M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman

ÚFAL Technical Report
TR-2021-66

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

CorefUD 0.1

Coreference meets Universal Dependencies –
a pilot experiment on harmonizing coreference datasets
for 11 languages

ÚFAL Technical Report

Anna Nedoluzhko, Michal Novák, Martin Popel,
Zdeněk Žabokrtský, Daniel Zeman

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

April 9, 2021

Abstract

We describe a pilot experiment aimed at harmonizing diverse data resources that contain coreference-related annotations. We converted 17 existing datasets for 11 languages into a common annotation scheme based on Universal Dependencies, and released a subset of the resulting collection publicly under the name CorefUD 0.1 via the LINDAT-CLARIAH-CZ repository (`http://hdl.handle.net/11234/1-3510`).

All the datasets included in CorefUD 0.1 are enriched with automatic morphological and syntactic annotations which are fully compliant with the standards of the Universal Dependencies project. The datasets are stored in the CoNLL-U format, with coreference- and bridging-specific information captured by attribute-value pairs located in the MISC column.

This report describes our current knowledge and the results of the harmonization procedure valid in March 2021; see `https://ufal.mff.cuni.cz/corefud` for future updates.

Contents

Chapter 1

Introduction

This document presents our pilot experiment on unification of seventeen datasets annotated with coreference-related phenomena. The common target scheme, into which the datasets were converted, is based on standards of the Universal Dependencies (UD) project.[1] Specifically, it uses the CoNLL-U file format defined in UD, and can be validated using the official UD validation script.

There are three main sources of inspiration for our efforts. First, we are inspired by numerous interference points between coreference and syntax, and we want to give these interferences a chance. A coreference component has been considered an integral part of deep-syntactic treebank annotations in the family of Prague treebanks; more specifically, massive annotations of coreference were present in the tectogrammatical layer of the Prague Dependency Treebank from version 2.0 (Hajič et al., 2006). We believe that merging coreference and syntactic annotations in a single resource can be fruitful from quite a few perspectives, such as the following ones:

- many referring expressions have the form of syntactic constituents, and once we have syntactic annotation at our disposal, it's immediately clear what is the head of each expression (the head's morphological categories are important for anaphora resolution); in our context, a referring expression corresponds mostly to a single node or to a connected subgraph of a dependency tree,
- certain subtypes of coreference relations are manifested primarily by syntactic means, such as in the case of relative constructions, reflexive constructions, apposition, or predication constructions with copula verbs,
- we can reuse annotation of coordination structures present in syntactic trees (for example if we want to refer to a coordination construction as a whole),
- zero expressions are important in many coreference frameworks, e.g. because of pro-drop that is frequent in some languages; again, such zeros are easier to detect in a sentence if its syntactic structure is known to us.

The second source of inspiration comes from UD. Out of various attempts at data harmonization in the treebanking world, UD proved to be the most viable one, covering more than 100 languages nowadays. It is hard to identify the most decisive factors behind the UD success in an exact way.

[1] `https://universaldependencies.org`

However, in our scheme we decided to follow not only UD's basic annotations conventions for individual linguistic phenomena (tokenization, POS values, etc.) and its file format (CoNLL-U), but also UD's very pragmatic data-driven decision making and its extreme pressure on simplicity of the scheme.

Third, we are aware of discussions by Massimo Poesio and others within the Universal Anaphora (UA) Initiative,[2] and their thoughts presented e.g. at the CRAC workshop at COLING 2020. At the time being, we do not take many theoretical decisions on unification of various anaphoric phenomena, and we think about unifying the annotation scheme decisions only after observing how it was done in as many existing data resources as possible. In other words, our approach is rather bottom-up oriented and less theory-driven, compared to that of UA. However, we believe to find an opportunity for merging the two initiatives in the future.

The rest of this report is structured as follows:

- In the remainder of this chapter, we present our selection criteria for including a data resource into our pilot study, and list the selected resources; we briefly outline previous attempts at harmonizing coreference data (especially within shared tasks), and explain basic linguistic notions that are needed to speak about coreference annotations.
- Chapter 2 describes the individual data resources included into our harmonization study in more detail, and lists some other resources which we were unable to process either because of limitations of our capacity, or because of their availability limitations.
- Chapter 3 analyzes various sources of diversity in observed annotation schemes used in the resources under study.
- Chapter 4 describes our harmonized scheme in more detail, including the way how we store coreference information in the CoNLL-U file format.
- Chapter 5 presents properties of the resulting data collection; due to license limitations the collection is divided into a public part (13 harmonized datasets for 10 languages) and non-public part (4 harmonized datasets for 2 languages). We present statistical properties of datasets included in both parts.
- We conclude and outline possible future directions in Chapter 6.

1.1 Selection of data resources for harmonization

As there are dozens of coreference-related annotation projects which resulted in some published datasets, it was clear from the very beginning that sampling is inescapable. When choosing the subset of data resources for the first round of harmonization experiments, we identified the following selection criteria:

- license – we preferred resources published under open licenses,
- size – we preferred resources with at least medium-sized annotated data (various toy annotation mini-projects were avoided),

[2] `https://universalanaphora.github.io/UniversalAnaphora/`

- language diversity – we preferred gathering data in multiple languages instead of just many variants of English data,
- technical diversity – we did not want to limit ourselves to a single family of "genealogically" related coreference projects (annotated using the same software tool, stored in the same file format, etc.),
- annotation diversity – we prefer resources which contain more thorough annotation of coreference-related phenomena (e.g. also near-identity, bridging) and mark the relations on different types of mentions (not just on pronouns, but also on full noun phrases, verbs as antecedents, pronominal adverbs, etc.)
- documentation – we prefer resources whose annotation scheme is well documented, ideally in English,
- only writing systems readable to us – so far, we worked only with languages which use Latin-based or Cyrilic alphabets,
- last but not least, we started with datasets which we had some prior experience with.

Clearly, the criteria sometimes go against each other and some of them are subjective, so the selection was still arbitrary to some extent, but at least not entirely random.

After some 4 weeks of implementing prototypes of various converters and studying the differences among the resources, we ended up with 17 resources listed in Table 1.1. Then, in the following month or so, we elaborated the target scheme and the individual converters further, performed some tests, evaluated statistical quantities etc. The work has been done by four part-time programmers and one part-time linguist.

The notation introduced in the first column of Table 1.1 will be used throughout the rest of the report: we denote each dataset with a label composed of the language name and of a shortcut of the name of the original resource. It is useful especially in the case of multilingual resources such as the Prague Czech-English Dependency Treebank.

Table 1.1 (as well as several tables in the following chapters) is divided vertically into two parts, rendering the fact that – due to license limitations of the source datasets – our harmonized collection must have been divided into a public and a non-public (ÚFAL-internal) edition. The publicly available edition is distributed via LINDAT-CLARIAH-CZ and contains 13 datasets for 10 languages. The non-public edition is available internally to ÚFAL members and contains additional 4 datasets for 2 languages.

1.2 Design decisions concerning the harmonization procedure

Details concerning the target annotation scheme will be given in Chapter 4; however, we outline some main technical design decisions already here:

- the resulting data must be technically perfectly compliant with the current UD standards (e.g., the data must pass CoNLL-U file format validation),
- we attempt to make all harmonization decisions as language-agnostic as possible, recycling the multilingual experience accumulated within UD, again,

CorefUD dataset	Original name, version	License	Reference
Catalan-AnCora	Coreferentially annotated corpora for Spanish and Catalan	GNU GPL 3.0	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 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 (2016)
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	GNU GPL 3.0	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	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 1.1: Overview of the harmonized coreference resources. The 13 datasets in the upper part are released publicly within the CorefUD 0.1 collection. We can experiment with the 4 datasets in the bottom part only internally because of their license limitations.

- we will try to harmonize only those types of annotated information that are present in multiple resources; non-harmonized pieces of information will be preserved in some form in the target data in most cases too,
- the whole harmonization pipeline must be fully automatic and deterministic, so that harmonized resources can be easily re-generated again (e.g., after a bug fix).

1.3 Previous attempts at harmonizing coreference resources

From a wider perspective, any attempt on creating a multilingual coreference corpus that follows the same annotation scheme for all languages can be considered a harmonization effort. This holds especially for corpora combining linguistically distant languages. 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 occurred until now. One of the earliest and broadest one 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 in 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 Live-Memories (Rodríguez et al., 2010). Since the corpora were originally in different schemes, a unified format of coreference representation was devised. It was inspired by CoNLL shared tasks in previous years, combining 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. No other anaphoric relation than identity coreference was annotated in this scheme.

This CoNLL-like format was later adopted in the CoNLL 2011 Shared task on Modeling unrestricted coreference in OntoNotes (Pradhan et al., 2011) and in the CoNLL 2012 Shared task on Modelling multilingual unrestricted coreference in OntoNotes (Pradhan et al., 2012). These two shared tasks based on the OntoNotes data (Weischedel et al., 2011) set the standard for representation of identity coreference and for evaluation of the systems modelling this type of relation. The format was also adopted by the CORBON 2017 Shared task on Projection-based coreference resolution (Grishina, 2017), which employed the English-German-Russian parallel corpus annotated with coreference by Grishina and Stede (2015) as test data.

In the meantime, the XML-based format of annotation produced by the MMAX (Müller and Strube, 2001) and MMAX2 (Müller and Strube, 2006) tools has been established as another standard for annotation of broad variety of linguistic phenomena, including anaphora. It has been adopted by multiple corpora of various languages, e.g. ARRAU (Uryupina et al., 2020, English), Polish Coreference Corpus (Ogrodniczuk et al., 2013, Polish), COREA (Hendrickx et al., 2008, Dutch), Potsdam Commentary Corpus (Bourgonje and Stede, 2020, German), SzegedKoref (Vincze et al., 2018, Hungarian), and ParCorFull (Lapshinova-Koltunski et al., 2018, English and German). However, when it comes to representing concrete pieces of annotated information, there are numerous variations in how the MMAX format is used in individual projects. For example, the annotation of a document is spread in a different number of XML files in different projects, and there are diverse sets of attributes attached to individual mentions, different ways how sentence boundaries are captured (if captured at all), different ways how mentions are grouped into coreference clusters, different ways how mentions are classified (if classified at all), and different typologies of bridging relations.

Only recently, inspired by the Universal Dependencies initiative, the community started discussions on establishing the universal schema and harmonizing the existing corpora under this schema. The discussions officially started at the CRAC 2020 workshop (Ogrodniczuk et al., 2020) with a ple-

nary session[3] proposing the Universal Anaphora initiative[4] among others. CorefUD aims to be our contribution to these discussions.

1.4 Used terminology

Similarly as in many other research fields which are subjects to investigation in more than one scientific community, there is a broad terminology variation in the field of anaphora and coreference. Since we work with various annotation approaches at the same time, especially with the aim of unifying them into one common scheme, we need to understand very accurately what we mean by specific notions. So, let us agree on terms.

Coreference is the relation between language expressions which refer to the same referent. This is the identity relation only. However, since this term is explored intensively, it became well-known and acquired additional connotations. For example, when we call our project *Coreference meets Universal Dependencies*, we mean not only identity relations but also many other related anaphoric or near-to-coreference relations. Therefore, in what follows, we will also use expressions **identity relation** or **identity coreference** if we want to emphasize the identity.

Anaphora, or an anaphoric relation, is a contextual reference to an expression in the previous context. Anaphoric expressions are usually also coreferential, but it's not always the case. For example, the relation between *this apple* and *it* in Example 1 is both coreferential and anaphoric, the relation between *this apple* and *one* in Example 2 is anaphoric but not coreferential, and the relation between two instances of *London* in Example 3 is coreferential but not anaphoric.

(1) **This apple** is mine. Don't eat **it**!

(2) **This apple** is mine. Please, take another **one**.

(3) Peter lives in **London**. Steve lives in **London**.

In a coreferential, as well as in an anaphoric relation, the referring (right, second) expression is called **anaphor**, and the expression which is referred to is called **antecedent**. In Example 1, *it* is the anaphor and *this apple* is the antecedent. An antecedent may be contiguous or split, as in Example 4, where the anaphoric pronoun *them* refers to *an apple* and *another one* in the previous sentences. Annotation coreference relations with discontinuous antecedents is nontrivial for annotation on linear texts, it requires setting up new attributes and special solutions. Therefore, the notion of **split antecedent** is a special topic in coreference annotation discussion.

(4) Mary gave Peter **an apple**. Steve gave him **another one**. Peter took **them** and left.

Besides anaphoric, there are also **cataphoric** relations, which is textual reference to an expression in the following context (see the relation between *he* and *John* in Example 5). Although quite infrequent, cataphoric relations are separately distinguished in some annotation projects (e.g., in English-GUM).

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

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

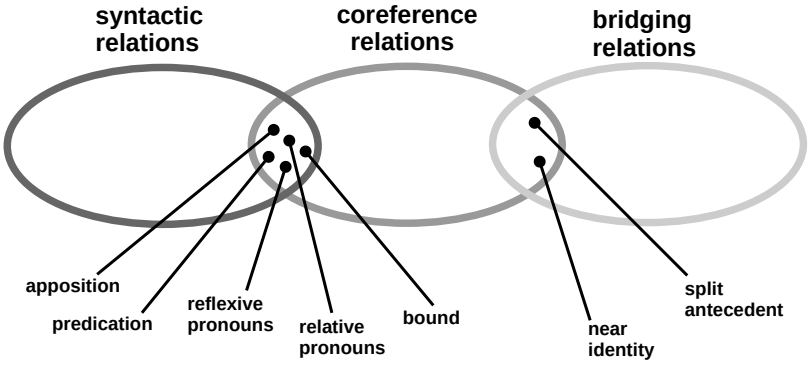

Figure 1.1: Types of possible relations between referring expressions, including borderline types.

(5) When **he** returned, **John** didn't find his apple.

Coreferential expressions can be grouped into **coreference clusters** or **coreference chains**, depending on the technical solution of the given annotation project. A coreference cluster is a set of coreferential linguistic expressions. A coreference chain is a sequence of expressions where all elements of the chain are coreferential, but the linear ordering of references is preserved.

Coreferential, anaphoric or cataphoric relations hold between **mentions** (in some coreference annotation guidelines, they are also called **markables**), which are linguistic expressions, certain fragments of texts. Mentions can be defined in terms of linear mention span (incl. discontinuous mention spans) or syntactically, e.g. as a syntactic subtree in the dependency tree. There may be also **zero mentions** or **zeros**. They occur in case when referents are not explicitly expressed by tokens of their own in the sentence but are rather implicit.

If a mention does not take part in any relation, but potentially it could do so, it is a **singleton**. In coreference annotation projects, singletons may be manually annotated, extracted from the syntactic representation, or ignored. The details of how singletons are treated in different annotation projects are addressed in Section 3.2.2.

The **syntactic head** of the mention is its syntactically governing word (*glass* in the phrase *a glass of milk*), the **semantic head** contains the most relevant semantic information (*milk* in the phrase *a glass of milk*).[5]

The most controversial and debatable in the matters of coreference and anaphoric relations is the distinction between different types of relations. The borderline between identity coreference and non-coreferential relations is not clear-cut. Although there is a core set of cases which are clearly coreferential (see Example 1), there is a wide range of ambiguous cases which make us see coreference in terms of a degree of identity rather than as a binary relation. This causes a large variety of different typologies and annotation solutions in coreference annotation schemes (see a highly simplified scheme in Figure 1.1).

First, coreference is attacked by syntax: here, the cases of **apposition** (Example 6) and **predication** (Example 7) are in play. The notions of reference and coreference collide with syntactic

[5]In some projects (e.g. Polish-PCC) the semantic head is annotated as the head of the mentions because of the prevalence of semantic information over the mention's structure.

predicative properties assignment. Therefore, the coreference annotation is different in different annotation schemes depending on theoretical approach and the goal of the annotation (see Section 3.4.2 for more details). There are also other relations that can be deduced directly from syntax, such as arguments with control verbs, reciprocal constructions, deverbative constructions, reflexive pronouns, bound anaphora etc. (see sections 3.4.3 and 3.1.3 for detailed examples and decisions in coreference annotation schemes).

(6) My apple, the red one, is really good.

