The entity/cluster ID is thus a required attribute of each mention, besides the mention's span.[6] Mentions have additional attributes, some of which pertain to the whole cluster.

5.1. File Format

Our main objective is maximum compliance with the current UD standards. We avoid decisions that would prevent our data from becoming part of a regular UD release.[7]

We adhere to the specification of the **CoNLL-U format**[8] (as opposed to the CoNLL-U Plus extension,[9] which would allow for extra columns for the coreference-related attributes, but unfortunately would disqualify the data from UD releases). We make sure that the harmonized data pass the official UD validation at level 2 (passing the higher levels may not be possible with automatically predicted POS tags and dependency relations).[10]

[4] Link-related features are annotated also in some cluster-based corpora for a mention, usually meaning the link to its nearest antecedent, for example distinguishing cataphora from anaphora.

[5] By contrast, many of them do not capture original untokenized sentence text.

[6] Cluster IDs are unique across one corpus within CorefUD. For example, `e1` refers to the same cluster everywhere in Czech-PDT but it is not related to `e1` in Czech-PCEDT. This is mainly to prevent confusion when interpreting the data. For coreference purposes it would be sufficient to make the IDs unique within one document, if the corpus has internal document boundaries.

[7] Note however that UD has additional requirements, which only some of our datasets comply with. Most notably, a UD-released treebank must have manually checked POS tags and dependency relations; in most of our datasets, this kind of annotation has been assigned automatically.

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

[9] https://universaldependencies.org/ext-format.html

[10] https://universaldependencies.org/validation-rules.html#levels-of-validity

From the perspective of the CoNLL-U format, coreference is additional annotation that belongs to the MISC column (column 10). While we deliberately avoid the CoNLL-U Plus file format, we argue that this option is very close to it, and users who prefer additional columns for coreference annotation can easily extract the coreference-related attributes from MISC and place them in separate columns, using tabs instead of the MISC column's pipe separators.

The main attribute that we add to the MISC column is called Entity and it identifies all mentions that begin or end at the current word. In the value of the attribute, each mention has an opening or closing bracket, accompanied by the entity/cluster ID. Additional mention attributes are specified at the opening bracket, i.e., at the first word of the mention. For example, Entity=(e8-place-1)e9) means that the current word is the entire span of one mention of entity e8, the corresponding entity type is a place, and the first (and only) word of the mention is also its syntactic head; furthermore, the attribute says that this is the last word of a larger mention belonging to cluster e9, which started at one of the previous words.

In case of a discontinuous mention, each part has its number and the total number of parts in square brackets after the cluster ID: Entity=(e10[1/2] ... Entity=e10[1/2]) ... Entity=(e10[2/2] ... Entity=e10[2/2]).

For an example of the CoNLL-U representation see Figure 1.

Zeros. Universal Dependencies provide a mechanism for inserting **empty nodes** (which may or may not have lexical values assigned to them) in the enhanced dependency graph. We use the empty nodes to represent reconstructed zeros.

Singletons. Both singletons and non-singletons are treated as clusters; a singleton cluster contains just a single mention. As a result, there are substantially more unique cluster IDs for the annotation projects that include annotation of singletons. In future versions, we may add singletons to datasets which did not have them originally, using the UD annotations and/or entity recognition tools.

Bridging. In the current version, bridging relations are understood very broadly as all relations annotated in the source schemes that cannot be considered types of identity coreference. To record bridging relations, we use the MISC attribute Bridge. It connects identity clusters, where one cluster may be part of more than one bridging relation. For example, Bridge=e173<e188:subset,e174<e188:part says that cluster e188 is related to cluster e173 with the subset bridging relation, and to cluster e174 with the part-whole bridging relation. The annotation appears at a selected mention of cluster e188; it is not repeated at the other mentions of that cluster.

Split antecedents. The MISC attribute SplitAnte points from a cluster to two or more other clusters. For example, SplitAnte=e5<e61,e10<e61 means that cluster e61 anaphorically refers to clusters e5 and e10. The attribute is a property of clusters, saying that the entity with a given cluster ID is equivalent to the union of the smaller entities whose IDs are listed in the value of the attribute. The annotation appears at a selected mention of cluster e61; it is not repeated at the other mentions of that cluster.