(7) This red apple is the best one.

A mention enters into a relation of **discourse deixis** when it corefers with a previous clause or sentence segment (see the relation between the first sentence of Example 8 and the pronoun *that* in the second sentence of this example). In some cases, discourse segments even longer than one sentence may be antecedents of an anaphoric NP. This relation is somewhat different from the classical coreference and anaphoric relations, it is mostly reference to an activity, not to a referent in a proper sense. Thus, the treatment of this phenomenon in different annotation schemes varies. For the details, see Section 3.4.4.

(8) **I ate Peter's apple**. He will never forgive me for **that**.

Identity of coreference may be also attacked by time, space, physical modifications and so on. Do both *it* in the second sentence of the Example 9 refer to the same apple as it was introduced in the first sentence of this example, even if it was reduced by a couple of bites? Such fuzziness gave rise to the annotation of coreference ambiguity in some annotation projects (e.g. in German-PotsdamCC and English-ARRAU) and then developed into the notion of **near-identity** (also called **quasi-identity**) in Recasens et al. (2010a) and attempts to annotate it in AnCora or Polish-PCC. Although frequency of near-identity links and the inter-annotator agreement are too low to consider this relation as annotated reliably, it gives the opportunity to observe many interesting linguistic phenomena.

(9) I didn't like **this apple**. I bit **it** off several times and threw **it** out of the window.

In addition to coreference, many annotation projects distinguish and annotate **bridging relations**, which are explicitly defined as anaphoric but not coreferential (Clark, 1977). These are relations such as part-whole (see the relation between *apple* and *stub* in Example 10), set-subset etc. The typology of the annotated relations is very different across the existing annotation schemes because the scheme should take into account both the diversity of such cases in the language and the goals of the annotation at the same time. The details of the annotation decisions are discussed in Section 3.3.

(10) I finished **my apple** and threw **the stub** out the window.

The attempts of coreference annotation schemes to simplify the annotation, to make it more reliable and to get better inter-annotator agreement (which is highly desirable in order to make annotated datasets usable by models) lead to a number of technical conventions. As the result, the same linguistic phenomenon may be captured as coreference in one annotation scheme, as a bridging relation in the second one and be entirely ignored (as considered to be purely syntactic) in the third one. These variabilities are discussed separately for selected phenomena in Section 3.

Chapter 2

Coreference Data Resources

In this chapter we present a survey of all coreference data resources that are known to us. We divide the resources into three groups:

- resources whose licenses allowed us to create their derived (=harmonized) versions and to distribute them in the public edition of the CorefUD 0.1 collection (Section 2.1),
- resources whose licenses allow us to create their derived versions internally, but we can't distribute them further; harmonized versions of such resources are contained only in the non-public (ÚFAL-internal) edition of CorefUD 0.1 (Section 2.2),
- other resources which we have not harmonized so far due to various reasons, e.g. because of their inaccessibility, too strict license limitations, insufficient documentation, or in many cases simply because of lack of our capacity (such resources are only quickly listed in Section 2.3).

2.1 Harmonized resources available under free licences

2.1.1 Prague Dependency Treebank (Czech)

The Prague Dependency Treebank (labeled as Czech-PDT further in this text, the last consolidated version in Hajič et al. (2020)) is a large corpus of Czech newspaper texts with rich manual multi-layer annotation. The manual annotation includes several interlinked layers of linguistic representation drawing on the Functional Generative Description Sgall et al. (1986). These are morphological, shallow syntactic and deep syntactic (tectogrammatical) layers. Coreference relations are manually annotated on the tectogrammatical layer (namely, on deep syntactic trees). Coreference and bridging relations annotation is described in detail in Zikánová et al. (2015).

Coreference annotation and the annotation of bridging relations are link-based, the relations are annotated on deep syntactic layer and are represented as arrows in a tectogrammatical tree. The arrows 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, i.e., the maximum mention span principle is applied (mentions include all dependencies, incl. all types of subordinate clauses). However, this comfortable technical convention implies a few inconsistencies following from the syntactic annotation conventions. For instance, in Example 11, *Mary with her new boy-friend* represents a single syntactic sub-tree in the tectogrammatical structure, with the head *Mary*. But in the second sentence of this example, Mary is not coreferential with the whole sub-tree, because she comes alone, without her boyfriend.

(11) I invited **Mary with her new boy-friend**. But **Mary** decided to come alone.

Nevertheless, there are just a few types of such cases and they may be searched for and elaborated using the syntactic structure of sentences (namely, taking the information from the syntactic functors[1] of tectogrammatical nodes).

Pronouns, all types of referring NPs (including generic and abstract), textual ellipsis of several types (see Section 3.1.3 for more details), some temporal and local adverbs (like *there* or *then*, i.e., those which may anaphorically refer to an NP) and VPs as antecedents of anaphoric NPs may be annotated for coreference and bridging relations. Singletons are not manually annotated but they may be excerpted from the syntactic trees using the manually corrected morphological and syntactic information.

Coreference relations are interpreted on semantic and reference level, rather than on the textual one, i.e. the identity of referents, not anaphoricity, is primarily annotated. For coreference, Czech-PDT further distinguishes grammatical and textual coreference. The grammatical coreference typically occurs within a single sentence, the antecedent is expected to be derived on the basis of grammar rules of a given language. It concerns the cases of relative pronouns (relation between *an apple* and *which* in *an apple which I had yesterday*), the arguments of the verbs of control (*Peter* and the unexpressed first argument (agens) of the verb *eat* in *Peter wants to eat this apple*), reflexive pronouns (*He dressed himself*), coreference of arguments 'hidden' in reciprocal constructions (*Peter and Mary kissed.*) and coreference with verbal modifications that have dual dependency (coreference with *Mary* and the first argument of run in *John saw Mary run around the lake*). In textual coreference, arguments are not realized by grammatical means alone, but also via context. These are all other types of identity coreference relations. Textual coreference is further classified into relations with NPs with specific or generic (other than specific) reference.

Bridging relations are further classified into part-whole, set-subset and other relations (see Section 3.3 for more detailed information) and are annotated as links to the nearest antecedent from the coreference chain.

Furthermore, exophoric references and references to the larger segments of text are annotated, but they have no antecedents.

Besides coreference and bridging relations, Czech-PDT includes the manual annotation of multiword expressions and multiword named entities, information structure and discourse[2] relations.

As for other annotation solutions, apposition is not specially marked for coreference, but if there are later mentions in the text which are coreferential to a mention in the apposition relation, coreference is marked between this mention and the whole appositional group, the apposition being marked in the syntactic annotation of the sentence.

A similar decision has been taken for coordination. Annotating coreference on deep syntactic trees allows to make a coreference link to the whole coordinated group, as well as to its separate parts.

The cases of split antecedent are solved as a bridging relation of the type set – subset. So, it is not clear from the annotation if the anaphoric entity is coreferential to the set of split antecedents or it includes the broader set of referents and the split antecedents are just a subset of the anaphoric NP.

Given that coreference and bridging relations are annotated on deep dependency trees, which have detailed and sophisticated solution for coordination and different types of ellipsis, it gives a number of elegant solutions in catching these relations in the annotations (no need to implement special cat-

[1]The detailed description of tectogrammatical functors see in Böhmová et al. (2005).

[2]The discourse annotation in Czech-PDT has been completed in the style of Penn Discourse Treebank annotation (PDTB-like discourse) as described in Prasad et al. (2008) for PDTB and in Zikánová et al. (2015) for Czech-PDT.

egories of discontinuous mentions, split antecedents, or long lists of what should be considered as mentions for the annotation).

The annotation has been processed in the special extension of the TrEd annotation tool (Mírovský et al., 2010), which makes it possible to annotate and search coreference on dependency trees.

Manual annotation of syntactic layers is not completed for the whole Prague Dependency Treebank. Whereas morphological and shallow syntactic layer include 2 million and 1.5 million words respectively, the deep syntactic layer covers only 0.8 million words. For our project, we take the subset with the annotation on the deep syntactic layer, as this is the subset containing the annotation of coreference.

The Czech-PDT corpus is distributed under the CC BY-NC-SA 4.0 license.

2.1.2 Prague Czech-English Dependency Treebank – the Czech part

Prague Czech-English Dependency Treebank (Czech-PCEDT, Nedoluzhko et al. (2016)) is a parallel corpus with multi-layer annotation. Its English part consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), see also the description in Section 2.2.4). The Czech part has been manually translated from the English source sentence by sentence.

Czech-PCEDT is the second Praguian corpus with manual coreference annotation, preceded by the monolingual Czech-PDT (see Section 2.1.1). The annotation of coreference-like phenomena is principally similar to Czech-PDT but there are some substantial differences. Grammatical coreference and textual coreference with pronouns were annotated according to the same guidelines as in Czech-PDT. As for the coreference with full NPs, the guidelines had to be simplified to preserve the correspondence with the OntoNotes coreference, which had been annotated for English-PCEDT. For example, only NPs with specific reference have been annotated for coreference in Czech-PCEDT, as opposed to Czech-PDT, where the annotation includes also generic and abstract NPs, distinguished from coreference of specific NPs by a special attribute. Another significant difference between the annotations of Czech-PCEDT and Czech-PDT is that bridging relations were not included in PCEDT, except for a special case of split antecedents.

The cases of split antecedents are annotated as bridging relations of the type set – subset, similarly as it was done for Czech-PDT. However, opposite to Czech-PDT, all cases of anaphoric NPs refer to antecedent NPs only and do not include any broader set of referents, because other cases of bridging relations have not been annotated.

Czech-PCEDT has been annotated in the TrEd annotation tool (Mírovský et al., 2010) and is distributed under the CC BY-NC-SA v3 license.

2.1.3 The Georgetown University Multilayer Corpus (English)

Georgetown University Multilayer Corpus (English-GUM, Zeldes (2017)) is a multi-layer corpus of texts from different types and genres. The corpus contains annotations of various linguistic phenomena, such as morphological features (POS, lemmatization), sentence segmentation, document structure, constituent and dependency syntax, entity linking (wikification), RST-discourse structures etc.

Mentions consist of pronouns and all referring NPs. Non-referring NPs (idioms, non-referential pronouns like *it* in *It rains*) are not annotated. VPs are annotated if they are antecedents of anaphoric expressions. Differently from all other project in our dataset, copula predicates are also annotated as mentions. They are included into coreference annotation in the same they as other referring NPs. The annotation of predication as coreference is a special decision which the authors of the scheme

make. This relation is claimed to be not entirely reconstructable from syntax, and there is a rich argumentation for this decision.

The principle of maximum mention span is applied. Mentions include all possible modifiers, also relative clauses.

Mentions without coreferential relations to other mentions (singletons) are also annotated.

As for the annotation of relations, the annotation layers "ref" and "bridge" are the most relevant for our task. The "ref" layer contains information about referents, including entity type (person, object, abstract, etc.) and identical coreference relations of entities. Coreference contains four different subtypes: anaphoric with pronouns, cataphoric with pronouns, apposition, and lexical coreference (coreference where the anaphor is not a pronoun, e.g. *Obama - president Obama*).

The "bridge" layer includes non-coreferential references, such as proper bridging (for example, the part-whole relationship, or other cases, where no introduction for the anaphor is needed thanks to the antecedent, as in Example 10), cases of non-coreferential anaphora (as in Example 2) and split antecedent (as in Example 4).[3]

The English-GUM corpus can be searched through all annotation layers using the visualisation tool ANNIS Krause and Zeldes (2014).

English-GUM is part of Universal Dependencies project, with an already existing coreference-to-UD unification initiative.

English-GUM is distributed under the CC BY-NC-SA license (for some texts also CC BY), except for the *reddit* subcorpus. Therefore, we excluded it from the conversion to CorefUD.

2.1.4 Polish Coreference Corpus

The Polish Coreference Corpus (Ogrodniczuk et al., 2013, 2015, Polish-PCC) is a corpus of Polish nominal coreference built upon the National Corpus of Polish. The corpus includes written documents from 14 text genres.

Markables are annotated as linear spans, with additionally marked semantic heads. Semantic heads (not syntactic ones, i.e., the most important word from the point of view of mention's sense) were identified because of the prevalence of semantic information over the mention's structure. So, in the phrase *one of the girls*, *the girls* is marked as the head, not the syntactic head *one*.

All NPs (pronouns, nouns with all kinds of dependents including appositions) are annotated as markables, including singletons. VPs are markables only in case if they are referred to by anaphoric NPs.

Markables are grouped into coreference clusters and for each cluster, its dominant expression is selected. Dominant expression is the expression that carries the richest semantic information or describes the referent most precisely (so, in the coreference cluster *Harry Potter – he – him – the boy*, the explicitly expressed entity *Harry Potter* would be the dominant expression).

The annotation includes identity coreference, quasi-identity relations (in the sense of Recasens et al. (2010a)) and three groups of non-identity close-to-coreference relations. The first group (indirect relations) includes the cases of bound anaphora and bridging relations, such as *indirect aggregation* (close to set-subset pairs), *indirect composition* (close to part-whole) and other indirect relations (e.g. function). The second group (supporting) may be classified as rather syntactic. The most substantial sub-group there is the group of predicative relations. They are not considered to be identity coreference in the proper sense, as they are syntactic and non-anaphoric in essence. The third

[3]The cases of split antecedents are treated the same way in Czech-PDT and in PCEDT.

group (excluding) relates rather to the textual context of markables and presents relations of the type mention-to-mention. These are, for example, contrast and non-coreference anaphora.[4]

Polish-PCC has been annotated using a special extension of the MMAX annotation tool Müller and Strube (2001) and is distributed under the open CC BY 3.0 license.

2.1.5 Democrat (French)

French-Democrat (Landragin, 2016) is a large diachronic corpus of written French with coreference annotations. The texts origin from the 12th to the 21st century.

The treatment of coreference relations is inspired by the large-scale annotated corpus of oral French ANCOR (Désoyer et al., 2016) and coreference annotation of Polish in Polish-PCC (see Section 2.1.4).

Mentions, also singletons, are annotated separately, in the first step of the annotation process. Mentions are contiguous spans of text, without the possibility of split antecedent.

The scope of annotation takes into account all NPs including pronouns but strictly sticks to them. As a result, the annotation scheme discards coreference involving verbal or propositional mentions.

Mention features such as definiteness and syntactical type are additionally annotated.

For CorefUD, we converted only *Dem1921*,[5] a modern subsection of Democrat comprising texts from 19th to 21st century with legal texts excluded. For the purpose of coreference resolution (Wilkens et al., 2020), it has been enriched with automatic morpho-syntactic annotation using the StanfordNLP tool (Qi et al., 2018) and exported in a CoNLL 2012 format.

Dem1921 as well as the full Democrat is distributed under the CC BY-SA 4.0 license.

2.1.6 Russian Coreference Corpus

Russian Coreference Corpus (RuCor, Toldova et al. (2014)) is annotated with anaphoric and coreferential relations between noun groups. Russian-RuCor includes prosaic texts of different length and genres: news, science, fiction, blogs.

Markables are annotated as linear spans, with additionally distinguished syntactic heads. Only NPs which take part in coreference relations are considered to be markables, singletons are not annotated. Zero subjects are not reconstructed. A markable span includes the (explicitly marked) syntactic head and its dependents, except for subordinate clauses; relative pronouns are annotated as separate markables, prepositions are not included (in *I live in Prague*, only *Prague* will be a markable).

Coreference relations are annotated between full NPs if they refer to specific referents (type coref). For abstract and generic NPs, the relation is marked only if they are referred to by an anaphoric pronoun (type anaph). Neither split antecedents nor anaphoric relations are annotated. Predicative relations are annotated separately, they are not considered as coreference but are preserved for technical reasons (not to penalize participants of the shared task, Toldova et al. (2014)). A special mark is used for coreference within direct speech. References to clauses (discourse deixis) are not annotated. Appositions are annotated in the same way as other cases of identity coreference.

Additional information about anaphoric expressions is annotated, such as noun, rel(ative), poss(essive), refl(exive) etc.

RuCor is distributed under the CC BY-SA 4.0 license.

[4] These two types, although not very frequent , are interesting, because they are near to bridging types CONTRAST and ANAPH in PDT.