Attributes of clusters and mentions. There are three "standardized" attributes: eid (entity/cluster ID), etype (entity type) and head (index of the head word), stored as a hyphen-separated list. Other attributes may follow. In CorefUD version 1.0, we just copy these additional attributes from the original annotation schemes. In future versions, we anticipate adding a number of modifications to unify the data further, for example the distinction between specific and generic NPs.

5.2. Adding UD Annotations

Some of the original corpora, especially those that have already been part of UD, contain all morpho-syntactic annotation required by the CoNLL-U format (e.g. English-GUM) or such annotation can be obtained by already available conversion (e.g. Czech-PDT). Where this is not possible, we enrich the corpora with additional annotation automatically, employing UDPipe 2.0[11] (Straka, 2018) and its models trained on UD 2.6. The automatic processing includes lemmatization, part-of-speech tagging (including morphological features), and dependency parsing.

5.3. Train/dev/test Splits

We divide each CorefUD dataset into a training section, a development section, and a test section (train/dev/test for short) in order to facilitate reproducibility and comparability of future machine learning experiments. Technically, each CorefUD dataset consists of three CoNLL-U files containing disjoint sets of documents; boundaries between the three sections can be placed only on document boundaries.

If such a division was indicated already in the original resource, then we preserved the division. Otherwise, we iterated along the sequence of documents present in the original dataset and repeatedly put 8 documents into train, 1 document into dev, and 1 into test.The resulting division, as well as total sizes of all CorefUD datasets in terms of the number of documents, sentences, and words is summarized in Table 4 in Appendix B.

5.4. Releasing and Licensing Policy

We have divided the harmonized data into two parts: The larger part is public and contains only resources whose original versions come with free licenses (recall the last column in Table 1) that allow modification and

[11] https://ufal.mff.cuni.cz/udpipe/2

```
# global.Entity = eid-etype-head-minspan-infstat-link-identity
# sent_id = GUM_academic_art-3
# text = Claire Bailey-Ross xxx@port.ac.uk University of Portsmouth,
         United Kingdom
1   Claire      Claire      PROPN  NNP  Number=Sing  0  root   0:root
    Entity=(e5-person-1-1,2,4-new-coref|Discourse=attribution:3->57:7
2   Bailey      Bailey      PROPN  NNP  Number=Sing  1  flat   1:flat
    SpaceAfter=No|XML=<w>
3   -           -           PUNCT  HYPH _            4  punct  4:punct
    SpaceAfter=No
4   Ross        Ross        PROPN  NNP  Number=Sing  2  flat   2:flat
    Entity=e5)|XML=</w>
5   xxx@port.ac.uk xxx@...  PROPN  NNP  Number=Sing  1  list   1:list
    Entity=(e6-abstract-1-1-new-sgl)
6   University  University  PROPN  NNP  Number=Sing  1  list   1:list
    Entity=(e7-organization-1-3,5,6-new-sgl-University_of_Portsmouth
7   of          of          ADP    IN   _            8  case   8:case    _
8   Portsmouth  Portsmouth  PROPN  NNP  Number=Sing  6  nmod   6:nmod:of
    Entity=(e8-place-1-3,4-new-sgl-Portsmouth|SpaceAfter=No
9   ,           ,           PUNCT  ,    _            11 punct  11:punct  _
10  United      unite       VERB   NNP  Tense=Past|... 11 amod 11:amod
    Entity=(e9-place-2-1,2-new-coref-United_Kingdom
11  Kingdom     Kingdom     PROPN  NNP  Number=Sing  1  list   1:list
    Entity=e9)e8)e7)
```

Figure 1: Example of the CoNLL-U encoding of English-GUM in CorefUD.

redistribution (at least for non-commercial purposes). This public edition is available as CorefUD 1.0 in the LINDAT/CLARIAH-CZ repository.[12]

The other part composed of the remaining resources is non-public: we can include it in our internal experimentation and can report statistics collected from the data, but we cannot redistribute it to users who do not have access to the underlying data.