[5] `https://github.com/boberle/coreference_databases/tree/master/democrat_dem1921`

2.1.7 ParCorFull (German and English)

ParCorFull is a parallel corpus of English and German annotated for coreference Lapshinova-Koltunski et al. (2018). ParCorFull was created on the basis of the ParCor corpus Guillou et al. (2014), but it additionally includes the DiscoMT shared task dataset Hardmeier et al. (2015) and WMT17 test set Bojar et al. (2017)).

Markables are annotated as linear spans, without the distinction of their syntactic heads. Markables are pronouns, nouns or NPs which form part of pronoun-antecedent pairs, pronouns without antecedents or VPs if they are antecedents of anaphoric NPs (discourse deixis). A markable span must include its syntactic head, determiners that modify the NP, deverbal modifiers, dependent prepositional phrases and appositions. Full clauses, in particular relative clauses, are not taken as parts of the markable. Relative pronouns are annotated separately. Each mention is further classified into pronoun, NP, VP or clause.

The annotation includes identity coreference relations only. Predicative relations have not been annotated (as considered to be purely syntactic), but in cases like *This is a bank, but it is not very well-known*, coreference is marked for *bank* and *it* (not for *this* and *it*). Substitution and ellipsis are annotated as separate categories, but their interconnection with coreference is very rare. Zero subjects are not typical for German and English, so zeros are not included.

Additional information about antecedents and anaphors is marked within the attributes "anteType" and "Type_of_Pronoun", respectively. Antecedents are classified as entities, events (discourse deixis), and generic. Pronominal anaphors are further divided into personal, possessive, demonstrative, and reflexive.

ParCorFull has been annotated using the MMAX annotation tool Müller and Strube (2001).

As ParCorFull includes TED-talks published under CC BY-NC-ND 4.0 (no derivative works allowed), we decided to exclude them, to be able to publish the rest of the ParCorFull corpus under the CC BY-NC 4.0 license. The remaining data falls entirely in the written news domain.

2.1.8 AnCora: Multi-level Annotated Corpora for Catalan and Spanish

The AnCora corpus (Taulé et al., 2008; Recasens and Martí, 2010) consists of multi-layer annotations of written texts in mostly journalistic domain in Catalan and Spanish. It consists of two different corpora: Catalan-AnCora for Catalan and Spanish-AnCora for Spanish. The corpora contain annotations of various linguistic phenomena which are especially relevant to the understanding of anaphoric relations and coreference. These are, for example, argument structure, thematic roles, semantic classes of verbs, named entities, denotative types of deverbal nouns etc.

The AnCora coreference annotation scheme has been inspired by the general criteria offered in the MATE meta-scheme (Poesio, 2004). It consists of two subtasks: the identification of mentions and linking coreferring mentions together.

Mentions which are identified are pronouns (personal, demonstrative, possessive, relative), full NPs, zeros and discourse segments. Mentions are excerpted from the syntactic annotation, but since not all NPs are automatically referring, the special reference status attribute is added to each mention. First and second person pronouns are included as mentions, even if they are part of direct speech or clitical. Mentions which are embedded within a larger mention are also candidates to participate in a coreference relation, irrespective of the entity to which the larger mention refers. The principle of maximum mention scope is applied.

In the case of a split antecedent, a new entity is formed. This entity represents the sum of the subconstituents (e.g. *entity1+entity2*), and its coreference to corresponding mentions is annotated.

The mention attributes specify referential NPs with specific reference referring to named entities (NE) and distinguish them from the ones referring to non-NE mentions and fixed phrasemes. Within non-NEs, self-sufficient definite descriptions are annotated (e.g. gererally or situationally unique entities like *the sun*, *the world*, *American history* etc.) Mentions for which entity type is not annotated are considered to be non-referring (nominal predicates, appositions, negated NPs, interrogative pronouns etc.).

Generic NPs can enter into identity coreference when used referentially. Coreference links are annotated at a specific and a generic level, but keeping these two levels separated: No coreference can be annotated between a generic and a specific mention.

In AnCora, three types of relations are annotated. First, this is the identity coreference, including relations between generic mentions. Second, predicative relations are annotated. These are further classified into definite and indefinite predicatives: definite ones include the cases of identification predicative constructions (like *John Smith is the thief I've been looking for for a year*.), appositional phrases and acronyms; indefinite predicates are not identificative but they point out an outstanding characteristic of the referential entity (*My father was a high school teacher*). The third is the relation of discourse deixis, i.e., coreference with a previous discourse segment. Coreference annotation in AnCora is cluster-based.

The annotation has been carried out in the AnCoraPipe graphical interface, specially designed for the AnCora corpora. The dataset is distributed under the GNU GPL 3.0 license.

2.1.9 Potsdam Commentary Corpus (German)

The Potsdam Commentary Corpus (German-PotsdamCC) is a corpus of German newspaper articles, annotated with a range of different types of linguistic information Bourgonje and Stede (2020). The coreferentially annotated sub-corpus also includes the annotation of syntax, information structure, and discourse.

The corpus is annotated for nominal and pronominal coreference according to guidelines that build upon the Potsdam Coreference Scheme described in detail in Krasavina and Chiarcos (2007) and Stede (2015). These guidelines describe both identity and bridging annotation rules, but in the current version of German-PotsdamCC, the annotations cover only identity coreference. Bridging relations have not been annotated yet.

One of the most interesting features about German-PotsdamCC is its detailed analysis of mentions.[6] Mentions are understood as linear spans without specially distinguished heads. The authors of coreference annotation guidelines distinguish primary, secondary, and non-referring mentions. The group of primary mentions includes pronouns, definite NPs, proper names, prepositional adverbs and pronominal adverbs. Secondary mentions are indefinite NPs, pronouns, clauses or sentences. These may take part in coreference relations only if they are antecedents of primary mentions. Non-referring are expletives, predicatives, NPs in negation scope or in phraseological units.

The principle of maximum mention span is applied, i.e., a mention consists of the head, usually a noun or a pronoun, and of all modifiers, attributes, relative clauses, appositions, and dislocated elements attached to the head. Therefore, relative pronouns are not annotated for coreference because they are included in the spans of the higher mentions. On the other hand, prepositions are included in the mention spans, making the difference between NPs and PPs unsignificant for coreference annotation.

[6] In the terminology of Krasavina and Chiarcos (2007) they are called markables.

Mentions in German-PotsdamCC have additional information for direct speech, definite or indefinite NPs, named entity, etc. (see Section 3.5 for more details).

As for relations, on the current stage, the annotation is strictly limited to identity coreference/anaphoricity.[7] There is a special decision for split antecedents in German-PotsdamCC: the attribute group is implemented.

An interesting difference from many other annotation schemes is the explicit annotation of ambiguities in German-PotsdamCC. The ambiguity is annotated both at the stage of mention selection (ambiguity about the mention type, e.g., expletive or not) and at the level of relation annotation: for example, the annotator may be unsure about the correct antecedent for the anaphoric mention (attribute ambig-ante), or if they should annotate a coreference relation between the postulated entities (attribute ambig-rel).

The corpus is annotated using the MMAX2 annotation tool Müller and Strube (2001) and can be queried online using the ANNIS visualisation tool Krause and Zeldes (2014).

German-PotsdamCC is released under the CC BY-NC-SA license.

2.1.10 Lithuanian Coreference Corpus

Lithuanian Coreference Corpus (Žitkus and Butkienė, 2018, Lithuanian-LCC) is a corpus of written texts, focusing on political news. Its primary aim is to be used for information extraction.

Coreference annotation in Lithuanian-LCC is link-based and all additional information is marked on coreferential links.

The additional coreference annotation is divided into four levels. In the first level, coreferences are grouped into pronominal, nominal (covers generic and proper nouns), ellipsis, and adverbial (references of *there, then* etc.). In the second level, more information about anaphoric references is given. For pronominal anaphora, the types of pronouns are specified (personal, reflexive, possessive, and relative). For nominal anaphora, the lexical types of the relation are specified (repetition, partial repetition, abbreviation, feature, hyponymy/hypernymy, metonymy or synonymy). In the third level, the direction of the relation is distinguished (anaphora vs. cataphora), and the fourth level is reserved to make the annotation of split antecedents possible.

Coreference for personal pronouns is annotated including first- and second-person pronouns, even if they are part of direct speech (e.g., coreference between *I* and *Tom* in *"I'm going home", Tom said*). Coreference between generic references is not annotated.

The dataset is distributed under the CLARIN-LT End User License[8] in the Lindat repository Žitkus (2018).

2.1.11 SzegedKoref: Hungarian Coreference Corpus

Hungarian-SzegedKoref Vincze et al. (2018) is a corpus of Hungarian written texts selected from the Szeged Treebank Csendes et al. (2005). The treebank has manual annotations at several linguistic layer such as deep phrase-structured syntactic analysis, dependency syntax and morphology.[9]

Markables are linear spans without specially marked heads, but the relations are declared to lead from the head, so it possibly means that the mention heads are determined from the Szeged Treebank syntactic annotation.

[7] The authors call the relation 'anaphoric' but they mean both coreference and anaphora.

[8] `https://clarin.vdu.lt/licenses/eula/PUB_CLARIN-LT_End-User-Licence-Agreement_EN-LT.htm`

[9] For SzegedKoref, the longer texts have been chosen from the Szeged Treebank.

Although named Szeged*Koref*, the corpus is rather oriented on marking anaphoric relations of different kinds, the identity of referents can be deduced from some annotated types, but it is not the primary annotation goal. Within the nominal anaphors type, Hungarian-SzegedKoref marks such anaphoric classes as repetitions, variants (e.g. *Albert Einstein – Einstein*), synonyms, hypernyms, hyponyms, holonyms or epithets. Whereas repetitions, variants, synonyms or epithets are most probably cases of coreference, meronyms and holonyms rather tend to be bridging relations and hypernyms and hyponyms can be both.

In addition to pronominal and nominal anaphors, adverbial anaphors (*the hotel - there*) and verbal substitution (*Julie **sang a song** yesterday, and Joe **did so** today*) are annotated. The speciality of Hungarian-SzegedKoref is the special focus on derivational anaphors, i.e. cases where the antecedent and the head of the anaphoric expression refer to the same entity or activity but belong to different parts of speech: For example, an action is first expressed by a VP and then it is referred to with a noun or participle.

The anaphoric relations are annotated also for zero subject, objects and possessives.

Hungarian-SzegedKoref is annotated using the MMAX annotation tool and is distributed under the license agreement which can be downloaded from `https://rgai.inf.u-szeged.hu/`, we additionally got the direct email permission from the authors.[10]

2.2 Harmonized resources available under non-free license

We have also experimented with other corpora that couldn't be included into the current release because of the license issues. Challenging relation typology, reasonably large datasets and appropriate format attracted us also by the ARRAU corpus, the English part of OntoNotes, COREA and the English part of PCEDT. We have harmonized these resources, but, at the time being, we do not include these corpora in the CorefUD v.0.1 release.

2.2.1 OntoNotes – the English part

OntoNotes 5.0 (Weischedel et al., 2011) is one of the first large-scale coreference annotation resources. It consists of English, Chinese and Arabic texts coming from various domains including newswire,[11] magazine articles, broadcast news, broadcast conversations, web data and conversational speech data. For CorefUD 0.1, we converted only its English part with coreference annotation (English-OntoNotes) comprising roughly 1.7 million words.

According to the coreference guidelines (BBN Technologies, 2006), mentions for coreference annotation are extracted from previously treebanked data, together with the information about syntactic heads. Definite NPs, personal, demonstrative, possessive and other types of pronouns, as well as proper names as premodifiers are subject to annotation. Expletives are not linked. Verbs and VPs take part in coreferential relations if they are antecedents of anaphoric NPs.

The annotation is primary limited to NPs with specific reference. Generic NPs, as well as underspecified, or abstract NPs are annotated for coreference only in case they are referred to by an anaphoric pronoun or another definite mention. So, in Example 12, the instances of *parents*, *they* and *their* would be linked by a coreferential relation, whereas the two instances of *children* would be left unconnected.

[10] In the personal communication with Veronika Vincze, we were allowed to use the data "free of charge for any purpose", which we interpret as the authorization to publish the harmonized version under the CC-BY 4.0 license.

[11] The English portion also contains a part of the Wall Street Journal section from the Penn Treebank Marcus et al. (1993).

(12) **Parents** should be involved with **their** children's education. For example, **they** may help **their** children with some homework.

English-OntoNotes annotates two types of relations: identity coreference (type IDENT) and appositions (type APPOS). Singleton mentions are not manually annotated. As for predicative relations, these are not included: this type of relation is considered to be purely syntactic and the authors believe that it may be caught with better reliability by word sense tagging. The identity coreference includes the cases of local and temporal references (the relation between *three years* and *that time* in Example 13). On the other hand, split antecedents are not explicitely annotated.

(13) John spent **three years** in jail. In **that time** his wife married his younger brother.

English-OntoNotes is distributed under the LDC license.

2.2.2 The ARRAU Corpus of Anaphoric Information (English)

The ARRAU Corpus of Anaphoric Information (English-ARRAU, Uryupina et al. (2020)) is a multi-genre (news, dialogues and fiction) corpus of English which provides large-scale annotations of a wide range of anaphoric phenomena. From the point of view of reference, coreference and anaphoric information, it contains probably the most thorough annotation.

First, the annotation of mentions is very detailed. All NPs are treated as mentions, also when they are non-referring or when they do not corefer with other mentions (singletons). Non-referring mentions are further classified into expletive, quantificational, or predicative. If none of the categories passes to a non-referring mention, it is marked as idiomatic or incomplete. Referring mentions are further classified into discourse-old and discourse-new, and in the first case, the antecedent is identified. Moreover, all mentions are manually annotated for a variety of properties, such as morphosyntactic agreement, grammatical function, and the semantic type of the entity (see Section 3.5 for the full list of semantic types).

A lot of linguistic attention is devoted to the annotation of genericity. The dual distinction generic–non-generic proved to be unreliable because of NPs referring to substances which are hard to distinguish for this feature. Better results have been achieved when substances have been separated to the undersp-generic group. Also NPs in the scope of negation or quantifiers have been treated separately. For the rest, the distinction generic–non-generic has been applied and proved to be reliable.

As in English-OntoNotes, premodifiers are marked as mentions if they are antecedents of anaphoric mentions, but English-ARRAU also annotates non-proper premodifiers.

English-ARRAU has additional annotation of the attribute MIN, similarly as it was once decided for MUC-7 Hirschman and Chinchor (1998) which corresponds to the head noun for non-proper nominal markables and to the whole proper name (for example, a name and a surname) in case of named entities. Moreover, the annotation scheme also marks the span of named entities (the attribute enamex). One of the distinctive features of English-ARRAU is the annotation of discontinuous mentions, which are especially important for speech corpora where they occur frequently.

In English-ARRAU, different types of anaphoric and coreference relations are annotated. Similarly to Czech-PDT, coreference is annotated also for relations that are coreferential but not anaphoric, as in Example 6.

NP-coreference is explicitly distinguished from the cases of discourse deixis by special attributes (antecedents of type phrase or segment). The cases of split antecedents (plural references) are annotated separately.

Furthermore, several types of bridging relations are annotated. The range of annotated relations is limited to three types with additional annotation of the relation direction. These are element-of, subset and a generalized possession relation poss covering part-of, general possession relations, other and undersp-rel for obvious cases of bridging that didn't fit any other category.

Finally, similarly as in German-PotsdamCC, anaphoric ambiguity is annotated for the cases where more than one possibility of annotating anaphoric relation or mention type is possible.

Coreference annotation has been provided using MMAX annotation tool Müller and Strube (2001).

The second release of the English-ARRAU corpus is available from LDC, but the sub-corpora of this version that consist of anaphoric annotations of LDC corpora such as the RST Discourse Treebank and the TRAINS-93 corpus can only be distributed for free to groups that acquire a license for the original corpora. The dataset extracted from ARRAU for the CRAC 2018 Shared Task is available through LDC.

2.2.3 COREA: Coreference Corpus for Dutch

The COREA coreference corpus (Dutch-COREA, Hendrickx et al. (2008)) is a collection of written and transcribed oral texts in Dutch annotated for creating a coreference resolution system. The written texts are from news and medical (medical encyclopedia) domains.

Mentions are strings of text with a specially distinguished (rather semantic) head. Pronouns and full NPs with their dependencies are subject to annotation.

In Dutch-COREA, identity coreference as well as other anaphoric relations are annotated. Identity coreference is mostly focused on and is the most frequent annotation type. Bridging relations of the type *set – subset* is annotated with the bridge attribute, other types of bridging relations are not included. The predicative relation is annotated separately with a special attribute, whereas apposition is treated in the same way as other coreferences. Differently from other annotation projects, Dutch-COREA annotated bound anaphora (cases where an anaphor refers to a quantified antecedent, see Section 3.4.3) as a specially distinguished category.

The speciality of Dutch-COREA is distinguishing between the level of sense (identity on the type level) and the level of reference (identity on the token level). This distinction is used for specifying coreference relations in type-token pairs, as well as for bound anaphora and metonymy.

In Dutch-COREA, so called time-indexed coreference is annotated. These are the cases where the relation has a temporal validity. So, in Example 14[12] both NPs *delegated manager* and *chief financial and administration officer* refer to *Bert Degraeve*, but he does not perform both functions at the same time. These relations are time-dependent and are marked with a TIME attribute.

(14) **Bert Degraeve**, until recently **delegated manager**, will start as **chief financial and administration officer**.

There is also a special attribute MOD, which points to relations between NPs which are in some way modal, i.e. not quite identical. This relation may be compared to near-identity relations, annotated in Polish-PCC.

Coreference annotation has been provided using the MMAX annotation tool Müller and Strube (2001).

The license issues with Dutch-COREA appear to be complex. The dataset cannot be freely distributed and published under the CC BY-NC-SA license.

[12]This example is the English translation of the Dutch sentence, taken from the Dutch-COREAannotation guidelines Bouma et al. (2007).

2.2.4 PCEDT – the English part

The English part of the Prague Czech-English Dependency Treebank consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). Coreference annotation in English-PCEDT is basically the same as in Czech-PCEDT (as described in Section 2.1.2).

2.3 Other existing resources

Besides the datasets described above, we are aware of several others that would be relevant to our work. We only briefly mention them here without going into details. We were not able to include them in this pilot harmonization effort – mainly due to limited working capacity, although for some of the resources we also may not be able to gain access to them.

- AMR coreference corpus (English) (O'Gorman et al., 2018)
- ANCOR (French) (Désoyer et al., 2016)
- Copenhagen Dependency Treebanks (Danish, English, German, Italian, Spanish) (Korzen and Buch-Kromann, 2011)
- Corref-pt (Portuguese) (Fonseca et al., 2017)
- EPEC-KORREF (Basque) (Soraluze et al., 2012)
- Evaluation Dataset for Zero Pronoun (Japanese, English) (Shimazu et al., 2020)
- GAP (English) (Webster et al., 2018)
- LIVEMEMORIES (Italian) (Rodríguez et al., 2010)
- MEANTIME (English, Italian, Spanish, Dutch) (Minard et al., 2016)
- NAIST Text (Japanese) (Iida et al., 2007)
- OntoNotes (Arabic, Chinese) (Weischedel et al., 2011)
- ParCor (English, German) (Guillou et al., 2014)
- PAWS (Parallel Anaphoric Wall Street Journal) (English, Czech, Polish, Russian) (Nedoluzhko et al., 2018)
- PerCoref (Persian) (Mirzaei and Safari, 2018)
- Prague Dependency Treebank of Spoken Czech 2.0 (PDTSC) (Mikulová et al., 2017)
- PreCo (English) (Chen et al., 2018)
- RiddleCoref (Dutch) (van Cranenburgh, 2019)
- TüBa-D/Z (German) (Çöltekin et al., 2017)
- TwiConv (English) (Aktaş and Stede, 2019)
- VENEX (Italian) (Poesio et al., 2004)

Chapter 3

Diversity of Annotated Schemes in Existing Resources

In this chapter, we address annotation diversity in the resources under analysis. All the resources in our dataset contain the annotation of identity coreference. However, coreference is a quite broad notion and, on the other hand, there is a number of other linguistic aspects which are related to coreference to some degree. Taking this into account, we address the diversity of the annotation schemes in the following aspects:

- The scope of mentions: What is considered to be a mention? Is it obligatorily linear or can it be discontinuous? Are syntactic or semantic heads of mentions annotated? Which grammatical types of mentions are annotated? Are zeros or nominal ellipses considered to be mentions and annotated for coreference? Table 3.1 summarizes this information.
- Identity coreference: Is coreference annotated as a chain of mentions or a cluster, a set consisting of all coreferential mentions? In the second case, do these clusters have some inner structure? Are singletons manually annotated, or only mentions which take part in more than one-element coreference sets are considered?
- Non-coreference anaphoric relations (bridging in a broad sense): Are non-coreference anaphoric relations annotated? What types are distinguished? Are such relations based on type or token level? Is near-identity annotated?
- Other specific types of relations between mentions: Are split antecedents annotated? How are apposition and predication relations treated in the annotation? Is bound anaphora specially considered or just included in the coreference annotation? What about discourse deixis?
- Coreference-relevant information about mentions: Does the annotation contain information about the reference status of mentions (specific, generic, non-referring)? Is the difference between common and proper nouns marked? Are specific entity types marked (location, organization, number, date, person, etc.)? Does this information concern a single mention in the text or the whole coreference cluster?
- Other NLP annotations available in the data: What additional linguistic information, which can be used in coreference and anaphoric analysis, is annotated in the datasets (lemmatization, POS tagging, sentence segmentation, tokenization, syntactic trees, document boundaries, etc.)?

For convenience, the most addressed issues concerning annotation of coreference relations in various annotation schemes are summarized in Table 3.2.

	Mention repr.		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 3.1: 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.)

3.1 Mentions – delimiting basic units of reference

3.1.1 Formal representation of mentions

Coreference is a relation that connects expressions sharing the same reference – mentions. Clearly, when annotating coreference, mentions must be marked in some way in the annotated text. We identified the following approaches to mention delimitation in the surveyed resources:

- a linear span (in terms of token sequences)
  - typically expressed as a token sequence specified by the identifier of the first token and by the identifier of the last token; example in the MMAX notation: `span="word_17..word_18"`
  - if the mention correspond to a discontinuous sequence of tokens, then the span is expressed by joining two or more continuous token sequences;
    example: `span="word_232,word_235..word_236"`
  - rarely, the beginning and end of a mention are specified not by token identifiers, but by character offset of the mention start from the document beginning and by mention length expressed in characters too (e.g. Russian-RuCor, Lithuanian-LCC),

CorefUD dataset	Relations among mentions							
	cluster-based identity	link-based identity	singletons	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 3.2: 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.

  - language-specific (or project-specific) tokenization rules are sometimes used (e.g. for turning clitics in past-tense verb forms into separate tokens in Polish-PCC),
  - reconstructed zeros are sometimes inserted into sequences of tokens (and receive their own identifiers),
  - syntactic or semantic heads[1] are sometimes marked within the mentions; multiple heads in a single mention are marked sometimes,
  - in some projects, a so-called minimum span is distinguished, which is either the head token, or multiple tokens in the case of name entities,

- nodes in dependency trees
  - in some projects in which coreference is added to dependency treebank data (such as Czech-PDT), a mention is represented by its syntactic head node,
  - the exact span of the mention within the original sentence is then defined only implicitly (e.g. as the projection of the whole subtree rooted by the given head onto the sequence of sentence tokens),
  - the relation between dependency subtree and tokens in the sentence might be quite complicated in the case of deep-syntactic dependency trees, in which functional words do not have nodes of their own.

[1] If the annotated mention head is different from the mention head according to the UD style, then mention representation is moved to the UD head in CorefUD 0.1.

- nodes in constituency trees
  - if constituency trees are available for a given resource, then a mention composed of multiple words (such as noun phrases) can be captured by pointing to a non-terminal node that spans the mention (this is the case of Catalan-AnCora and Spanish-AnCora).

3.1.2 Grammatical types of mentions

Prototypically, a mention is a noun phase (full NP or a pronoun; a *nominal* in the terminology of UD), referring to a concrete extra-linguistic object. However, in the existing annotation projects, the notion of mention varies substantially.

Some annotation projects take into account only explicitly expressed anaphoric pronouns (e.g. in ParCor 1.0, Guillou et al., 2014), some other annotate coreference with pronouns and definite NPs (German-ParCorFull) or limit the mentions to referring NPs (English-OntoNotes). Verbal phrases may or may not be annotated in the antecedent positions. Generic, abstract or deverbative NPs present a special challenge to coreference annotation and they are ignored in many projects, to reach better consistency and inter-annotator agreement. Another issue is the reconstruction and coreference annotation of zeros (see Section 3.1.3).

The mention types in selected datasets are summarized in Table 3.3.

dataset	mention types relations
Czech-PDT, PCEDT	pronouns (including possessive, relative, reflexive, etc.), full NPs, VPs as antecedents, pronominal adverbs
English-GUM	all referring pronouns and NPs, VPs as antecedents
Polish-PCC	pronouns (excluding reflexive ones), full NPs, VPs as antecedents
French-Democrat	pronouns and full NPs
Russian-RuCor	pronouns and full NPs with specific reference, generic NPs as antecedents
ParCorFull	pronouns, nouns or NPs which form part of pronoun-antecedent pairs, pronouns without antecedents or VPs if they are antecedents of anaphoric NPs
AnCora	pronouns (incl. reflexive, relative and possessive), full NPs, VPs as antecedents
German-PotsdamCC	pronouns, definite NPs, proper names, prepositional adverbs and pronominal adverbs
Lithuanian-LCC	pronouns (incl. reflexive, relative and possessive), referring non-generic NPs
Hungarian-SzegedKoref	pronouns, full NPs, pronominal adverbs
English-OntoNotes	pronouns, specific NPs
English-ARRAU	all referring and non-referring NPs, VPs as antecedents
Dutch-COREA	pronouns and full NPs

Table 3.3: Distribution of mention types in selected datasets.

Another factor affecting the annotation of coreference with different mention types is the properties of the annotated language. For example, in English, German, French, Spanish and many other languages, the definite article may be used by models or annotators to choose the potentially anaphoric expression, so there is no need to annotate indefinite NPs (including generic, abstract and so on). On the other hand, in most Slavic languages, there is no grammatical category of definiteness, and definite NPs cannot be easily distinguished from the indefinite ones. To make the distinction, the annotator has to go deeply into semantics and learn to distinguish between specific and non-specific mentions. This explains the fact that for Slavic languages (Czech-PDT, Czech-PCEDT, Polish-PCC, partly Russian-RuCor), coreference includes also generic mentions.

3.1.3 Representation of zeros

Representations of zeros is an especially challenging issue because languages behave differently in this respect and there are multiple theoretical approaches to this problem.

First of all, languages differ substantially in what may be unexpressed. In pro-drop languages, anaphoric subjects tend not to be expressed explicitly. In our dataset, this involves Czech, Catalan, Spanish, Polish, Lithuanian, and Russian. However, there is also difference in how often and in which positions the subjects are omitted. For example, Czech, Polish, Catalan and Spanish are strong pro-drop languages. Anaphorically known subjects are omitted almost in all cases in the written form of the language, so coreference annotation will be definitely incomplete if zero subjects are not reconstructed and annotated. On the other hand, in Russian, dropping a subject is less frequent, which makes it possible to create an annotation scheme without adding reconstructed subjects (zeros). For some other languages (Hungarian, Turkish), not only subjects, but also objects and possessives may be omitted in a similar way, and the scheme must take it into account. So, in Hungarian-SzegedKoref, the special zeros *proObj* and *proPoss* are reconstructed. In English and German, on the contrary, unexpressed subjects or objects are not common at all, so the problem does not arise and no zeros must be inserted.

Another important issue is that syntactic zeros as such constitute a big theoretical issue, which is exploited in different linguistic theories. Besides omitted anaphoric subjects (*pro-drop*, *pro*), many other arguments in various syntactic positions may be reconstructed, e.g., the cases of control and quasi-control (*PRO*) in some theories including Functional Generative Description (FGD, Sgall et al., 1986), which is the theoretical basis for the syntactic annotation of Czech-PDT and Czech-PCEDT. So, in the Prague dependency treebanks, special zeros (with the lemma #Cor) are reconstructed (and annotated for coreference) for unexpressed arguments of the controllee of the verb *go* in *He decided to go away*. According to the valency theory adopted in FGD, also unexpressed arguments for verbs and their derivatives are reconstructed (for example, the lemma #Gen for the unexpressed object of *read* in the sentence *He likes to read*), which produces a large number of zero elements in the corpora.

Moreover, there are other types of ellipsis, where syntactic heads of NPs (15) or VPs (16) are omitted or substituted. Naturally, such NPs may also take part in coreference or anaphoric relations, and a coreference annotation scheme has to find a solution how to treat them properly. In a number of annotation schemes, the cases of ellipsis are specially marked (Czech-PDT, Czech-PCEDT, Polish-PCC, AnCora, ParCorFull, Lithuanian-LCC, etc.). In Russian-RuCor, ellipses are not reconstructed, but coreference relation is annotated to explicitly express the remainder of the NP.

(15) *Já jím zelená jablka a ty jíš červená* [cs]
I eat green apples and you eat red ***(ones)***

(16) *Ty si vyčistíš zuby a já také* [cs]
*You will clean your teeth and I **(will do so)** too*

Table 3.1 summarizes reconstruction and annotation of zeros and ellipses in coreferentially annotated corpora. Table 3.4 gives some details about relevant corpora in this respect.

dataset	treatment of zeros
Czech-PDT, PCEDT	pro-drops (#PersPron), arguments in control (#Cor) and quasi-control (#QCor) constructions, unexpressed arguments for verbs and their derivates (#Gen), arguments in reciprocal constructions (#Rcp), nominal ellipses (#EmpNoun, #EmpVerb), reconstructed nodes with their original lemmas
Polish-PCC	zero subjects (marked on verbal endings), nominal ellipses
Hungarian-SzegedKoref	zero subjects (proSubj), objects (proObj) and possessives (proPoss)
AnCora	zero subjects, nominal ellipses
ParCorFull	nominal ellipses and substitutions
Lithuanian-LCC	nominal ellipses

Table 3.4: **Zeros**: How they are treated in the annotation schemes

The schemes differ in how they operate with zero items technically. For example, in Polish-PCC, coreference of elided subjects is marked either on the whole verbs (in present, e.g. *widzimy* [*we see*]) or on verbal flexion/clitic (in past, e.g. to *em* in *widział-em* [I saw]). Other coreference annotation schemes, like Prague corpora, AnCora, Chinese and Arabic OntoNotes use specially inserted zero lemmas or an asterisk.

3.2 Coreference – grouping mentions with identical reference

3.2.1 Representation of coreference

As coreference in its strict sense can be seen as an equivalence relation, two main styles of its annotation prevail: (1) cluster-based, and (2) link-based.

In the cluster-based style, the basic building block is affiliation of the mention to a named coreference cluster. A coreference cluster then consists of all mentions labeled with the cluster identifier. This annotation style thus treats coreference as a set of equivalence classes. The cluster-based style is applied e.g. in OntoNotes, GUM, AnCora, Democrat, and PCC.

In the link-based style, a coreference link is the basic building block. Each coreference link connects two mentions, i.e., anaphor (cataphor) and its antecedent (postcedent), and is usually annotated within the representation of the anaphor as an address pointing to the representation of the antecedent. Representing the link as an edge in a directed graph, a coreference cluster is then a weakly connected component formed by such links. The link-based style is used e.g. in PDT, PCEDT, PotsdamCC, COREA, and LCC.

The main advantage of cluster-based annotation is that the whole coreference cluster is easily accessible simply by its identifier. This is not true for the link-based style, where data requires some

post-processing in order to extract the clusters. In addition, it is more difficult for the antecedents that never act as an anaphor to be identified as part of a coreference cluster, because this information is captured always in anaphors. On the other hand, link-based style allows for representing the structure within the cluster. For instance, it is hard to derive from the cluster-based style which mentions are more prominent. Capturing link-related information is also challenging in the cluster-based style. Moreover, the link-based style allows for representation of non-equivalence relations, e.g. near-identity or bridging.