We distribute the public resources under the same licenses that the original resources came with. As a result, the CorefUD 1.0 package has a mixed license, with different terms applying to different datasets (a license file is stored with each dataset).

6. Applications of CorefUD

CorefUD harmonization is far from being finished. For instance, when looking into basic quantitative characteristics of the individual datasets in Tables 5 and 6 in Appendix B, one can find some correlates of general linguistic expectations (for example, languages without determiners tend to have shorter mentions on average). There are also differences among the datasets which can be attributed rather to design choices of the original resources (such as substantially different distributions of mention lengths across resources for the same language, or amounts of singletons).

However, CorefUD is becoming useful for linguistically interpretable theoretical, typological and NLP research already now. The first proof-of-concept version of CorefUD described in the technical report (Nedoluzhko et al., 2021a) has been used for a study on the relation between UD trees and independently annotated mention spans (Popel et al., 2021), and in a study focused on differences between UD-induced heads and heads of mentions annotated independently in some coreference resources (Nedoluzhko et al., 2021b). In addition, the preliminary version of CorefUD has also been employed in pilot experiments with multilingual coreference resolution (Pražák et al., 2021). We are organizing a shared task on multilingual coreference resolution,[13] and hope that it will boost the field, similarly to the impact of dependency parsing tasks on parser development in the past.

7. Conclusions

The most important contributions of this work are the following: (1) we presented a survey of coreference-related resources, emphasizing their diversity from various viewpoints; to the best of our knowledge, no comparably broad and detailed survey has been published to date, and (2) we designed a common scheme and implemented automatic converters of source datasets into this unified scheme, and released a part of the collection publicly under the name CorefUD 1.0; again, this is the widest coreference data collection we are aware of, and a first push in the direction of exposing multilingual coreference data to users in a carefully crafted, unified format, adhering to and compatible with UD design principles. We plan to continue work on adding new languages and datasets, enhancing the data to represent as much information as possible, and harmoniz-

[12] http://hdl.handle.net/11234/1-4698

[13] https://ufal.mff.cuni.cz/corefud/crac22

ing that information to promote convergence and standardization in the area of coreference.

8. Acknowledgements

This work was supported by the Grants GA19-14534S and 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; and EC/H2020/825303 (Bergamot) of the European Commision.

9. Bibliographical References

Bourgonje, P. and Stede, M. (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, May. European Language Resources Association.

Clark, H. H. (1977). Bridging. In P. N. Johnson-Laird et al., editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, London and New York.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Dobrovolskii, V. (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, November. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., and Štěpánková, B. (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.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J., and Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Hinrichs, E. W., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 13–20, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Hoste, V. and De Pauw, G. (2006). KNACK-2002: a richly annotated corpus of Dutch written text. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Kirstain, Y., Ram, O., and Levy, O. (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, August. Association for Computational Linguistics.

Kübler, S. and Zhekova, D. (2011). Singletons and coreference resolution evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria, September. Association for Computational Linguistics.

Kunz, K. and Lapshinova-Koltunski, E. (2015). Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14:258–288, 03.

Landragin, F. (2021). Le corpus Democrat et son exploitation. Présentation. *Langages*, 224:11–24, December.

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). ParCorFull: a Parallel Corpus Annotated with Full Coreference. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Lüdeling, A., Ritz, J., Stede, M., and Zeldes, A. (2016). Corpus linguistics and information structure research. In Caroline Féry et al., editors, *The Oxford Handbook of Information Structure*, pages 599–617. Oxford University Press, Oxford.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Müller, C. and Strube, M. (2001). Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. In *Corpus Tech-*

nology and Language Pedagogy: New Resources, New Tools, New Methods.

Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., and Mírovský, J. (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, May. European Language Resources Association (ELRA).