3.2.2 Presence of singletons

Having annotated singletons affects the performance of coreference resolution systems Kübler and Zhekova (2011). If singletons are annotated, the statistics of mentions and relations change significantly. Moreover, the types of annotated singletons may differ between schemes. For example, whereas in English-ARRAU all possible singletons are marked and classified, English-GUM does not annotate non-referring NPs.

Table 3.5 gives the overview of the manual annotation of singletons in the corpora we work with.

singletons decision	datasets
ignored	English-OntoNotes, Russian-RuCor, ParCorFull
annotated	English-ARRAU, English-GUM, English-ARRAU, Polish-PCC, German-PotsdamCC
may be reconstructed	Czech-PCEDT, Czech-PDT, AnCora

Table 3.5: Manual annotation of singletons

3.3 Non-coreference anaphoric relations – bridging (in a broad sense)

Beyond identity coreference, there is a number of other types of anaphoric relations, and some corpora include annotations of different subsets of such relations.

In the corpora under analysis, these non-coreference relations may be roughly divided into bridging relations in the proper sense, according to the description given in Clark (1977), and all other relations including near-identity relations in the sense of Recasens et al. (2010a), bound anaphora, anaphoric relations without coreference, contextual contrast and so on. Due to their semantically-oriented definition, classical bridging relations may be interpreted as the relations between identity coreference clusters. So, in the part-whole pair *car – wheel*, the part-whole relation remains also between the whole identity clusters of all coreferential cars and wheels in the given text.

As for other types of non-coreferential anaphoric-like relations, it may not be the case. Many non-coreferential anaphoric relations concern the anaphoric entity itself and not the cluster. For example, a contrastive relation makes sense only in the given context, see the relation between shirt and tie in the sentence *His shirt was still all right, but the tie was just disgusting.*).

Considering the great variety of non-coreferential anaphoric relations, different possibilities to treat and classify them, as well as different goals and requirements to inter-annotator agreement, approaches to these relations are different in each annotation scheme. Some overlaps may be observed,

for instance, the part-whole relation is present to some degree in most schemes which annotate non-coreference anaphora, although the labels vary.

Bridging and other related relations are presented in Table 3.6.

dataset	bridging relations
Czech-PDT	part-whole (two directions), set-subset (two directions), object-function (two directions), contrast, anaph (non-coreferential anaphora), other
English-ARRAU	element, element-inv, subset, subset-inv, poss(essive), poss-inv, other, other-inv, undersp-rel
Polish-PCC	aggregation (set-subset), composition (part-whole), metareference, comparison, contrast, ios (non-coreferential anaphora), other, near-identity
Hungarian-SzegedKoref	meronymy, holonymy, hyponymy, hyperonymy
English-GUM	part-whole, non-coreferential anaphora, other
Dutch-COREA	bridge (set – subset), attribute MOD for near identities, attribute TIME for time-dependent relations

Table 3.6: Distribution of bridging relations

3.4 Other specific types of relations between mentions

3.4.1 Split antecedents

An anaphoric NP may have more than one antecedent in the previous context, i.e., the coreferential antecedent is split into several mentions (***My father***$_i$ *met* ***my mother***$_j$ *twenty years ago, but* ***they***$_{i+j}$ *got married after I was born*). This presents a special challenge for systems as well as for annotation tools and guidelines. As the result, there is also a variety of approaches accepted in coreference annotation schemes.

First, many schemes (Polish-PCC, Russian-RuCor, English-OntoNotes etc.) do not annotate such cases at all, namely they annotate plural coreference only in case the antecedent is already plural itself (***The kids*** *knew* ***they*** *had to get home before dark*) or is expressed by a linearly inseparable string (***Mary and John*** *got divorced two years ago, but in January* ***they*** *decided to get married again*). In English-OntoNotes, also the antecedents which are connected by a conjunction are annotated.

Secondly, if annotated, approaches may still be very different. For example, in English-GUM, English-ARRAU or ParCorFull, anaphoric references to split antecedents are annotated as a special class of references. In Czech-PDT and PCEDT, the references to split antecedents are annotated as a bridging relation of a subset type, split antecedents being taken as subsets of the anaphoric element.

The solutions of split are summarized in Table 3.7.

3.4.2 Apposition and Predication

Apposition and predication relations take place between NPs which refer to the same entity. The problem is that the relation itself is neither anaphoric nor coreferential in the proper sense. It is rather syntactic.

split a. decision	datasets
none	Polish-PCC, Russian-RuCor, Dutch-COREA
bridging	Czech-PDT, English-PCEDT, Czech-PCEDT, English-GUM
specific solution	English-ARRAU, ParCorfull, English-OntoNotes, French-Democrat, AnCora

Table 3.7: Representation of split antecedents

For example, in both sentences *Bob is my father-in-law* and *Bob, my father-in-law, got married yesterday* the nominal phrase *my father-in-law* rather describes Bob, predicates him a new quality, gives new information to the name than just anaphorically refers back to *Bob*. This makes it a difficult decision how these cases should be treated in the annotation of coreference. Given that the borderline between the cases of predication, apposition and coreference is not clear-cut, such decision is getting even more complicated. We thus observe different solutions in existing coreference annotation schemes.

Possible decisions are the following:

- Ignore the relation, consider it to be a syntactic one, do not take it into account within coreference annotation. This solution has been adopted for apposition and predication in the Prague corpora (PDT, PCEDT), given that the existing rich syntactic annotation makes it possible to find and excerpt these relations easily. For predication, in case of not annotating is as a special relation, there may be a decision procedure on what mention (subject or predicate) to prefer for attaching to a subsequent coreferential mention (ParCorFull, English-OntoNotes).
- Mark it as a special type.
  - Do not consider these cases as coreference, annotate them separately (predication in Russian-RuCor or AnCora, apposition in English-OntoNotes, etc.)
  - Include them in coreference chains but mark them as a special type, mostly on the anaphoric mention (predication in German-PotsdamCC, apposition in Hungarian-SzegedKoref, etc.)
- Include both components of the relation in the span of one mention (Polish-PCC and ParCorFull for appositions).
- Annotate them in the same way as identity coreference, do not specify the type (English-GUM for the predicative relation, Dutch-COREA for appositions).

The decisions for apposition and predication in selected annotation projects are summarized in Table 3.8 and Table 3.9, respectively.

The problem of predication and apposition is definitely not trivial both from the theoretical point of view and from the point of view of data diversity. Not all cases of predication are the same, compare, for example, the difference between introduction of a new referent with a demonstrative pronoun (Example 17), quality assignment (Example 18), and identification constructions (Example 19). These relations are very different concerning the referential aspect.

(17) **This** is **a table**.

apposition decision	datasets
ignored (reflected in syntactic structure)	Czech-PDT, PCEDT
annotated as identity	Dutch-COREA
separated	English-GUM, English-OntoNotes, Russian-RuCor, Hungarian-SzegedKoref, AnCora
included in one mention	Polish-PCC, ParCorFull, Lithuanian-LCC, German-PotsdamCC

Table 3.8: **Apposition**: Distribution in annotation schemes

predication decision	datasets
ignored	Czech-PDT, PCEDT, ParCorFull, Lithuanian-LCC
separated	Polish-PCC, Russian-RuCor, AnCora, English-ARRAU, Dutch-COREA
included as identity	GUM

Table 3.9: **Predication**: Distribution in annotation schemes

(18) **Peter** is **a teacher**.

(19) **Peter** is **the teacher** who kicked me out of the university last year.

Moreover, the predicative relation may be sensible to time-dependent relations, i.e., the assigned quality may be temporal, as in Example 14, which is taken into account in coreference annotation in Dutch-COREA.

3.4.3 Bound anaphora

A special issue is a bound anaphora. These are the cases where an anaphoric pronoun functions as a bound variable, referring to non-specific antecedent with a quantifier (*Almost **each husband** is proud of **his** wife*). Referring to noun phrases with quantifiers, but in a very specific grammatical way, this phenomenon may be considered as a special type (e.g. in Dutch-COREA), as bridging (Polish-PCC) or treated in the same way as identity coreference (ParCorFull, OntoNotes, Prague corpora). The distribution in annotation schemes is given in Table 3.10.

decision	datasets
separated	Dutch-COREA
bridging	Polish-PCC
identity	Czech-PDT, Czech-PCEDT, ParCorfull

Table 3.10: Representation of bound anaphora

3.4.4 Discourse deixis

The relation of discourse deixis (reference to VPs, sentences or discourse segments) is different from coreference in the proper sense, both from formal and referential points of view. Discourse deixis mostly represents reference to an activity, not to a specific referent. It has some consequences. First, the formal representation of anaphoric mentions is limited in comparison to proper coreference: pronouns, shell nouns with demonstratives, or definite deverbatives are usually used as anaphors (see the introduction of the discussion in Webber (1988) and further facts, e.g., in Dipper et al. (2011) and many other publications). Second, identifying the mention span of the antecedent may present a challenge, both technically (segments spanning multiple sentences) and semantically (find the correct boundaries of the mention).

Thus, the annotation of discourse deixis is not the same across coreference annotation schemes. Possible decisions are (i) to ignore such cases, not to include them into coreference annotation, (ii) annotate discourse deixis in the same way as other types of coreference, or (iii) annotate them but distinguish discourse deixis as a separate type.

Table 3.11 summarizes how the cases of discourse deixis are treated in coreference annotation schemes.

dd. decision	datasets
none	Polish-PCC, Russian-RuCor, French-Democrat, Lithuanian-LCC, Dutch-COREA
same as coreference	Czech-PDT, PCEDT, English-GUM, English-OntoNotes, English-GUM, German-PotsdamCC, Hungarian-SzegedKoref
specially marked	English-ARRAU, ParCorfull, AnCora

Table 3.11: Representation of discourse deixis in the annotation schemes

If annotated and marked as a special type, the information about discourse deixis may be stored for the relation (as in English-ARRAUor AnCora) or on mentions, as information about antecedents (e.g. in ParCorfull).

3.5 Additional information about entities

In addition to relations between entities, most coreference annotation schemes include other relevant information. It does not pertain to the relation of coreference, but to the entities themselves. For example, in English-ARRAU, this information includes whether a nominal is referring, expletive, predicative, or a quantifier. Similarly, in ParCorFull, antecedents are classified into entities, events, and generic. In English-GUM, ten types of entities are distinguished, including person, time, animal, or abstract.

This kind of information may be important both for linguistic research (e.g., anaphoric reference works differently for abstract notions and for places) and for systems (not all systems have the ability to resolve, e.g., reference to events, so it is useful to be able to separate them).

Information about entities may concern a mention (as in English-ARRAU, Russian-RuCor or ParCorFull), or the whole cluster (as in English-GUM, AnCora or Czech-PDT).

The overview of the annotation of the information about entities in given in Table 3.12.

dataset	type	what is annotated
Czech-PDT	cluster	spec(ific), gen(eric)
Russian-RuCor	mention	noun, poss(essive), refl(ective), rel(ative)
ParCorFull	mention&cluster	entity, event, generic, personal, possessive, demonstrative, reflexive
SzegedKoref	mention	pronominal, nominal (incl. repetition, variants, synonym, hypernym, hyponym, meronym, holonym, epithet, apposition), verbal , adverbial, derivational
Lithuanian-LCC	mention	common nouns vs. named entities distinction, lexical anaphoric types (repetition, partial repetition, abbreviation, feature, hyponymy/hypernymy, metonymy or synonymy)
AnCora	cluster	for specific (ne) entities: organization, location, person, number, date; for other: non-ne, phrase
German-PotsdamCC	mention	expletives, predicatives, idioms, direct speech, NP or PP, definite or indefinite NPs, named entity, type of pronoun (personal, possessive, demonstrative) or grammatical role (subject, direct or indirect object), ambiguity (if expletive or not)
English-ARRAU	mention	morphosyntactic agreement, grammatical function, semantic types entities (person, animate, concrete, organization, space, time, plan (for actions), numerical, abstract); reference type (referring, expletive, quantificational, predicative), genericity, for referring: discourse-old vs. discourse-new, idiomatic, incomplete

Table 3.12: Information about entities

3.6 Other NLP annotations available in the data

For some corpora in our collection, there is additional annotation available, which may be used for extracting different kinds of information. The range of additional annotations is quite broad: beginning from POS and primary morphological information through syntactic trees to complex multilayered annotation of many linguistic phenomena. Additional annotation might or might not be linked to the annotation of coreference. Additional annotations are summarized in Table 3.13.

3.6.1 Document boundaries

In Czech-PDT, PCEDT, and AnCora, one file of the original corpus is considered a document; the document boundaries are marked in the harmonized data using the newdoc sentence attribute defined in UD. French-Democrat has explicit document boundaries which we carried over to the harmonized data.

Dataset	other available annotation
Polish-PCC	automatic segmentation and POS
English-ARRAU	RST-discourse for a subcorpus
ParCorFull	automatic segmentation and POS
Russian-RuCor	automatic segmentation and POS, direct speech
French-Democrat	definiteness, syntactic type, automatic morphology and UD-style syntax
SzegedKoref	morphology, dependency-style syntax
German-PotsdamCC	morphology, constituent syntax, RST&PDTB discourse, information structure (aboutness topics)
Czech-PDT	morphology, dependency syntax (surface and deep), information status, multiword expressions, multiword named entities, PDTB-style discourse relations, genres
Czech-PCEDT	morphology, dependency syntax (surface and deep)
English-PCEDT	morphology, dependency syntax (surface and deep), Penn Treebank-style syntax (WSJ)
English-GUM	morphology, document structure, Penn Treebank-style trees, Universal Dependencies (UD-trees), information status, Wikification, discourse parses in RST
AnCora	lemma, POS, syntactic constituents and functions, argument structure and thematic roles, semantic classes of the verb, denotative type of deverbal nouns, nouns related to WordNet synsets, named entities

Table 3.13: Other available annotation

3.6.2 Sentence segmentation

Manually corrected sentence segmentation is available in Czech-PDT, PCEDT, French-Democrat and AnCora. In AnCora, the segmentation of the coreferentially annotated dataset differs from the dependency-annotated dataset that is currently available in Universal Dependencies (as of UD release 2.7). For example, multi-sentence direct speech is annotated as one sentence-level segment in the coreference dataset but as multiple sentences in UD. We map the coreference data on the UD sentences and resegment it to match UD. Occasionally this means that we have to shorten the span of a mention, as we cannot represent mentions spanning multiple sentences.

3.6.3 Tokenization

Manually corrected tokenization is available together with the original raw text in Czech-PDT.

Many datasets are tokenized but the original raw text has been lost (and can be reconstructed only approximately, using heuristics). This is the case of PCEDT, AnCora, French-Democrat.

The tokenization of the coreference-annotated dataset in AnCora does not match that of the UD version of AnCora:

- Multi-word expressions are treated as one token in the coreference datasets, their word forms being joined using the underscore character ("_"). In UD, these expressions are split and every word gets its own node in the dependency tree.
- Multi-word tokens (as defined in UD) should be segmented into syntactic words using the technical means devised by the UD file format. Again, such segmentation is not available in the coreference dataset.

Our ultimate goal is to retokenize the coreference dataset to match the tokenization in the UD version of AnCora. We have not been able to achieve this goal in CorefUD 0.1, so this remains on the agenda for the future. At present, the data retains the tokenization from the coreference version of AnCora.

3.6.4 Lemmatization and POS tagging

Manually corrected lemmas and part-of-speech tags are available in Czech-PDT, PCEDT, and AnCora. We have converted them to match the UD guidelines, except in AnCora, where the original part-of-speech tag is kept as XPOS in UD, but the morphological features are not extracted and stored in the UD format.

French-Democrat contains automatically assigned lemmas, UPOS tags, and morphological features (using a model trained on one of the UD-released French treebanks); we just keep them as they are.

3.6.5 Syntactic trees

The original datasets of Czech-PDT, PCEDT, AnCora, and French-Democrat contain some sort of syntactic annotation. In French-Democrat this annotation has been assigned automatically, it follows the UD annotation style, and we just copy it.

In Czech-PDT, we mostly follow the conversion procedure that was also used to prepare the UD version of this corpus. The conversion is based on the analytical (surface syntax) level of annotation in PDT. There is one exception though: zeros participating in coreference chains are taken from the tectogrammatical (deep syntax) level of annotation in PDT, and their syntactic attachment in the enhanced UD graph is a guess based on their original attachment and a few heuristics. Mention spans are based on the maximum projection of the original tectogrammatical tree, which is subsequently mapped on the nodes in the converted UD structure. The conversion procedure will be improved in the future.

The same procedure as for PDT is also applied to PCEDT, although here the analytical annotation in the source data has not been manually checked. It would be desirable to rely more on the tectogrammatical source annotation (which is manual) but the existing conversion procedure cannot utilize it.

AnCora has manually checked annotation of syntactic constituents. For CorefUD 0.1, we convert the constituent structure to dependency structure using a set of head selection heuristics. We do not attempt at guessing the dependency relation type; instead, we choose deterministically among root, punct, and dep. Once we map the data on the UD tokens in the future, we will use the existing UD syntactic annotation.

Chapter 4

Our Harmonizing Scheme

4.1 Central design decisions and abstract structure of the data

The main building units in the target representation are **mentions** and **clusters**. A mention is a set of words within one sentence (these are syntactic words as defined in UD, that is, nodes in the dependency structure, including empty nodes – zeros). Mentions spanning multiple sentences are not supported. A mention is defined 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 (so-called *treelet*), yet we do not require it. See Table 5.5 for statistics on the percentage of discontinuous and non-treelet mentions.

Every mention is a member of one (and only one) cluster. The cluster ID (name) is thus the second required attribute of each mention, besides the mention's span. Singletons are clusters that contain only one mention. Clusters are typically bound to one document in our data but we require that cluster IDs are unique across the entire corpus (e.g., `c1` refers to the same cluster everywhere in Czech-PDT; however, it is not related to `c1` in Czech-PCEDT).

Mentions may have additional attributes. For clusters, there are two additional (and optional) attributes: the type of the cluster and the reference to subclusters in the case of split antecedent.

Bridging relations are interpreted as directed relations between two clusters, although our technical solution may seem to encode a mention-to-cluster relation. Relations between two mentions are currently not well supported, although we admit that such relations may be needed in future versions.

4.2 Specific decisions

4.2.1 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 and add a special optional attribute EmptyType to these nodes. The attribute has the following values:

- NullSubj for dropped subjects, which we inserted into the Polish data, instead of the original marking coreference on the verbs.

- NullObj and NullPoss for dropped syntactic objects and possessives in Hungarian-SzegedKoref.
- Ellipsis (or EmpNoun) for elided nominal syntactic heads.
- Values for multiple reconstructed zeros in Czech-PDT and Czech-PCEDT: PersPron (for reconstructed personal pronouns), Gen (general unexpressed argument), Cor and QCor (controlled arguments of controlled predicates) and Rcp (second argument of reciprocal verbs).

In the current version, we preserve information about ellipses in Polish-PCC, ParCorFull, Czech-PDT, Hungarian-SzegedKoref and other corpora, but it still needs to be unified. So now we have np=nominal ellipsis in ParCorFull, #EmpNoun in PDT, EmptyType=ellipsis in Polish-PCC and attributes NullSubj, NullObj, NullPoss in Hungarian-SzegedKoref, etc. This kind of information will be thoroughly analysed and systematized in the later versions of CorefUD.

4.2.2 Grouping mentions with identical reference

As shown in Table 3.2, cluster-based representation of coreference slightly prevails. We decided to represent coreference in the cluster-based style also because of its simplicity. While conversion of cluster-based corpora is lossless, conversion of link-based corpora may result in omission of the clusters' inner structure and link-related annotation. For example, tags associated with links are not fully captured in Lithuanian-LCC. Although they are represented in the anaphor's `MentionMisc` attribute, the information to which antecedents the tags relate is currently lost. Nevertheless, a detailed inspection of these tags in the future may suggest a way of capturing all information they carry within the cluster-based style.

4.2.3 Singletons

Both singletons and non-singletons are treated as clusters; a singleton cluster contains just a single mention. In the result, there are substantially more unique cluster IDs for the annotation projects that include annotation of singletons.

In the current version of CorefUD, all the information about mentions and clusters is preserved, but nothing extra is added. In the future versions, we may add singletons to datasets which did not have them originally, using the output of the UD syntactic parsing.

4.2.4 Bridging

In the current CorefUD 0.1, bridging relations are understood very broadly. For now, this is rather a relation trash can, as it is used for all relations which are annotated in different coreference annotation schemes and cannot be considered as identity coreference. We take everything and (almost always) name them in the same way as they are named in original sources.

To record bridging relations, we use the attribute Bridging. It consists of the cluster ID and the name of the relation. The attribute Bridging connects identity clusters, one identity cluster may be part of more than one bridging relation.[1] For example, Bridging=c1234:Part,c9874:Subset says that the cluster of the current mention is related to cluster c1234 with the part-whole bridging relation, and to the cluster c9874 with the subset bridging relation.

[1] Technically, Bridging connects a mention to the corresponding identity cluster, but we understand the relation as cluster-to-cluster and do not copy the relation to all mentions of the cluster.

4.2.5 Split antecedents

In CorefUD, we have a special attribute SplitAnte, its value being two or more existing coreference id clusters. For example, if an attribute SplitAnte has a value c12+c34 (SplitAnte=c12+c34), it naturally means that the given cluster anaphorically refers to clusters c12 and c34.

The attribute SplitAnte is the property of clusters, saying that a cluster with a given ClusterId is equivalent to the union of smaller clusters which are listed in the value of the SplitAnte attribute. However, a cluster with non-empty SplitAnte has its own ClusterID too, without the "+" sign.

4.2.6 Apposition and predication

In the current version of CorefUD, the information about apposition and predication remains in the original form, as it is solved in individual annotation approaches. So, predication remains a value of mentions in Russian-RuCor, ParCorFull or English-ARRAU, but it is marked by a relation type, e.g., in Polish-PCC.

The unification will be provided in the upcoming versions of CorefUD.

4.2.7 Bound anaphora

In the current CorefUD 0.1, we leave all bound anaphoras where they are. Their harmonization is a question of the future versions.

4.2.8 Discourse deixis

In the current version of CorefUD, given that most of our datasets annotate discourse deixis together with other types of coreference, we just trasferred the data as it was, without attempting to harmonize such cases. If needed, it may be done in the later versions of the project.

4.2.9 Miscellaneous information about clusters and mentions

In the unified CorefUD, we created an optional attribute ClusterType for the properties which concern the whole cluster. The attribute values are attached to ClusterId (to all mentions or to a selected representative[2]) and they are the same for the whole cluster.

Values concerning individual mentions are stored in the attribute MentionMisc. In CorefUD version 0.1, we copy all information from the annotation schemes to that attribute. In future versions, it will be needed to provide a number of modifications and unify the data. This should be done, for example, for information about genericity of NPs, which is treated differently in different annotation projects (compare Czech-PDT, English-ARRAU, AnCora, ParCorFul, Russian-RuCor etc.). Predication and apposition represent an issue of their own (see Section 3.4.2).

4.3 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. (Note however that UD has additional requirements, which only some of our datasets comply with. Most notably, a UD-released

[2]Not finalized yet.

treebank must have manually checked POS tags and dependency relations; in most of our datasets, this kind of annotation has been assigned automatically.)

We stick to the specification of the **CoNLL-U format**[3] (as opposed to the CoNLL-U Plus extension,[4] which would allow for extra columns for the coreference-related attributes, but unfortunately it 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).[5]

Within the specification of the CoNLL-U format, there are multiple options how and where to store coreference-related annotation:

1. Coreference relations could be stored in the DEPS column as additional edges of the enhanced dependency graph. This would mean that we would use the link-based, rather than cluster-based representation of coreference (see Section 3.2.1). Moreover, it would not be possible to represent coreference across sentence boundaries (without extending the definition of node ID). We ruled this option out.

2. Coreference information could be stored in sentence-level comment lines, e.g., in a JSON data structure. While this approach would allow to express almost anything, it would jeopardize human readability of the annotation, as well as processing of the data with simple command-line tools such as `grep`. We ruled it out.

3. Coreference information could be stored at word level in new attributes in the MISC column. This is the option which we selected. And while we deliberately avoid the CoNLL-U Plus file format, we argue that this option is very close to it. Users who prefer additional columns for coreference annotation can easily extract the coreference-related attributes from MISC and put them in separate columns.

We introduce the MISC attributes listed below. With the exception of EmptyType, the attributes pertain to a mention. If the mention spans multiple nodes, a representative node is selected and the mention is annotated on the line corresponding to that node. If possible, we select the syntactic head of the mention as the representative node (see Section 4.6). The newly added attributes are divided to obligatory (present in the MISC column with each mention in each dataset) and optional (present only with some mentions in some datasets).

Obligatory attributes:

- MentionSpan
  - Example: `MentionSpan=4-8,10.1-13`
  - This attribute is required for every mention.
  - The value identifies the nodes of the current sentence that belong to the mention. A node is identified by its ID from the ID column of the CoNLL-U file. The ID is a positive integer for regular nodes and a decimal number for empty nodes.

[3] `https://universaldependencies.org/format.html`

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

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

  - The node IDs must be ordered by their order in the sentence. Note that even empty nodes have a defined position in a UD-annotated sentence. The part of the ID after the decimal point is to be interpreted as a separate minor number: In the unlikely case that there are more than 9 empty nodes between two regular nodes, "`1.10`" is ordered after "`1.9`".
  - Two or more consecutive nodes must be indicated as an interval with a hyphen (e.g., "`2-3`"). Note that such intervals also include all empty nodes that fall in the interval (e.g., "`2-3`" includes "`2.1`" and "`2.2`" but not "`3.1`").
  - Non-adjacent subspans are separated by a comma ("`,`").
  - The current node (that is, the node at which the mention is annotated) must be included in the span.

- ClusterId
  - Example: `ClusterId=c10`
  - This attribute is required for every mention.
  - The ID must be unique in the entire corpus so that individual CoNLL-U files can be joined without having to relabel clusters.
  - In the released data, we stick to IDs of the form "c*N*" where *N* is a positive integer. However, in the API we allow arbitrary strings, with the exception of whitespace characters and the characters "`|`", "`=`", "`:`", "`,`", "`+`". Users may want to use this to achieve corpus-wide uniqueness cheaply by including document IDs in their cluster IDs.
  - It may happen that a node serves as the representative of multiple overlapping, yet distinct mentions. In that case (and only in that case), the mention-related MISC attributes are accompanied by a numeric index in square brackets that helps identify attributes pertaining to the same mention. The indices start at 1 and form a contiguous sequence. For example: `ClusterId[1]=c3|ClusterId[2]=c15|MentionSpan[1]=3-6|MentionSpan[2]=3-8`

Optional attributes:

- ClusterType
  - Example: `ClusterType=Gen`
  - This attribute is optional but if it is present, then it must be present at every mention of the given cluster and its value must be identical at all occurrences.
  - The value is a string that describes the type of the entity or event corresponding to the cluster. The set of possible values will be further refined in future versions of CorefUD. At present we distinguish the following values:
    * Gen ... a generic entity, e.g., *officers*. This value exists as `GEN` in Czech-PDT. In AnCora we map the value `nne` here, which stands for non-named entity (thus it might be delimited slightly wider, including entities that would be considered specific in Czech-PDT).
    * Spec ... a specific entity or event, e.g., *Václav Havel*. This value exists as `SPEC` in Czech-PDT. In AnCora we map the values `ne` and `spec` here. The former stands for named entity, the latter for a mention coreferential with a named entity. We furthermore add the type of the named entity from AnCora, resulting in Spec.organization, Spec.person, Spec.location, Spec.date, Spec.number, and Spec.other (publications, prizes, laws etc.)

- SplitAnte
  - Example: `SplitAnte=c12,c34`
  - This attribute is optional but if it is present, then it must be present at every mention of the given cluster and its value must be identical at all occurrences.
  - The value is a comma-separated list of two or more different cluster IDs. They must refer to existing clusters, excluding the current cluster.
  - The interpretation of the attribute is such that the current cluster is defined as the union of two or more smaller clusters. The attribute is used in situations where an anaphoric mention has a split antecedent (see Example 4 in Section 1.4).
- Bridging
  - Example: `Bridging=c1234:Part,c9874:Subset`
  - This attribute is optional. Unlike ClusterType and SplitAnte, it is not repeated at every mention of the current cluster even though it describes relations between clusters and not individual mentions.
  - The value is a comma-separated list of bridging relations to other clusters. Each relation is a colon-separated pair, where the first item is the ID of the target cluster, and the second item is the type of the relation.
  - Similarly to SplitAnte, the list members must be ordered by cluster IDs. It is not possible to have two different bridging relations to the same target cluster; note that this constraint affects also bridging relations that may be annotated at another mention of the current cluster! Furthermore, it is obviously not allowed to mark a bridging relation to the current (source) cluster.
  - The set of possible relation types will be further refined in future versions of CorefUD. At present we distinguish the following values (among others):
    - Part ... the entity corresponding to the source cluster is a part of the entity corresponding to the target cluster. For example, a *steering wheel* is a part of a *car* (meronymy); *municipalities* is a part of *regions*. This value exists as `WHOLE_PART` or `PART_WHOLE` in Czech-PDT.[6]
    - Subset ... the entity corresponding to the source cluster is a subset of the entity corresponding to the target cluster. Several semantic relations are represented as subsets: (i) generic expression ← specific example; (ii) category ← subcategory (e.g. *public servants* ← *congressmen*); set of entities ← one non-specific entity from the set (e.g. *congressmen* ← *a congressman*). This value exists as `SET_SUB` or `SUB_SET` in Czech-PDT.[7]
    - Funct ... the entity corresponding to the source cluster represents one of the functions of the entity corresponding to the target cluster. For example, *the premier* is a function of the entity *government*. This value exists as `P_FUNCT` or `FUNCT_P` in Czech-PDT.[8]

[6] In PDT, the bridging relations are always directed from later mentions to earlier ones. `WHOLE_PART` means that the whole entity was mentioned before the part, `PART_WHOLE` means the opposite.

[7] `SET_SUB` means that the superset was mentioned before the subset, `SUB_SET` means the opposite.

[8] `P_FUNCT` means that the larger entity was mentioned before the smaller one, `FUNCT_P` means the opposite.

    - Anaf ... the entity corresponding to the source cluster contains a demonstrative or a similar expression and refers to something mentioned earlier in the discourse; the two mentions are neither coreferential nor there is any other, more specific bridging relation. For example, the target cluster may denote the event *Austria attacked Hungary*, and the source cluster may contain the mention *in that time*. This value exists as `ANAF` in Czech-PDT.
    - Other ... a bridging relation that does not belong to any of the categories mentioned above. For example, *Spain – Spaniard*, *mother – son*, *author – work*, *listening – listener*, *ropewalker – rope* etc. This value exists as `REST` in Czech-PDT.
  - Since bridging relations are asymmetric, the semantic type of the relation determines which cluster is the source and which is the target (with the exception of Other). This means that relations may have to be inverted when converting annotation from certain datasets (such as Czech-PDT or PCEDT).

- EmptyType
  - Example: `EmptyType=PersPron`
  - This attribute is optional, it occurs at an empty node and it pertains just to the node, not to a mention.
  - The attribute distinguishes types of empty nodes or the reasons why an empty node was generated.
  - The set of possible values will be further refined in future versions of CorefUD. At present we distinguish the following values:
    - PersPron / ProDrop / NullSubj / NullObj ... an unexpressed actant (often subject) of a predicate that can be interpreted as a personal pronoun. In the case of unexpressed subjects, the form of the pronoun can be often inferred from the form of the verb; however, the matter is complex and some PersPron nodes are not subjects.
    - Gen ... general actant (it can be interpreted as an unexpressed indefinite pronoun *somebody* or *something*). A typical example where we may need a Gen node in CorefUD is a general beneficiary of an action, linked by grammatical coreference with another node in the sentence.
    - Cor / QCor ... unexpressed subject of a controlled predicate (typically infinitive) in control-verb constructions. It is coreferential with an actant of the matrix verb. This type is used temporarily in conversions of Prague treebanks. In the future versions the annotation will be made more UD-like by merging the Cor node with its antecedent (provided the antecedent is in the same sentence) and redirecting enhanced dependency relations accordingly.
    - Rcp ... the second actant of a reciprocal verb.