Nedoluzhko, A., Novák, M., and Ogrodniczuk, M. (2018). PAWS: A multi-lingual parallel treebank with anaphoric relations. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 68–76, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., and Zeman, D. (2021a). Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. Technical Report 66, ÚFAL MFF UK, Praha, Czechia.

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., and Zeman, D. (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.

Novák, M. and Nedoluzhko, A. (2015). Correspondences between Czech and English Coreferential Expressions. *Discours: Revue de linguistique, psycholinguistique et informatique.*, 16:1–41.

Ogrodniczuk, M., Glowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (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.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Maciej Ogrodniczuk, et al., editors. (2020). *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, Barcelona, Spain (online), December. Association for Computational Linguistics.

Osborne, T., Putnam, M., and Groß, T. (2012). Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Pagel, J. and Reiter, N. (2020). GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 55–64, Marseille, France, May. European Language Resources Association.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Popel, M., Žabokrtský, Z., Nedoluzhko, A., Novák, M., and Zeman, D. (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, November. Association for Computational Linguistics.

Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007). Ontonotes: a unified relational semantic representation. *International Journal of Semantic Computing*, 1(4):405–419.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (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, July. Association for Computational Linguistics.

Pražák, O., Konopík, M., and Sido, J. (2021). Multilingual coreference resolution with harmonized annotations. *arXiv preprint arXiv:2107.12088*.

Recasens, M. and Martí, M. A. (2010). AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Lang. Resour. Eval.*, 44(4):315–345, December.

Recasens, M., Hovy, E., and Martí, M. A. (2010a). A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010b). 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.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6:47–88.

Rodríguez, K. J., Delogu, F., Versley, Y., Stemle, E. W., and Poesio, M. (2010). Anaphoric annotation of

Wikipedia and blogs in the live memories corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague/Dordrecht.

Straka, M. (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, October. Association for Computational Linguistics.

Taulé, M., Martí, M. A., and Recasens, M. (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, May. European Language Resources Association (ELRA).

Toldova, S., Roytberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., Sim, G., Gorshkov, D., Ivanova, A., Nedoluzhko, A., and Grishina, Y. (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.

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., and Poesio, M. (2020). Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering*, 26(1):95–128, January.

Vincze, V., Hegedűs, K., Sliz-Nagy, A., and Farkas, R. (2018). SzegedKoref: A Hungarian coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122, Buffalo, New York, USA, June. Association for Computational Linguistics.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. (2011). Ontonotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63, New York. Springer-Verlag.

Wilkens, R., Oberle, B., Landragin, F., and Todirascu, A. (2020). French coreference for spoken and written language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 80–89, Marseille, France, May. European Language Resources Association.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of ACL 2013*, pages 1–6, Sofia, Bulgaria.

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

Zeldes, A. (2022). Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Zhang, S. and Zeldes, A. (2017). GitDOX: A linked version controlled online XML editor for manuscript transcription. In *Proceedings of FLAIRS-30*, pages 619–623, Marco Island, FL.

Žitkus, V. and Butkienė, R. (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. Diversity of coreference representation in original resources

	Mention representation		Reconstructed zeros	
original corpus	linear span	syn/sem. head	null subj.	nom. ellips.
Catalan-AnCora	✓	✓	✓	✓
Czech-PCEDT	×	✓	✓	✓
Czech-PDT	×	✓	✓	✓
English-GUM	✓	(✓)	×	×
English-ParCorFull	✓	×	×	✓
French-Democrat	✓	(✓)	×	×
German-ParCorFull	✓	×	×	✓
German-PotsdamCC	✓	×	×	×
Hungarian-SzegedKoref	✓	(✓)	✓	×
Lithuanian-LCC	✓	×	×	✓
Polish-PCC	✓	✓	✓	✓
Russian-RuCor	✓	✓	×	×
Spanish-AnCora	✓	✓	✓	✓
Dutch-COREA	✓	✓	×	×
English-ARRAU	✓	×	×	×
English-OntoNotes	✓	(✓)	×	×
English-PCEDT	×	✓	✓	✓

Table 2: Diversity of coreference-related annotations in the original corpora: properties of mentions. Brackets around the check mark mean that this kind of information has not been completed manually within the annotation of coreference-related phenomena, but it can be obtained from other annotation layers (mostly, from the syntactic annotation.)

	Relations among mentions							
CorefUD dataset	cluster-based identity	link-based identity	single-tons	appos.	pred.	split antec.	disc. deixis	bridg.
Catalan-AnCora	✓	×	✓	✓	✓	✓	✓	×
Czech-PCEDT	×	✓	(✓)	(✓)	(✓)	✓	✓	×
Czech-PDT	×	✓	(✓)	(✓)	(✓)	✓	✓	✓
English-GUM	✓	×	✓	✓	✓	✓	✓	✓
English-ParCorFull	✓	×	×	✓	(✓)	✓	✓	×
French-Democrat	✓	×	✓	×	×	×	×	×
German-ParCorFull	✓	×	×	✓	(✓)	✓	✓	×
German-PotsdamCC	×	✓	✓	✓	✓	×	✓	×
Hungarian-SzegedKoref	✓	×	×	✓	?	×	✓	✓
Lithuanian-LCC	×	✓	×	×	×	✓	×	×
Polish-PCC	✓	×	✓	✓	✓	×	✓	✓
Russian-RuCor	✓	×	×	✓	✓	×	×	×
Spanish-AnCora	✓	×	✓	✓	✓	✓	✓	×
Dutch-COREA	×	✓	✓	✓	✓	×	✓	✓
English-ARRAU	✓	✓	✓	✓	✓	✓	✓	✓
English-OntoNotes	✓	×	×	✓	×	×	✓	×
English-PCEDT	×	✓	(✓)	(✓)	(✓)	✓	✓	×

Table 3: Diversity of coreference-related annotations: types of relations among mentions. Brackets around the check sign mean that this kind of information has not been completed manually within the annotation of coreference-related phenomena, but it can be obtained from other annotation layers (mostly, from the syntactic annotation.

B. Statistical properties of CorefUD 1.0

CorefUD dataset	total size				division [%]		
	docs	sents	words	empty	train	dev	test
Catalan-AnCora	1550	16,678	546,665	6,377	78.5	10.6	10.9
Czech-PCEDT	2312	49,208	1,155,755	43,054	80.9	14.2	4.9
Czech-PDT	3165	49,428	834,721	32,617	78.3	10.6	11.1
English-GUM	175	9,130	164,392	92	75.9	11.9	12.1
English-ParCorFull	19	543	10,798	0	81.2	10.7	8.1
French-Democrat	126	13,054	284,823	0	80.1	9.9	10.0
German-ParCorFull	19	543	10,602	0	81.6	10.4	8.1
German-PotsdamCC	176	2,238	33,222	0	80.3	10.2	9.5
Hungarian-SzegedKoref	400	8,820	123,968	4,857	81.1	9.6	9.3
Lithuanian-LCC	100	1,714	37,014	0	81.3	9.1	9.6
Polish-PCC	1828	35,874	538,885	470	80.1	10.0	9.9
Russian-RuCor	181	9,035	156,636	0	78.9	13.5	7.6
Spanish-AnCora	1635	17,662	559,782	8,112	80.9	9.5	9.6
Dutch-COREA	844	9,270	140,063	0	78.6	10.0	11.4
English-ARRAU	413	9,540	228,901	0	81.2	4.3	14.5
English-OntoNotes	3493	94,269	1,631,995	0	79.6	10.0	10.4
English-PCEDT	2312	49,208	1,173,766	36,115	80.9	14.2	4.8

Table 4: Data sizes and train/dev/test split (in words) of CorefUD data sets. If this division was already present in an original resource, then we preserved the division, otherwise iteratively divided the dataset's documents in 8/1/1 fashion (see Section 5.3 for details). 'words' is the number of non-empty UD nodes (corresponding to syntactic words). 'empty' is the number of empty UD nodes.

CorefUD dataset	clusters				distribution of lengths				
	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	69,239	127	101	1.6	74.6	14.1	4.7	2.2	4.4
Czech-PCEDT	52,743	46	247	3.4	1.4	62.7	15.6	6.8	13.4
Czech-PDT	78,880	94	186	2.5	35.3	38.9	11.0	5.2	9.5
English-GUM	24,801	151	131	1.9	74.5	13.9	4.8	2.1	4.7
English-ParCorFull	180	17	38	4.0	6.1	55.0	13.9	6.7	18.3
French-Democrat	40,937	144	895	2.0	81.8	10.6	3.0	1.3	3.2
German-ParCorFull	259	24	43	3.5	6.2	64.9	11.6	5.0	12.4
German-PotsdamCC	3,752	113	15	1.4	76.5	13.9	5.0	1.8	2.7
Hungarian-SzegedKoref	5,182	42	36	3.0	8.0	51.1	19.0	9.1	12.9
Lithuanian-LCC	1,224	33	23	3.7	11.2	45.3	11.8	8.2	23.5
Polish-PCC	127,688	237	135	1.5	82.6	9.8	2.9	1.4	3.2
Russian-RuCor	3,636	23	141	4.5	3.3	53.7	15.6	6.9	20.5
Spanish-AnCora	73,210	131	110	1.7	73.4	14.8	4.7	2.4	4.7
Dutch-COREA	28,455	203	31	1.2	88.3	8.3	2.0	0.6	0.8
English-ARRAU	48,333	211	163	1.5	83.0	8.9	3.2	1.5	3.4
English-OntoNotes	51,557	32	217	4.1	0.4	58.3	15.4	7.4	18.5
English-PCEDT	54,514	46	258	3.4	1.2	62.4	15.9	7.0	13.5

Table 5: Statistics on coreference clusters. The total number of clusters and the average number of clusters per 1000 tokens in the running text. The maximum and average cluster "length", i.e., number of mentions in the cluster. Distribution of cluster lengths. Note that certain amount of singleton clusters (length = 1) occur even in datasets that do not target singletons. It is because we create clusters also for mentions that participate in bridging.

	mentions				distribution of lengths					
CorefUD dataset	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	62,416	114	141	4.8	10.2	28.2	21.7	7.9	5.3	26.8
Czech-PCEDT	178,376	154	79	3.5	23.0	28.6	16.1	8.3	4.0	20.0
Czech-PDT	169,545	203	99	2.9	17.2	36.4	18.7	8.5	4.0	15.1
English-GUM	28,054	171	95	2.6	0.0	55.6	20.0	8.1	3.9	12.4
English-ParCorFull	718	66	37	2.1	0.0	58.9	24.5	6.0	2.9	7.7
French-Democrat	47,172	166	71	1.7	0.0	64.2	21.7	6.4	2.5	5.3
German-ParCorFull	896	85	30	2.0	0.0	64.8	17.5	6.2	4.0	7.4
German-PotsdamCC	2,519	76	34	2.6	0.0	34.8	32.4	15.6	6.4	10.9
Hungarian-SzegedKoref	15,165	122	36	1.6	15.1	37.4	32.5	10.2	2.6	2.2
Lithuanian-LCC	4,337	117	19	1.5	0.0	69.1	16.6	11.1	1.2	2.0
Polish-PCC	82,804	154	108	2.1	0.3	68.7	14.9	5.2	2.7	8.2
Russian-RuCor	16,193	103	18	1.7	0.0	69.1	16.3	6.6	3.5	4.6
Spanish-AnCora	70,664	126	101	4.8	11.4	31.6	18.8	7.2	4.5	26.3
Dutch-COREA	8,623	62	60	2.6	0.0	42.6	33.2	8.6	4.0	11.6
English-ARRAU	31,895	139	75	2.9	0.0	45.4	26.9	10.7	4.2	12.8
English-OntoNotes	209,425	128	94	2.5	0.0	56.3	19.8	8.1	4.2	11.7
English-PCEDT	183,836	157	91	3.6	19.3	28.1	16.9	10.7	4.8	20.2

Table 6: 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 non-empty nodes in the mention. Distribution of mention lengths.