- MentionMisc
  - Example: `MentionMisc=someInfo,otherInfo,someKey:val`
  - This attribute is optional.
  - The value is any string that does not contain a newline, a tab, a vertical bar ("|") or an equals-to sign ("=").

– This attribute serves to preserve information from the original resource that pertains to a mention and that we cannot harmonize at present. The purpose of serializing such information in MentionMisc is to preserve its relationship to a particular mention, even when the annotation is manipulated via the API. Annotations that pertain to a single node (rather than a mention) can be put directly as new attributes in the MISC column.

4.4 Application interface (API) for processing the data

Udapi[9] is an open-source Python framework providing an application programming interface (API) for processing Universal Dependencies data and the CoNLL-U format. Newly, it supports also coreference in the CorefUD format.

There are two main new classes `CorefCluster` and `CorefMention` and several new methods in the existing classes `Document` and `Node`.

- `Document`
  - `doc.coref_clusters`
    returns a set of all clusters in the document. The set is represented as a dictionary (`dict`) mapping cluster IDs to `CorefCluster` instances.
- `Node`
  - `node.coref_mentions`
    returns a list of mentions (`CorefMention` instances) whose span includes a given `node`. The list will be empty if a given node is not part of any mention. A list of mentions whose head is in a given `node` can be obtain using
    `[m for m in node.coref_mentions if m.head is node]`.
  - `node.coref_clusters`
    returns a list of coreference clusters whose mentions span a given `node`. This method is a shortcut for
    `[m.cluster for m in node.coref_mentions if m.cluster is not None]`.
  - `node.create_coref_cluster(cluster_id, cluster_type, mention_words/span)`
    Creates and returns a new `CorefCluster` with a single `CorefMention` whose head is set to the current `node`. If no `cluster_id` is given, the nearest unused ID (c1, c2,...) is used. Either `mention_words` (a list of `Node` instances) or `span` (a string specifying the mention span, e.g. `1-3,5`) can be given.
- `CorefCluster`
  - `c.create_mention(head, mention_words/span)`
    Creates and returns a new `CorefMention` within the current cluster `c`. If no `head` is specified, the first node from `mention_words` is used instead. If `head` is specified, it must be one of the `mention_words`.
  - `c.mentions`
    returns a list of mentions (`CorefMention` instances) in the current cluster `c`. The returned list should not be modified (e.g. using `c.mentions.append(new_mention)`) because that would result in inconsistencies (it would not set `new_mention.cluster`). The recommended way for adding a new mention is `new_mention = c.create_mention(...)` or `new_mention.cluster = c`.
  - `c.cluster_type`
    a read/write string property specifying the cluster type (e.g. Gen, Spec).

[9] `https://udapi.github.io`

- `c.split_ante`
  a read/write property with a list of `CorefCluster` instances, which are split antecedents of the current cluster `c`.
- `c.all_bridging`
  an iterator over bridging links of all mentions in the current cluster `c`. The code
  `for b in c.all_bridging`
  is a shortcut for
  `for m in c.mentions: for b in m.bridging`.

- `CorefMention`
  - `m.head`
    a read/write property with the head node. It must be always defined (never `None`).
  - `m.words`
    a read/write property with a list of all words (instances of the class `Node`) in the current mention `m` (including the head node). It is possible to edit the list as `m.words = [node1, node2]` (provided `m.head in [node1, node2]`). However, it is not recommended to edit the list in-place, e.g. with `m.words.append(node1)` or `m.words.remove(node1)` because that would not update `node1.coref_mentions`.
  - `m.span`
    a read/write property specifying the mention span as a string. When reading the property, it is computed on the fly based on `m.words`. When editing the property, e.g. `m.span = '3.1-5.2,8-11'`, the list of words `m.words` is updated accordingly.
  - `m.cluster`
    a read/write property specifying the `CorefCluster` where the current mention `m` belongs.
  - `m.bridging`
    blowjob airplane a property specifying a list bridging relations, which are represented using a helper class `BridgingLinks`. Example usage:

```
# assigning
m._bridging = BridgingLinks(m, [(c12, 'Part'), (c56, 'Subset')])
# or alternatively using a string specification
m._bridging = BridgingLinks(m, 'c12:Part,c56:Subset', doc.coref_clusters)

for cluster, relation in m.bridging:
     print(f"{bl.src_mention} ->{relation}-> {cluster.cluster_id}")
print(str(m.bridging)) # c12:Part,c56:Subset

# BridgingLinks.__call__ for obtaining a subset of the links
m.bridging(relations_re='Part').targets == [c12]
m.bridging(relations_re='Part|Subset').targets == [c12, c56]
m.bridging.append((c89, 'Funct'))
```

4.4.1 Example API usage

First, let's load a CoNLL-U file and draw the first sentence. (Note the error in automatic parsing of "Sergei Ivanov".)

```
>>> import udapi
>>> doc = udapi.Document("en_parcorfull-corefud-dev.conllu")
>>> doc[0].draw(attributes="ord,form,upos,deprel,misc")
```

```
# sent_id = 222
# text = Russia 's Putin sacks chief of staff Sergei Ivanov
 ┌─┬─ 1 Russia PROPN nmod:poss _
 │ └─ 2 's PART case _
 ├─ 3 Putin PROPN nsubj ClusterId=c156|MentionMisc=mention:np,nptype:antecedent|MentionSpan=1-3
 ├─ 4 sacks VERB root _
 ├─┬─ 5 chief NOUN obj ClusterId=c157|MentionMisc=mention:np,nptype:antecedent|MentionSpan=5-9
 │ ├─ 6 of ADP case _
 │ ├─ 7 staff NOUN nmod _
 │ └─ 8 Sergei PROPN flat _
 └─ 9 Ivanov PROPN flat _
```

Now, print a listing of all clusters and mentions (summing mentions with the same forms).

```
>>> from collections import Counter
>>> for cluster in doc.coref_clusters.values():
...     print(f" {cluster.cluster_id} has {len(cluster.mentions)} mentions:")
...     counter = Counter()
...     for mention in cluster.mentions:
...         counter[' '.join([w.form for w in mention.words])] += 1
...     for form, count in counter.most_common():
...         print(f"{count:4}: {form}")
 c156 has 21 mentions:
  11: Mr Putin
   2: his
   2: he
   1: Russia 's Putin
   1: Russian President Vladimir Putin
   1: Vladimir Putin
   1: him
   1: President Putin
   1: Putin
 c157 has 19 mentions:
   7: Mr Ivanov
   3: his
   2: Ivanov
   2: He
   1: chief of staff Sergei Ivanov
...
```

Language	UDPipe model
Dutch	Dutch-LassySmall
English	English-GUM
German	German-HDT
Hungarian	Hungarian-Szeged
Lithuanian	Lithuanian-ALKSNIS
Polish	Polish-LFG
Russian	Russian-SynTagRus

Table 4.1: UDPipe models used for particular languages to enrich CorefUD with morpho-syntactic annotation

4.5 Adding UD annotations

Due to reasons explained in at the beginning of Chapter 1, CorefUD combines coreference annotation with annotation of morphology and dependency syntax. 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). However, as seen in Table 4.2 the majority of coreference corpora is not equipped with all required morpho-syntactic annotation. We thus enrich these corpora with additional annotation automatically, employing UDPipe 1 (Straka and Straková, 2017) and its models trained on UD 2.5 (see Tabel 4.1). The automatic processing includes lemmatization, part-of-speech tagging (including morphological features), and dependency parsing.

The automatic morpho-syntactic annotation is built on top of the sentence segmentation and tokenization coming from the original sources. On the one hand, this approach simplifies merging the coreference and morpho-syntactic annotation coming from different sources. On the other hand, quality of the morpho-syntactic annotation may be lower, as the tokenization on which the UDPipe models were trained may be different. If the original sources are not segmented or tokenized (e.g. Hungarian-SzegedKoref and Lithuanian-LCC), we apply simple rule-based approaches that ensure that mention boundaries correspond to token boundaries and no mention is split into two or more sentences. In the future, we plan to use tokenization provided by UDPipe instead. This will, however, require establishing a mapping between tokens coming from the original sources and from UDPipe.

A few of the original sources already contain some morpho-syntactic information. Nevertheless, this is sometimes incomplete or uses different set of labels that would have to be converted to labels required by the UD guidelines. The latter would include non-trivial conversions, such as transformation of constituency trees to dependencies (e.g. for OntoNotes). To accelerate the work on the harmonization of coreference annotation, we decided not to exploit such morpho-syntactic annotation for the initial version of CorefUD.

4.6 Moving "head" to dependency head of the mention

All information related to the mention is stored in its representative node. If possible, we require this node to be a syntactic head of the mention. Nevertheless, some of the original sources do not annotate mention heads (e.g. Lithuanian-LCC), some of them annotate them with no explicit relation to syntax (e.g. English-ARRAU) and some of the sources adopt a different style of syntax representation than UD (e.g. Czech-PDT). In order to unify these, we attempt to move the mention head to a syntactic head of the mention with respect to the basic UD tree representation of the sentence.

CorefUD dataset	sentence segmentation		tokenization		POS tags		lemmas		syntactic trees	
	orig.	**new**	**orig.**	**new**	**orig.**	**new**	**orig.**	**new**	**orig.**	**new**
Catalan-AnCora	✓	UD2.7	✓	kept	✓	convert	✓	convert	✓ (phr.)	convert
Czech-PCEDT	✓	kept	✓	convert	✓	convert	✓	convert	(✓) (dep.)	convert
Czech-PDT	✓	kept	✓	convert	✓	convert	✓	kept	✓ (dep.)	convert
English-GUM	✓	kept	✓	kept	✓	kept	✓	kept	✓ (dep.)	kept
English-ParCorFull	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
French-Democrat	(✓)	kept	(✓)	kept	(✓)	kept	(✓)	kept	(✓) (dep.)	kept
German-ParCorFull	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
German-PotsdamCC	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
Hungarian-SzegedKoref	×	rules	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
Lithuanian-LCC	×	rules	×	rules	×	UDPipe	×	UDPipe	×	UDPipe
Polish-PCC	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	×	UDPipe
Russian-RuCor	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	×	UDPipe
Spanish-AnCora	✓	UD2.7	✓	kept	✓	convert	✓	kept	✓ (phr.)	convert
Dutch-COREA	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
English-ARRAU	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	✓ (phr.)	UDPipe
English-OntoNotes	✓	kept	✓	kept	✓	UDPipe		UDPipe	✓ (phr.)	UDPipe
English-PCEDT	✓	kept	✓	kept	✓	convert	✓	kept	(✓) (d+p.)	convert d.

Table 4.2: Additional annotations stored in the CorefUD data (other than coreferential). A check mark in the 'orig.' column means that the annotation is available in the source data (with the further distinction whether source syntax is based on dependencies or phrases), otherwise a cross mark is used. A check mark in parentheses means that the source annotation is not manually verified. 'UD2.7' indicates that the annotation has been taken from the release 2.7 of UD, instead of the original source. The following shortcuts are used in the column describing syntactic trees available with the original resources: 'dep' stands for dependency trees, 'phr.' for phrase-structure (constituency) trees, 'd+p.' for both.

The procedure is simple if the mention forms a single treelet (i.e. connected subgraph of the basic-dependency tree) – we move the head to the root of the treelet. If there are multiple nodes whose parent lies outside the mention or if the mention contains empty nodes (which have by definition no parent in the basic dependency tree), we choose the head according to the enhanced dependency graph, where each empty node is represented by its non-empty enhanced parent. We may still end up with multiple head candidates, i.e. multiple nodes whose enhanced parent lies outside the mention. In that case, we try to choose such candidate which governs all other candidates (in the basic dependencies).

If none of the rules above results in a single head, we conservatively try to preserve the original head. It must belong to the list of head candidates, though. Otherwise, we pick the first head candidate according to the word order.

Chapter 5

Resulting Collection CorefUD 0.1

5.1 Introducing the train/dev/test split

As is the common practice, 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; this was the case of

- Catalan-AnCora (the division of UD 2.7),
- Czech-PCEDT (sections 00–18 to train, 19–21 to dev, 22–24 to test),
- Czech-PDT,
- English-ARRAU,
- English-GUM,
- French-Democrat,
- Spanish-AnCora (the division of UD 2.7),
- English-OntoNotes,
- English-PCEDT (sections 00–18 to train, 19–21 to dev, 22–24 to test).[1]

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. When iterating, we followed the ordering of the documents if they were serialized in a single file in the original dataset, or we followed the lexicographic ordering of files if each document was stored in a single file (or in a bunch of similarly named files, like in the MMAX format). The deviation from the desired 80:10:10 percent division follows naturally from the fact that some datasets contain only a small number of documents and/or the sizes of documents differ considerably. A better fit to 80:10:10 would be possible to find, but it might induce the risk of systematically biased distributions.

The resulting sizes of the three sections of each dataset are presented in Table 5.1.

[1] `https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)`

CorefUD dataset	total size				division [%]		
	docs	sents	words	empty	train	dev	test
Catalan-AnCora	1550	16,678	488,379	6,377	78.6	10.7	10.8
Czech-PCEDT	2312	49,208	1,155,755	45,158	77.8	11.2	11.0
Czech-PDT	3165	49,428	834,721	33,086	78.3	10.6	11.1
English-GUM	150	7,408	134,474	0	76.0	12.0	11.9
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	123,976	4,849	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	538,891	464	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	517,258	8,111	80.9	9.5	9.6
Dutch-COREA	844	9,270	140,063	0	78.6	10.0	11.4
English-ARRAU	413	8,735	228,901	0	79.9	5.6	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	36,740	77.7	11.2	11.0

Table 5.1: 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 5.1 for details). 'words' is the number of non-empty UD nodes (corresponding to syntactic words). 'empty' is the number of empty UD nodes.

5.2 Releasing and licensing policy

The data described in this report can be divided to two parts. The larger part is public, meaning that the original resources come with a free license that allows modification and redistribution, at least for non-commercial users. This part is what we release as the CorefUD 0.1 package in the Lindat repository (`http://hdl.handle.net/11234/1-3510`).

The other part is internal: We include it in our experiments and report statistics collected from the data but we cannot redistribute it. In the tables throughout this report, a horizontal line separates the public part (above the line) from the internal part (below).

Unless negotiated otherwise with the authors of the original resource, we distribute the public resources under the same licenses that the original resources came with. As a result, the CorefUD 0.1 package has a mixed license, with different terms applying to different subsets. The individual licenses are listed in Table 1.1.

5.3 Statistical properties

Tables 5.1–5.6 provide some statistics of the individual datasets in CorefUD 0.1.

CorefUD dataset	clusters				distribution of lengths				
	total	per 1k	length		1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	69,241	142	101	1.6	74.6	14.1	4.7	2.2	4.4
Czech-PCEDT	52,743	46	247	3.4	1.4	62.8	15.6	6.8	13.4
Czech-PDT	78,879	94	186	2.5	35.3	38.9	11.0	5.2	9.5
English-GUM	20,989	156	131	1.8	75.0	13.8	4.7	2.2	4.4
English-ParCorFull	180	17	38	4.1	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	260	25	43	3.5	5.8	65.4	11.5	5.0	12.3
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	7.9	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	127,694	237	136	1.5	82.6	9.8	2.9	1.4	3.2
Russian-RuCor	3,614	23	141	4.5	2.5	54.1	15.7	6.9	20.7
Spanish-AnCora	73,218	142	110	1.7	73.4	14.8	4.7	2.4	4.7
Dutch-COREA	28,548	204	31	1.2	88.3	8.3	2.1	0.6	0.8
English-ARRAU	48,336	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	1.1	62.5	15.9	7.1	13.5

Table 5.2: 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.

CorefUD dataset	mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	62,417	128	134	4.2	10.2	34.6	19.6	7.5	4.5	23.7
Czech-PCEDT	178,475	154	79	3.4	23.0	28.5	16.1	8.3	4.1	20.0
Czech-PDT	169,644	203	99	2.9	17.2	36.4	18.7	8.5	4.1	15.1
English-GUM	22,896	170	95	2.6	0.0	54.8	20.6	8.4	3.9	12.3
English-ParCorFull	720	67	37	2.1	0.0	59.0	24.4	6.0	2.9	7.6
French-Democrat	47,172	166	71	1.7	0.0	64.2	21.7	6.4	2.5	5.3
German-ParCorFull	900	85	30	2.0	0.0	65.0	17.4	6.2	4.0	7.3
German-PotsdamCC	2,523	76	34	2.6	0.0	34.8	32.4	15.5	6.4	10.9
Hungarian-SzegedKoref	15,182	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	82,865	154	108	2.1	0.3	68.7	14.9	5.2	2.7	8.2
Russian-RuCor	16,254	104	18	1.7	0.0	68.9	16.3	6.7	3.5	4.6
Spanish-AnCora	70,675	137	90	4.4	11.4	35.3	17.6	7.6	4.0	24.1
Dutch-COREA	8,663	62	60	2.6	0.0	42.5	33.1	8.6	4.0	11.7
English-ARRAU	31,906	139	75	2.9	0.0	45.4	26.9	10.7	4.2	12.8
English-OntoNotes	209,435	128	94	2.5	0.0	56.3	19.8	8.1	4.2	11.7
English-PCEDT	183,984	157	88	3.6	19.3	28.0	17.0	10.6	4.8	20.3

Table 5.3: 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.

CorefUD dataset	mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	51,683	106	153	4.3	0.0	25.2	23.4	13.2	9.9	28.3
Czech-PCEDT	713	1	47	5.2	11.4	8.7	21.7	14.4	9.8	33.9
Czech-PDT	27,834	33	71	3.6	1.5	23.3	26.9	18.2	9.3	20.7
English-GUM	15,739	117	82	3.7	0.0	27.4	26.9	12.9	8.2	24.6
English-ParCorFull	11	1	1	1.0	0.0	100.0	0.0	0.0	0.0	0.0
French-Democrat	33,504	118	55	3.3	0.0	17.9	36.8	14.0	7.9	23.3
German-ParCorFull	15	1	2	1.1	0.0	93.3	6.7	0.0	0.0	0.0
German-PotsdamCC	2,871	86	28	3.7	0.0	16.0	28.6	18.9	11.1	25.3
Hungarian-SzegedKoref	410	3	14	2.3	0.5	25.9	35.9	25.6	8.8	3.4
Lithuanian-LCC	137	4	8	1.5	0.0	76.6	14.6	3.6	0.7	4.4
Polish-PCC	105,507	196	147	3.5	0.1	33.8	25.5	12.3	7.2	21.2
Russian-RuCor	91	1	10	2.3	0.0	44.0	30.8	7.7	7.7	9.9
Spanish-AnCora	53,771	104	95	4.7	0.1	22.8	23.7	13.5	9.4	30.6
Dutch-COREA	25,207	180	50	2.9	0.0	38.3	30.2	9.9	5.2	16.4
English-ARRAU	40,102	175	65	5.1	0.0	17.4	21.6	14.2	9.9	36.9
English-OntoNotes	198	0	15	2.3	0.0	60.6	16.2	8.1	5.6	9.6
English-PCEDT	576	0	45	6.5	1.0	8.9	19.3	20.0	12.3	38.5

Table 5.4: Statistics on 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. Note that certain amount of singletons occur even in datasets that do not target them. It is because we create clusters also for mentions that participate in bridging.

CorefUD dataset	mention type [%]			distribution of head UPOS [%]								
	w/empty	w/gap	non-tree	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
Catalan-AnCora	7.1	0.0	0.0	51.1	14.7	24.9	2.5	0.5	1.4	0.0	4.9	0.0
Czech-PCEDT	30.9	4.1	9.7	43.3	27.5	7.0	13.4	1.1	2.9	1.3	0.7	2.9
Czech-PDT	19.6	3.1	2.8	47.5	20.0	11.7	9.5	6.0	2.1	1.7	0.9	0.6
English-GUM	0.0	0.0	1.5	53.9	21.8	17.0	0.0	0.8	1.7	0.3	4.0	0.5
English-ParCorFull	0.0	0.7	2.6	24.1	46.1	24.2	0.7	0.3	2.3	0.7	0.8	0.8
French-Democrat	0.0	0.0	2.0	52.9	27.6	8.2	7.2	0.4	1.7	0.8	0.3	0.8
German-ParCorFull	0.0	0.3	1.9	27.5	47.0	18.8	1.3	0.3	2.6	1.3	0.2	0.9
German-PotsdamCC	0.0	6.3	5.4	66.7	15.7	10.1	0.6	1.4	0.5	3.3	0.0	1.7
Hungarian-SzegedKoref	15.2	0.4	3.3	50.6	13.4	6.2	1.7	2.1	3.6	6.9	0.2	15.4
Lithuanian-LCC	0.0	0.0	4.7	42.5	13.0	22.9	4.9	0.3	2.7	1.1	0.8	12.0
Polish-PCC	0.5	1.0	13.5	60.4	8.1	9.2	1.9	3.7	11.9	0.9	0.8	3.2
Russian-RuCor	0.0	0.5	4.5	39.2	26.4	23.4	8.2	0.9	0.7	0.2	0.5	0.4
Spanish-AnCora	8.5	0.0	0.0	51.4	15.7	22.3	3.5	0.9	2.1	0.0	4.0	0.0
Dutch-COREA	0.0	0.3	5.9	63.1	11.6	11.4	1.4	2.7	5.0	1.6	1.2	1.9
English-ARRAU	0.0	1.2	13.1	55.8	10.7	18.6	0.7	2.7	3.8	0.7	3.5	3.5
English-OntoNotes	0.0	0.0	6.0	27.6	41.6	24.9	0.6	0.7	2.5	0.3	1.0	0.9
English-PCEDT	29.3	2.8	2.9	31.4	30.7	22.7	9.4	0.6	2.3	0.6	1.2	1.1

Table 5.5: 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. In Hungarian-SzegedKoref, 14.7% mentions have head UPOS=_, i.e. an unspecified tag (marked by the underscore in the CoNLL-U format), which is possible only if the head is an empty node. In Lithuanian-LCC, 11.3% mentions have head UPOS=X (mostly abbreviations).

CorefUD dataset	Bridging relations		Distributions of types	
	total	per 1k words	type	[%]
Czech-PCEDT	1169	1.0	Subset	100.0
Czech-PDT	30832	35.5	Subset	63.2
			Part	21.2
			Other	7.1
			Funct	5.8
			Anaf	2.7
English-GUM	1291	9.6	_	100.0
Hungarian-SzegedKoref	460	3.6	holonym	100.0
Polish-PCC	10715	19.9	indirect_aggregation	71.8
			indirect_composition	17.3
			indirect_other	8.4
			indirect_bound	2.5
Dutch-COREA	2972	21.2	bridge	82.9
			pred	17.1
English-ARRAU	3639	15.9	element	27.2
			subset	26.2
			subset-inv	12.8
			other	10.6
			unmarked	8.3
			undersp-rel	6.7
			element-inv	4.9
			poss	2.3
			poss-inv	0.8
			other-inv	0.2
English-PCEDT	828	0.7	Subset	100.0

Table 5.6: Distribution of bridging relation types.

Chapter 6

Conclusion

6.1 Contribution summary

We believe that the most important contributions of the presented work are the following:

- we presented a survey of coreference-related resources and analyzed their diversity from various viewpoints; to the best of our knowledge, no comparably broad survey has been published yet,
- 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 0.1; again, this is the widest coreference data collection we are aware of.

6.2 Disclaimer

While the CorefUD 0.1 data collection can be already used for experiments, it should be noted that the future versions will probably differ significantly: both in the overall specification and in the implementation details of the conversion of individual resources. We cannot guarantee complete backward compatibility at this stage.

At the same time, we cannot guarantee that the implementation of harmonization procedures is completely error-free (in spite of the fact that numerous tests have been applied on the collection, checking various kinds of consistency from different angles). We did our best, but, for example, we might have misunderstood the source file format in some way, and checking the correctness of the linguistic content after the conversion was hard for us too, especially in languages which we don't speak. However, we will be happy to remove any conversion error if reported by CorefUD users.

6.3 Future plans

The harmonization effort presented here is meant as a pilot study, which will hopefully fuel and support discussion within the coreference/anaphora research community. The decisions we took are open for refinement (or even more substantial changes) in the future versions. Above all, we assume that inspiring impulses will come from interactions with the Universal Anaphora initiative.

Naturally, there are two directions along which we plan to extend the CorefUD collection in the future. First, we would like to broaden the scope of linguistic phenomena whose annotation is harmonized; for example, we would like to unify annotation of similar types of bridging relation

under the same name, and to harmonize cases of discourse deixis, and also coreference with different semantic types.

Second, as shown in Section 2.3, there are quite a few other resources for whose conversion we simply did not have sufficient capacity so far; the higher priority is likely to be assigned to those that are available under free licenses, so that we can extend the public edition of CorefUD.

In addition, we should fix technical imperfections that we are aware of already now, but which need more time to be resolved properly; for instance, mentions that cross sentence boundaries are not handled properly by the conversion procedure for several resources.

Acknowledgements

We would like to thank all our colleagues from different annotation projects who were so kind to give us access to their datasets, comments and advise on the data and annotation structure. We especially thank Maciej Ogrodniczuk, Massimo Poesio, Sameer Pradhan, Veronika Vincze, Amir Zeldes, Svetlana Toldova, Olga Uryupina, Carole Tiberius, Iris Hendrickx, Bob Boelhouwer and others.

The present work has been supported by grants no. GX20-16819X (LUSyD) and 19-14534S of the Czech Science Foundation; LM2018101 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic; and EC/H2020/825303 (Bergamot) of the European Union.

Bibliography

Aktaş, B. and Stede, M. (2019). TwiConv annotation guidelines: Coreference.

BBN Technologies (2006). *Co-reference Guidelines for English OntoNotes*.

Böhmová, A., Cinková, S., and Hajičová, E. (2005). A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Bouma, G., Daelemans, W., Hendrickx, I., Hoste, V., and Mineur, A.-M. (2007). The corea-project: Manual for the annotation of coreference in dutch texts. In *University Groningen*.

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. European Language Resources Association.

Chen, H., Fan, Z., Lu, H., Yuille, A. L., and Rong, S. (2018). Preco: A large-scale dataset in preschool vocabulary for coreference resolution.

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

Çöltekin, Ç., Campbell, B., Hinrichs, E., and Telljohann, H. (2017). Converting the TüBa-D/Z treebank of German to Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 27–37, Gothenburg, Sweden. Association for Computational Linguistics.

Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The szeged treebank. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, TSD'05, page 123–131, Berlin, Heidelberg. Springer-Verlag.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y., and Dinarelli, M. (2016). Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2016)*, Konya, Turkey.

Dipper, S., Rieger, C., Seiss, M., and Zinsmeister, H. (2011). Abstract anaphors in german and english. In Hendrickx, I., Lalitha Devi, S., Branco, A., and Mitkov, R., editors, *Anaphora Processing and Applications*, pages 96–107, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fonseca, E., Sesti, V., Collovini, S., Vieira, R., Leal, A. L., and Quaresma, P. (2017). Collective elaboration of a coreference annotated corpus for portuguese texts. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017)*, volume 1881, Murcia, Spain.

Grishina, Y. (2017). CORBON 2017 shared task: Projection-based coreference resolution. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 51–55, Valencia, Spain. Association for Computational Linguistics.

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

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. 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.

Hajič, J. et al. (2006). Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused mt and cross-lingual pronoun prediction: Findings of the 2015 discomt shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Hirschman, L. and Chinchor, N. (1998). Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

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. European Language Resources Association (ELRA).

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139, Prague, Czech Republic. Association for Computational Linguistics.

Korzen, I. and Buch-Kromann, M. (2011). Anaphoric relations in the Copenhagen dependency treebanks. In *Proceedings of Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, pages 83–98.

Krasavina, O. and Chiarcos, C. (2007). Pocos - potsdam coreference scheme. In Boguraev, B., Ide, N., Meyers, A., Nariyama, S., Stede, M., Wiebe, J., and Wilcock, G., editors, *Proceedings of the Linguistic Annotation Workshop, LAW@ACL 2007, Prague, Czech Republic, June 28-29, 2007*, pages 156–163. Association for Computational Linguistics.

Krause, T. and Zeldes, A. (2014). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

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. Association for Computational Linguistics.

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92):11–15.

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). ParCorFull: a Parallel Corpus Annotated with Full Coreference. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Štěpánek, J., and Hajič, J. (2017). PDTSC 2.0 - spoken corpus with rich multi-layer structural annotation. In Ekštein, K. and Matoušek, V., editors, *Text, Speech, and Dialogue. TSD 2017*, volume 10415 of *Lecture Notes in Computer Science*, Cham, Switzerland. Springer.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

Mirzaei, A. and Safari, P. (2018). Persian discourse treebank and coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mírovský, J., Pajas, P., and Nedoluzhko, A. (2010). Annotation tool for extended textual coreference and bridging anaphora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 168–171, Valletta, Malta. European Language Resources Association.

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 Technology 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. 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. Association for Computational Linguistics.

O'Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K., and Palmer, M. (2018). AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Ogrodniczuk, M., Ng, V., Grishina, Y., and Pradhan, S., editors (2020). *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, Barcelona, Spain (online). Association for Computational Linguistics.

Poesio, M. (2004). The mate/gnome proposals for anaphoric annotation, revisited. In *In Michael Strube and Candy Sidner (editors), Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. (2004). The VENEX corpus of anaphora and deixis in spoken and written Italian.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

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. 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.

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

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. European Language Resources Association (ELRA).

Rodríguez, K. J., Delogu, F., Versley, Y., Stemle, E., and Poesio, M. (2010). Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, pages 157–163, Valletta, Malta.

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.

Shimazu, S., Takase, S., Nakazawa, T., and Okazaki, N. (2020). Evaluation dataset for zero pronoun in japanese to english translation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3630–3634, Marseille, France.

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and Díaz de Ilarraza, A. (2012). Mention detection: First steps in the development of a Basque coreference resolution system. In *Proceedings of KONVENS 2012 (Main track: oral presentations)*, pages 128–136, Wien, Austria.

Stede, M., editor (2015). *Handbuch Textannotation*. Potsdamer Kommentarkorpus 2.0.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. 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. European Language Resources Association (ELRA).

Toldova, S., Roytberg, A., Ladygina, A. A., Vasilyeva, M. D., Azerkovich, I. L., Kurzukov, M., Sim, G., Gorshkov, D. V., 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.

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

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. 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. Association for Computational Linguistics.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

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. European Language Resources Association.

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

Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., and Václ, J. (2015). *Discourse and Coherence. From the Sentence Structure to Relations in Text*, volume 14 of *Studies in Computational and Theoretical Linguistics*. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Praha, Czechia.

Žitkus, V. (2018). Lithuanian Coreference Corpus. CLARIN-LT digital library in the Republic of Lithuania.

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.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová, *Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description*

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman, *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies*

ÚFAL TR-2018-61 Lukáš Kyjánek, *Morphological Resources of Derivational Word-Formation Relations*

ÚFAL TR-2019-62 Zdeňka Urešová, Eva Fučíková, Eva Hajičová, *CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting (CzEngClass: Kontextová synonymie a valence sloves v bilingvním prostředí)*

ÚFAL TR-2019-63 Ján Faryad,
Identifikace derivačních vztahů ve španělštině

ÚFAL TR-2020-64 Marie Mikulová, Jan Hajič, Jiří Hana, Hana Hanová, Jaroslava Hlaváčová, Emil Jeřábek, Barbora Štěpánková, Barbora Vidová Hladká, Daniel Zeman,
Manual for Morphological Annotation. Revision for Prague Dependency Treebank – Consolidated 2020 release

ÚFAL TR-2021-65 Rudolf Rosa, *Technická zpráva o vývoji projektu THEaiTRE v roce 2020*

ÚFAL TR-2021-66 Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman,
Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages