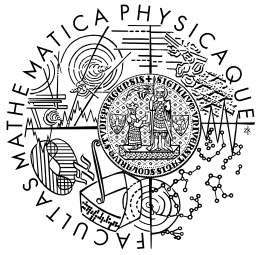

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Michal Novák

Coreference from the Cross-lingual Perspective

Institute of Formal and Applied Linguistics

Supervisor of the doctoral thesis: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Study programme: Informatics
Study branch: Mathematical Linguistics

Prague 2018

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Title: Coreference from the Cross-lingual Perspective

Author: Michal Novák

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. Ing. Zdeněk Žabokrtský, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The subject of this thesis is to study properties of coreference using cross-lingual approaches. The work is motivated by the research on coreference-related linguistic typology. Another motivation is to explore whether differences in the ways how languages express coreference can be exploited to build better models for coreference resolution. We design two cross-lingual methods: the bilingually informed coreference resolution and the coreference projection. The results of our experiments with the methods carried out on Czech-English data suggest that with respect to coreference English is more informative for Czech than vice versa. Furthermore, the bilingually informed resolution applied on parallel texts has managed to outperform the monolingual resolver on both languages. In the experiments, we employ the monolingual coreference resolver and an improved method for alignment of coreferential expressions, both of which we also designed within the thesis.

Keywords: coreference, anaphora, coreference resolution, anaphora resolution, cross-lingual processing, bilingually informed coreference resolution, coreference projection, word alignment

I would like to thank my supervisor Zdeněk Žabokrtský for his guidance. Although I was sometimes pessimistic, he has a gift for finding positive aspects anywhere, which I really admire.

Thanks to ÚFAL, the Institute of Formal and Applied Linguistics, that you are such a great place to work. Great thanks to some of my colleagues for their helpful comments. Especially to Anja Nedoluzhko, who is my sister-in-arms for the research related to coreference and she is also a wonderful person. And thanks to Eda Bejček for all those years we spent laughing in the office.

I would like to thank my family, especially to my mother and sister, who were always supportive despite having hard times in their lives along the way.

Many thanks to my friends that they never forgot to remind me about that rock I was pushing in front of me.

And, finally, I am grateful for all those places that kept me inspired while working on this thesis.

Contents

1. Introduction

The subject of this thesis is to study properties of coreference using cross-lingual approaches.

Before we start discussing the particular topics that this thesis deals with, let us put this work into the context. The research on coreferential and anaphoric relations at our institute dates back to mid 1980s [Hajičová et al., 1985, Hajičová, 1987, Panevová, 1992], continued with building coreference-annotated corpora in Czech [Hajič et al., 2006], and also collecting the parallel Czech-English data [Hajič et al., 2011, Nedoluzhko et al., 2016a]. Currently, we are involved in a research project that attempts to collect the multilingual parallel data of English, Czech, Russian and Polish [Nedoluzhko et al., 2018] in order to cross-lingually study the typological similarities and differences of the languages with respect to coreferential and anaphoric relations. The aim of the research is to explore the ways how coreference is expressed in different languages. The traditional language typology is based on general, mainly morphological and syntactic, similarities and differences of languages. Nevertheless, they do not necessarily accord with the similarities and differences in the ways how coreference is realized across languages. For instance, one of the aspects which is strongly related to coreference is the dropping of pronouns. The languages that can be considered pro-drop (to various degrees, e.g. Czech, Russian, Spanish, Italian, Japanese, Chinese, Arabic, Turkish, Swahili and even English) span across different types of languages in terms of classical typologies. A similar divergence could be observed for other aspects related to coreference, such as functions of reflexive and possessive pronouns, and the degree of nominalization and using deverbatives. The present thesis considerably contributes to this research by exploring these aspects on Czech and English.

Although the objectives of the project are rather theoretical, we adopt computational methods to reach them. Particularly, we make use of projection and bilingually informed resolution techniques, both of which aim at measuring the similarity or difference in languages. However, each of them utilizes different means to achieve it: (i) *Cross-lingual projection* of any linguistic phenomena from a source language to a target language is generally considered to work better for closely related languages. (ii) *Bilingually informed resolution*, in contrast, takes advantage of the information from the source language to help identify and disambiguate a particular linguistic phenomenon in the target language. It appears to be beneficial if the languages do not share many similarities. This project tries to apply these techniques to coreference relations.

The linguistic objectives affect the choice of the algorithms for the methods. For this purpose, we did not expect the proposed methods to outperform current state of the art. Instead, we implement a simple but interpretable solutions in order to help reveal the individual linguistic aspects that contribute on differences and similarities. Nevertheless, if even such simple method works well, i.e. the bilingually informed system gains a lot of beneficial information from the other language, it opens the door for being used also for natural language processing. And this is, apart from the motivation related to linguistic typology, the other motivation of the present work – to explore the possibilities of using the bilingually

informed system to improve coreference resolution.

While conducting a research on cross-lingual methods for a given task, it is natural to raise the following questions. Is English as a language with most resources always the best choice for a source language? Or does there exist a trade-off between the size of resources and relatedness of the languages in question? Is any language that is seemingly related according to the morphology-based typology also appropriate as a source language for cross-lingual techniques addressing a given task? And is it possible to combine multiple sources?

Availability of resources for many various languages are necessary to answer these questions. Nevertheless, conditions for a multilingual study on coreference are far from excellent. Compared to the situation in dependency parsing, which currently enjoys growing popularity as regards the cross-lingual approaches, the situation in coreference resolution is dramatically different. While the project of Universal Dependencies [Nivre et al., 2016] encompasses over 60 languages, Onto Notes 5.0 [Pradhan et al., 2013], the biggest multilingual coreference-annotated corpus with unified annotation, consists of data in only 3 languages. A similar disproportion between resources for parsing and coreference occurs also for parallel corpora. It is thus very challenging to develop cross-lingual methods on coreference resolution or to undertake cross-lingual studies on coreference, in general.

As a consequence, this thesis focuses only on two languages – Czech and English. These languages are one of the few that supply multiple coreference-annotated corpora, including the parallel ones.

Czech and English are actually a good choice of language also from the linguistic point of view. The way how they realize coreference relations on the surface almost could not differ more. Contrast the following example[1] of the English original sentence and its Czech translation from the PCEDT corpus [Hajič et al., 2011]:

(1.1) ∅ *přešla* *na* *bezkofeinovou* ***recepturu,*** ***kterou*** *používá* *pro* ***svojí*** *kolu.*
it switched to a caffeine-free formula [which] [it uses] [for] [self] Coke.
It switched to a caffeine-free **formula** [[$\varnothing_{ACT}$] using [**its** new Coke] [in 1985]].
V roce 1985 přešla na bezkofeinovou recepturu, kterou používá pro svojí novou kolu.

Let us look at coreferential means represented in this sentence pair. The first difference between English and Czech can be seen at the subject of the main clause. While expressed by the personal pronoun "*it*" in English, the subject in Czech is elided. Such correspondence is common for these two languages as Czech is a typical pro-drop language, which omits the subject if it can be easily reconstructed from the previous context using the information from subject-verb agreement. Second, we have a participle construction "*using its new Coke*" that is

[1]Many examples of the similar form can be encountered throughout the thesis. In the majority of cases, they are structured as follows. The first line represents the important excerpt of the Czech sentence as it appears in the corpus, with possibly inserted zeros. The second line is an English gloss of the Czech excerpt (the expressions in the square brackets do not appear in the original sentence). The third line is the original English sentence in its full length as it appears in the corpus. The fourth line is the Czech translation in its full length as it appears in the corpus. If necessary, an embedded square bracketing visualizing the dependency structure is introduced (except for the second line). Finally, the **anaphor** and the **antecedent** may be highlighted in the sentences.

translated to Czech as a relative clause with a relative pronoun "*který*" ("*which*"). The last pronoun correspondence in this sentence is the possessive pronoun "*its*", which, is translated here to Czech with the reflexive possessive pronoun "*svůj*", a category missing in English.

To have a better insight into coreference-related correspondences between Czech and English, we collect many of such examples from the parallel corpus. We accompany the examples with the statistics that quantify the frequencies of occurrences for individual pairs of expression types.

The example shows that it is advisable to count on ellipses (or zeros) that often appear in a language and participate in coreferential relations. It is absolutely vital to address them somehow in Czech, as Czech is a pro-drop language and zero subjects thus contribute to a substantial number of coreferential expressions. Existence of zeros in English becomes clear if it is contrasted with another language. The example shows that the zeros, which can be reconstructed to represent unexpressed arguments of a non-finite verbal form may have its clear counterpart in Czech relative pronouns. If we ignored these cases, the coreference projection, for instance, would not be able to discover coreference relations for many relative pronouns. In this thesis, we therefore work with a coreference represented on the so-called tectogrammatical layer, which is a deep-syntax dependency tree consisting almost exclusively of the content words and the reconstructed ellipses important for the meaning of the sentence.

In both cross-lingual methods that we deal with in this work, word alignment plays a central role. Without the alignment, it would be difficult to project coreference links or extract the important information from the other language. To ensure alignment also for zeros, we utilize a variant that identifies correspondences between nodes in the tectogrammatical trees in two languages.

1.1 Aims of the Thesis

The aims of this thesis are twofold:

- *Linguistic typology:* to design and test cross-lingual computational methods that will be able to quantify the similarities and differences of languages with respect to how they what means they use to express coreferential relations. In the end, the methods will serve as the tool to build a coreference-related linguistic typology.

- *Coreference resolution:* to explore the ways how to take advantage of differences of languages to build a better model for coreference resolution. We will particularly inspect the bilingually informed resolution as a means to obtaining better automatic coreference annotation on parallel corpora in comparison to using independent monolingual resolvers for each of the languages. Examples from such automatically resolved corpus might be in the future utilized in a semi-supervised learning.

1.2 Structure of the Thesis

The thesis is structured as follows. In Chapter 2 we introduce the important theoretical concepts including coreference, anaphora and Prague tectogrammatics. We also specify the expressions that are often involved in coreferential relations and highlight their interesting properties in both Czech and English. Chapter 3 presents the works related to this thesis, including the approaches to monolingual as well as cross-lingual coreference resolution In Chapter 4 we introduce all the datasets employed throughout the thesis. In addition, we describe the preprocessing pipeline required by our coreference resolver and the coreference resolution systems which our monolingual resolver is compared to. In Chapter 5, our own work begins with collecting the statistics on correspondences between Czech and English coreferential expressions. Chapter 6 devises a supervised method for aligning coreferential expressions trained on the data also described in this chapter. In Chapter 7, we propose our coreference resolver, which can be used in the monolingual as well as the bilingually informed setting, and test its quality in experiments. Chapter 8 contains our experiments with coreference projection. Finally, we summarize our main findings in Chapter 9.

2. Theoretical Fundamentals

In this chapter, we provide a theoretical background that is necessary for the rest of the work. In Section 2.1, we start by explaining phenomena related to the text linguistics, including coherence, cohesion, anaphora and coreference. We contrast anaphora and coreference and discuss why we opt for using mainly the term coreference in this work. Section 2.2 describes the Prague tectogrammatics, the theory underlying the way how we approach and represent linguistic information, reflected in the corpora we employ and the tools we use. Section 2.3 presents the way how the tectogrammatics represents coreferential relations and what types of coreferential relations it distinguishes. Lastly, in Section 2.4, we examine Czech and English expression types that are involved in coreference relations and discuss the most interesting linguistic aspects related to them, which might be important from monolingual and cross-lingual perspectives. We also delimit the expressions which belong to the core of our cross-lingual research.

2.1 Anaphora and Coreference

Text or *discourse* is a written or spoken passage that constitutes a unified whole.[1] A text thus cannot be formed of a mix of completely unrelated sentences. To ensure that the text is a unified whole, it must be coherent and cohesive inter alia. *Coherence* is a conceptual, functional and semantic unity of the text. *Cohesion* subsumes the grammatical and lexical means that link the components of the text and help to maintain its coherence. The following example[2] shows a coherent passage mutually tied with a lot of cohesive relations (not everything is highlighted).

(2.1) **President Trump**$_1$ told **British Prime Minister Theresa May**$_2$ that **she**$_2$ should "sue" **the European Union**$_3$ for a quicker Brexit, **May**$_2$ said **Sunday**$_4$. "**He**$_1$ told me **I**$_2$ should sue **the E.U.**$_3$ – not go into negotiations. Sue **them**$_3$. Actually, no, **we**$_5$'re going into negotiations with **them**$_3$," **May**$_2$ told the BBC in an interview that published **Sunday**$_4$. It is unclear how such a lawsuit would work for **Britain**$_5$, a member of **the European Union**$_3$, but **Trump**$_1$ has often threatened lawsuits in dealmaking.

The text in 2.1 features also several examples of anaphora. *Anaphora* is a cohesive relation that points back to an expression in the context. The pointing expression is called an *anaphor* whereas the expression which the anaphor links to is its *antecedent*.[3] The anaphora is for instance the relation between the anaphor "*them*" and the antecedent "*the European Union*" in the example text. The anaphor "*such a lawsuit*" signals another anaphoric relation, where "*such a lawsuit*" cannot be interpreted without the antecedent clause "*I should sue the E.U.*", and that in turn is not interpretable without the previous context.

[1] For more details on the theory of discourse, anaphora and text linguistics in general, see [Halliday and Hasan, 1976, De Beaugrande and Dressler, 1981, Zikánová et al., 2015].

[2] Taken from `https://www.washingtonpost.com/politics/trump-told-britain-to-sue-european-union-to-speed-brexit-prime-minister-says/2018/07/15/1b5178a0-8817-11e8-8b20-60521f27434e_story.html` on 15 July 2018.

[3] Analogously, *cataphora* denotes the relation that points forward to the following context. Its arguments are called *cataphor* and *postcedent*.

Let us now look at the notion of coreference. While a text is being perceived by a reader or hearer, he is gradually creating a *discourse model* in his mind. *Discourse entities* that often correspond to the real-world entities are the main building blocks of the model. The relation that links discourse entities to components of the text is called *reference*. Two expressions that refer to the same discourse entity are consequently in a relation of *coreference*. Such expressions are called *mentions* (of the same entity). Example 2.1 shows coreference between many co-indexed mentions, which refer to five discourse entities.

The given definitions of anaphora and coreference imply that they are not fully identical. Here are the main differences:

- *Cohesion vs. coherence.* While anaphora is a means of cohesion and operates on the textual level, coreference is determined by a discourse model, which is more related to the text coherence.

- *Interpretability.* Whereas anaphor cannot be interpreted without the antecedent, there is no such limitation on coreferential mentions. The mentions "*the European Union*" and "*the E.U.*" in Example 2.1 are both self-interpretable but still coreferential. On the other hand, none of the two occurrences of the pronoun "*them*" makes the other one interpretable, but they are still coreferential.

- *Direction.* Due to the interpretability requirement, anaphora is strictly directed. Conversely, no direction is specified for coreference.

- *Identity vs. any relation.* Coreference is a relation of identity. Two coreferential mentions refer to an identical discourse entity. In contrast, there can be basically any relation between the anaphor and its antecedent. Apart from identity-of-referent, it can be identity-of-sense anaphora (e.g. the anaphora connecting the expression "*such a lawsuit*"), verb anaphora, various types of associative anaphoric relations (or bridging) etc.[4]

- *Classes of equivalence.* Due to the identity relation, coreference is reflexive, symmetrical and transitive relation, i.e. the equivalence relation. Mentions belonging to the same entity thus form a class of equivalence, called the *coreferential chain*. It does not necessarily hold for expressions connected by anaphoric relations.

Nevertheless, these two phenomena often intersect. And mainly this intersection is in the focus of this thesis.

With respect to the aforementioned definitions of anaphora and coreference, the tasks of automatic *anaphora resolution* and *coreference resolution* (CR) attempt to automatically disclose the respective relations in the text.

Anaphora or Coreference? As this thesis concentrates mainly on the intersection of anaphora and coreference phenomena, we could denote the focused relations by any of the terms. We decided to use the terms coreference and coreference resolution for several reasons. Firstly, we want to emphasize that we are

[4]Please refer to any of [Mitkov, 2002, Zikánová et al., 2015, Poesio et al., 2016] for different types of relations accompanied with examples.

interested only in identity relation. Secondly, all the corpora that we employ are annotated with coreference, no matter how the specifications of annotated coreference relations differ across corpora (see Section 4.1). The corpora are in fact equipped with full-fledged coreference even for expressions that we so far do not target with our resolver. Thirdly, although the name of the evaluation measure that we propose – Prague anaphora score – suggests the opposite, the measure accepts if a mention is linked to any of the remaining mentions coreferential with it (see Section 4.4). The word *anaphora* in its name has been chosen only to emphasize that this measure does not evaluate whole cluster at once and can be decomposed over all mentions. Finally, despite the limitation of our resolver Treex CR to pronouns and zeros only (see Section 7.1), it is ready to be extended with additional models for other mention types.[5]

As it is usual in research on coreference, we borrow the terminology of anaphora. Having two coreferential mentions, one of them will be often denoted as anaphor and the other one as antecedent. Which one of the mentions is the anaphor will be always clear from the context. We will remain consistent with this terminology even in the rare cases when the anaphor would precede the antecedent in the text. We will call an expression anaphoric, if it belongs to a coreferential chain and happens to stand as an anaphor, i.e. there is a mention in the previous context referring to the same entity or it is clearly anaphoric in its original linguistic sense.

2.2 Prague Tectogrammatics

add an example of tectogrammatical tree, possibly also the English one

Most of the research in this work is carried out on the data that adhere to the principles of the *Prague tectogrammatics*. Also the coreference resolution system that we design in this thesis operates on the data that must at least to some extent comply with these principles.

Prague tectogrammatics is originally based on the theory of Functional Generative Description [Sgall, 1967, Sgall et al., 1986]. Since then, it has been implemented in multiple data resources. The real implementations, however, required the original theory to be slightly modified, limited or extended, e.g. for written Czech in the Prague Dependency Treebank series [Hajič et al., 2006, PDT], for spoken Czech in Prague DaTabase of Spoken Czech [Hajič et al., 2017], or for other languages including English in the Prague Czech-English Dependency Treebank series [Hajič et al., 2011, PCEDT], Arabic in Prague Arabic Dependency Treebank [Hajič et al., 2009], and recently for Russian and Polish in PAWS [Nedoluzhko et al., 2018]. Another bunch of adjustments was required if tectogrammatics-like annotation was provided by automatic machine approaches (e.g. in the Treex framework [Popel and Žabokrtský, 2010]) for particular applications (e.g. machine translation by the TectoMT system [Žabokrtský et al., 2008], or coreference resolution by Treex CR presented in Section 7.1). In the rest of this section, we thus describe tectogrammatics in its basic principles as implemented in PDT 2.0 [Hajič et al., 2006]. If any modification or extension to the annotation frame-

[5]In fact, models for German and Russian nominal groups have been already incorporated to Treex CR in coreference projection experiments [Novák et al., 2017].

work described here is needed at some place in the following sections, it will be presented there.

A fundamental principle of the theory of tectogrammatics is annotation stratification. That is, the annotation is split into multiple layers corresponding to the depth of linguistic description. In its basic version, it contains three layers:[6] *morphological*, *analytical*, and *tectogrammatical*.

Morphological layer (*m-layer*). It represents the surface form of the sentence as a sequence of tokens. In addition, every token is assigned morphological information, e.g. part-of-speech tag, lemma and grammatical categories.

Analytical layer (*a-layer*). The sentence is represented here as a surface syntax dependency tree, where each node (*a-node*) corresponds to one token on the morphological layer. All dependencies are labeled with a type of dependency relation.

Tectogrammatical layer (*t-layer*). Again, each sentence is realized as a dependency tree, though this time its nodes stand for the content words only. Tectogrammatical nodes (*t-nodes*) are linked to corresponding lexical and associated auxiliary tokens (e.g., prepositions, auxiliary verbs) in the analytical layer. Apart from that, selected types of ellipses are restored here. In that case, a new node is created, which either has no visible surface counterparts (e.g. unexpressed pronouns), or shares its surface counterpart with another node. It allows for an elegant treatment of zero subjects, which are very frequent in Czech as a pro-drop language, or zeros in non-finite clauses, which are, on the other hand, frequent in English.

Whereas every edge connecting two nodes is assigned a semantic role (e.g. actor (ACT), patient (PAT), addressee (ADDR), benefactor (BEN)),[7] to every tectogrammatical node various attributes are attached. These include:

- *tectogrammatical lemma* (*t-lemma*): a generalized variant of the surface lemma. For instance, all verbal forms are represented by infinitive. Importantly for this work, all personal, possessive and reflexive pronouns also share the same tectogrammatical lemma. The features discriminating individual pronouns are then stored as grammatemes (see below).

- *grammatemes*: attribute-value pairs representing semantic part-of-speech[8] and semantically indispensable morphological features, e.g. gender and number for nouns, tense for verbs, degree for adjectives, sentence modality, etc.

- *valency frame*: it is defined by a link to the valency dictionary. The valency frame of a given verb occurrence disambiguates a sense of the verb, and specifies how the verb binds with its arguments and modifiers.

[6] In fact, there is one more non-annotation layer called *word* layer (*w-layer*), which represents the raw text.

[7] We use these abbreviations of semantic roles in several examples in the following text. Please refer to the manual of tectogrammatical annotation [Mikulová et al., 2007] for details.

[8] Its four categories – semantic nouns, semantic adjectives, semantic adverbs and semantic verbs correspond to the basic onomasiological categories – substances, properties, circumstances and events [Mikulová et al., 2007].

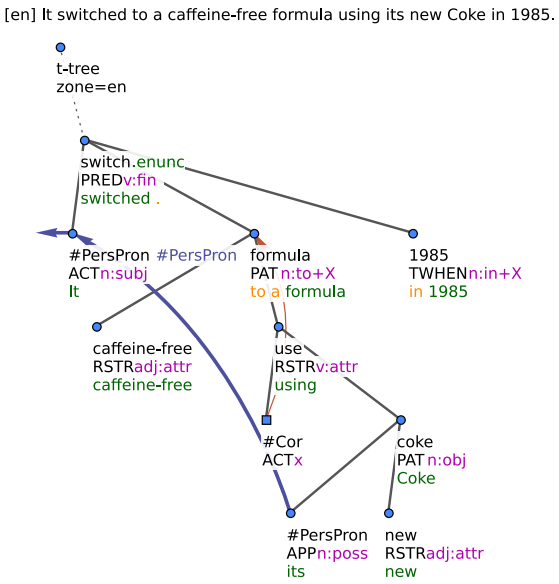

Figure 2.1: The tectogrammatical representation of the English sentence from Example 1.1.

- *topic-focus articulation*: it is captured by two phenomena. First, the property of *contextual boundness* indicates whether the expression associated with the node is used by the speaker as given (for recipients). Second, *communicative dynamism* reflects a relative degree of importance of the expression in comparison with other expressions in the sentence, as attributed by the speaker.

- *coreference and anaphora*: various types of coreferential and anaphoric relations. A relation is specified in the anaphor's head node by a reference to the antecedent's head node and additional information. See Section 2.3 for more details.

The tectogrammatical representation of the English sentence from Example 1.1 is depicted in Figure 2.1.

2.3 Coreference in Tectogrammatics

The tectogrammatical layer is also a place where coreference and other anaphoric relations can be annotated. It allows for capturing coreference relations also for zeros, which is essential, especially for Czech. In the following, we describe the coreference annotation which is exploited by the present work and which corresponds to the annotation of the Czech-English corpus PCEDT 2.0 Coref [Nedoluzhko et al., 2016a].[9]

[9]The corpus also contains annotation of some types which are not in the scope of this work, e.g. split antecedents, exophora and discourse deixis. In addition to all these types, PDT 3.0 [Bejček et al., 2013], which is among the corpora based on tectogrammatics the one with the richest annotation, includes the annotation of coreference of nominal groups with

In all the implementations of the Prague tectogrammatics, coreference is technically annotated as an oriented link connecting two mentions. The mention is explicitly defined only by its head, but its span is implicitly specified as a full subtree of the mention's head. A coreference link thus always connects two nodes on a tectogrammatical layer – the heads of the mentions. The orientation of the link determines the anaphor and the antecedent.

The representation of coreference in Prague tectogrammatics remarkably differs from the majority of other annotation styles, especially those originally tailored to English (e.g. the OntoNotes [Pradhan et al., 2013] style). They usually represent mentions as continuous spans of a surface text with specified boundaries and coreference by co-indexing the mentions that belong to the same chain.

Tectogrammatics distinguishes two types of coreference relations: *grammatical* and *textual coreference*.

Grammatical coreference includes the following subtypes of relations, which appear as a consequence of language-dependent grammatical rules:

- *Reflexive pronoun coreference.* In this case, the anaphoric pronoun mostly refers to the closest subject, as in the following example where the reflexive pronoun "*herself*" corefers with the subject "*daughter*".

  (2.2) **My daughter** likes to dress **herself** without my help.

- *Coreference with relative elements.* Relative pronouns and pronominal adverbs introducing relative clauses are linked to their antecedent in the governing clause. See the next example where the relative pronoun "*who*" corefers with the noun "*boy*" modified by the dependent relative clause.

  (2.3) Alex is the **boy who** kissed Mary.

- *Coreference of zeros in non-finite clauses.*[10] It may concern unexpressed arguments of verbal modifications with a so-called *dual dependency* (e.g., present and past participles, infinitives). This is, for example, the case of the unexpressed subject "$\varnothing_{ACT}$" of the present participle "*laughing*", which is coreferential with the subject "*John*" of the governing clause in Example 2.4. Furthermore, a *control relation* also belongs to this category. It arises with certain verbs, called control verbs, e.g. "*begin*", "*let*" and "*want*". In Example 2.5, it links the unexpressed subject governed by the infinitive "*sleep*" and the subject "*Peter*".

  (2.4) **John** cannot stop [[$\varnothing_{subj}$] laughing].

  (2.5) **Peter** wants to [[$\varnothing_{subj}$] sleep].

generic reference and various types of bridging relations, e.g. set-subset, part-whole, contrast [Nedoluzhko, 2011, Zikánová et al., 2015].

[10] To be correct, we should rather use the term *non-finite verbal construction*, because it does not have to be considered a clause in some cases (e.g. in Example 2.5). However, we will use the term *clause* for the sake of simplicity.

- *Coreference in reciprocal constructions.*[11] It appears for instance in the following sentence as an unexpressed object.

  (2.6) [[**John and Mary**] kissed [$\varnothing_{obj}$]].

Textual coreference. Its arguments are not realized by grammatical means alone, but also via context. Several types of expressions may serve as an anaphor in such a coreference relation, e.g. personal, possessive and demonstrative pronouns, anaphoric zeros, and nominal groups. All pairs of the highlighted coindexed mentions in Example 2.1 are examples of textual coreference.

2.4 Types of Expressions

In order to ensure coherence, readability and naturalness of the text, coreference must be manifested in many different forms. The variety of possible means of expressing coreference even grows if we take into account multiple languages. In the following sections we introduce the types of coreferential expressions in Czech and English. We highlight the most interesting linguistic aspects important from both monolingual and cross-lingual perspectives.

Moreover, we specify the criteria that we use to partition the space of coreferential expressions into individual types. The partitioning over mentions must be defined for several reasons concerning the architecture of our CR system and the way how we perform the evaluation. The criteria are mostly based on lexical and morphological aspects, such as surface-level and tectogrammatical lemmas and grammatical categories. Since one of the tasks of this thesis is to address automatic coreference resolution, we cannot rely on coreference information while selecting candidates on mentions. Coupled with the frequent lexical ambiguity of mainly pronouns, this inevitably leads into including expressions that are not coreferential (and some not even potentially). It gets even worse, if the mention candidates are to be selected using the information from the automatically acquired tectogrammatical representation. The criteria thus need to be simple to work properly on automatic annotations, they need to embrace as many true mentions as possible and, at the same time, avoid introducing too many spurious mentions.

All the categories that this thesis distinguishes are listed in Table 2.1. The table is horizontally split into the three groups of expression types that form the core of our cross-lingual research and the last group aggregating the remaining expression types. All these groups are elaborated in the following four sections: central pronouns in Section 2.4.1, relative pronouns in Section 2.4.2, zeros in 2.4.3, and other types in Section 2.4.4. The columns of the table indicate which types are deemed to be anaphors by our CR system and in the datasets we employ. This topic is further commented in Section 2.4.5.

[11] Reciprocal constructions are the only expressions involved in grammatical coreference which are not in the scope of this thesis.

2.4.1 Central Pronouns

Central pronouns is a term coined by Quirk et al. [1985] encompassing English *personal* (e.g., "*he*", "*she*", "*him*", "*her*"), *possessive* (e.g., "*his*", "*her*", "*mine*"), and *reflexive pronouns* (e.g., "*myself*", "*themselves*"). By using this term for Czech pronouns we mean the category consisting of *personal* (e.g., "*on*", "*jemu*", "*ní*"), *possessive* (e.g., "*jeho*", "*jejich*"), *(basic) reflexive* (e.g. "*se*", "*sebe*"), and the *reflexive possessive pronouns* ("*svůj*").

Role of gender. In both languages, gender and number are defined for the majority of central pronouns and they often agree with the gender and number of the pronoun's antecedent if there is any. However, the nature of gender is not identical in Czech and English. Whereas in English gender is notional, in Czech it is rather grammatical.

Notional gender in English is associated with entities, not with the linguistic expressions. As a consequence, most of the entities are referred to with a pronoun whose gender is neuter (e.g. "*it*", "*its*"). The feminine and masculine gender are mostly exclusively associated with persons. English gender is thus strongly connected to the aspect of animacy.

In contrast, genders in Czech are associated with nominal expressions, not the entities which the expressions denote. It is no exception if the same entity is denoted by nouns of different genders. For example, "*děvče*" is in the neuter and "*dívka*" in the feminine gender, yet both may refer to the same entity representing a girl. Assignment of genders to nouns appears to be more or less arbitrary and genders are thus distributed relatively evenly over nouns.

Functions of "*it*". The English pronoun "*it*" serves multiple functions. One of several possible ways to categorize its uses is as follows:

- *referring to an entity*: anaphora to a nominal group in the preceding context or a deictic reference to an entity,

  (2.7) John bought **a new car**. **It** cost him a fortune.

- *referring to an event*: anaphora to a clause or a larger discourse segment or a deictic reference to an event,

  (2.8) John **isn't** coming. **It**'s sad!

- *non-referential (pleonastic)*: its presence is imposed only by the syntactic rules of English.

  (2.9) **It** is possible that John isn't coming.

Only the first function is shared with the rest of the English personal pronouns. And it is also the first function (referring to an entity) that this work focuses on.

Rich features of Czech possessive pronouns. Information on two genders and numbers can be encoded in the form of a Czech possessive pronoun: those of the possessed object and the possessor's ones. For instance, in Example 2.10 the pronoun "*jejího* /her/" carries the information on masculine gender of the possessed objected in its suffix "*-ho*" and the feminine gender of the possessor in its stem "*její-*". Nevertheless, the same forms are in fact shared among different combinations of genders and numbers, thus making many of them underspecified (see below).

(2.10) *před rokem* ∅ *nechal* *zastřelit* *svou* ***manželku***$_{\mathbf{f.s}}$ *a* ***jejího***$_{\mathbf{m.s\text{-}f.s}}$ *milence*$_{\text{m.s}}$
a year ago he let shot his wife and her lover
He had his wife and her lover shot a year ago.

Multifarious reflexive pronouns. The main distinction between Czech and English reflexive pronouns is that Czech employs a special type of reflexive possessive pronouns that does not exist in English. In fact, it consists of a single expression – "*svůj*".

Czech basic reflexive pronouns appear in two equivalent forms: long (e.g. "*sebe*", "*sobě*") and short ("*se*", "*si*"). While the short form is used by default, the long form is used after a preposition, and it is preferred in case of emphasis or showing contrast. The short forms, however, frequently appear in Czech in other functions than a reflexive pronoun. Svoboda [2014] distinguishes three broad functional domains of reflexive markers "*se*"/"*si*":

- *pronoun*: either reflexive or reciprocal;

  (2.11) ***Jan*** ***si*** *koupil* *nové* *auto.*
  John to himself bought new car.
  John bought himself a new car.

- *derivational particle*: it merges with a non-reflexive verb to form a new lexical entry. This function also includes the cases of *reflexivum tantum*, i.e. the reflexive verb which does not exist without the reflexive marker (e.g. "*smát se* /to laugh/");

  (2.12) *Jan* ***se*** *zajímá* *o* *koupi* *nového* *auta.*
  John REFL be_interested in buying new car.
  John is interested in buying a new car.

- *grammatical marker*: used in *reflexive passives*, i.e. syntactical constructions which resemble the passive.

  (2.13) *Takové* *auta* *se* *už* *neprodávájí.*
  Such cars – anymore don't sell.
  They$_{\text{generic}}$ don't sell such cars anymore.

Only the first pronominal use of a reflexive marker is interesting for this work.

As for English reflexive pronouns, they have two distinct uses according to Quirk et al. [1985, p. 356]:

- *basic*: the reflexive pronoun serves in the clause as an object or a complement and its antecedent is the subject (Example 2.2 and the translation of Example 2.11);

- *emphatic*: it is in apposition with its antecedent. The function of the emphatic reflexive is to put special stress on its antecedent.

  (2.14) It is **John himself** who bought a new car.

Both uses are in the scope of the present work.

Underspecification of Czech pronouns. Unlike English reflexive pronouns, all Czech reflexive pronouns are morphologically underspecified. They share the

same form whatever gender, number and person they represent.[12] The gender and number of agreeing with the possessed object in possessive pronouns is often underspecified. But the more problematic for coreference resolution is, when the underspecification occurs for the possessor's gender. It happens e.g. for the word "*jeho* /him, it, his, its/", which may anaphorically point to a nominal expression of either masculine or neuter gender. Moreover, the same form represents the personal pronoun of the masculine/neuter gender in the genitive/accusative case.

Selection criteria. T-nodes corresponding to central pronouns are easy to identify by their t-lemma `#PersPron`. In addition, such nodes need to have its own surface representation. To mitigate inclusion of spurious mentions on the automatically acquired t-layer, we reinforce the selection criteria by checking the surface lemma and part-of-speech subcategory. Due to the reasons which will be explained in Section 2.4.5, we also distinguish if the central pronouns are either of (1) the 3rd or unspecified person, or (2) the 2nd or 3rd person. Whereas within the former category we will differentiate between the subtypes and will not further specify the constraint on their person, the latter category of central pronoun will be treated as a whole and the reader will be always reminded about its person.

If present, gold t-layer annotation helps filter out expressions homonymous with anaphoric central pronouns. For instance, the pleonastic use of "*it*" and non-pronominal uses of Czech reflexive pronouns should not be represented by their own node on the t-layer. Czech reciprocal pronouns, on the other hand, are represented on the t-layer, but the t-nodes are labeled by a specific t-lemma `#Rcp`. As we do not try to resolve such homonymy on the system t-layer, these types of mentions will be included in the selection.

2.4.2 Relative Pronouns

Relative pronouns are used to introduce relative clauses. English relative pronouns are e.g. "*which*", "*who*", "*whose*" and "*what*". Another expression that appears as a relative pronoun, and which we put in its own category at some places in the thesis, is "*that*". Czech relative pronouns are e.g. "*který*", "*jenž*", "*kdo*", "*co*", and "*jaký*". A special position is held by the pronoun "*což*".

Gender vs. animacy. Some of the Czech relative pronouns (e.g. "*který*", "*jenž*" and "*jaký*") carry the gender and number information. Similarly to central pronouns, the pronoun usually agrees in gender and number with its antecedent. English makes the distinction with respect to animacy, where "*who*" refers to persons and "*which*" to the rest.

Relative, interrogative, or fused? In both languages, many of the forms of relative pronouns are shared with interrogative pronouns (e.g. wh- words in English as shown in Example 2.15, "*který*", "*kdo*" etc. in Czech). English also uses some of these forms in fused relative constructions (Example 2.16). Such a construction consists of a relative clause, which is not attached to an external antecedent, but fused with it.

[12] We mean the possessor's gender and number in the case of the reflexive possessive pronoun "*svůj*".

(2.15) He was then told **who** had been shot.

(2.16) There are different versions of **what** actually happened.

Overloaded "*that*" The English word "*that*" is heavily overloaded with functions. Examples 2.17 through 2.20 show its four main functions, respectively: demonstrative pronoun, relative pronoun, subordinating conjunction, and adverb.

(2.17) I've got **that** pain in my back again.

(2.18) There are lots of **things that** I need to buy before the trip.

(2.19) The fact **that** he is your brother-in-law should not affect your decision.

(2.20) It was quite a large fish – about **that** long.

Referring to an entity or an event? The Czech pronoun "*což*" keeps a special position because it can refer to both entities and events (see Examples 5.16 and 5.17 in Section 5.3). It is analogous to the two referential functions of English pronoun "*it*" and demonstrative pronouns in both languages. Unlike these expressions, "*což*" cannot be deictic. In English, a similar role is played by the pronoun "*which*".

Selection criteria. In the datasets we use (namely, PDT and PCEDT, see Section 4.1), coreference is annotated also for adverbs that act like relative pronouns (e.g., "*how*", "*where*" in English, "*kde* /where/", "*kdy* /when/" in Czech). Our criteria for relative pronouns are thus defined more broadly to cover also these adverbs.

There is no straightforward way to distinguish relative pronouns on the t-layer. Furthermore, in both languages some relative pronouns (e.g. in correlative pairs, see Example 5.24 in Section 5.4) are not represented by its own node on the t-layer. Only surface-level constraints, such as surface lemma and part-of-speech subcategory, are therefore imposed to select relative pronouns. Nevertheless, such criteria do not suffice to filter out conjunctions, interrogative and fused pronouns. We thus implemented one more criterion exploiting the syntax to improve the selection precision. It filters out such a node representing the word "*that*", which has some children in a surface dependency tree.[13] Despite the additional criterion, precision of covering truly coreferential English relative pronouns by these criteria is only 65% on the PCEDT data and it thus must be accounted for in the design of coreference resolver.

2.4.3 Zeros

All *zeros (or ellipses)* share the same feature which declares them to be zeros – they do not have its own surface word representation. According to the tectogrammatical manual [Mikulová et al., 2007, p. 413], several different types of ellipses have been annotated. Coreference is concerned mostly with the ellipsis of dependent elements. This includes the following coreferential types, which are denoted by a specific t-lemma based on the coreferential relation they enter:

[13] In an a-tree, i.e. in the surface tree respecting the principles of Prague tectogrammatics, relative pronouns cannot have children.

`#PersPron` (textual), `#Cor` (control, dual dependency), `#QCor` (quasi-control), `#Rcp` (reciprocity), and `#Benef` (control for non-obligatory arguments). It also comprises the types not entering the coreference relation as anaphors: `#Gen` (general argument), `#Unsp` (unspecified agent), and `#Oblfm` (ellipsis of an obligatory free modification).[14]

In this work, we propose another categorization of zeros. It is motivated by automatic coreference resolution, and thus maybe not so linguistically correct, but simple. *Zero subjects* are unexpressed subjects of finite verbs and are typical for Czech. No such zeros exists in English. *Zeros in non-finite clauses* are formed by unexpressed arguments in non-finite clauses. They usually correspond to the zeros entering the control and dual dependency grammatical coreference as an anaphor. The category of other zeros is then constituted by remaining zeros that figure in a coreferential relation.

Is the anaphoric zero present? The main problem related to zeros is whether they should be reconstructed on the t-layer or not. This should be driven by the underlying principles of tectogrammatics, specifically by a valency dictionary. But compiling such a dictionary is not a quick process and it may not be exhaustive for all possible cases. Consequently, even the human annotators are sometimes inconsistent in a question whether the zero should be present, let alone the automatic method for zero reconstruction. This may happen especially for English present and past participles. For example, two almost identical phrases from PCEDT 2.0 have ended up with different manual annotations of the word "*closing*" – once annotated as a present participle with two zero arguments (2.21), and once as a deverbative adjective with no arguments (2.22).

(2.21) an expected premium of 21/2% to [[the **closing**$_{verb}$ [$\varnothing_{PAT}$] [$\varnothing_{ACT}$]] [share] price [when terms are fixed Tuesday]] (wsj_0271-s44)

(2.22) a 5% premium over [[the **closing**$_{adj}$] [share] price [Tuesday, when terms are scheduled to be fixed]] (wsj_0125-s44)

The variety of original types of ellipses is another problem that concerns especially the automatic processing of a text. Some of them are anaphoric, some not. And some of the types are rare and at the same time difficult to unveil, so it does not pay off to attempt to address them. Therefore, we come up with a simplified categorization of zeros presented above.

Manifesting grammatical features. Grammatical features of a Czech zero subject are usually manifested in a form of the governing verb. The features may still remain underspecified, though. For instance, it is impossible to determine its gender without the context if the verb is in the present tense.

Selection criteria. The category of zero subjects consists of such `#PersPron` t-nodes that they: (1) have no surface word representation, (2) are governed by an expressed verb, (3) satisfy the semantic role of an actor if the verb is in active voice and a patient if the verb is passive, and (4) are of 3rd or unspecified person. The last condition is related to how our coreference resolver works and will be further explained in Section 2.4.5. The category of zeros in non-finite clauses

[14] Please refer the tectogrammatical manual for further explanation and examples.

consists of all `#Cor` t-nodes. The remainder of zeros is outside the scope of this thesis.

Precision and recall of selecting anaphoric zeros strongly depends on whether they are restored. This is a topic especially for an automatic pre-processing of a text to the tectogrammatical representation, which is further discussed in Section 4.2.1.

2.4.4 Other Expressions

The remaining group of potentially coreferential expressions that do not belong to the core of our cross-lingual research accounts for a relatively large number of mentions. It comprises yet another three big subgroups of expressions: demonstrative pronouns, nominal groups, and named entities.[15] Although we do not address them primarily in this work, we count on them at the places where it makes sense (e.g. in evaluation of some CR systems and projection of gold coreference). *Demonstrative pronouns* are selected using a lemma ("*ten*" in Czech, "*this*", "*that*" etc. in English) and the semantic part-of-speech saying that the pronoun acts like a noun. *Nominal groups* are all phrases governed by a semantic noun excluding all the pronouns and zeros covered by the categories described above and *named entities*, which constitute its own category.

The rest of the coreferential expressions includes adjective, verbs, coordination roots etc. In most of the tables throughout the thesis, we merge these expressions with the category of other zeros (see above) and represent them in a single Other category.

Two specific types stand out within the expressions which are never coreferential but may correspond to a coreferential expression in the other language. The first category of *definite articles* appears only in English and is represented by a single word "*the*". On the other hand, the other category *sám / samotný* is defined only in Czech. It comprises only the pronoun "*sám* /alone, personally/" and the adjective "*samotný* /alone, very/", which both often occur as counterparts of English reflexive pronouns in their emphatic use.

2.4.5 Delimiting the Mentions

Table 2.1 lists all types of expressions introduced so far. Furthermore, it indicates if particular expressions are defined in the languages, if they are annotated with coreferential relations in manually annotated datasets that we employ in the thesis, and finally, if the expressions belong to the core of our cross-lingual research. For the core expressions, we have both manually and automatically collected the statistics of their cross-lingual correspondences (Chapter 5). We have also designed a supervised method that improves their cross-lingual word alignment (Chapter 6). Last but definitely no least, it is exactly these expressions that are represented as anaphors in the scope of the coreference resolver and the

[15] Although named entities in fact belong to nominal groups, we will treat named entities and nominal groups without named enities separately in the evaluation. For the sake of simplicity, in the rest of the thesis we will be using the term "*nominal groups*" and mean nominal groups without named entities, if the term *named entities* appears in the same context.

experiments on cross-lingual CR (Chapter 7).[16]

In its Theory columns, Table 2.1 indicates if an expression type is defined in a given language. Apparently, Czech does not know definite articles and the category of the relative pronoun "*that*". English, on the other hand, does not define reflexive possessive pronouns, zero subjects, the relative pronoun "*což*" and the words "*sám*" and "*samotný*".

The column Tecto shows for which mention types coreference is manually annotated in the newest versions of the tectogrammatical corpora used in this thesis. No surprise that the set of coreferentially annotated mentions is a subset of expressions existing in a language. In addition, definite articles and the "*sám/samotný*" category are excluded from the respective sets for Czech and English. The CoNLL column in the English part of the table indicates which types are coreferentially annotated in the CoNLL dataset (see Section 4.1.4). We employ this dataset that adheres to the annotation specifications for OntoNotes [Pradhan et al., 2013] for evaluation. Out of the three groups constituting the core of our research, this dataset supports only the central pronouns.

Finally, the Core column indicates which anaphor types belong to the core of the cross-lingual research in this thesis. The core comprises almost entirely the three main groups of expressions. Nevertheless, it excludes the category of other zeros and the expressions in the 1st or 2nd person and does not include a large set of nominal groups, named entities and demonstrative pronouns.

Let us briefly explain why this work mostly limit to these categories of expressions. The category of the other zeros is excluded as we wanted to restrict zeros only to those which are really important from both monolingual and cross-lingual perspective. Even without the other zeros, we have to deal with inconsistencies in annotation and problems of their automatic reconstruction. Demonstrative pronouns have been excluded because of their frequent extra-linguistic reference. In addition, they appeared to be rather rare in journalistic texts of our corpora.

The reason for excluding a huge set of nominal groups and named entities is that we do not consider them interesting from a cross-lingual perspective. We expect that the prevailing majority of such expressions does not change the type during their translation and the grammatical aspects of language are not as important for noun coreference as it is for pronouns and zeros.

Finally, the expressions in the 1st and 2nd person as well as the nominal groups have not been included also from a reason concerning the process of annotating coreference in PCEDT, the key corpus in this work (see Section 4.1.2). It proceeded in two stages. The core of our cross-lingual study focuses roughly on the expressions that were annotated in the first stage published in PCEDT 2.0 [Hajič et al., 2012]. The core covers more than 95% of anaphors in this first-stage corpus. Although the annotation works continued and concluded with the release of PCEDT 2.0 Coref [Nedoluzhko et al., 2016b], the labor-intensive corpus study on correspondences between coreferential expressions [Novák and Nedoluzhko, 2015], which constitutes the basis for our cross-lingual approaches, had been already conducted on the core of expressions and was not further extended.

[16]It is important to emphasize the phrase "as anaphor" because such restrictions are not imposed on antecedent candidates, which can belong to practically any of the potentially coreferential categories.

Expression type	Czech			English			
	Theory	Tecto	Core	Theory	Tecto	CoNLL	Core
Central pron. (3rd/unsp. pers.)							
Personal pron.	•	•	•	•	•	•	•
Possessive pron.	•	•	•	•	•	•	•
Reflexive pron.							
Reflexive possessive pron.	•	•	•				
(Basic) reflexive pron.	•	•	•	•	•	•	•
Central pron. (1st/2nd pers.)	•	•		•	•	•	
Zero							
Zero subj. (3rd/unsp. pers.)	•	•	•				
Zero in non-finite clause	•	•	•	•	•		•
Other	•	•		•	•		
Relative pron.							
That				•	•		•
Což	•	•	•				
Other	•	•	•	•	•		•
Demonstrative pron.	•	•		•	•	•	
Named entities	•	•		•	•	•	
Nominal group	•	•		•	•	•	
Definite article				•			
Sám / Samotný	•						
Other	•	•		•	•	•	

Table 2.1: Types of expressions distinguished in this work. The table indicates for each type, if it theoretically exists in the language (Theory), if it is annotated for coreference in the PDT/PCEDT/CoNLL dataset (see Section 4.1), and if it is addressed by our Treex CR system (see Section 7.1).

3. Related Work

In this chapter, we overview the literature related to our work. Section 3.1 surveys monolingual approaches to coreference resolution with an emphasis to the design questions that have been raised during development of our coreference resolver Treex CR. The second part in Section 3.2 then discusses the three main cross-lingual approaches to coreference resolution: coreference projection, delexicalized CR, and multilingually informed and joint multilingual CR.

3.1 Monolingual Coreference Resolution

Anaphora and coreference resolution has been one of the important tasks in computational linguistics and natural language processing for many years. However, advent of data-driven methods has been observed only for last two decades, with prevalence of supervised learning methods.

In Section 3.1.1, we will give a brief overview of milestones in supervised methods for monolingual CR.[1] The following sections elaborate further upon aspects of coreference revolvers' design that are particularly interesting for the present work: mention-pair and mention-ranking models (Section 3.1.2), treatment of non-anaphoric mentions (Section 3.1.3), and addressing different types of coreference with separate models (Section 3.1.4). Finally, we will provide a historical summary of coreference resolution for Czech in Section 3.1.5.

3.1.1 A Historical Overview of Supervised Coreference Resolution

The evolution of supervised[2] methods for CR was framed by three major evaluation campaigns. The rise of the interest in data-driven methods started with extending the data of the Message Understanding Campaign [MUC-6, 1995] with coreference annotation and proposing the first coreference measure, the MUC score [Vilain et al., 1995]. Despite a relatively small size of the data, it opened space for first systematic evaluation as well as machine-learning approaches to CR, e.g. mention-pair models [Soon et al., 2001, Ng and Cardie, 2002a]. The second evaluation campaign [NIST, 2003] introduced much larger ACE datasets released in years 2002–2005, in which coreference was restricted to only pre-defined entity types (e.g. Person, Organization, Location). New datasets encouraged researchers to address major weaknesses of the mention-pair model by improving the mention-based approach, e.g. mention-ranking models [Denis and Baldridge, 2007b], as well as by introducing entity-based approaches, e.g. entity-mention [Luo et al., 2004, Yang et al., 2004, Culotta et al., 2007] and cluster-ranking mod-

[1] Overviews with more details or given from different perspectives can be found in overview papers [Ng, 2010, 2017, inter alia], PhD theses [Kobdani, 2012, Martschat, 2017, Tuggener, 2016, inter alia], and books [Poesio et al., 2016, inter alia].

[2] Although it is not the topic of this work, we also need to mention unsupervised approaches, which saw its peak of popularity around 2010. All such works designed a generative model for resolution of general coreference [Ng, 2008, Haghighi and Klein, 2010] or just pronouns [Charniak and Elsner, 2009].

els [Rahman and Ng, 2009]. Simultaneously, shortcomings of the MUC score drove the development of new evaluation measures, e.g. B^3 [Bagga and Baldwin, 1998], CEAF [Luo, 2005]. Nevertheless, the unclear situation of two not very convenient datasets (MUC was small, ACE were restricted) with undefined train/test partitions and multiple independent evaluation measures with their variants provoked building a new robust coreference-annotated corpus, OntoNotes [Hovy et al., 2006, Pradhan et al., 2013] and the first official evaluation script setting the average of three metrics as the official CoNLL score [Pradhan et al., 2014]. The new dataset and unified evaluation approach was then utilized in a third major evaluation campaign, the CoNLL 2011 and 2012 Shared Tasks [Pradhan et al., 2011, 2012]. The winner of the CoNLL 2011 Shared Task, a carefully designed rule-based system by Lee et al. [2011], encouraged the community to reconsider their approaches to CR. At the CoNLL 2012 Shared Task, the winning system [Fernandes et al., 2012] triggered with its antecedent trees a new approach to CR – structured prediction. From that time on, structured prediction has been employed to both mention-based, e.g. [Durrett and Klein, 2013, Lassalle and Denis, 2015, Wiseman et al., 2015], and entity-based models, e.g. [Björkelund and Kuhn, 2014], [Clark and Manning, 2015]. Even though coreference resolution has not been left untouched by a renewed interest in neural networks, e.g. [Wiseman et al., 2015, Clark and Manning, 2016, Lee et al., 2017, 2018], the change has not been until now as dramatic as in the task of machine translation for instance.

3.1.2 Mention-pair and Mention-ranking Models

Mention-pair model was the first supervised learning approach to CR. Although it had been applied to specific domains by Aone and Bennett [1995] and McCarthy [1996], it gained in popularity primarily thanks to the works of Soon et al. [2001] and Ng and Cardie [2002a]. It is designed to make binary classification of two candidate mentions, an anaphor and antecedent candidate. That is, for each pair of mentions it determines whether the two mentions are or are not coreferential. The estimate is based on a set of features describing both mentions in isolation as well as the relation between them and their attributes. In order to obtain coreference chains that obey the transitivity property, the independent pairwise estimates must be consolidated in a following clustering stage. Two clustering algorithms are generally used for this purpose: *closest-first clustering* [Soon et al., 2001] and *best-first clustering* [Ng and Cardie, 2002a]. Whereas the former algorithm selects the antecedent candidate which is closest to the anaphor, the latter clustering selects the one that has been labeled as the most probable.

The mention-pair model later emerged in various modifications that concerned construction of training examples (to hinder skewed distributions of positive and negative examples), classification methods, and clustering algorithm (see [Ng, 2010] for an overview). There appeared also attempts to avoid independent nature of the classification and the clustering step, for instance by guaranteeing transitivity in pairwise decisions using integer linear programming [Finkel and Manning, 2008].

Mention-ranking model as proposed by Denis and Baldridge [2007b] alleviates the weakness of two isolated steps in mention-pair models in a more elegant way, though. Instead of running a binary classifier for every pair of mentions

independently, the mention-ranking model looks at antecedent candidates for a given anaphor simultaneously and assigns a ranking score to each of them. As the model directly represents competition between antecedent candidates, the candidate that is ranked the highest can be marked as the antecedent. Unlike in the case of the best-first clustering algorithm for mention-pair models, the competition between candidates is captured during training.

Superiority of the mention-ranking in comparison with the mention-pair model has been shown several times [Denis and Baldridge, 2007b, Nguy et al., 2009, i.a.]. Surprisingly, Martschat and Strube [2015] has shown that it also outperforms some of the approaches based on antecedent trees and structured prediction [Fernandes et al., 2014, Björkelund and Kuhn, 2014].

Both mention-pair and mention-ranking are examples of mention-based models, which, however, suffer from another shortcoming – the lack of expressiveness. The lack of expressiveness occurs when the anaphor or the antecedent candidate is underspecified, i.e. it does not carry enough evidence to discard the pair from potentially coreferential. For example, imagine a document that consists of three mentions: *"Mr. Clinton"*, *"Clinton"*, and *"she"*. A mention-based model would probably link *"Mr. Clinton"* with *"Clinton"* based on a string-matching feature. A possible short distance and a lack of gender information in *"Clinton"* may cause this mention to be linked with *"she"*. Due to transitivity property, *"she"* and *"Mr. Clinton"* consequently end up in the same coreferential chain, even though they strongly violate gender agreement of two coreferential mentions. As reported by Tuggener and Klenner [2014], the underspecification issue occurs also for German pronouns, e.g. personal pronoun *sie (she/they)* and possessive pronoun *sein (his/its)* has ambiguous number and gender, respectively. Furthermore, the same morphological underspecification happens also for Czech pronouns and zeros (see Sections 2.4.1 and 2.4.3, respectively).[3]

Despite its limitations, approaches taking advantage of mention-based models remain dominant, as they are relatively fast, scalable, easy to implement and simple to train.

3.1.3 Treatment of Non-anaphoric Mentions

Mention-ranking model could seem to have another potential shortcoming – it forces every anaphor to be linked with an antecedent. Therefore, apart from antecedent selection, a procedure for *anaphoricity detection (determination)*, which solely filters potentially anaphoric mentions, must be run. Two strategies of how to combine anaphoricity detection and antecedent selection are distinguished: (1) pipeline, and (2) joint.

In the *pipeline* strategy, a system for anaphoricity detection is run prior to antecedent selection. Although the mention-pair model is able to treat non-anaphoric mentions implicitly, this strategy has been first introduced for this kind of model by Ng and Cardie [2002b] in order to reduce the number of pairwise

[3] *Entity-based models* [Luo et al., 2004, Yang et al., 2004, Raghunathan et al., 2010, Björkelund and Kuhn, 2014, Clark and Manning, 2015] aim to rectify this issue. Instead of linking two mentions, they try to assign a mention to an already collected cluster of mentions or merge intermediate clusters and step by step arrive to the final clustering. Cluster features, e.g. prevailing number or gender in a cluster, are used to mitigate underspecification.

comparisons. Later, it was reused also for a mention-ranking model [Denis and Baldridge, 2008], which on the other hand requires special treatment of non-anaphoric mentions. An obvious drawback of this strategy is that errors incurred by the anaphoricity detector tend to propagate to the next step, which in turn requires careful tuning of the anaphoricity threshold [Ng, 2004].

As its name suggests, in the *joint* strategy the anaphor detection and antecedent selection are modeled jointly. Denis and Baldridge [2007a] perform joint inference of the two subsystems using integer linear programming, with antecedent selection provided by a mention-pair model. In contrast, Rahman and Ng [2009] involve joint treatment of the two subtasks already at train time. Instead of processing solely the anaphoric mentions, their ranker learns from both anaphoric and non-anaphoric mentions. For each such mention – a potential anaphor, the set of antecedent candidates is augmented with an additional *dummy candidate* representing the potential anaphor itself. Selecting the dummy candidate is equivalent to labeling the potential anaphor as non-anaphoric. Joint strategy is adopted also in other approaches to CR, for instance by Lassalle and Denis [2015] who extended the latent trees approach [Fernandes et al., 2014, Björkelund and Kuhn, 2014].

3.1.4 Specialized Models

Different types of anaphors are characterized by different coreference properties. For instance, while a reflexive pronoun and its antecedent usually appear in the same clause, the antecedent of a personal pronoun can lie further away, and the antecedent of an anaphoric nominal group even more further away. Likewise, while the relative pronoun *who* refers almost exclusively to persons, the antecedent of *which* is not even restricted to a nominal group or pronoun.

One has to first define along which dimension to split the anaphors. Ng [2005] proposed specialized classifiers for different types of pronouns split along their lexical value. That is, separate models are learned for pronouns *his* and *her* and pronouns not seen in a training data are handled by a backoff model trained on all pronouns. This approach is, however, not possible for open classes, e.g. nouns.

Denis and Baldridge [2008] rather posit partitioning of anaphors and corresponding specialized models to five types: third person pronouns, speech pronouns, proper names, definite descriptions, and others. Since these types are mainly motivated by theoretical linguistic studies in salience [Ariel, 1988, Gundel et al., 1993], the specialized ranking models may differ in the size of a scope from which they select antecedent candidates. Nonetheless, the types differ also in other aspects. For instance, string matching features would receive much higher weight in a model for proper names or definite descriptions than in a model for pronouns.

A concept similar to specialized models is adopted by *easy-first approaches* to CR. Coreference chains are built sequentially by applying a battery of models specialized at subsets of mention pairs. It proceeds from models with highest precision to models with lowest precision. This approach was employed in a system by Lee et al. [2013], which outperformed all the others in the CoNLL 2011 Shared Task [Pradhan et al., 2011].

3.1.5 Coreference Resolution in Czech

Most works on anaphora or coreference resolution in Czech relate to the theory of Prague tectogrammatics. Nevertheless, until the first data annotated with coreference within the project of Prague Dependency Treebank [Hajič et al., 2011, PCEDT] appeared, the task of coreference modeling in Czech had been proposed only theoretically. It mainly included the works on activation (salience) models considering syntax and topic-focus articulation [Hajičová, 1987, Hajičová et al., 1990].

The first experiments on automatic anaphora resolution by Kučová et al. [2003] targeted grammatical coreference. Presented within the proposal of the annotation schema for coreferential relations in PDT, their set of high-precision rules were primarily designed to facilitate manual labor of the annotators. Kučová and Žabokrtský [2005] followed up with an approach aimed at pronominal textual coreference. The proposed a sequence of filters gradually applied to antecedent candidates selected for the current and the previous sentence. Later, it has been outperformed by another rule-based approach [Nguy and Žabokrtský, 2007]. While a machine learning approach (namely, decision trees) was first taken for Czech coreference using a mention-pair model in [Nguy, 2006], it could not outperform the heuristic-based methods unless a mention-ranking approach was adopted in [Nguy et al., 2009]. Reported performance of the methods for pronominal coreference have gradually increased from 60 F-score points in [Kučová and Žabokrtský, 2005] to nearly 80 F-score points in [Nguy et al., 2009]. Nevertheless, all these experiments share one major shortcoming – they were carried out on gold tectogrammatical trees. Such experiments not only take unfair advantage of perfect syntactic and shallow-semantic structures of the sentences, but they also often exploit manually disambiguated features that would remain ambiguous without the information coming through the coreference (e.g. gender and number for some expressions).

The Saara framework [Němčík, 2006, 2009] was the first anaphora resolution system that could operate on texts containing no manual linguistic annotation. This modular system addresses pronouns and at its core re-implements several classical algorithms that can be alternated (e.g. Hobbs algorithm [Hobbs, 1978] and activation models based on topic-focus articulation [Hajičová, 1987]). In [Bojar et al., 2012], the authors used the same ranker and the feature set as [Nguy et al., 2009], this time extracted from the data automatically analyzed to the tectogrammatical representation though. Unreliability of information on tectogrammatical gender and number as well as uncertainty of reconstructed zero subjects resulted in a dramatic drop in performance to 50 F-score points.

Ongoing annotation work on extended textual coreference and bridging anaphora [Nedoluzhko, 2011][4] encouraged research on textual coreference resolution of nominal groups. Novák [2010] carried out the first experiments on coreference of nominal groups in Czech. The approach of maximum entropy ranking was further elaborated in Novák and Žabokrtský [2011], where the authors compared systems based on classification and ranking approaches. As a result, the best system achieved 44 F-score points on coreference with specific reference. Novák

[4]Later released in Prague Discourse Treebank 1.0 [Poláková et al., 2012] and PDT 3.0 [Bejček et al., 2013].

[2010] also pioneered resolution of coreference with generic reference and some bridging relations. Inspired by Hearst patterns [Hearst, 1992] addressing distributional semantics of words, they proposed a morphology-based feature targeting the part–whole relation.

A closely related task to coreference resolution in Czech is the task of identifying an unexpressed subject. Nguy and Ševčíková [2011] utilize a classifier to identify a place for introducing a zero subject and to determine its linguistic type (e.g. anaphoric, general and unspecified). Veselovská et al. [2012] focused on the same task, this time with a rule-based method. In addition to monolingual approach, they presented a cross-lingual method where identification of Czech zero subject is aided by information from English.

3.2 Cross-lingual Approaches to Coreference Resolution

Cross-lingual approaches to natural language processing have received special attention with the advent of multilingual resources. It is especially obvious for tasks such as part-of-speech tagging and dependency parsing, where considerable effort has been invested in lowering barriers to such research by collecting data in multiple languages, seeking for language-universal annotating standards, and adjusting the data collections to satisfy these standards. Probably due to unclear standards in coreference annotation and evaluation even within English (see Section 3.1.1), let alone the situation across languages, the research of cross-lingual approaches to coreference resolution lags behind the other tasks. Nonetheless, there are still some works that the present work can relate to.

We divide the cross-lingual approaches to three categories. A special section is devoted to each of them. Namely, Section 3.2.1 surveys works on coreference projection, Section 3.2.2 introduces delexicalized approaches, and Section 3.2.3 presents attempts to bilingually informed coreference resolution.

3.2.1 Coreference Projection

Approaches to cross-lingual projection are usually aimed to bridge the gap of missing resources in the target language. So far, they have been quite successfully applied to part-of-speech tagging [Täckström et al., 2013], syntactic parsing [Hwa et al., 2005], semantic role labeling [Padó and Lapata, 2009], opinion mining [Almeida et al., 2015], etc. Projection techniques are generally grouped into two types with respect to how they obtain the translation to the source language, i.e. the language for which sufficient amount of language resources exists. *MT-based approaches* apply a machine-translation service to create synthetic data in a source language. *Corpus-based approaches* take advantage of the human-translated parallel corpus of the two languages. Let us describe these approaches in a greater detail with the focus on coreference.

MT-based Approaches. The workflow of these approaches is as follows. Starting with a text in the target language to be labeled with coreference, it first must be machine-translated to the source language. A coreference resolver for the

source language is then applied on the translated text and, finally, the newly established coreference links are projected back to the target language.[5] Flexibility of this approach is in the fact that it can be applied at both train and test time, and no linguistic tools for the target language are required.

Rahman and Ng [2012] seem to have published the earliest work that implements this workflow. They presented three possible settings based on availability of tools for analysis of the target language: (1) no tools, (2) a mention extractor, or (3) all necessary tools available. In Setting (1), target-language mentions are determined by the source-language ones and word alignment. To ensure contiguity, each target-language mention is formed by a minimal span that covers all words in the mention. In Setting (2), the source-language mentions, on which a coreference resolver operates, are predefined by projection of target-language mentions identified by the mention extractor. Finally, in Setting (3), the target-language coreference annotation acquired as in Setting (2) serves as a training dataset to build a coreference resolver for the target language. The resolver takes advantage of the additional linguistic information provided by the available target-language tools. Rahman and Ng [2012] show that in projection from English to Spanish and Italian, Setting (3) outperforms the other settings, achieving the CoNLL F-scores of 50.6 and 57.4 for Spanish and Italian, respectively, which is 93% and 90% of the CoNLL F-scores reached by supervised systems.

Ogrodniczuk [2013] presents an approach that combines Rahman and Ng's settings (1) and (2). Instead of enforcing mention boundaries in one language by mentions from the other language, they posit that mentions should be identified by the extractor for the particular language. Assignment to coreference cluster is then projected only between mentions with their heads aligned. In projection from English to Polish, they achieve surprisingly high CoNLL F-score of 70.3, which is comparable to supervised systems.

Corpus-based Approaches. Here, a human-translated parallel corpus of the two languages is available and the projection is performed within this corpus. Coreference annotation in the source-language side of the corpus may be both labeled by humans or a coreference system. The target-language side of the corpus then serves as a training dataset for a coreference resolver. This approach thus must be applied at train time and, moreover, it demands availability of necessary linguistic tools for the target language. On the other hand, human translation and gold coreference annotation, if available, should increase the quality of the projected coreference.

Postolache et al. [2006] followed this approach using a small English-Romanian corpus of 638 sentence pairs in order to create a bilingually-annotated resource. They projected manually annotated coreference, which was then post-processed by linguists to acquire high quality annotation in Romanian. A Romanian mention is formed as a minimal span covering all the words, whose counterparts belong to the corresponding English mention. The head of the Romanian men-

[5]Note that there is a possible modification to this approach, where an MT service is applied in the opposite direction. The advantage is that in such scenario, a gold annotated coreference corpus in the source language can be exploited to obtain a synthetic translation to the target language with projected gold coreference annotation. Such data can then serve as a training dataset. To the best of our knowledge, this approach has never been tested on coreference resolution.

tion is identified using constraints based on part-of-speech tags. If the head does not align with the head of its source English mention, the projected Romanian mention is discarded. Based on the gold coreference annotation of the Romanian side of the corpus, they evaluated the F-scores of mention heads' matching, mention spans' overlapping, and coreference clusters on all as well as only correctly projected mentions. The error analysis they carried out shows that the majority of errors in coreference projection stems from a lower recall (around 70%) caused by missing alignment due to alignment errors or language differences introduced in the translation.

Unlike the previous work, de Souza and Orăsan [2011] adopted a similar projection mechanism for a parallel English-Portuguese corpus in order to build a coreference resolver for Portuguese. Their resolver trained on projected links with its performance of 7.1 MUC and 14.4 CEAF F-score failed to surpass a baseline system that classifies two mentions as coreferential only if their heads match. The authors saw the reason for the system's low performance in insufficient quality of the output by the English coreference resolver.

Just recently, Wallin and Nugues [2017] presented corpus-based projection experiments for English to Swedish and German that follows the same approach as de Souza and Orăsan [2011]. Although they propose target-language heuristics to improve mention identification, their coreference resolvers trained on projected links and system-detected mentions fall considerably behind the resolvers trained on gold mentions (CoNLL F-score of 13.1 vs. 37.0 on system vs. gold mentions, respectively).

Martins [2015] extended this approach by learning coreference with a specific type of regularization at the end. Their system performs with CoNLL F-score of 38.8 and 37.2 using the automatic coreference annotation projected from English to Spanish and Portuguese, respectively, gaining 4-6 points over the standard projection mechanism as introduced by de Souza and Orăsan [2011]. The gains result from ability of their method to recover links missing due to projection via inaccurate alignment. Their projection approach achieves 88% and 94% of CoNLL F-score reached by a supervised system for Spanish and Portuguese, respectively.

Yulia Grishina with her colleagues also investigate possibilities of corpus-based coreference projection. In [Grishina and Stede, 2015], they introduced a "generalizable" annotation schema that they tested on parallel texts of three languages (English, Russian and German) and three genres (newswire articles, short stories, medical leaflets). Using this dataset, they conducted experiments on projection from English to the two other languages. Unlike Postolache et al. [2006], they adopted a knowledge-lean approach using no language-dependent tools or linguistic knowledge about the target language. In other two works [Grishina and Stede, 2017, Grishina, 2017], they pursue a goal of multi-source projection of manual and automatic annotation, respectively. They propose several strategies of combining projections from multiple languages, with some of them slightly improving the F-score of the best-performing projection source. Despite using two sources, they appear to overlap considerably as the authors see only small improvements in recall. They also provide a qualitative analysis suggesting that pronouns have much higher projection accuracy[6] than nominal groups. They justify their unsatisfac-

[6]As far as we are concerned, it is misleading to call it accuracy as there is another aspect of how many target mentions are not covered by projected links at all. We rather denote this

tory results especially for German nominal groups by problems with including unaligned German determiner of a definite description. This harmonizes with observations in [Wallin and Nugues, 2017], where CR quality in German drops dramatically after a switch from gold to system-detected mentions.

3.2.2 Delexicalized Approaches

Delexicalization is another way to exploit data from different languages. Similarly to projection techniques, the motivation behind delexicalization is to acquire coreference information in a resource-poor language. Such systems are based on features, which strictly ignore lexical or other language-dependent information. It allows them to be trained on one language and applied to another, or trained on several different languages at once. In many natural language processing tasks, a related work on delexicalized approaches would deserve a separate section. Nonetheless, it is not the case in coreference resolution, where this area has been somewhat neglected, so far.

Bodnari [2014] suggests a joint approach to named entity recognition and coreference resolution based on a factorized hidden Markov model. Their model incorporates merely two observed variables: a universal part-of-speech tag and a dependency head. This delexicalized nature of the model allows them for training on multiple languages, including English, German, Dutch, French, Spanish and Catalan.

Although the main idea of [Martins, 2015] is an enhanced projection-based coreference resolution, they compare their system with delexicalized approaches. Their delexicalized systems utilize universal part-of-speech tags [Petrov et al., 2012], cross-lingual word embeddings and a universal representation capturing the gender, number and person of a pronoun. Even though the best delexicalized system applied to the same language on which it was trained performs on par with the lexicalized system, in cross-lingual scenario it is outperformed by both the basic and the enhanced projection-based CR.

3.2.3 Multilingually Informed and Joint Multilingual Resolution

The previous two approaches with the motivation in obtaining annotation for resource-poor languages should benefit from relatedness of the source and the target language. Multilingually informed resolution could, on the contrary, improve with growing diversity of the languages involved. The systems adopting this approach are trained on parallel corpora and some of them must be given parallel texts also at test time. Similarly to projection, the parallel data may be either human- or machine-translated. Whereas multilingually informed resolution aims at processing only a single language and the other languages only guide its decisions, joint multilingual resolution attempts to build a joint model for multiple languages.[7] Note that most of the previous works in fact handle only two languages, i.e. they utilize a bilingual approach. This is also the case of the

score as precision in our experiments in Chapter 8.

[7]We adopt this terminology from Haulrich [2012, p. 42–43], who made this distinction specifically for parsing.

present work. However, the truly multilingual approaches are currently becoming more and more popular, e.g. [Tiedemann, 2018].

Although numerous NLP tasks has been so far addressed by multilingual resolution, it has attracted much attention in word sense disambiguation (WSD) and parsing. As for the WSD, the popularity can be attributed to the fact that the systems may benefit just from the raw translations with no need to linguistically analyze them. The reason is that different senses of a word might end up with different translations to another language and various languages may indeed differ in how words and senses are mutually distributed. This idea even gave birth to a specific task of *Cross-lingual WSD*,[8] in which the sense is defined by its translations to multiple languages. A multilingual approach to the cross-lingual WSD was for instance utilized by Lefever et al. [2011], who augmented their system with bag-of-word features capturing the translation of a word to other languages. Gonen and Goldberg [2016] approached the task of traditional WSD (for English prepositions) by exploiting representations induced by two Long Short-Term Memory networks [Hochreiter and Schmidhuber, 1997], while predicting a corresponding German, French and Spanish preposition in a multilingual parallel corpus.

The popularity of multilingual approaches in the parsing task may be attributed to availability of parallel texts annotated with syntax. Joint parsing of both the source and the target text along with searching for the best alignment between the trees has been approached in a more [Burkett et al., 2010] or less [Smith and Smith, 2004, Burkett and Klein, 2008] integrated approach. Much closer to our work is the research on bilingually informed parsing by Haulrich [2012], in which English trees are used to enrich the feature set for a Danish parser and vice-versa. Rosa et al. [2012] explored the same approach on the Czech-English language pair. Moreover, they adapted this technique to parse a machine-translated text.

At the same time, development of multilingually informed coreference resolution has been hindered probably by the lack of parallel data annotated with coreference. To the best of our knowledge, the only parallel corpora containing coreference annotation are Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2011, PCEDT 2.0] with its successors and derivations (e.g. PAWS [Nedoluzhko et al., 2018]), German-English texts in ParCor 1.0 [Guillou et al., 2014], and the German-Russian-English data collected by [Grishina and Stede, 2015]. It is therefore surprising that most of the works that clearly make use of parallel corpora with manual annotation of coreference (as we do in Section 7.3) date back to the time before all the mentioned datasets were released.

Harabagiu and Maiorano [2000] present an heuristics-based approach to coreference resolution. The set of heuristics is expanded by exploiting the transitivity property of coreferential chains in a bootstrapping fashion. Moreover, they expand the heuristics even more, following mention counterparts in translations of source English texts to Romanian with coreference annotation. To this end, the authors translated the English corpora MUC-6 [1995] and MUC-7 [1998] to Romanian and provided them with manual coreference annotation.

A study by Mitkov and Barbu [2003] is the most related approach to the work presented in Section 7.3. They adjust a rule-based pronoun coreference resolution

[8] Also organized as shared tasks [Lefever and Hoste, 2010, 2013].

system to work on a parallel corpus. After providing a linguistic comparison of English and French pronouns and their behavior in discourse, the authors distill their findings into a set of cross-lingual rules to be integrated into the CR system. In evaluation, they observe improvements in resolution accuracy of up to 5 percentage points compared to the monolingual approach. The scores were measured on test data, which comprise English technical texts manually translated to French and annotated with coreference on pronouns.

The work of Veselovská et al. [2012], in which the author of this thesis was involved, did not address coreference resolution but two related tasks – identification of types of the English personal pronoun "*it*" and Czech types of unexpressed subject. The authors first built isolated monolingual systems for the tasks. Taking advantage of PCEDT 2.0, they subsequently extended the system for Czech subjects with the features from English. It resulted in improvement of 8 points in F-score.

Chen and Ng [2014] employed translations with coreference annotation in a source language to address the task of coreference resolution of Chinese pronouns expressed on the surface. Their bilingually informed model exploits Chinese features as well as features, extracted from the Chinese texts machine-translated to English. It allows for taking advantage of English nouns' gender and number lists, which as the authors claim correspond to the distribution of genders and numbers over Chinese nouns. In their experiments, they combined this model with a monolingual and projection-based model, performing 2-3 F-score points above the best monolingual system. Furthermore, Chen and Ng [2014] conducted experiments where synthetic translations were replaced with the manual ones. The replacement earned only one additional F-score point, suggesting that either their machine translation system did a great job on pronouns, or pronoun translation from Chinese to English is not a big issue.

4. Data Sources, Tools and Evaluation

In this section, we introduce all data sources, tools and evaluation measures that we utilize in the remainder of this work. Section 4.1 briefly presents and gives basic statistics on the data sources used in this work for training and testing. These include Czech and English monolingual as well as Czech-English parallel corpora.

Description of tools that we employ is divided into two parts. In the first part in Section 4.2, we thoroughly present the pre-processing analysis pipeline of our CR system, which automatically builds the representations that to some extent satisfy the requirements of Prague tectogrammatics (see Section 2.2 for more). Each text must undergone this analysis in order to be correctly processed by our coreference resolution.

The second part in Section 4.3 introduces CR systems that we directly use in some of the experiments to compare with the approaches presented in this thesis. For both Czech and English it includes the in-house predecessor of our system. In addition, our CR system for English is also contrasted with three variants of the Stanford CR: rule-based, statistical and neural one.

Section 4.4 presents the evaluation measure for coreference that is employed in this thesis and compares it with some other existing measures.

4.1 Data Resources

Let us introduce all the corpora and data resources that we use throughout the thesis. Some of these corpora are exploited for training the models, some of them for testing the approaches and analyzing the decisions the models made, and one only to show its statistics. As the topic of this work is related to cross-lingual studies, apart from monolingual corpora of Czech and English we employ also some Czech-English parallel corpora. Most of these corpora are in fact treebanks of dependency trees, whose linguistic annotation includes the *tectogrammatical layer*, i.e. the layer of deep syntax that follows the theory of Functional Generative Description [Sgall, 1967, Sgall et al., 1986]. Nature of the annotation varies across the corpora, from fully manual, through hybrid to fully automatic. The corpora also differ in whether they include coreference annotation, and what is its style.

Namely, this work takes advantage of the following corpora:

- *Prague Dependency Treebank 3.0* [Bejček et al., 2013, PDT 3.0],
- *Prague Czech-English Dependency Treebank 2.0 Coref* [Nedoluzhko et al., 2016a, PCEDT 2.0 Coref] and its subsection denoted as *PAWS* [Nedoluzhko et al., 2018],
- *CzEng 1.0* [Bojar et al., 2011],
- *CoNLL 2012 test set* [Pradhan et al., 2012].

Data Source	Sents	Langs	Parallel	Domains	Tecto gold	Coref gold
PDT 3.0	49k	CS	×	journalistic	✓	✓
PCEDT 2.0 Coref	49k	CS, EN	✓	journalistic	✓	✓
PAWS	1k					
CzEng 1.0	14,573k	CS, EN	✓	mixed	×	×
CoNLL 2012	9.5k	EN	×	journalistic	×	✓

Table 4.1: Basic characteristics of the data sources that we utilize throughout this work. The columns represent from left to right the following properties: size in sentences; languages; is the corpus parallel?; domains; does the corpus contain manual tectogrammatical annotation?; does the corpus contain manual annotation of coreference?

The aforementioned characteristics for all these corpora are clearly listed in Table 4.1. The following sections elaborate on all of these corpora further.

You might notice in Table 4.1 that some corpora comprise manual annotation of trees with respect to Prague tectogrammatics. Nevertheless, we decided to ensure the real-world scenario, where no gold annotations are available. We thus stripped all the annotations off the text and replaced them with automatic ones obtained by the pre-processing pipeline described in Section 4.2. The only gold information that remains there for the purpose of training and testing is coreference. In fact, as required by our CR system, all the presented datasets must be pre-processed with this automatic pipeline. Most of the experiments you may encounter in the rest of the thesis are thus conducted on such *automatically annotated trees*.[1] We will warn you explicitly, if experiments employ the manual trees, instead.

There is too many mention types and datasets to show all data characteristics in a single table. We rather split split it into three tables. First, Table 4.2 shows the basic statistics on all train and test data used in this work. It is calculated on automatically annotated trees. Table 4.3 then shows statistics only on the evaluation test sets, however, this time with detailed numbers of covered and coreferential nodes within each of the expression categories. Finally, the statistics of the CzEng 1.0 is displayed separately in Table 4.4, as we use this dataset only in Attachment A.

4.1.1 Prague Dependency Treebank

The Prague Dependency Treebank 3.0 [Bejček et al., 2013, PDT, PDT 3.0][2] is a collection of Czech journalistic texts, augmented with manual linguistic annotation ranging from morphology to semantics and discourse. PDT consists of articles published in four newspapers and journals in the early 90s. These articles have been linguistically annotated following an extended variant of the principles of the Prague tectogrammatics as described in Section 2.2. Annotation of all the phenomena has been carried out manually. Not all the texts are covered by all

[1] We call the also *system trees* or *automatically analyzed trees*.

[2] `http://ufal.mff.cuni.cz/pdt3.0`

	Train			Eval					
	PCEDT 2.0 Coref		PDT 3.0	PCEDT 2.0 Coref		PAWS		PDT 3.0	CoNLL 2012
	Czech	English		Czech	English	Czech	English		
Sentences	38,219		38,727	5,462		1,078		5,476	9,479
Words	927 k	890 k	649 k	126 k	132 k	25 k	27 k	92 k	173 k
T-nodes	643 k	681 k	497 k	103 k	94 k	22 k	20 k	71 k	97 k
Coref. nodes	99 k	86 k	86 k	12 k	14 k	3 k	3 k	14 k	15 k

Table 4.2: Basic statistics of the train and the evaluation test data collected on system trees. Note that the PAWS dataset is both train and evaluation (using 10-fold cross-validation).

	PCEDT 2.0 Coref		PAWS		PDT 3.0	CoNLL 2012
	Czech	English	Czech	English		
Personal pron.	270 / 232	1,760 / 1,550	64 / 64	354 / 301	554 / 502	4,138 / 3,406
Possessive pron.	397 / 379	1,000 / 984	107 / 107	241 / 234	379 / 362	1,278 / 1,186
Refl. poss. pron.	549 / 515	0 / 0	90 / 90	0 / 0	369 / 361	0 / 0
Reflexive pron.	209 / 116	52 / 50	44 / 23	12 / 12	290 / 219	141 / 131
Demonstr. pron.	781 / 378	696 / 155	108 / 56	112 / 16	774 / 432	1,960 / 214
Zero subject	3,128 / 1,878	0 / 0	483 / 406	0 / 0	2,194 / 1,087	0 / 0
Zero in nonfin. cl.	698 / 639	4,647 / 3,178	130 / 130	869 / 600	685 / 629	5,368 / 0
Relative pron.	1,816 / 1,568	1,223 / 855	349 / 310	239 / 156	1,362 / 1,103	2,096 / 16
1st/2nd pers. pron.	727 / 351	865 / 388	94 / 38	113 / 39	1,497 / 1,177	5,691 / 3,133
Named entities	8,466 / 121	11,353 / 538	25 / 11	2,576 / 0	6,780 / 711	11,400 / 3,267
Nominal group	41,812 / 5,258	46,658 / 6,144	8,536 / 1,356	9,983 / 1,332	26,768 / 5,573	41,053 / 3,818
Other	43,705 / 576	25,291 / 326	11,694 / 482	5,012 / 96	29,682 / 1,593	23,862 / 64
Total	102,558 / 12,011	93,545 / 14,168	21,724 / 3,073	19,511 / 2,786	71,334 / 13,749	96,987 / 15,235

Table 4.3: Number of all t-nodes and coreferential t-nodes (all/coref) for all node types collected on the system trees in the evaluation sets.

		Czech		English	
	Sent. pairs	Tokens	T-nodes	Tokens	T-nodes
News	201 k	4,280 k	3,167 k	4,737 k	3,054 k
Fiction	4,335 k	57,177 k	43,129 k	64,264 k	42,049 k
Subtitles	3,077 k	19,572 k	14,709 k	23,354 k	15,046 k
EU	3,993 k	78,022 k	57,252 k	87,489 k	55,057 k
Complete	15,136 k	206,442 k	154,042 k	232,691 k	150,377 k

Table 4.4: Basic statistics of CzEng 1.0 in the four selected domain and over the complete corpus.

the three layers, but those which contain more than 833,000 tokens in almost 50,000 sentences formed in over 3,000 documents.

The principles of the Prague tectogrammatics as we defined them in Section 2.2 basically reflect the extent of annotated phenomena in the Prague Dependency Treebank 2.0 [Hajič et al., 2006, PDT 2.0].[3] For the version 3.0, PDT was extended with annotations of additional phenomena. These include multiword expressions, clause segmentation, genres, discourse relations, etc.

Among all corpora based on Prague tectogrammatics, PDT 3.0 is the treebank with the richest annotation of coreference and anaphoric relations. It contains coreference of all the types listed in Section 2.3, including coreference of nominal groups, coreference of nouns with generic reference, and coreference of pronouns in the 1st and 2nd person. To the best of our knowledge, it is also the only corpus of Czech with the annotation of bridging relations.

PDT comes with data divided into three sets: training (`train`), development test (`dtest`), and evaluation test set (`etest`). During the experiments, we exploited the sets as expected. The training set was used for training the models, while the development test set served for development or tuning purposes. The evaluation test data were set aside for the final evaluation. Evaluation scores that we report are all measured on the `etest` unless stated otherwise.

4.1.2 Prague Czech-English Dependency Treebank

The Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2011, 2012, PCEDT 2.0][4] is a Czech-English parallel corpus of journalistic texts, whose linguistic annotation draws on the adjusted version of the Prague tectogrammatics. The English part contains the entire Wall Street Journal Section of the Penn Treebank [Marcus et al., 1999]. The Czech part consists of translations of all the texts. Altogether, it comprises over 1.2 million words for each of the languages in almost 50,000 sentence pairs. Figure 4.1 depicts a sample pair of sentences from PCEDT and their tectogrammatical representations.

The data from both language parts have been annotated on all the three layers following the theory of Prague tectogrammatics. Yet unlike the PDT, the annotation is not completely manual. English dependency trees on the analytical layer have been obtained by automatic transformation of the original manual phrase-structure annotation of the Penn Treebank. The original phrase-structure trees are in fact included in PCEDT as a *p-layer*. Czech translations have been processed automatically to the morphological and analytical layer. But importantly, tectogrammatical trees have been built completely manually and independently for each of the languages.

In comparison to the principles of Prague tectogrammatics implemented in PDT 2.0, PCEDT is slightly simplified. The most visible limitation is missing annotation of topic-focus articulation. The range of annotated grammatemes in Czech is very limited. On the other hand, a morphosyntactic tag called *formeme* is introduced for every tectogrammatical node, showing how the node is realized on the surface (see Section 4.2.1 for more). It can serve as a practical shortcut

[3] `http://ufal.mff.cuni.cz/pdt2.0`

[4] `http://ufal.mff.cuni.cz/pcedt2.0`

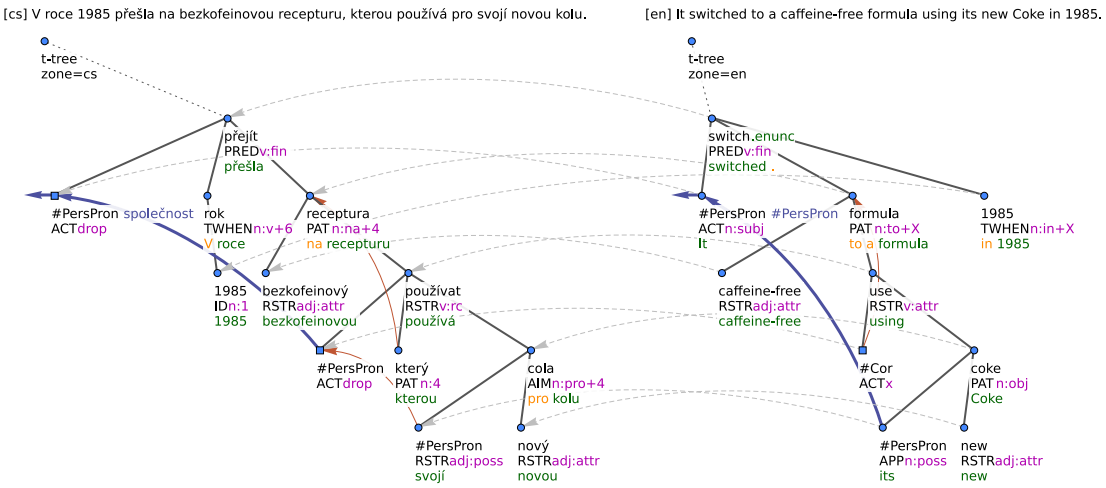

Figure 4.1: A tectogrammatical representation of a sample sentence pair from PCEDT with grammatical and textual coreference links and node alignment links denoted by normal solid, bold solid, and dashed arrows, respectively.

for search queries and it is a key factor of the translation through the tectogrammatical layer with TectoMT [Dušek et al., 2012].

In fact, we do not use the original PCEDT 2.0 in this work. Instead, we employ its successor Prague Czech-English Dependency Treebank 2.0 Coref [Nedoluzhko et al., 2016a,b, PCEDT 2.0 Coref],[5] which offers extended annotation of coreference an alignment on coreferential expressions. Coreference annotation for PCEDT 2.0 Coref has proceeded in two stages.[6] In the first stage resulting in PCEDT 2.0, grammatical coreference was automatically pre-annotated by heuristics and then manually corrected. Whereas Czech pronominal textual coreference has been annotated completely manually, the English one was first transformed from the BBN Pronoun Coreference and Entity Type Corpus [Weischedel and Brunstein, 2005, BBN-PCETC] and then corrected and complemented with the coreference of zeros. The second stage has introduced nominal coreference. This has been performed by transformation from BBN-PCETC for some English mentions, but mostly manually for the rest.

For PCEDT 2.0, the nodes of the Czech and English trees had been automatically aligned on analytical as well as tectogrammatical layer using GIZA++ and some heuristics (more on the original PCEDT alignment in Section 4.2.3). Later for the PCEDT 2.0 Coref release, the annotation of alignment was slightly modified. The modification concerned a selected set of potentially coreferential expressions in both languages. In a section of the data comprising around 1,000 sentences, alignment between such expressions was annotated manually. The rest was annotated automatically by a supervised aligning method trained on the manually annotated portion. For experiments on PCEDT with the supervised alignment, we replace the manual alignment in the particular section with the

[5] `http://ufal.mff.cuni.cz/pcedt2.0-coref`

[6] Coreference in PCEDT and the process of its annotation is described in detail in [Nedoluzhko et al., 2016b].

supervised one acquired in a 10-fold cross-validation fashion. This ensures that all PCEDT data are aligned consistently.

PCEDT does not come with its own split into a training and test set. Consequently, we needed to divide the data ourselves and we did it as follows. We declared sections 00–18 to be a training set, sections 20–21 a development test set, and sections 22–24 an evaluation test set. In addition, we reserved the section 19, half of which has been annotated with manual alignment, only for training and testing the supervised aligner.

The author of this thesis co-authored PCEDT 2.0 Coref. Concretely, he collaborated on manual annotation of alignment and developed the supervised aligning method for coreferential expressions. We will thus elaborate on these topics further in Chapter 6. The author of this thesis was also responsible for assembling all the treebank's components and releasing the data.

PAWS. Recently, the section of PCEDT 2.0 Coref containing the manual annotation of alignment was extracted, extended and released as the PAWS treebank [Nedoluzhko et al., 2018].[7] PAWS is a multilingual coreference-annotated parallel treebank. Besides its Czech-English part, which was copied from PCEDT without any change, it currently contains the Russian and Polish part translated for the English original. These translations have then been automatically analyzed to the tectogrammatical representation and manually post-edited in an extent that allows for annotation of coreferential relations. Coreference has been annotated manually separately on each of the languages. Furthermore, Russian coreferential expressions have been manually aligned with their English and Czech counterparts.

In several places of the thesis, we use the manually aligned part of PCEDT. For simplicity, we will be referring to it with the name PAWS, even though we never employ its Russian and Polish part in this work.

4.1.3 CzEng

CzEng 1.0 [Bojar et al., 2011, 2012] is a massive Czech-English parallel treebank, annotated with respect to the simplified version of the Prague tectogrammatics. Unlike all the other described corpora, CzEng embraces texts from multiple domains including (1) Fiction, (2) EU Legislation, (3) Movie Subtitles, (4) News, etc. CzEng comprises more than 200 million tokens for each language in 15 million sentence pairs. Note that the size of CzEng is 300-times bigger than PCEDT. Moreover, PCEDT is not a subset of CzEng. Although CzEng is split into documents, these are solely artificial. The longest sequence of continuous utterance (called *CzEng block*) in fact never exceeds 15 sentences. Blocks have been shuffled so that it is impossible to reconstruct the original text sources.

Due to its size, it is no surprise that the linguistic annotation is completely automatic. But still, it tries to comply with the standards of Prague tectogrammatics, even though some phenomena are ignored and some annotations are inevitably wrong. The annotation pipeline that originally produced CzEng is very similar to the one required by our coreference resolver (see Section 4.2). However, there are some differences, e.g. in the extent of reconstructing zeros in non-finite

[7] The name stands for the Parallel Anaphoric Wall-Street journal.

clauses. Hence, to avoid any issues caused by possible incompatibilities between these two variants of tectogrammatical annotation, we decided to replace the original annotation with the one expected by our CR system. The annotation pipeline has processed both language sides of the corpus independently.

Application of the new pipeline also deleted the original coreference annotation. The original coreference annotation in CzEng contains grammatical coreference and pronominal coreference with anaphors in 3rd person. In order to acquire it, we applied an approach described in Section 4.3.1 We mention it here, because the modules that provided the original coreference annotation in CzEng are employed as a baseline system in experiments with monolingual coreference resolution in Chapter 7.1.

Alignment for CzEng was obtained in a way similar to the original alignment for PCEDT 2.0. The GIZA++ tool was used to find alignments between surface words. These alignments were then projected to the tectogrammatical layer and complemented by a heuristic-based alignment for some reconstructed nodes.

Data in the CzEng 1.0 release are split into training, development test as well as evaluation test part. In our experiments, we make use solely of the training set, which still accounts for 98% of all CzEng data.

In the present work, we employ CzEng only to collect the statistics on distributions of coreferential expressions and their counterparts and contrast them with the other parallel corpora as well as within the CzEng corpus betweem four selected domains. We present these statistics in Attachment A.

The author of this thesis is one of the co-authors of CzEng 1.0.[8] He was responsible for automatic coreference annotation in both languages of the corpus.

4.1.4 CoNLL 2012 Test Set

This dataset is an official testset for CoNLL 2012 Shared Task [Pradhan et al., 2012] to evaluate English systems. It comprises around 170,000 tokens in 9,500 sentences that have been sampled from texts in OntoNotes 5.0 [Pradhan et al., 2013] corpus. Except for coreference, it contains no other manual annotation.

The texts come from different domains including news and magazine articles, broadcast news and conversations but also telephone conversations, web data and excerpts from the New Testament. Nevertheless, the texts with a non-conversational journalistic style prevail, accounting for 85% of documents. There is around 25% of documents in the CoNLL dataset which come from the Wall Street Journal part of the Penn Treebank. These documents are also included in PCEDT. Fortunately, all of the documents belong to the section `23`, so they overlap only with the evaluation test set of PCEDT. Blind evaluation on all the evaluation test sets is thus ensured.

OntoNotes, and consequently CoNLL 2012 testset, differ from the Prague treebanks in the following main aspects that relate to coreference: (1) as it is common for many coreference datasets especially for English, coreference is annotated on the surface; mentions of the same entity are represented as co-indexed spans of consecutive words with fixed mention boundaries, (2) it contains no zeros and relative pronouns are not annotated for coreference.[9]

[8] And also of the follow-up version CzEng 1.6 [Bojar et al., 2016].

[9] Reasons for ignoring relative pronouns in OntoNotes are unclear. They might have been

4.2 Treex Pre-processing Pipeline

The aim of the pre-processing pipeline is to automatically annotate a text, so that this annotation to some extent satisfies the principles of the Prague tectogrammatics. In other words, it should perform the linguistic analysis up to the tectogrammatical layer. It is essential, because the coreference resolver that we design in this work requires a tectogrammatical representation of the text to correctly operate.[10] We utilize the Treex framework to produce such annotation.

Treex [Popel and Žabokrtský, 2010] is a modular open-source NLP framework. Originally, it had been designed for machine translation from English to Czech through the tectogrammatical layer with the TectoMT system [Žabokrtský et al., 2008], but it has gradually developed to a multi-purpose framework, isolating TectoMT as one of its applications. Treex Coreference System (to be introduced in Section 7.1) and the CzEng production pipeline are other examples of Treex applications.

Modularity of Treex is ensured by *blocks*. Block is the smallest reusable processing unit, usually with a well-defined input and output and linguistically interpretable functionality. Blocks may be concatenated into sequences, called *scenarios*, which in combination with conversion of the input and output data and other supporting tools form an already mentioned application.

From its very beginning, Treex has been tailored to process texts in a way that resembles the Prague tectogrammatics as much as possible. The inner representation of documents reflects this theory. A sentence is thus represented as a *bundle* of trees. Each tree in a bundle is identified by its layer (analytical,[11] tectogrammatical), its language and possibly by an additional identifier, which may come in handy if, for instance, one wants to represent both the manual and automatic tectogrammatical tree of a sentence. Expectedly, trees consist of nodes, which are attribute-value structures containing the attributes associated with a given layer.

The pre-processing pipeline may consist of four scenarios: (1) Czech analysis, (2) English analysis, (3) monolingual alignment of manually and automatically analyzed trees, and (4) cross-lingual alignment. Not all the four scenarios must be necessarily run, it depends on the type of the data. In the following, we elaborate on all of these scenarios.

4.2.1 Czech and English Analysis

The analysis scenario is responsible for enriching the surface representation of a text with all necessary information that should be represented on the morphological, analytical and tectogrammatical layer. The principles of Prague tectogrammatics are not achieved entirely. On the one hand the processing pipeline ignores some annotation (e.g., topic-focus articulation), but on the other hand it adds the annotation of other phenomena (e.g., named entities, formemes, anaphoricity of the pronoun "*it*").

---

seen as so tied up with rules of grammar and syntax that annotation of such cases is too unattractive to deal with.

[10] The reasons for this requirement are listed in Section 7.1.1.

[11] Morphological layer does not have its own representation in Treex. Instead, morphological attributes are stored in an analytical node.

The analysis scenario consists of tokenization, morphological analysis, part-of-speech tagging and lemmatization, named entity recognition, dependency parsing, surface-to-tecto tree transformation, semantic role labeling, and some language-specific annotation. If necessary, the whole pipeline can be preceded with a sentence splitter, e.g. the one in Treex (`W2A::{CS,EN}::Segment`). The rest of this section gives more details on these individual steps of the Czech and English pipeline.[12]

Tokenization. Sentences are first split into tokens. This is ensured by tokenization blocks implemented in Treex (`W2A::{CS,EN}::Tokenize`). All these blocks are built on the same set of basic rules and extended with additional rules tailored to the orthography of a particular language.

Morphological analysis, part-of-speech tagging, lemmatization and dependency parsing. Subsequently, the tokens are enriched with morphological information including part-of-speech tags, morphological features as well as lemmas, and a dependency tree is build on top of this annotation. The output of this stage is a representation of the sentence on the analytical layer. We use different tools for different languages.

For English, The Morče tool by Spoustová et al. [2007] is being used for part-of-speech tagging and lemmas are collected using a rule-based lemmatizer (`W2A::EN::Lemmatize`). Even though a MorphoDiTa [Straková et al., 2014a] model for English is also available, Morče appeared to cooperate better with the rest of the pipeline, producing less translation errors in Czech-to-English translation with TectoMT. Finally, maximum spanning tree (MST) parser [McDonald et al., 2005] is utilized to construct dependency trees and its output is adjusted for the next processing by a sequence of several rule-based blocks.

For Czech, the MorphoDiTa tool is able to provide all the morphological information. Dependency parsing is provided by the MST parser adapted to Czech [Novák and Žabokrtský, 2007].

Named entity recognition. Named entity recognition is carried out by the NameTag [Straková et al., 2014b] tool for both languages.

Surface-to-tecto tree transformation. Both analytical and tectogrammatical trees are dependency trees. The tectogrammatical tree thus can be constructed by transforming the analytical tree and enriching it with some new information.

The tecto-transformation pipeline consists of several steps. These prevailingly rule-based steps aim at filling most of the tectogrammatical attributes as defined in Section 2.2. Here is a list of the most important steps with additional details for some of them:

- **Hiding function words.** Unlike in the analytical tree, only content words are represented as nodes in the tectogrammatical tree. Prepositions, auxiliary verbs, particles, some punctuation etc. need to be hidden. This also

[12]Similar or more simplified scenarios exist also for other languages, e.g. German, Dutch, Spanish, Portuguese and Basque [Dušek et al., 2015], and recently also for Russian and Polish. The pipelines for the latter two languages were employed in an automatic pre-annotation of the PAWS parallel treebank [Nedoluzhko et al., 2018].

holds for conjunctions, except for the coordinating ones, which often play a role of a coordination root.

- **Reconstructing nodes.** Some expressions elided on the surface should be reconstructed by its own tectogrammatical nodes. While in gold annotation of tectogrammatics newly established nodes can be either copied or generated, the automatic processing creates new nodes only in the latter way. As reconstructing elided nodes is a key factor for CR, more space is devoted to this topic further in this section.

- **Semantic role labeling.** Dependency relations between tree nodes should be assigned semantic roles. We utilize a system [Bojar et al., 2012] trained on Czech and English data, namely on PDT 2.0 and PCEDT 2.0, respectively.

- **T-lemma construction.** Tectogrammatical lemma is a generalized variant of the surface lemma. For instance, all central pronouns share the same t-lemma. Therefore, the rule-based block not only sets the t-lemma, but it also possibly stores discriminative features that would otherwise get lost as grammatemes.

- **Formeme construction.** *Formeme* is a morpho-syntactic tag that does not belong to the standard set of attributes prescribed by the Prague tectogrammatics. Assigned to every t-node, it shows how the t-node is realized on the surface (e.g. `n:subj`, `n:to+X` and `v:fin` for nodes representing a subject noun, a noun introduced by the preposition "*to*" and a finite verb, respectively).Although it is rather a surface concept, it sneaked into the tectogrammatical layer due to practical reasons: to enhance modeling of grammar transfer in machine translation with TectoMT [Dušek et al., 2012], and to simplify the search queries.

- **Grammatemes filling.** Grammatemes are set by a complex set of rules.

- **Valency frame linking.** Automatic linking of verbs to items in valency dictionary [Dušek et al., 2014] is in our case conducted only for Czech. Valency frames specify how a verb is connected with its arguments and modifiers in a particular sense. It can then be used to check if the tectogrammatical structure complies with the estimated valency frame.

- **Coreference resolution.** Some rule-based CR blocks are also included into the tecto-transformation, e.g. resolution of relative and reflexive pronouns. They are required for filling some of the grammatemes. We could possibly circumvent them by extracting the related parts of the grammateme filler to a separate block and shifting the block after the Treex CR system is applied. Nevertheless, due to high complexity of the grammateme filler, we decided to keep the rule-based CR blocks and rather delete all the coreference links before Treex CR is applied. Note that some traces of rule-based coreference still remain hidden in the grammatemes and Treex CR can take advantage of them.

Additional processing. The following additional modules augment the text with information that none of the layers in Prague tectogrammatics originally supports, but it is required for our coreference resolver.

The NADA tool [Bergsma and Yarowsky, 2011] is applied to help distinguish referential and non-referential occurrences of the English pronoun *it*. Every occurrence is assigned with a probability estimate based on n-gram features.

Lexical ontology is employed to collect possible senses for nouns. A noun is assigned all senses that are associated with its lemma, and this set is extended also with all hypernymous senses. Note that no word sense disambiguation is performed afterwards. WordNet [Fellbaum, 1998] and EuroWordNet [Vossen, 1998] databases are used as ontologies for English and Czech, respectively.

Ellipsis reconstruction. To mimic the style of the tectogrammatical annotation in automatic analysis some nodes that are not present on the surface must be reconstructed. Out of all types of zeros annotated in tectogrammatics (see Section 2.4.3), the automatic reconstruction mostly focuses on the cases that directly relate to coreference – anaphoric zeros. It is thus one of the key factors for coreference resolution.

As Czech is a pro-drop language, zero subjects in finite clauses (see Example 5.1 in Section 5.1) are the prevailing anaphoric zeros in Czech. Their restoration is therefore crucial for Czech CR. A subject is restored as a child of a finite verb if the verb has no children in subject position or in nominative case. Grammatical person, number and gender are inferred from the verb form.

Another type of Czech anaphoric zeros appears in control constructions (see Section 2.3). Control constructions arise with infinitives governed by a certain group of verbs, denoted as control verbs, e.g. "*začít* /to start/" and "*nechat* /to have sth done/". The subject of the infinitive, which is unexpressed, then corefers with one of the control verb's arguments. To reconstruct such subjects, we employ a rule-based method proposed by Nguy and Ševčíková [2011]. It also assigns a semantic role to the new zero. The role is determined by a finite verb that governs the infinitive.

Perhaps surprisingly, English uses zeros as well. The coreferential ones can be found in relative clauses with a zero relative pronoun (see Example 5.19 in Section 5.3) and *non-finite clauses*, e.g. in participles and infinitives including control constructions. We seek for all such constructions and add a zero child with a semantic role corresponding to the type of the construction.[13]

Table 4.5 shows the precision and recall scores of the zero reconstruction subtask. The scores are measured on two corpora that we use throughout this work: Prague Dependency Treebank (PDT) and Prague Czech-English Dependency Treebank (PCEDT); see Section 4.1 for details on the corpora. Regardless the recall differences between the corpora, all the scores move around 90%, except for the recall of Czech zeros in non-finite clauses. Despite achieving a high precision, the reconstruction method is unable to restore 30-50% of such zeros.

[13]In fact, the original Treex module for English zeros' generation treated only infinitives. This work extends it for present and past participles.

	Czech		English
	PDT	PCEDT	PCEDT
Zero subjects	86 / 93	87 / 84	—
Zeros in non-fin. cl.	99 / 50	100 / 68	90 / 93

Table 4.5: Success rates (precision in % / recall in %) of automatic zero reconstruction measured on Prague Dependency Treebank (PDT) and Prague Czech-English Dependency Treebank (PCEDT) as introduced in Section 4.1.

4.2.2 Monolingual Alignment

Monolingual alignment serves to find correspondences between the products of two separate analyses of a single sentence. Its most usual use case is to interlink the nodes in manually annotated tree structures with the nodes yielded by the automatic pre-processing pipeline. Interlinking manually and automatically annotated data has two main reasons. It allows for: (1) comparison of automatic annotation with its manual equivalent, and (2) transferring gold information about a particular phenomenon on top of the machine-produced annotations in order to train a ML model for this phenomenon. For instance, without monolingual alignment it would not be possible to calculate the scores of automatic zero reconstruction in Table 4.5.

The scenario for monolingual alignment is the same for Czech and English. It first attempts to align a-nodes corresponding to surface tokens. The links are then projected to the tectogrammatical layer and alignment for generated t-nodes is established. The task is trivial for a considerable proportion of node pairs – those, which correspond to the same token in the surface sentence. However, there are two main obstacles that may hinder the alignment from being found.

Firstly, two tokenizations of the same sentence may slightly differ. This would cause problems already in the alignment between the analytical trees. The solution is simple. At first, the shortest text span where the tokenizations differ must be found. Alignment links are then made as a Cartesian product of the tokens that belong to each of the tokenizations of the span.

Another problem naturally arises with alignment of generated t-nodes, i.e. t-nodes which have no corresponding surface token. A standard alignment link is created between two yet unaligned generated nodes, which share the same semantic role and their parents are aligned. Yet, many of the zeros in the manual as well as in the automatic annotation still remain unaligned. The reasons for missing alignments are closely related to the heuristic method of zero reconstruction (see Ellipsis Reconstruction in Section 4.2.1). They are twofold. First, incorrect automatic analysis causes a heuristic method to generate zero in an incorrect position. The second reason concerns the unclear boundary between deverbative adjectives and participles and is further discussed in Section 2.4.3. If the missing alignments are not fixed, projection of some gold coreference and alignment annotation to automatically analyzed trees would fail. As having such gold annotation in automatically analyzed trees is necessary for building supervised resolvers for these

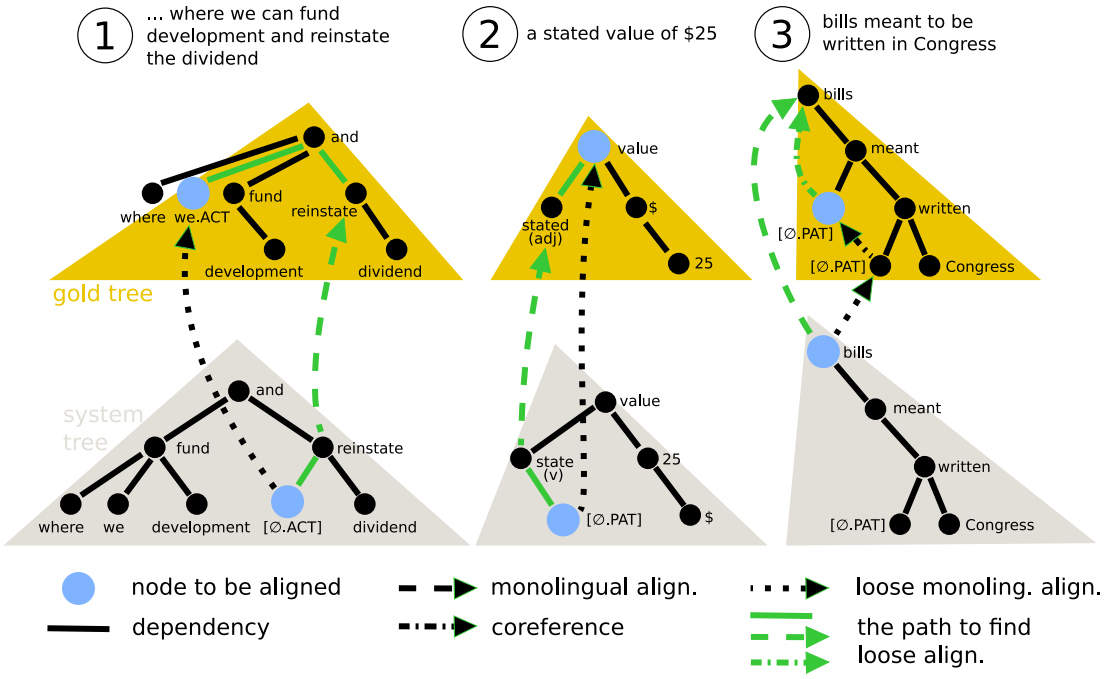

Figure 4.2: The three heuristic-based methods to align zeros with loose monolingual alignment.

phenomena, the resolvers would thus learn to find no coreference and alignment links in many cases.

Example 4.1 illustrates a typical case. Imagine this sentence is to be included to the data for training a CR system. The true subject of the verb "*se nezměnila* /hasn't changed/" is the word "*hodnota* /value/". However, due to incorrect syntactic parsing a new zero subject is generated. Both the true and the spurious subject are correctly labeled as feminine singular – the label of the zero subject is inferred from the form of the verb. The problem is that if we did not find a counterpart to the zero subject in the manually annotated tree, the zero would remain non-anaphoric for the training data. This definitely introduces undesirable noise, since a feminine singular zero subject is very unlikely to be non-anaphoric in Czech.

(4.1)

***Hodnota*$_{\text{subj.f.sg}}$**	*toho, co*	*jste vlastnili*	*a*	*měli rádi,*	~~∅$_{\text{subj.f.sg}}$~~	*se nezměnila.*
Value	of what	you owned	and	you liked	–	hasn't changed

The value hasn't changed if you owned it and liked it.

To alleviate these negative effects, the aligning procedure introduces *loose monolingual alignment*, which attempts to connect the zero with a node that plays the same role in the aligned tree. It may connect an unaligned generated node with an already aligned node. In Example 4.1, if we align the zero with the manually annotated counterpart of the word "*hodnota*", it will allow for introducing an artificial but formally correct anaphoric link between the two expressions to the training data. To find a loosely aligned monolingual counterpart, the three following heuristics (also illustrated in Figure 4.2) are run one after another.

1. The first heuristic aims at aligning spurious zeros, which were created as a consequence of erroneous syntactic analysis (as shown in Example 4.1). A loose alignment link is made from an unaligned automatically generated node to a node in the manually build tree, if the two nodes share both their governing nodes[14] and semantic roles. Furthermore, the two nodes must not belong to the same clause.

2. Heads of non-finite clauses may be deemed either as a non-finite form of a verb or a non-verbal part-of-speech (adjective or noun). Yet, zeros are usually not generated in the second case. Another heuristic thus aims at aligning zeros in a non-finite clause, if its counterpart is a non-verbal phrase. It creates a loose alignment from an automatically reconstructed zero to the parent of the manually annotated counterpart of the zero's parent.

3. The last heuristic exploits coreference and loosely interlink an unaligned zero with the aligned counterpart of its antecedent. This heuristic usually takes advantage of the manually annotation of coreference. However, in the case of extracting the data for training the supervised alignment (see Section 6.2.2), the output of a CR system is also exploited.

Note that due to the errors in system trees and the errors of the CR method sometimes required by the third heuristic, monolingual alignment may not be always correct.

4.2.3 Original Cross-lingual Alignment

The purpose of the cross-lingual alignment scenario is to find correspondences between words (and associated nodes) in two languages, Czech and English in our case. We denote it as the *original alignment*, as this is the approach that has been originally adopted for aligning two main parallel corpora that we employ, PCEDT and CzEng (see Section 4.1). Later in Chapter 6, we will introduce an upgrade to this method, which addresses coreferential expressions by means of supervised learning.

The alignment method here assumes the text is already aligned on the sentence level. It can be achieved by instructing the translators to produce sentences in 1:1 correspondence (PCEDT). Otherwise, the translations must be paired with a specialized tool (e.g., for CzEng).[15]

The original alignment interlinks not only surface tokens (represented by nodes on the analytical layer), but also tectogrammatical nodes. Having alignment directly between t-nodes is convenient, as all the cross-lingual CR approaches we will introduce operate on the t-layer. Apart from this technical advantage, we also see a more important, linguistic benefit. Tectogrammatical layer aims to represent a sentence in a way that is closer to its meaning than its surface representation. It is assumed that such representation should be also less language-dependent, thus making the representations of a sentence and its translation to

[14]Given a node, its governing node is usually its parent in the tree. Nevertheless, it can be a different node in some cases where dependencies do not correspond to edges in the tree, e.g. coordinations (see Figure 4.2).

[15]A commonly used tool for sentence alignment is Hunalign [Varga et al., 2005].

another language more similar. This property has been verified by Mareček et al. [2008]. They measured Inter-Annotator Agreement (IAA) on word alignment labeled by two annotators on 515 sentence pairs. Furthermore, they measured IAA on tectogrammatical representations of the sentences, where the alignment was obtained by projecting the manual word alignment. The study showed that whereas the annotators agreed on 82.1% links between words, it increased to 94.7% for tectogrammatical nodes (in terms of F-measure). The improvement can be justified by the fact that function words, which contribute the most on the disagreement rate, are almost never represented as tectogrammatical nodes.

The process of getting the original alignment proceeds in three stages: (1) unsupervised alignment of surface words, (2) transfer of the surface alignment to the tectogrammatical layer, and (3) rule-based alignment for reconstructed zeros.

Unsupervised word alignment. Majority of methods for alignment of surface tokens are based on unsupervised induction taking advantage of co-occurrence statistics. This is also the case of GIZA++ [Och and Ney, 2000], a tool that has become enormously popular especially in the community of statistical machine translation.[16] GIZA++ implements a cascade of IBM models (from 1 to 5) [Brown et al., 1993] where each model refines the translation probability distribution of the previous one. While in IBM Model 1 this distribution boils down to lexical translation probabilities, further models take word position and fertility (number of source language words aligned to a target language word) into account. Probability distributions are estimated by running Expectation-Maximization algorithm on the training data.

Alignment as produced by IBM models does not allow for one-to-many alignment links. Therefore, GIZA++ is usually run in both directions and the obtained alignment is symmetrized. In our work, we use two symmetrization strategies: *intersection* of both alignment directions and *grow-diag-final-and*, which is based on the intersection but extends it by the links from the union using heuristics (see [Och and Ney, 2003] for details).

GIZA++ in its basic version does not allow for applying the pre-trained alignment distribution on new data. If necessary, we thus replace it with MGIZA++ [Gao and Vogel, 2008], a multi-threaded version of GIZA++ that allows such usage. For instance, the alignment present in CzEng was acquired by running GIZA++ solely on this data. On the other hand, in automatic pre-processing of PCEDT we successfully utilize MGIZA++ and models trained on CzEng, as its 300-times greater size should increase reliability of collected alignment.

Transfer of surface alignment to the t-layer. In the present work, we employ a simple method designed by Mareček et al. [2008] to transfer alignment from the surface to the tectogrammatical layer. Only the links that align content words are projected.

Alignment of zeros. Tectogrammatical nodes also comprise reconstructed nodes, which have no surface counterpart, e.g. zero subjects. The method used for PCEDT 2.0 [Hajič et al., 2012] addresses such nodes in a way that a node that

[16] With the rise of neural machine translation, it is slowly being pushed to the fringe by the methods implementing *attention mechanism* [Bahdanau et al., 2014].

fulfills the following two conditions is labeled as a counterpart of the reconstructed node:

- Their parents must be aligned.
- Both nodes must stand in the same semantic role[17] with regards to their parents.

4.3 Coreference Systems to Compare

Experiments on monolingual coreference resolution in Section 7.1 contrast our coreference resolver with other CR systems for Czech and English. Let us briefly introduce these systems.

Since to the best of our knowledge there is currently no other publicly available system for Czech, we compare it with the set of CR modules in Treex originally used to annotate coreference in CzEng 1.0 [Bojar et al., 2012]. A similar set of modules was applied to acquire coreference annotation also on the English side of CzEng, so we can compare performance of Treex CR with its predecessor also on English. Section 4.3.1 gives a short description of this "CzEng coreference resolver".

We want to compare Treex CR with some of the best-performing systems for English. There are several third-party CR systems that are ready to use with not too much effort. Eventually, we opted for the Stanford CoreNLP toolkit that embraces even three coreference resolvers representing three different approaches. Each of them claimed to improve over the state of the art at the time of its release. Section 4.3.2 gives more details.

4.3.1 CzEng CR

This CR system consists of mostly rule-based Treex blocks that originally served to create the coreference annotation on both language sides of CzEng 1.0 [Bojar et al., 2012]. It is a direct predecessor of Treex CR and, therefore, it might have been denoted by the same name in some previous works (e.g. by Novák et al. [2015]). In this work we strictly distinguish the older CzEng CR and the newer Treex CR, which this thesis concentrates on.

The Czech set of blocks focuses on relative, reflexive, personal and possessive pronouns, and zero subjects. Basic reflexive, reflexive possessive and relative pronouns in Czech are addressed by rule-based methods similar to those presented in [Kučová et al., 2003] and [Nguy, 2006]. They exploit the dependency structure and semantic role annotation of t-trees. Personal pronouns, possessive pronouns and zero subjects (all in the 3rd or unknown person) are targeted with a reimplementation of Nguy et al. [2009]'s ranker. Originally in the CzEng 1.0 release, the model was trained by an averaged perceptron [Collins, 2002] adapted to the ranking scenario. Due to compatibility issues, we had to replace it with the implementation of a cost-sensitive one-against-all classifier within the Vowpal Wabbit toolkit.[18] Czech zeros in non-finite clauses have not been addressed

[17] Semantic roles are more appropriate for this purpose than surface dependency labels, as they should remain the same no matter if the clause is in active or passive voice.

[18] `https://github.com/JohnLangford/vowpal_wabbit/wiki`

by this approach. The pre-processing pipeline for creating CzEng does not even reconstruct such Czech zeros in t-trees.

The English set of blocks tries to resolve coreference for relative, reflexive, personal and possessive pronouns, and zeros in non-finite clauses. All the blocks are analogous to their Czech counterparts including the method for personal and possessive pronouns, which is the only one employing machine learning. The only exception is a rule-based block for zeros in non-finite clauses, which had no Czech analogy in Treex at the time when CzEng 1.0 was released. The inventory of cases where such a zero is reconstructed was also narrower in the pre-processing pipeline for CzEng 1.0, achieving only 34% in the reconstruction recall. Some of the English CR blocks were originally created to serve in the English-to-Czech translation by the TectoMT system [Žabokrtský et al., 2008].[19]

4.3.2 Stanford CR

Stanford's coreference resolvers are represented by three systems integrated in the Stanford CoreNLP toolkit [Manning et al., 2014]. Each system adopts a completely different approach: deterministic, statistical and neural. Nevertheless, all the approaches are implemented within the same schema of an entity-based model. Starting from individual mentions, coreference chains are being built up incrementally by agglomerative clustering performing one merge of two partially formed coreferential clusters at a time.

All the three approaches share the same deterministic mention detection algorithm by Lee et al. [2011]. Trying to achieve high recall, it identifies mentions as defined by OntoNotes annotation specifications [Hovy et al., 2006]. As mentioned in Section 4.1.4, annotation of coreference in OntoNotes differs from the annotation in Prague treebanks in two main aspects.

First, relative pronouns are not considered mentions and zeros are not annotated at all. The same thus holds for Stanford CR systems. At the same time, they can handle coreference of nominal groups and coreference of pronouns in first or second person, which has not been addressed by Treex CR so far.

Second, mentions in OntoNotes are represented as continuous spans of surface text, and the Stanford CR systems obey this. Consequently, the output of the systems is not compatible with our evaluation schema designed primarily for tectogrammatical trees (see Section 4.4.3). The surface mentions thus must be transformed to the tectogrammatical style of coreference annotation, i.e. mention heads connected by links. We may use the information on mention heads provided by the Stanford system itself. However, by using this approach we observed completely contradictory results on different datasets. Manual investigation on a sample of the data revealed that the Stanford system often identified a correct antecedent mention, but it selected a head different to the one in the data. Most of these cases, e.g. company names like "*McDonald's Corp.*" or "*Walt Disney Co.*", have no clear head, though. Therefore, we decided to use the gold tectogrammatical tree to identify the head of the mention labeled by the Stanford system. Even though employing gold information for system's decision is a bad practice, here it should not affect the result so much and we use it only for the

[19]For experiments on utilizing coreference resolution in the TectoMT system, see [Novák et al., 2015].

third-party systems, not for our Treex CR.

Deterministic. Raghunathan et al. [2010] (later extended by Lee et al. [2011]) developed an entity-based rule-based CR system. After mentions have been detected by a recall-oriented component, construction of coreference clusters proceeds in multiple precision-oriented steps, denoted as *sieves*, sorted from the highest to the lowest precision. In the experiments we use the version of the system by Lee et al. [2011], which won the CoNLL 2011 Shared Task [Pradhan et al., 2011]. It consists of twelve sieves including the sieve for pronominal mentions in quotations, sieves for exact and relaxed string match, head match, proper head noun match, and the pronoun match inter alia.

Statistical. Clark and Manning [2015] proposed a structured prediction approach that employs simple mention-based logistic regression models. They provide the information combined in cluster features that guide each merge operation of the agglomerative clustering. In their implementation, there are two models:

- *the mention-pair classifier* tries to identify all antecedents of an anaphor. It targets for instance nominal groups that often overlap, such as "*President Clinton*", "*the president*", and "*Bill Clinton*". In addition, it serves to prune the search space of possible merge operations;
- *the mention-ranking model* predicts a single antecedent. It targets expressions bound with a single antecedent, not the whole coreferential chain (e.g. pronouns).

To train the system, the authors applied imitation learning using the DAgger method [Ross et al., 2011].

Neural. The CR system presented by Clark and Manning [2016] adopted a cluster-ranking structured prediction approach, represented by a single feed-forward neural network that consists of four components:

- *the mention-pair encoder* produces a distributed representation of a mention pair exploiting word embedding features, position and distance features, speaker features, mention type and document genre;
- *the cluster-pair encoder* produces distributed representations for pairs of partially built cluster by applying the pooling operation[20] over a set of mention-pair representations. This set corresponds to a Cartesian product of the two clusters of mentions;
- *the cluster-ranking model* scores the cluster-pair representations;
- *the mention-ranking model* scores the mention-pair representations. It is used to initialize the cluster-ranking model. In addition, it prunes the list of candidates that the cluster-ranking model considers.

The whole model is trained by a learning-to-search algorithm inspired by SEARN [Daumé et al., 2009].

[20]For explanation of the terminology, please refer to a literature on deep learning, e.g. [Goodfellow et al., 2016].

4.4 Coreference Evaluation Measures

In order to quantify the performance of proposed approaches and to compare them, it is essential to have a method that can evaluate the quality of produced output. In this section, we introduce evaluation measures that we utilize for coreference resolution and word alignment. We discuss most relevant measures that have been used for these tasks and give reasons why we did not simply adopted any of them, but instead took inspiration for proposing our own measures. In the following, we will denote manual coreference links or chains as *key* and the ones produced by a system as *response*.

4.4.1 Standard Measures

After a turbulent period of multiple measures with different variants having been employed by authors in various combinations, the community managed to agree on a single evaluation schema introduced for the CoNLL 2011 Shared Task [Pradhan et al., 2011]. It applies five acknowledged metrics (MUC, B^3, $CEAF_m$, $CEAF_e$, and BLANC) and calculates an average of three of them (MUC, B^3, $CEAF_e$) as the main score, known as the CoNLL score [Pradhan et al., 2014]. All the participating metrics have in common that they view the evaluation as a clustering evaluation task, i.e. they measure how good is the matching between key and response entity clusters. Let us briefly overview the metrics.

- *MUC* [Vilain et al., 1995] is a link-based measure. Recall is complementary to the proportion of key links that must be inserted to the response so that all key links are covered. Precision is complementary to the proportion of response links that must be deleted so that the key partitioning is not violated.
- B^3 [Bagga and Baldwin, 1998] calculates precision and recall for each mention and then averages them over all key and response mentions, respectively. Recall of a mention expresses what proportion of mentions in the key chain corresponding to the mention belongs also to the response chain of the mention. To calculate precision, the key and response chains must be swapped.
- *CEAF* [Luo, 2005] aims at rectifying the issue that each chain may be used more than once in calculation of B^3 scores. It is ensured by finding an optimal bijection of key and response chains. Whereas $CEAF_m$ is the mention-based variant, $CEAF_e$ is entity-based.
- *BLANC* [Recasens and Hovy, 2011] is based on the Rand index used to measure similarity of two clusterings. Precision and recall are calculated separately for anaphoric and non-anaphoric links. The final scores are then a result of averaging a pair of precision/recall.

4.4.2 Addressing the Issues of Standard Measures

The standard coreference metrics all view coreference chains as unordered clusters of generic items [Chen and Ng, 2013]. Tuggener [2016] pointed out three main issues of such treatment:

- *Interpretability.* Even the measures which are easy to compute (e.g. MUC) are not so easy to interpret. Interpretation gets more difficult with more complex definitions (e.g. CEAF and BLANC). The icing on the cake is the final averaging of some of the measures to form the CoNLL score.

- *Informativeness.* The standard measures are linguistically agnostic, they do not distinguish between mention types. But various mention types differ considerably in their properties related to coreference, which may also affect how difficult it is to address them automatically (cf. relative pronouns and nominal groups). The standard measures thus cannot offer any fine-grained qualitative insight into the performance of a resolution method.

- *Differentiability.* Neither the final score nor an intermediate result of any of the measures gives information that could facilitate differentiate between two resolvers that perform with the same score.

Furthermore, the scores measured by the standard metrics usually do not correlate [Holen, 2013]. They behave as different dimensions of a single coreference score, and therefore most of the works on CR report their results in several of these metrics. Multi-dimensionality of the score thus hinders ranking of the CR systems by their quality. This is the reason why the CoNLL score was established.

Adjusted standard measures. To rectify some of the standard measures' issues, some authors decided to adjust them.

Chen and Ng [2013] attempted to tackle the informativeness issue. They proposed a unified schema under which they reformulated the MUC, B^3 and CEAF measure, while incorporating into the schema the parameters that can change the weight of different mention types.

It does not need to be necessarily mention types what is meant by informativeness. Zeldes and Simonson [2016] explored how a CR system performs for various sentence types (declarative, question, subjunctive, fragment etc.). To measure it, they proposed the *p-link* score. It extends the link-based MUC metric in a way that the credit/blame for a correct/incorrect link is shared between both entities participating in the link. P-link allowed the authors to evaluate only on specific portions of data.

Application-Related Coreference Scores. Instead of fixing the standard cluster-oriented measures, Tuggener [2014, 2016] proposed his mention-oriented score – the Application-Related Coreference Scores (ARCS) evaluation framework.

In ARCS, four scores are aggregated over all key and response anaphors:

- *True positive* (TP): both the key and the response are anaphoric and the response link is correct.

- *False positive* (FP): anaphoric in the response but not in the key.

- *False negative* (FN): anaphoric in the key but not in the response.

- *Wrong linkage* (WL): both the key and the response are anaphoric but the response link is incorrect.

Using these scores, precision (P), recall (R) and F-score (F) are calculated as follows:

$$P = \frac{TP}{TP + FP + WL} \quad R = \frac{TP}{TP + FN + WL} \quad F = \frac{2PR}{P + R}$$

Tuggener suggested three strategies of determining whether the response is correct. He considers each of the following strategies to be tailored to a specific higher-level application purpose. The response link is correct if its target is:

- *immediate antecedent*. This evaluation strategy should be chosen if the final application of coreference resolution is discourse modeling or event sequence modeling.
- *closest nominal antecedent*. This strategy is well suited for application, such as text summarization, and machine translation.
- *anchor mention*. It is the most representative surface mention within the entity. The author decided to pick the first mention of an entity in order to represent its anchor mention, as it is usually most informative. This evaluation strategy suits the applications such as sentiment analysis, and text mining.

ARCS deals with all issues that Tuggener highlighted. As the precision and recall are calculated from simple scores aggregated over anaphors, its interpretation seems to be straightforward: *How many of the key/response anaphors are resolved correctly?* for recall/precision, respectively. ARCS is informative, as it is decomposable on the level of mentions, so different mention types can be examined separately. Furthermore, information about the distribution over concrete types of errors are directly accessible from the aggregated scores. This decomposability also facilitates differentiability of the evaluation framework. Outputs yielded by multiple systems on the same dataset can thus be compared mention by mention and help to identify strong and weak points of the systems.

4.4.3 Prague Anaphora Score

Prague Anaphora Score is the evaluation framework for anaphora and coreference resolution that is used in experiments throughout this thesis. It was developed by the author of this thesis, based on the measures used for some previous CR experiments on Czech [Nguy et al., 2009, Novák and Žabokrtský, 2011, inter alia] and refined over years.

The most important requirement for the design of the metrics was that it should be able to score only a subset of mentions. The reasons for this requirement were twofold. First, even if it might seem violating common practices at first sight, this requirement was partially driven by the coreference resolver we built – Treex CR. Treex CR consists of multiple modules, each of which targets a specific anaphor type, e.g. personal pronouns, relative pronouns, anaphoric zeros (see Section 7.1.2). To tune our system, we needed to evaluate our modules separately on the anaphor types they target. Second, experiments in this thesis are mainly focused on pronouns and zeros. Especially for cross-lingual experiments we believed that pronouns and zeros are more engaging to research than nominal

groups. And we preferred the measure that does not penalize the CR systems for not discovering the relations which they do not even address.

Another requirement imposed on the measure is specific to the tectogrammatical representation. The measure should not be too strict as regards decisions made on spurious zeros, which are ellipses that should not have been reconstructed because they are in fact not anaphoric. Example 4.1 in Section 4.2.2 shows the typical example. The system's decision on the spurious Czech zero subject will be correct, if it labels it as non-anaphoric. However, the spurious zero should be deemed to be correct also in the case, if it is linked with the word "*hodnota* /value/" or any other mention coreferential with it. We allow such relaxed handling only for the nodes in the automatically pre-processed tree, which are connected with a gold tree by a loose monolingual alignment (see Section 4.2.2).

Even though it was designed completely independently of Tuggener's works, Prague anaphora score is in fact very similar to ARCS. Let us show you how it operates with an emphasis on the differences with ARCS.

Five scores are aggregated over all key and response anaphors – TP, FP, FN and WL, all known from ARCS, and:

- *Spurious zero positive* (SZP): the only key counterpart of the response mention is accessible by loose monolingual alignment and this counterpart belongs to the same entity as the response antecedent.

Using all these scores, precision (P), recall (R) and F-score (F) are calculated as follows:

$$P = \frac{TP + SZP}{TP + SZP + FP + WL} \quad R = \frac{TP}{TP + FN + WL} \quad F = \frac{2PR}{P + R}$$

Due to the relaxed handling of spurious zeros the precision and recall may be unusually calculated using different numbers in the numerators.

We follow the relaxed approach also in a strategy of determining whether the response is correct. We use none of the strategies proposed by Tuggener. Instead, we deem the response link to be correct if it targets:

- *any mention in the key chain* except for the anaphor.

Although we did not know Tuggener's strategies at the time when we designed ours, there are several reasons why adopting some of his strategies would be problematic. Due to reconstructed zeros, discrepancies between key and response mentions appear frequently. The immediate antecedent strategy would thus penalize an anaphor whose immediate antecedent is a zero which failed to be reconstructed by the automatic analysis, even if its response antecedent belongs to the same key chain as the anaphor. Since coreference of nominal groups is not addressed by our CR system, response chains are inevitably full of gaps. The key anchor (first) mention would thus often be inaccessible in a response chain. The closest nominal antecedent strategy thus remains the only one that seems to be fair in our experimental setting.

In order for the Prague anaphora score to work correctly,[21] both the key and response coreference chains should be formed in a sensible and expectable way.

[21] And it almost certainly holds for ARCS, too.

Figure 4.3: Contrasting the outputs of two CR systems using a visual diagnostics in the Prague anaphora score evaluation framework. The screenshot shows instances corresponding to seven anaphor candidates. Each instance consists of the (1) anaphor's identifier, (2) decision category, (3) surface English text and its Czech translation, (4) linearized tectogrammatical representation of the English text with the coreference highlighted for each of both systems, (5) confidence level of the system's decision (negative number means wrong decision), and (6) the extent of a decision change (by which the instances can be sorted). The colors to highlight annotation have the following meaning: (yellow) the anaphor, (inverted) a response mention candidate, (green) correctly resolved antecedent, (red) incorrect response antecedent, and (cyan) correct key antecedent.

It means that the chain contains very few antecedents which are not anaphors at the same time. The ideal chain would contain only one such antecedent – the first mention of the entity. However, note that any chain looking in a non-standard way can be rearranged to satisfy this criterion.

Diagnostics. Like ARCS, Prague anaphora score also fulfills the requirements of informativeness and differentiability. Its design allows for looking up easily a system's decision for any anaphor. Different systems may be contrasted by laying the lists of decisions on the same dataset side by side.

Moreover, the Prague anaphora score framework provides additional machinery that can visualize the output of a coreference resolver. As illustrated in Figure 4.3, the visual diagnostics tool also allows to contrast the outputs of two systems. Furthermore, it was developed to examine the results of cross-lingual experiments. Therefore, the tool displays translations of focused sentences if available. The visual output is completely in a text format, so it can be easily processed for some further analysis.

These diagnostic techniques will be extensively used e.g. in Section 7.4.

5. Analysis of the Parallel Data

In this chapter, we explore the cross-lingual counterparts of coreferential expressions between Czech and English. We turn our attention only to expressions that belong to the core of our cross-lingual research as specified in Section 2.4 and analyze what their counterparts are in the other language or why they are missing. The presented analysis is based on the study co-authored by the author of this thesis [Novák and Nedoluzhko, 2015].[1]

The linguistic analysis is carried out on the gold trees in the PAWS section of the PCEDT treebank (see Section 4.1.2). The main advantages of this dataset are: (1) its annotation is based on Prague tectogrammatics and its tectogrammatical layer including the coreferential expressions is constructed completely by human annotators, and (2) in addition to the original automatic Czech-English word alignment contained in the dataset, we had provided it with manual annotation of alignment on selected coreferential expressions (see Section 6.1 for details on manual alignment). Most of all, the manual annotation of the t-layer ensures that the core expression categories, which have been specified without using coreference information (see Section 2.4), actually cover all the truly coreferential mentions that we want to focus, especially the anaphoric zeros (see Section 2.4.3). At the same time, we managed not to include too many non-anaphoric expressions, i.e. the precision of covering truly coreferential mentions is kept high, 92% and 89% for Czech and English, respectively. The only outlier is the category of English relative pronouns which inevitably includes many non-anaphoric interrogative and fused pronouns (see Section 2.4.2), achieving the precision 65%.

The gold trees and manual alignment allows us to conduct a proper and relatively accurate linguistic analysis of cross-lingual mappings with no or just a small risk of introducing noise by automatic methods.[2] On the other hand, the factor of manual alignment limits the size of the corpus to only around 1,000 sentences annotated with it. Another two disadvantages of the dataset are that it consists of a single domain of Wall Street Journal texts and that it comprises only texts translated in a direction from English to Czech. Both aspects may decrease the reliability of the final statistics that we collect.

In the analysis, we articulate differences in mapping mainly with respect to morphology and (deep) syntax. Such information may be possibly revealed by the automatic pre-processing pipeline (see Section 4.2.1) and then exploited by our coreference resolution.

In the following sections, we collect the frequencies of correspondences and show them in tables gradually for all the three big groups encompassing the core expressions: central pronouns, relative pronouns, and anaphoric zeros. The most frequent or interesting mappings are accompanied with examples extracted directly from the dataset. Note that the seemingly confused ordering of the following sections has been chosen on purpose in order to always start with the

[1] As we mention in Section 6.1, the annotation work of manual alignment was carried out by both co-authors. The analysis itself and the sections in the paper related to it are prevailingly authored by the author of this thesis.

[2] The charts in Figures 9.3 and 9.4 in Attachment 9 contrast for the PAWS and PCEDT datasets the distributions of potentially coreferential expressions and distributions of their counterparts.

EN\CS	Aligned								Not aligned		Total
	pers	zero	poss	refl poss	refl	demon	noun	other	noword	reword	
pers	34	135	2			1	7	2		6	187
"*it*"	15	55	1		1	20	11	5	29	10	147
poss	2	1	94	80	2		6	1	46	4	236
refl					3			8			11
Total	51	191	97	80	6	21	24	16	75	20	581

Table 5.1: Statistics on the correspondence of English central pronouns to their Czech counterparts. The last two Czech categories indicate the reason why there is no corresponding word in Czech for an English pronoun. The abbreviated names stand for the following: personal except for the pronoun "*it*" (pers), possessive (poss), reflexive (refl), reflexive possessive (refl poss), and demonstrative (demon) pronouns, missing Czech counterpart with (reword) or without (noword) a substantial clause rewording.

language, for which the particular category has a wider variety of counterparts.

In addition to the statistics on gold but small data, Attachment A provides analogous statistics visualized in bar charts, this time calculated on all parallel datasets described in Section 4.1. The statistics are collected also on the CzEng 1.0 corpus and its four selected domains. It thus offers interesting comparison of distributions of expressions and their counterparts across domains.

5.1 English Central Pronouns

Table 5.1 shows how frequently English central pronouns, particularly the personal, possessive, and reflexive pronouns, form alignment pairs with Czech nouns, anaphoric zeros, personal, possessive, reflexive possessive, reflexive, or demonstrative pronouns. For cases where the English central pronoun had no Czech counterpart, Table 5.1 also indicates if most of the other words in the clause containing the pronoun have their Czech counterparts, or the clause is substantially reworded in Czech. The pronoun "*it*" forms a separate category in the table due to its nature differing considerably from other personal pronouns.

Personal pronouns. As for English personal pronouns, most of them (57%) turn into Czech anaphoric zeros, as in Example 5.1 (99% of these cases occur in the subject position). Translations to Czech personal pronouns expressed on the surface account only for 15%. Even though these pronouns are mainly in non-subject positions, still over 35% of them are subjects. These are expressed in Czech mostly due to topic–focus articulation reasons or because they are coordinated.

(5.1) ∅ *zanechal* *zprávu*
He left a message
He left a message accusing Mr. Darman of selling out.
Zanechal mu zprávu, ve které viní Darmana ze zaprodanosti.

Pronoun "*it*". In Section 2.4.1, we distinguished three coarse-grained functions of the pronoun "*it*": (1) referring to an entity, (2) referring to an event, (3) non-referential (pleonastic). At the same time, we observe four prevailing ways of how "*it*" can be translated to Czech: (a) zero subject, (b) personal pronoun, (c) demonstrative pronoun, and (d) no lexical counterpart. Let us demonstrate what is the usual correspondence between these uses in English and Czech.

The cases when either Czech zero subjects or personal pronouns are aligned with "*it*" account for more than 50% of all occurrences of "*it*". And in these cases each instance of the pronoun "*it*" almost exclusively refers to an entity. Clear correspondence can be also found in such Czech translated sentences that have a slightly different syntactic structure than its English source, resulting in no Czech lexical counterpart of the English "*it*" (Example 5.2). In these cases, the pronoun "*it*" is often pleonastic.

(5.2) – *Nebylo známo, do jaké míry bylo zařízení poškozeno.*
It wasn't known to what extent was the facility damaged.
It wasn't known to what extent, if any, the facility was damaged.
Nebylo známo, do jaké míry, a jestli vůbec, bylo zařízení poškozeno.

Nevertheless, Czech demonstrative pronouns most often represented by the pronoun "*to* /it, this/"[3] can be aligned with any of the listed uses of the pronoun "*it*". Example 5.3 illustrates the case where the pronoun "*it*" was annotated as referring to an entity represented by the noun group governed by the verb "*strategy*". The pronoun "*to*" is commonly used in such cases especially when the referred entity is attributed some further characteristics, mostly in constructions with a verb "*to be*" like "*It is something.*"[4]

(5.3) *Ta přijala* ***strategii*** *v domnění, že je* ***to*** *cesta k vítězství.*
It endorsed **the strategy** [in the belief that] [is] **it** the way to victory.
It endorsed the White House strategy, believing **it** to be the surest way to victory.
Ta přijala strategii Bílého domu v domnění, že je **to** nejjistější cesta k vítězství.

However, without the context of surrounding sentences, the same sentence might also serve as an example of the Czech demonstrative pronoun "*to*" aligned to an event-referring "*it*" (the antecedent would be the clause governed by the verb "*endorsed*").

Cleft sentences (Example 5.4) and some other syntactic constructions are the case when the demonstrative "*to*" appears as a translation of a pleonastic "*it*". In some cases, both translations of pleonastic "*it*" are possible: neuter demonstrative "*to*" or a different syntactic construction with no lexical counterpart of "*it*". Compare the examples where "*it*" with similar syntactic function was translated by changing the syntactic structure in (5.5) and using a neuter "*to*" in (5.6):

(5.4) *je* ***to*** *Lane, kdo je posedlý*
is **it** Mr. Lane, who has been obsessed
But **it** is Mr. Lane, as movie director, who has been obsessed with refitting Chaplin's Little Tramp in a contemporary way.
Ale je **to** Lane jako filmový režisér, kdo je posedlý tím, že zmodernizuje Chaplinův film "Little Tramp (Malý tulák)".

[3]The Czech pronoun "*to*" is a form of a demonstrative pronoun "*ten*" in its neuter singular form.

[4]In fact, even "*he*"/"*she*"/"*they*" can be translated to this Czech demonstrative pronoun in such contexts.

(5.5) – *Bylo skvělé, že jsme měli dostatek času*
It was great [that we had] the luxury of time
"**It** was great to have the luxury of time," Mr. Rawls said.
"Bylo skvělé, že jsme měli dostatek času," řekl Rawls.

(5.6) ***to*** *vypadá zvrhle, když si vyhradím čas na masáž.*
it seems decadent [if] [I reserve to myself] time for a massage.
"On days that I'm really busy," says Ms. Foster, "**it** seems decadent to take time off for a massage."
"Ve dnech, kdy mám opravdu mnoho práce," říká paní Fosterová, "**to** vypadá zvrhle, když si vyhradím čas na masáž."

Possessive pronouns. Unlike personal pronouns, possessive pronouns often remain in the same class when translated to Czech. In 40% cases they are translated as possessive pronouns, in almost 35% they become the Czech reflexive possessive "*svůj*", a pronoun that shares some features with reflexive pronouns and substitutes Czech possessive pronouns in some positions when referring to the subject.[5] This category is missing in English, the pronoun "*svůj*" being translated to English with possessive pronouns "*his*", "*her*", "*my*", "*your*" (Example 5.7).

(5.7) ***svůj*** *podtitul* ***kniha*$_{\textbf{subj}}$** *dostatečně ospravedlňuje*
its subtitle **the book** amply justifies
While the book amply justifies **its** subtitle, the title itself is dubious.
Zatímco **svůj** podtitul kniha dostatečně ospravedlňuje, samotný název je zavádějící.

Interestingly, a substantial amount of possessive pronouns (20%) disappear in Czech (Example 5.8). The relation of possession is then understood intuitively from the context and as in case of reflexive possessive pronouns, it relates mostly to the subject of the sentence (37 out of the 46 instances).

(5.8) *Důsledkem — nemoci*
As a result of **their** illness
As a result of **their** illness, they lost $1.8 million in wages and earnings.
Důsledkem nemoci, přišli na mzdách a výdělcích o 1.8 milionu dolarů.

Besides, we found a few interesting cases where the benefactor entity of the predicate and the possessor entity of the direct object are identical (in Example 5.9, such entity is represented by the word "*residents*" and its Czech counterpart). Then, it is sufficient for a language to express only one of these positions explicitly. For instance, in Example 5.9, only the pronoun "*their*", which is a possessor of the direct object, is expressed in English. At the same time, only the reflexive pronoun "*si*", which fills the role of a benefactor of the governing predicate, is expressed in Czech. Consequently, there is a clear cross-lingual correspondence between the two.

(5.9) ***Obyvatelé*** *města* ***si*** *razili — cestu*
Residents [of the city] [to themselves] picked **their** way
Residents picked **their** way through glass-strewn streets.
Obyvatelé města **si** razili cestu ulicemi zasypanými sklem.

[5]The fact that their antecedent is usually the subject of the same sentence is the main reason why we divide them into a specific subcategory. The rules of use for the reflexive possessive "*svůj*" in Czech have been addressed in multiple linguistic studies [Daneš and Hausenblas, 1962, Pitha, 1992, inter alia].

CS\EN	Aligned						Not aligned	Total
	pers	poss	refl	"*the*"	zero	other		
pers	49	2			7	2	4	64
poss	3	94		3		4	3	107
refl poss		80		3		3	4	90
refl	1	2	3		1	4	14	25
Total	53	178	3	6	8	13	25	286

Table 5.2: The statistics on the correspondence of Czech central pronouns to their English counterparts. The abbreviated names are explained in the caption of Table 5.1.

Reflexive pronouns. As discussed in Section 2.4.1, English reflexive pronouns have two distinct uses: (1) basic and (2) emphatic. This distinction shows up nicely if moving to Czech: counterparts of basic reflexives are reflexive pronouns, but emphatic reflexives are usually expressed by different means in Czech, e.g., by the pronoun "*sám* /alone, of one's own/" or the adjective "*samotný* /alone/" (Example 5.10). In PCEDT, emphatic uses of English reflexive pronouns are annotated with coreference (the expression "*Mr. Bronner*" is the pronoun's antecedent) but the Czech expressions "*sám/samotný*" are not. Translation of the English reflexive pronoun may end up with both Czech reflexive pronoun and the pronoun "*sám*" occurring in the Czech sentence.

(5.10) *Jak říká* ***sám*** *pan Bronner*
As says **himself** **Mr. Bronner**
As Mr. Bronner **himself** says, the smell of "raw meat" was in the air.
Jak říká **sám** pan Bronner, ve vzduchu byl cítit zápach "syrového masa".

5.2 Czech Central Pronouns

The statistics of Czech central pronouns, namely the personal, possessive, reflexive possessive, and reflexive pronouns and their English counterparts are illustrated in Table 5.2. The most important counterpart categories are English personal, possessive, and reflexive pronouns, definite article "*the*", and anaphoric zeros.

English counterparts of Czech central pronouns are not as diverse as the counterparts of English central pronouns. The majority of personal and possessive pronouns remain in the same category and the reflexive possessive "*svůj*", which does not exist in English, is, not surprisingly, most often translated as a possessive pronoun (see Section 5.1).

Personal pronouns. While translation of personal pronouns to zeros is common in English-to-Czech direction, one expects it to be less frequent in the opposite direction. The collected data support this expectation, as we have found 10% of

such cases. A closer look at the individual examples reveals that Czech personal pronouns are realized as zeros in English mostly in the case of infinite clauses, where the argument occupied by the personal pronoun in Czech does not have to (or must not) be expressed in English (see the unexpressed argument of addressee for the verb "*ordered*" in Example 5.11).

(5.11) *dopisy* ∅ *napíše* *tak, jak* ***mu*** *bylo* *nařízeno.*
the letters he would write [in the way how] [to him] [it was] ordered.
Rep. Bates said he would write the letters as [[$\varnothing_{ACT}$] ordered [$\varnothing_{PAT}$] [$\underline{\varnothing}_{\boldsymbol{ADDR}}$]].
Poslanec Bates prohlásil, že dopisy napíše tak, jak **mu** bylo nařízeno.

Possessive and reflexive possessive pronouns. Czech possessive pronouns mostly translate as English possessives (94 of 107 instances). Among the cases where the translation is different, their co-occurrence with the definite article is especially interesting. Unlike in English, there is no grammatical category of definiteness in Czech. Determination in Czech is expressed by other means, e.g., demonstrative pronouns, intonation and word order. As we can see from our data, in a few instances, the Czech possessive and reflexive possessive pronouns are introduced for this purpose (Example 5.12). Whereas the Czech pronoun is coreferential with the nominal group governed by the word "*maloobchodník*", no coreference is annotated for definite articles.

(5.12) *Tento* ***maloobchodník*** *nebyl schopen* *najít* *pro* ***svoji*** *budovu* *kupce*
[This] retailer was unable to find for [his] building a buyer
The retailer was unable to find a buyer for **the** building.
Tento maloobchodník nebyl schopen najít pro **svoji** budovu kupce.

Basic reflexive pronouns. The majority of Czech basic reflexive pronouns remain unaligned. In 10 out of 14 such cases, the pronoun carries the semantic role of a benefactor or an addressee. In some of these cases, its missing counterpart can be attributed to the phenomenon shown in Example 5.9. While in Example 5.9, the English possessive pronoun is replaced by a Czech personal or reflexive pronoun in dative with the semantic role of a benefactor, in Example 5.13, the Czech sentence contains a reflexive pronoun occupying the benefactor role as well as a reflexive possessive pronoun, both referring to the same entity. Then, having aligned the possessive pronouns together, there is no node left to be aligned to the Czech reflexive pronoun. In such cases, Czech tends to be more pleonastic than English.

(5.13) ***reformátoři*** ***si*** *mohou* *ve* *své* *zemi* *připomenout* *ideály*
reformers [to themselves] can in their country recall the ideals
Czech reformers can recall the Wilsonian ideals of the same period in their country.
Čeští reformátoři **si** ve své zemi mohou ze stejné doby připomenout Wilsonovy ideály.

Finally, a Czech reflexive marker usually used in its reciprocal function (see Section 2.4.1) can be a part of some longer phrase which is translated into English by a completely different expression, e.g., "*po sobě (jdoucí)*" (lit. *going after one another*) and "*proti sobě (jdoucí)*" (lit. *going against each other*) to "*consecutive*" and "*contradictory*", respectively (see Example 5.14).

CS\EN	Aligned						Not aligned			Total
	"*that*"	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
"*což*"		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

Table 5.3: The statistics on the correspondence of Czech relative pronouns to their English counterparts. The last three English categories indicate the reason why there is no corresponding word in English for a Czech pronoun. The abbreviated names stand for wh-words used in relative clauses (wh-word relat), wh-words used in fused relative or interrogative constructions (wh-word inter & fused), roots of appositive constructions (appos), personal pronouns (pers), modifiers of a nominal group (NP modif), and verb phrase modifiers (VP modif).

(5.14) *míra* *chudoby* *klesala* *pátý* *po* ***sobě*** *jdoucí* *rok*
rate [of] poverty [declined] the fifth [going after one another] [year]

Last year's figure was down from 13.4% in 1987 and marked the fifth **consecutive** annual decline in the poverty rate.

Loňská hodnota klesla z 13.4% z roku 1987 a ukázala, že — míra chudoby klesala pátý **po sobě jdoucí** rok.

5.3 Czech Relative Pronouns

As for the relative pronouns, we start with the Czech ones since their English counterparts are more diverse. Table 5.3 gives a picture of how Czech relative pronouns and relative determiners are represented in English. Czech relative pronouns map to the English pronoun "*that*", wh-words used in relative clauses, wh-words used in fused relative or interrogative constructions, zeros, roots of appositive constructions, and (rarely) to personal pronouns. Some Czech relative pronouns have no English counterpart: most frequently relative clauses introduced by Czech relative pronouns are replaced with modifiers of a nominal group or with verb phrase modifiers.

As the anaphoric functions of Czech relative pronoun "*což*" differ from other relative pronouns (as discussed in Section 2.4.2, "*což*" can refer both to nominal groups and sentences), we cover it separately from the rest.

The relative pronoun *což*. The expression "*což*" is a specific relative pronoun frequently used in Czech to refer to a clause or a longer utterance. The wh-words aligned with it are exclusively instances of the pronoun "*which*", commonly used as an introducing element of the so-called *sentential relative clauses* [Quirk et al., 1985, p. 1118]. However, more often (44% cases) an apposition is used instead, as in Example 5.15 and Figure 5.1.

(5.15) *akcie* ***uzavřely*** *včera* *na* *28.75 dolaru* ***což*** *je* *pokles*
The stock closed yesterday at $28.75 **[which]** [is] [a decrease]

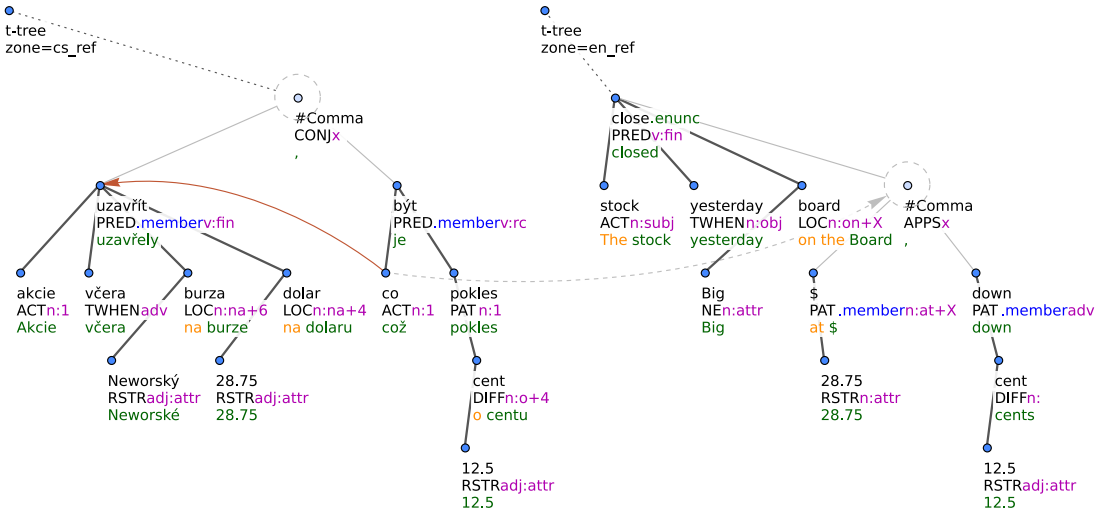

Figure 5.1: A tectogrammatical representation of the sentence pair from Example 5.15, where Czech "*což*" turns into an English root of apposition. The alignment is denoted by a dashed arrow. The solid arrow identifies the grammatical coreference.

The stock closed yesterday on the Big Board at $28.75**,** down 12.5 cents.

Akcie včera uzavřely na Neworské burze na 28.75 dolaru, **což** je pokles o 12.5 centu.

Another way of translating the relative "*což*" referring to a clause is using a non-finite or verbless clause [Quirk et al., 1985, p. 992–997] (Example 5.16).

(5.16) *Whitbread dala na prodej svoji divizi* ***čímž*** *rozpoutala boj*
Whitbread put up for sale its division [by which] it set off a scramble

Whitbread of Britain put its spirits division up for sale, setting off a scramble among distillers.

Společnost Whitbread z Británie dala na prodej svoji divizi lihovin, **čímž** rozpoutala boj mezi lihovary.

The relative pronoun "*což*" may also refer to nominal groups. This occurred in two cases in our data (see Example 5.17), where the relative clause introduced by this pronoun translates as a verbless clause postmodifying a nominal group.

(5.17) ***zvýšení*** *o 11.5%* ***<u>což</u>*** *je méně než doporučoval úředník*
increase [by] 11.5% [which] [is] lower than recommended an officer

The commission authorized an 11.5% rate increase at Tucson, lower than recommended by an officer.

Komise schválila společnosti Tucson zvýšení sazby o 11.5%, **což** je méně, než doporučoval úředník.

Other relative pronouns. Other Czech relative pronouns are used mainly within *adnominal relative clauses*, i.e., clauses post-modifying a nominal group. In 50% cases, the English counterpart is a relative pronoun (see Example 5.18).

(5.18) *mohou se objevit* ***síly,*** ***<u>které</u>*** *tento scénář pozdrží.*
may [appear] **forces** **<u>that</u>** this scenario would delay.

There may be forces **that** would delay this scenario.

Mohou se objevit síly, **které** tento scénář pozdrží.

Over 23% of the instances are translated to an anaphoric zero. The reason why this happens is twofold: Czech relative clauses introduced by a pronoun are replaced either with English relative clauses using a zero relative pronoun (Example 5.19), or with a non-finite clause, specifically with to–infinitive, "*-ing*" or "*-ed*" participles (see Example 5.20). In both cases, the PCEDT t-layer representation of the subordinate clause contains an anaphoric zero node coreferring with the modified noun.

(5.19) *To je* ***otázka*** ***na níž*** *nemůže Východní Německo odpovědět snadno.*
That's a question [which] can't East Germany answer easily.
That's [**a question** [[East Germany] can't answer [$\varnothing_{PAT}$] [easily]]].
To je otázka, **na níž** nemůže Východní Německo odpovědět snadno.

(5.20) ***zprávu*** ***ve které*** *viní Darmana*
a message [in which] [he accuses] Mr. Darman
He left [**a message** [[$\varnothing_{ACT}$] accusing [Mr. Darman] [of selling out]]].
Zanechal mu zprávu, **ve které** viní Darmana ze zaprodanosti.

In over 18% cases, an English counterpart could not be found. In the majority of these cases, the relative clause is transformed into a form not using a verb, thus not having a zero argument on the t-layer that could be aligned with the pronoun. These forms include premodifiers (adjectives, nouns, participles treated as adjectives) as in Example 5.21, prepositional post-modifiers and post-modifiers using a verbless clause[6] as in Example 5.22.

(5.21) *stádia* ***kdy*** *fakta se zjišťují.*
stage [when] [facts are being found]
The two that remain haven't yet reached the fact-finding stage.
Dvě zbývající dosud nedosáhly stádia, **kdy** se před líčením zjišťují fakta.

(5.22) *Dovoz* ***který*** *tehdy činil šest milionů barelů*
Imports [which] then [was] six million barrels
Imports, then six million barrels a day, came from Canada.
Dovoz, **který** tehdy činil šest milionů barelů denně, přicházel z Kanady.

We have not yet mentioned a special subclass of Czech relative pronouns which maps to the English pronouns introducing *interrogative* (see Example 5.23) and *fused (nominal) relative clauses* (Example 5.24).

(5.23) *Nebylo jasné* ***kdy*** *se obnoví tempo*
It wasn't clear **when** will resume the pace
It wasn't clear **when** the normal 750-car-a-day pace will resume.
Nebylo jasné, **kdy** se znovu obnoví normální tempo 750 vozů za den.

(5.24) *je hodně práce třeba udělat na tom,* ***co*** *máme.*
There is plenty of work [needed] to be done on [that, what] we have.
There is plenty of work to be done on **what** we have.
Na tom, **co** máme, je třeba udělat hodně práce.

[6]The post-modifiers using a verbless clause are in fact equivalent to apposition of nominal groups. Nevertheless, the PCEDT annotators decided not to represent these cases as apposition, producing a structure missing a an apposition root node that would otherwise become the alignment counterpart of the Czech relative pronoun.

EN\CS	Aligned				Not aligned	Total
	"*což*"	other relat	conj	other		
"*that*"		49		1	6	56
wh-words relat	7	102	2		7	118
wh-words inter & fused		23		14	6	43
wh-words conj			16		1	17
Total	7	174	18	15	20	234

Table 5.4: The statistics on the correspondence of English relative pronouns to their Czech counterparts. The abbreviated names are partly explained in the caption of Table 5.3, the rest stand for wh-words used as conjunctions (wh-words conj), Czech relative pronouns other than "*což*" (other relat), and conjunctions (conj).

While the pronoun in the former example does not have any antecedent, the pronoun in the latter is fused with its antecedent. However, it is often very difficult to distinguish which of the two categories a particular occurrence belongs to. All in all, from the computational point of view it is more important to find reliable formal differences between these two categories and the "real" relative pronouns in order to avoid looking for their antecedents in the task of coreference resolution.

5.4 English Relative Pronouns

In Table 5.4, we show the statistics of English relative pronouns, consisting of the pronoun "*that*" and wh-words used in adnominal and sentential relative clauses, interrogative and fused clauses, and as a conjunction. Their Czech counterparts have been categorized into four main classes: the Czech relative pronoun "*což*", other relative pronouns, conjunctions, and other expressions.

About 68% of all instances of English relative pronouns can be attributed to alignments between similar categories of true relative pronouns, i.e. the pronoun "*that*"[7] and relative wh-words on the English side, and the pronoun "*což*" and other relative pronouns on the Czech side (see Example 5.18).

The majority of wh-words that appear in interrogative or fused relative constructions turn into relative pronouns other than "*což*" on the Czech side. Over 43% of them are expressed using a so-called *correlative pair*, which in our case consists of a demonstrative pronoun and the following relative pronoun introducing a dependent clause. The antecedent of the relative pronoun is the demonstrative

[7]One would expect the numbers of English "*that*" translated to other relative pronouns in Table 5.4 and of the same case in the opposite direction in Table 5.3 to be the same. The disproportion (49 vs. 51 instances) came up due to incorrect part-of-speech tags assigned to two instances of "*that*", which prevented the automatic selection method described in Section 2.4.2 to include these examples.

EN\CS	Aligned				Not aligned	Total
	zero	relat	pers	other		
zero	263	75	7	28	329	702

Table 5.5: The statistics on the correspondence of English anaphoric zeros to their Czech counterparts. The abbreviated names stand for relative (relat) and personal (pers) pronouns.

pronoun itself, added to the sentence only for syntactic and stylistic reasons (see Example 5.24). The 13 occurrences of interrogative or fused pronouns not aligned to a Czech relative pronoun mostly contain the instances of wh-adverbs "*why*" and "*how*". While for English we included them in the class of relative pronouns, their Czech translations "*proč*" and "*jak*", which are never anaphoric in PCEDT, did not meet the specification of the class introduced in Section 2.4.2.

We also spotted 17 occurrences of wh-words, consisting solely of the adverbs "*when*" and "*where*" used as a subordinating conjunction (see Example 5.25). Since this class is irrelevant for the task of coreference resolution, they should be excluded from the set of English relative pronouns. However, to identify them we would have to include more syntax-based constraints into the specification of the class in Section 2.4.2.

(5.25) *V roce 1956* ***když*** *Británie Francie a Izrael napadly Egypt*
In [the year] 1956 **when** Britain France and Israel invaded Egypt
In 1956, **when** Britain, France and Israel invaded Egypt, Arab producers cut off supplies to Europe.
V roce 1956, **když** Británie, Francie a Izrael napadly Egypt, zastavili arabští výrobci dodávky do Evropy.

To sum up, let us recall the final remark from Section 2.4.2 on the precision of the criteria specifying the category of relative pronouns: 35% of the selected nodes are in fact non-anaphoric. Nonetheless, a deeper investigation summarized in Table 5.4 discloses that 72% of them are in fact used in interrogative and fused relative constructions or as a conjunction. The rest might be attributed to some special cases and annotation errors.

5.5 English Anaphoric Zeros

Unlike it was specified in Section 2.4.3, at the time of collecting these statistics we decided not to distinguish the two subtypes of the zeros that we target. Table 5.5 thus gives an overview of how all English nodes selected by our criteria on potentially anaphoric zeros map to their Czech counterparts.

Unsurprisingly, the most frequent aligned counterparts for anaphoric zeros in English are Czech anaphoric zeros. In most cases, missing valency arguments of a verbal predicate are aligned, cf. the unexpressed actor of the verbs "*do*" and "*ride*" in Example 5.26.

CS\EN	Aligned					Not aligned	Total
	zero	pers	pers 1st & 2nd	poss	other		
zero	263	190	40	1	84	278	856

Table 5.6: The statistics on the correspondence of Czech anaphoric zeros to their English counterparts. The abbreviated names stand for personal pronouns in the 3rd (pers), first and 2nd person (pers 1st & 2nd), and possessive pronouns (poss).

(5.26) *Jejich reakcí bylo* $\underline{\varnothing}_{\mathbf{ACT}}$ *nedělat nic a* $\underline{\varnothing}_{\mathbf{ACT}}$ *nechat to odeznít.*
Their reaction was $\underline{\varnothing}_{\boldsymbol{ACT}}$ to do nothing and $\underline{\varnothing}_{\boldsymbol{ACT}}$ ride it out.
Their reaction was to do nothing and ride it out.
Jejich reakcí bylo nedělat nic a nechat to odeznít.

About 10% of English anaphoric zeros correspond to Czech relative pronouns. These cases represent relative clauses with a zero relative pronoun or non-finite clauses in English (see the description in Section 5.3 and Examples 5.19 and 5.20).

Almost 50% of anaphoric zeros in English have no Czech counterparts. The most frequent reasons for such an absence are either substantial rewording in the translation, or the absence of corresponding verbal arguments from the t-layer annotation of Czech. Some of these unaligned cases have more or less technical reasons. For example, the verb "*chtít*" ("*want*") is considered to be modal in Czech, so it does not have its own node in the tectogrammatical representation. In English, the verb "*want*" is represented in t-layer as a separate node, so its arguments are reconstructed, but cannot have Czech counterparts (see Example 5.27).

(5.27) *"Já chci — vydávat takový který uspěje*
"I want $\underline{\varnothing}_{\boldsymbol{ACT}}$ to publish one that succeeds
"I want to publish one that succeeds," said Mr. Lang.
"Já chci vydávat takový, který uspěje," řekl Lang.

5.6 Czech Anaphoric Zeros

Table 5.6 shows a statistic of alignment counterparts for Czech anaphoric zeros. The cases where Czech zeros correlate to English anaphoric zeros have been exemplified in the previous section. The difference between two languages as concerns the use of anaphoric zeros is the pro-drop character of Czech, which results in a large number of zeros in a subject position. These positions in English are occupied by personal pronouns in the 3rd person (190 cases, see Example 5.1 in Section 5.1) or in the first and 2nd person (40 cases in our data, see Example 5.28).

(5.28) $\underline{\varnothing}$ *nemáme pasivní čtenáře.*
We don't have passive readers.
We don't have passive readers.
Nemáme pasivní čtenáře.

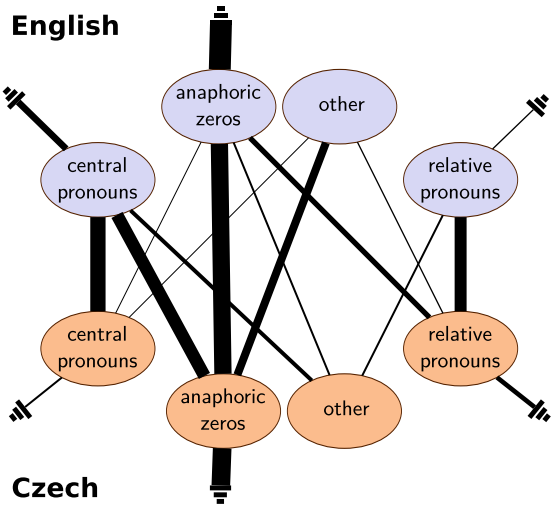

Figure 5.2: The overall schema of correspondences between Czech and English potentially coreferential expressions. The line width roughly represents a relative frequency of correspondences between the two groups of expressions. The ground symbol stands for the case that an expression has no counterpart in the other language.

Czech anaphoric zeros are not aligned in about 33% of cases. Similarly as in Section 5.5, the most frequent reasons for that are substantial rewording of the translation or missing arguments in the PCEDT t-layer annotation of English.

5.7 Summary

In this chapter, we collected the statistics of correspondences between Czech and English potentially coreferential expressions. The statistics is based on the annotation of manual alignment in the PAWS section of the PCEDT corpus (acquired also within this work), which consists of 1,000 sentences and covers over 1,300 coreferential expressions in each of the languages. Apart from delivering statistics, we also conducted a qualitative analysis of the most frequent or linguistically interesting correspondences and accompanied them with examples extracted from the data.

The overall picture of the correspondences between the groups of expressions is depicted in Figure 5.2. It suggests that Czech central pronouns and English relative pronouns are the groups with most straightforward counterparts, because the counterparts mostly belong to the same groups in the other language. Other groups exhibit more heterogeneous correspondences. This might have a major impact on cross-lingual techniques, especially on projection, which should work better on closely related languages. And the closer the languages are, the more straightforward the correspondence between the expression types should be.

Many English central pronouns, especially possessive pronouns, remain unaligned and a big proportion consisting of personal pronouns in subject position is mapped to Czech zeros. We have also shown that the English pronoun "*it*" and

English reflexive pronouns have several possible translations to Czech. A specific translation often depends on the function of the source pronoun. This can be leveraged to disambiguate the non-referential occurrences of "*it*" or to select the correct antecedent for two different uses of reflexive pronouns.[8] Another interesting case is that English possessive pronouns may be translated to possessive pronouns, reflexive possessive pronouns or they can disappear in Czech. The latter two cases might be helpful for coreference resolution as they often suggest that the antecedent is the subject.

A substantial proportion of Czech relative pronouns finds its counterparts in English zeros governed by infinitives and participles. Some of the relative pronouns do not have a counterpart at all, because the Czech relative clause is translated to a nominal group in English.

Zeros are most often mapped to zeros or have no counterpart at all. An many Czech zeros are aligned with 1st and 2nd person English personal pronouns, which we do not handle by our coreference resolver.

[8]Except for the work on bilingually informed CR (see Section 7.3) this variety of possible translations for these pronouns motivated our works on cross-lingual disambiguation of "*it*" [Veselovská et al., 2012] and machine translation [Novák et al., 2013a,b].

6. Cross-lingual Alignment of Coreferential Expressions

Before moving ahead to coreference resolution, let us take a little detour and elaborate more on improvements in aligning coreferential expressions. We already encountered the method that achieves the improvements in Chapter 5. The statistics on counterparts there would be less accurate without collecting them on the manual and supervised alignment. In this chapter, we describe these two types of alignment in detail.

A brief look at the data with the original alignment reveals that it tends to be less accurate on function words than on content words. This shortcoming also concerns pronouns. Their properties make it more difficult for unsupervised alignment algorithms, which heavily rely on co-occurrence statistic, to find their counterparts. First, pronouns act as placeholders in a text. Taken out of the context, they carry almost no meaning themselves. Next, because of their lack of meaning, their functions may vary greatly across languages. And last, partially due to the other two reasons, pronouns are more tied with grammatical rules of a particular language than for example nouns and verbs.

Other frequent coreferential expressions are zeros. An unsupervised alignment algorithm for surface words is not able to address them. Although they are partially treated in the original alignment by a heuristic, this heuristic is too simple to capture the complexity of context, in which the zero might appear.

In this chapter, we propose a method that aims at improving alignment between Czech and English coreferential expressions. Unlike the original alignment, the new method is based on supervised learning. It therefore requires that such an alignment is manually annotated in a portion of data, later used for training the method. Since, to the best of our knowledge, no satisfactory enough data exist we annotated them ourselves.

Section 6.1 introduces the data collection containing manual alignment between Czech and English coreferential expressions. We then exploit this data in Section 6.2. We create a supervised aligner for such expressions and compare its performance to the original alignment.

6.1 Manual Alignment

The task of alignment of coreferential expressions in the selected settings is too specific to find some third-party data for it, especially due to zeros.[1] We therefore opted for annotating such data ourselves [Novák and Nedoluzhko, 2015].

We chose PCEDT 2.0 as the data source for our annotation. Taking the size of PCEDT into account, it is understandable that annotating the entire corpus

[1] Mareček et al. [2008] presented a supervised method for aligning tectogrammatical trees, for which they prepared a dataset of 515 sentences sampled from PCEDT with manual annotation of alignment. However, the manual annotation was carried out on surface representations of the sentences and the alignment of t-nodes was then obtained just by automatic projection. Consequently, none of the elided expressions was covered by the alignment. As zeros account for a substantial part of coreferential expressions, we decided not to utilize this dataset.

would be extremely time-demanding. Therefore, we limited the dataset to only the first half of the PCEDT section 19, i.e., the 50 documents from `wsj_1900` to `wsj_1949` comprising 1,078 sentence pairs. This part of PCEDT including the newly annotated alignment has later developed into a multilingual parallel treebank PAWS [Nedoluzhko et al., 2018].

The alignment links were labeled by two annotators – the author of this thesis and Anna Nedoluzhko. Each instance has been annotated only once by one of the annotators, i.e. there is no instance with duplicate annotations. The original and heuristically refined alignment[2] served as pre-annotation to speed up the manual work.

Like in the monolingual case (see Section 4.2.2), the cross-lingual annotation supports *loose alignment*. Given an expression, the annotators may have aligned it loosely if they could not find a clear counterpart to the expression. Such alignment link then leads to a counterpart of the expression's antecedent. However, this alignment was allowed only in specific syntactic constructions, e.g., when the relative clause introduced by the Czech relative pronoun is in English expressed by a simple modifier depending on a noun, or by a predicative complement or other construction depending on a verb (see Examples 5.21 and 5.22 in Section 5.3, respectively). There are still many nodes that remain unaligned after having been annotated with the loose alignment.

The alignment has been manually annotated for the expression types that belong to the core of our cross-lingual research and which are thus targeted by our coreference resolver: personal, possessive, reflexive possessive (Czech only), reflexive and relative pronouns, zero subjects (Czech only), and zeros in non-finite clauses (see Section 2.4). Pronouns and zeros which are clearly in the 1st or 2nd person were excluded. Although the total sum of Czech and English expressions in the focused group approaches 3,000, many of them have counterparts that also belong to this group (see Chapter 5). By annotating one such an alignment link we cover two expressions. It thus sufficed to annotate roughly two thirds of them.

The data with manual alignment have been published in the PCEDT 2.0 Coref [Nedoluzhko et al., 2016a] and the PAWS treebank [Nedoluzhko et al., 2018].

6.2 Supervised Alignment

The supervised aligner tries to mimic the manual alignment as described in the previous section by post-processing the original alignment. It operates on the tectogrammatical layer and addresses the t-nodes representing selected coreferential expressions in English and Czech. The manually aligned data serves as a training data to build a composite model taking advantage of various features. Such model can be trained and run on both gold (manually annotated) and system (automatically annotated) trees.

In the following, we will describe in further details all the components of the aligner and provide some experiments on both gold and system trees. The

[2]Before we started annotating alignment manually, we had constructed some heuristics to address the cases for which the original alignment often failed. In the end, formulating the rules appeared to be too demanding, so we resorted to use them only as an automatic pre-annotation. The rules are described in [Novák and Nedoluzhko, 2015].

experiments should give us some evidence whether the supervised alignment is really the one we should use in cross-lingual experiments with coreference.

6.2.1 Design of the Aligner

Similarly to GIZA++ [Och and Ney, 2000], our aligner consists of two models, one for alignment links from English to Czech, the other one for the opposite direction. Each model is a ranker. For a given expression, all nodes from the aligned sentence are ranked at once to find a best-fitting counterpart. A dummy candidate is included to capture the option that the expression has no counterpart in the other language. The models have been trained using Stochastic Gradient Descent with L1 regularization in the Vowpal Wabbit[3] machine learning toolkit.[4]

At test time, the models are applied one after another on focused nodes in both languages, producing alignment links in both directions. They are subsequently symmetrized starting with the links that belong to the intersection of both directions. The remaining links are then processed in descending order of the aligners' decision confidences assigned to the links. A link is included only if neither the source nor the target node has already been covered by supervised alignment. Consequently, the aligner yields only 1:1 alignments, which might be restrictive. The manual annotation, however, shows that the subset of the nodes involved in more than one alignments accounts only for 1.6% of all focused nodes.

Features. For every coreferential expression and a potentially aligned node from the corresponding tree in the other language, the following types of features are extracted:

- *Original alignment:* presumably the most valuable set of features. It may be also considered the main input if we view the supervised aligner as a post-processor of the original alignment. It indicates if there is a link between the two nodes in the original alignment and if there is any between their parents.

- *Graph-based:* we designed these features to reflect the path between the nodes. Figure 6.1 illustrates how they work. The pair of aligned tectogrammatical trees is treated as a bipartite graph and a shortest path between the nodes is found using a sequence of dependency edges and a single alignment link. We applied the Dijkstra algorithm to find the shortest path. We ensure that it only uses a single alignment link by setting large weights to alignments and small weights to dependency edges, i.e., 100 and 1, respectively. The features then comprise the length of the shortest path and the sequence of edge labels (parent, child, alignment).

- *Grammatical:* these include lemmas, part-of-speech tags, reflexivity indicators, semantic role labels both for each of the nodes individually and as a concatenation of the two.

- *Expression types:* categories as introduced in Section 2.4.

[3] `https://github.com/JohnLangford/vowpal_wabbit`

[4] Note that we utilized almost identical modeling to CR (see Section 7.1).

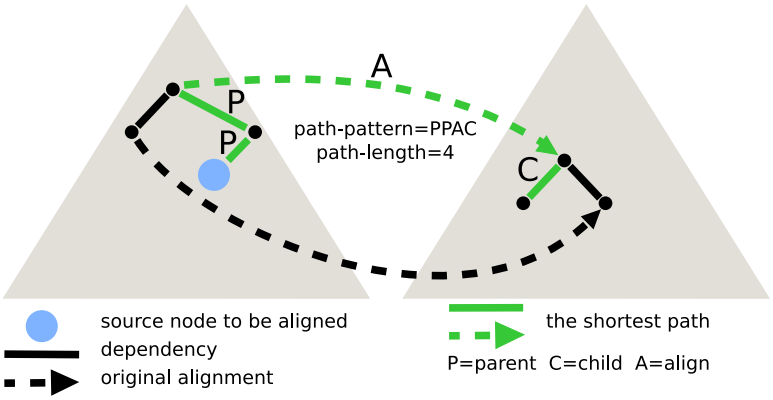

Figure 6.1: Extraction of graph-based features for alignment. They describe the length and the nature of the shortest path between two nodes lying in different trees that uses a single alignment link.

- *Combinations:* these features combine selected features from the types mentioned above. For instance, the concatenation of parents' alignment and semantic role correspondence mimics the heuristic for zeros used in the original alignment (see Section 4.2.3) and extends it to all candidate nodes in any of the two languages. Furthermore, features combining lemmas with direct alignment or alignment through parents are included.

6.2.2 Evaluation

The original objective of implementing the supervised aligning method was to refine alignment between manually annotated t-trees in PCEDT 2.0. Nevertheless, if the method is expected to be applicable to any parallel texts, it will need to operate on the t-trees automatically built by the pre-processing pipeline. To avoid performance decrease due to incompatibility of tectogrammatical representations, we decided to build models and evaluate them for both scenarios.

Accessing Manual Alignment from System Trees. Since manual alignment, on which the method should be trained and evaluated, is annotated only between manually annotated trees, we have to make sure that it is accessible from the trees created by the pre-processing pipeline. As illustrated in Figure 6.2, a pair of nodes from automatic trees in different languages is counted as positive instance only if there is no interruption in the path comprising three edges:

- *source monolingual alignment*: from the node in the source language parsed tree to its counterpart in the source language gold tree
- *gold (cross-lingual) alignment*: from the node in the source language gold tree to its counterpart in the target language gold tree
- *target monolingual alignment*: form the node in the target language gold tree to its counterpart in the target language parsed tree

This is where the monolingual alignment (see Section 4.2.2) plays a key role. Although monolingual alignment is trivial for the majority of nodes, problems may appear with reconstructed nodes. The method of monolingual alignment tries to mitigate these problems with three special rules introducing loose monolingual alignment, presented in Section 4.2.2 and illustrated in Figure 4.2.

Importance of loose monolingual alignment stands out for English zeros in non-finite clauses, whose annotation in manually and automatically analyzed trees differs considerably. While in the gold trees, 64% of such zeros have its manually aligned counterpart, it is 52% in the system trees.[5] If we switch off the three rules responsible for loose monolingual alignment, the proportion of positive examples drops to 34%, though.

Evaluation Measure. We employ both intrinsic and extrinsic measures to assess performance of the supervised aligner.

Given an expression, the task of finding its counterpart in the other language is technically similar to the task of finding its antecedent. We thus decided to measure the alignment quality with a similar intrinsic measure as we use for coreference resolution (see Section 4.4.3).

Running over all nodes potentially targeted by the aligner, we aggregate five counts: true positive (TP), true negative (TN), false positive (FP), false negative (FN), and wrong linkage (WL). The meaning of these counts is analogical to what they mean in case of coreference. The difference to the Prague anaphora score is that the spurious zero positive count is replaced by the true negative count. It is incremented whenever the focused node is correctly labeled as unaligned.

Interpretation of the counts may be a bit unclear in case of alignment between system trees. There are several options where a path between the nodes may fail. Instead of complicated explanations, we decided to clarify it in Figure 6.2 that pairs different types of errors with corresponding counts. There is one more aspect in which this alignment score for system trees differs from the Prague anaphora score. It does not incur a false negative error for a source-language node existing only in a gold tree and missing in the system tree. We focus solely on alignment of existing nodes.

Having all the counts aggregated, we can calculate the final scores in terms of precision and recall on aligned nodes, and accuracy on all focused nodes (including those correctly left unaligned).

$$A = \frac{TP+TN}{TP+TN+FP+FN+WL} \quad P = \frac{TP}{TP+FP+WL} \quad R = \frac{TP}{TP+FN+WL}$$

Apart from the intrinsic evaluation, we also used an approximate approach allowing for large-scale evaluation on full PCEDT 2.0, based on the following assumption: coreference is one of the means to maintain coherence in the text. If we assume that text coherence is not violated during translation, coreference chains representing an entity in each of the languages should correspond. Since language grammar differences have apparent effects on coreference, this is far from being true. Nevertheless, alignment improvements should lead to a higher rate of

[5] Nevertheless, not all the instances are guaranteed to be correct as the monolingual alignment may be wrong sometimes.

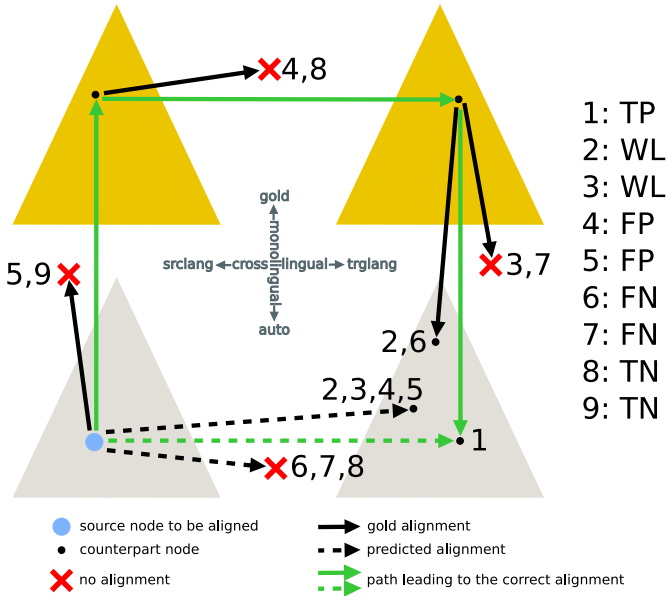

Figure 6.2: Accessing manual alignment from system trees brings several complications in training as well as in evaluation. We couple possible errors that might appear with the corresponding evaluation counts.

entity correspondence. To guarantee that the scores are not affected by the noise from automatic pre-processing, this evaluation of alignment was measured only on gold trees with manual annotation of coreference.

We measure this tendency by two correspondence scores: the *coreferring counterpart ratio*, and the *entity alignment rate*. The former is calculated as a proportion of the coreferring nodes targeted by the supervised aligner, whose counterparts in the other language are also coreferring. In contrast, the latter score takes the whole coreferential chain into account. It takes the proportion of the nodes referring to a common entity whose counterparts also refer to the same entity and averages it over all non-singleton entities.

Experiments and Analysis. Due to a small size of the dataset manually annotated with alignment, we carried out 10-fold cross-validation over the full dataset to intrinsically evaluate our supervised method. Table 6.1 shows its performance in comparison with the original alignment. Both approaches were evaluated on all the focused expressions in each of the languages (in terms of accuracy, precision and recall). In addition, we tested the approaches separately on the following coarse-grained classes of expressions (in terms of accuracy): central pronouns, relative pronouns, and zeros (see Section 2.4 for details on what these classes exactly include). Moreover, each cell of the table distinguishes two numbers corresponding to the performance on the system and the gold trees for a number on the left-hand and the right-hand side, respectively.

The key observation is that the supervised method outperforms the original alignment by a large margin, no matter of the language, the type of focused expressions, the type tree annotation, or the measure. The method achieves about the same quality of alignment for Czech and English expressions, 80% on

Method	Central	Relative	Zero	All		
	A	*A*	*A*	*A*	*P*	*R*
English						
Original	72 / 80	94 / 97	65 / 76	72 / 81	82 / 94	66 / 76
Supervised	82 / 91	98 / 98	70 / 85	79 / 89	93 / 96	72 / 87
Czech						
Original	89 / 88	55 / 67	60 / 79	65 / 78	76 / 95	71 / 74
Supervised	94 / 95	72 / 82	77 / 87	80 / 88	91 / 94	77 / 87

Table 6.1: Intrinsically evaluated performance of the original and supervised alignment. Each score cell is separated by a slash to two numbers. Performance on system trees is the number on the left-hand side while performance on gold trees is on the right-hand side. Unless stated otherwise for precision (P) and recall (R), all presented scores are accuracies (A; everything in %).

system trees and close to 90% on gold trees.

The task of t-node alignment therefore seems to be harder on system trees than on gold trees. It is not surprising as the errors on system trees tend to be propagated further. Such explanation is supported by the observation that the difference in performance is most pronounced for zeros and for the expressions which are often aligned with zeros in the other language (see Chapter 5). Nevertheless, possible inaccuracies in transferring the manual alignment to system trees may also play a role.

Precision and recall scores show that the supervised method improves mainly the recall on gold trees, i.e. it addresses the expressions which had no counterpart in the original alignment. System trees paint a different picture. Despite the doubtless improvement in recall, the precision increase seems to be more dominant. In this case the supervised method thus often fixes the false positives, i.e. it removes the spurious alignments added by the original alignment.

Let us contrast the results of the intrinsic evaluation on gold trees (but similar patterns are notable even on system trees) with the findings from the cross-lingual analysis of correspondences in Chapter 5 summarized in Figure 5.2. The groups of expression types that have mostly a straightforward counterpart of the same group in the other language (English relative pronouns, Czech central pronouns) are the ones, for which we observe the smallest difference in performance between the original and the supervised alignment. Yet, it is still 7 percentage points in the case of Czech central pronouns.

Table 6.2 compares the alignment types in terms of the coreference-related correspondence scores measured on the entire PCEDT 2.0 Coref as well as on its PAWS subset. We use the manual alignment available in PAWS to set the upper bounds for these scores. And this upper bound is really higher than the scores achieved with other alignment types, which confirms the correctness of the aforementioned assumption behind the correspondence measures. The numbers on supervised alignment accord with the scores from the intrinsic evaluation, performing better than the original alignment overall.

	English						Czech					
	orig		super		manual		orig		super		manual	
Coref. counter. ratio	57.4	55.3	66.4	62.2	71.3	—	54.9	55.2	62.9	62.4	68.5	—
Entity alignment rate	56.2	49.7	59.3	52.0	60.5	—	53.4	52.2	56.1	54.9	57.8	—

Table 6.2: The coreference-based metrics showing the quality of node alignment (in %), comparing the original, supervised and manual alignment. In each cell, the first number is measured on the sections `wsj_1900`–`49`, while the second one on the complete PCEDT 2.0 Coref.

6.3 Summary

In this chapter, we proposed a supervised alignment method that targets potentially coreferential expressions.

First, we manually annotated alignment on potentially coreferential expressions in around 1,000 sentence pairs. We focused only on the expression types that we want to address by bilingually informed CR. In order to increase the number of aligned expressions, some of the expressions that have no direct counterpart were annotated with loose alignment links to the counteparts of their close antecedents.

Second, we designed a supervised method to predict alignmnent on coreferential expressions. It was trained on the manual annotation of such alignment and consequently it targets the same set of expressions. Using a variety of features including the original alignment based mainly on GIZA++, graph-based and grammatical features, it substantially outperforms the original alignment on both gold and system trees. The highest improvement is achieved on expressions that do not have a straightforward mapping to the other language. As can be seen in Figure 5.2 in Chapter 5, these are actually the same that tend to find their counterparts among the zeros in the other language. Moreover, the increase in coreference-related correspondence scores suggests that supervised alignment should positively affect cross-lingual techniques to coreference resolution. The supervised aligner can also be viewed as a post-processor of the original alignment, augmenting its results on coreferential expressions.

7. Adding Cross-lingual Features to Coreference Resolution

In this chapter, we introduce the first out of the two cross-lingual approaches to coreference resolution presented in the thesis – bilingually informed CR. Before delving into the cross-lingual experiments, we need to describe our coreference system Treex CR in general and conduct experiments in a monolingual setting. The results of these experiments can then be compared with the cross-lingual approach.

Although there are multiple third-party coreference resolvers for English available (e.g. Stanford systems [Lee et al., 2011, Clark and Manning, 2015, 2016], the Berkley system [Durrett and Klein, 2014] and BART [Versley et al., 2008]), none of them has a support for Czech. Furthermore, they address neither zeros, nor relative pronouns. Both expression types play a key role in Czech-English coreferential correspondences, as can be seen in Chapter 5. Moreover, none of them is ready to be directly utilized for bilingually informed CR.

We therefore developed our own coreference resolver – Treex CR. Treex CR is a successor of the CzEng CR (see Section 4.3.1), which has been used to automatically annotate coreference in CzEng 1.0 [Bojar et al., 2011]. Unlike CzEng CR, the resolver presented here is entirely based on machine learning, which makes the resolver easily adjustable to a cross-lingual scenario. The component that is responsible for bilingually informed CR is able to reach information from the other language through the alignment (established in Chapter 6) and convey this information in the form of features to the resolver.

The results of the analysis on the parallel data (see Chapter 5) suggest that the aligned language may introduce some new information and thus improve the resolution. One of the indicators is that the space of counterparts of some potentially coreferential mentions is considerably heterogeneous. Some of the types in the aligned language then may be easier to resolve than their target-language counterparts. For example, the Czech reflexive possessive pronoun, usually coreferential with the sentence's subject, may help in finding the correct antecedent of the English possessive pronoun. Even if the types of the mention and its counterparts agree, other grammatical aspects of the language (see Section 2.4) may give some beneficial information. For instance, we believe that Czech genders, which are more evenly distributed over the nouns than the English genders, may help filter out English antecedent candidates that are unpropable due to gender disagreement in the Czech side. In the opposite direction, the English personal pronoun as a counterpart may facilitate resolution of the underspecified Czech zero subject.

The chapter is structured as follows. Treex CR along with its cross-lingual component is thoroughly described in Section 7.1. In Section 7.2, we carry out the experiments with Treex CR in the monolingual settings and compare its performance with the other systems for Czech and English introduced in Section 4.3. The cross-lingual experiments are all conducted in Section 7.3 and, finally, we conduct a detailed quantitative and qualitative analysis of the two approaches in Section 7.4.

7.1 Treex Coreference Resolver

Treex coreference resolver [Novák, 2017, *Treex CR*] is a coreference resolution system, whose main distinctive feature is that it operates on the tectogrammatical layer. As the tectogrammatics is inherently capable of representing some types of structural ellipsis (see Section 2.3), Treex CR may easily address zero anaphora. This is crucial for monolingual CR in pro-drop languages such as Czech. However, zero anaphora may be present in more latent form also in other languages, for example in English non-finite clauses.

The system is based on machine learning, thus making all the components fully trainable if appropriate training data is available. Although the system has been so far build for Czech, English, Russian, and German, in this work we concentrate only on Czech and English.

Treex CR takes inspiration in its architecture from a supervised resolver for Czech personal pronouns and zero subjects by Nguy et al. [2009]. It also implements some of the features they proposed. Some of the features are also inspired by rule-based approaches to CR introduced by Kučová and Žabokrtský [2005] and Nguy [2006], and later reimplemented in order to be used in translation with TectoMT [Žabokrtský et al., 2008]. A combination of these approaches has been applied to the original automatic annotation of coreference in the CzEng 1.0 corpus [Bojar et al., 2012], presented in Section 4.3.1. Treex CR cherry-picks the best of all these approaches, introduces some new features, enhances the ML-method and extends the resolver also to another anaphor types. All of it, as its name suggests, has been implemented as an integral part of the Treex NLP framework [Popel and Žabokrtský, 2010].

The training workflow of Treex CR in its monolingual setting is schematized in Figure 7.1. In the remaining parts of this section, we will describe the individual stages of the workflow, while referring to them in the schema. Each input text must be first pre-processed to form the system trees by the pipeline already introduced in Section 4.2 and denoted by no. 1 in the schema. In Section 7.1.1, we focus on the reasons why this pre-processing stage is essential. In the training stage, also the coreference annotation from gold trees is projected to the system trees and later transformed to gold labels in training examples (see no. 2 in the schema). As it is common for traditional ML, a set of descriptive features which the system uses to drive its decisions must be extracted from the underlying pre-processed text (see no. 3 in the schema). We discuss the features for monolingual resolution in Section 7.1.3. In Section 7.1.2, we present the overall architecture of the system and its models and the learning method, which takes advantage of extracted features and the gold coreference (see no. 4 in the schema).

The bilingually informed setting of the system differs from the monolingual in the set of features it extracts. We elaborate more on this cross-lingual extension in Section 7.1.4.

7.1.1 Tectogrammatical Analysis

Treex CR is a unified solution for finding coreferential relations on the t-layer. It requires the input texts to be automatically analyzed up to this level of linguistic annotation. There are several reasons for this requirement.

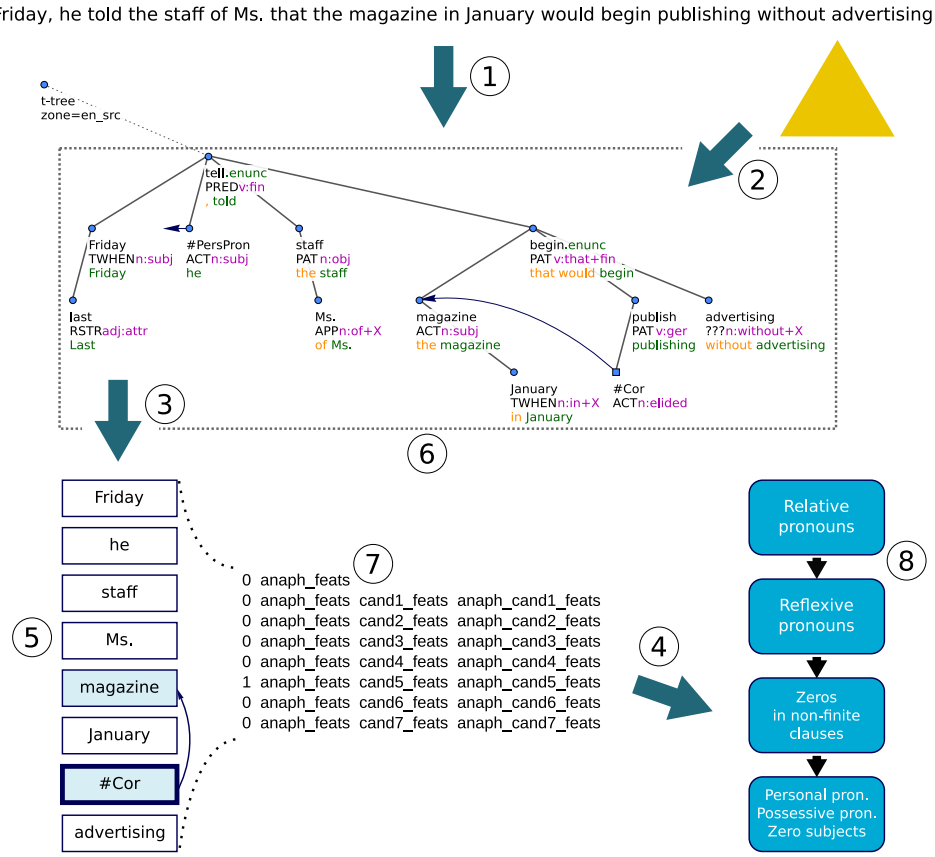

Figure 7.1: The architecture and the workflow of Treex CR in its monolingual setting.

Coreference is a phenomenon that is usually manifested on multiple linguistic layers. For example, anaphoric pronouns tend to agree with their antecedent in morphological gender and number, reflexive pronouns point to a sentence subject, or coreferential nominal groups should be semantically compatible. Rich annotation can then be exploited by a rich feature set, which significantly affects performance.

Furthermore, morphological information play an important role in the system design. They drive the selection of anaphor candidates and their partitioning by the anaphor type for multiple specialized models. They can also limit the number of antecedent candidates. These limits are further tightened by the t-layer and its property that it represents only content words. Last but not least, the possibility of the t-layer to represent expressions missing on the surface enables addressing zero anaphora.

The pre-processing pipeline that Treex CR builds on is the one that we introduced in Section 4.2 (and schematized in Figure 7.1, no. 1). Note that the pipeline is the same for the texts to be resolved at test time as well as for those exploited to train CR models. The pre-processing steps applied to the train and test data must be identical to guarantee the performance of the Treex CR system.

7.1.2 System Design

Treex CR models coreference in a way that can be easily optimized by supervised learning. Specifically, we use logistic regression with stochastic gradient descent optimization implemented in the Vowpal Wabbit toolkit.[1] In the training stage, the gold labels are extracted from the coreferential links in the gold trees via the monolingual alignment (see Figure 7.1, no. 2). The design of the model employs multiple concepts that have proven to be useful and simple at the same time (see Section 3.1 to refer to the related work).

Mention-ranking model. Given an anaphor and a set of antecedent candidates, *mention-ranking* models [Denis and Baldridge, 2007b] are trained to score all the candidates at once (Figure 7.1, no. 5). Competition between the candidates is captured in the model. Every antecedent candidate describes solely the actual mention. It does not represent a possible cluster of coreferential mentions built up to the moment.

Antecedent candidates for an anaphor are selected from the context window of a predefined size (Figure 7.1, no. 6). This is done only for the nodes satisfying simple morphological criteria (e.g. nouns and pronouns). Both the window size and the filtering criteria can be tuned as hyperparameters.

Joint anaphoricity detection and antecedent selection. What we denote as an anaphor in the model is, in fact, an anaphor candidate. There is no pre-processing that would filter out non-referential anaphor candidates. Instead, both decisions, i.e. (1) to determine if the anaphor candidate is referential, and (2) to find the antecedent of the anaphor, are performed in a single step. This is ensured by adding a fake "antecedent" candidate representing solely the anaphor candidate itself (see Figure 7.1, no. 7). By selecting this candidate, the model labels the anaphor candidate as non-referential.

A cascade of specialized models. The properties of coreferential relations are so diverse that it might be advantageous to model individual anaphor types separately rather than jointly [as shown in Denis and Baldridge, 2008]. For instance, while personal pronouns may refer to one of the previous sentences, the antecedent of relative and reflexive pronouns always lies in the same sentence. By representing coreference of these expressions separately in multiple specialized models, the abovementioned hyperparameters can be adjusted to suit the particular anaphor type. The processing of these anaphor types may be sorted in a cascade so that the output of one model is taken into account in the following models (Figure 7.1, no. 8). Currently, we do not take advantage of this feature. Models are thus independent of each other and can be run in any ordering.

7.1.3 Feature Sets

The pre-processing stage (see Section 7.1.1) enriches raw text with a substantial amount of linguistic information. Feature extraction stage then uses this material to yield *features* consumable by the learning method (see Figure 7.1, no. 3).[2]

[1] `https://github.com/JohnLangford/vowpal_wabbit`

[2] In addition, Vowpal Wabbit supports additional feature combination. The features must be first manually grouped into namespaces and Vowpal Wabbit then produces new features as

Most of the feature extraction mechanism is language-independent. The majority of feature templates is thus shared among the languages supported by Treex CR. Nevertheless, a language-dependent component of the feature extractor have to be plugged in if a feature is based on: (1) linguistic annotation with a form that depends on a language (e.g. Czech vs. English part-of-speech tags), or (2) linguistic annotation or a resource that has not been made available for some languages (e.g. anaphoricity estimate of an English pronoun *it*).

Features used in Treex CR can be categorized by their form. The categories differ in the number of input arguments they require. *Unary features* describe only a single node, either an anaphor or an antecedent candidate. Such features start with prefixes `anaph` and `cand`, respectively. *Binary features* require both the anaphor and the antecedent candidate for their construction. Specifically, they can be formed by agreement or concatenation of respective unary features, but they can generally describe any relation between the two arguments. Finally, *ranking features* need all the antecedent candidates along with the anaphor candidate to be yielded. Their purpose is to rank antecedent candidates with respect to a particular relation to an anaphor candidate.

Our features also differ by their content. They can be divided into three categories: (1) location and distance features, (2) (deep) morpho-syntactic features, and (3) lexical features. The core of the feature set was formed by adapting features introduced in [Nguy et al., 2009].

Location and distance features. Positions of anaphor and an antecedent in a sentence were inspired by [Charniak and Elsner, 2009]. Position of the antecedent is measured backward from the anaphor if they lie in the same sentence, otherwise it is measured forward from the start of the sentence. As for distance features, we use various granularity to measure distance between an anaphor and an antecedent candidate: number of sentences, clauses, and words. In addition, an ordinal number of the current candidate antecedent among the others is included. All location and distance features are bucketed into predefined bins.

(Deep) morpho-syntactic features utilize the annotation provided by part-of-speech taggers, parsers, and tectogrammatical annotation. Their unary variants capture the mention head's part-of-speech tag, morphological features,[3] e.g. gender, number, person or case. As the gender and number are considered important for resolution of pronouns, we do not rely on their disambiguation and work with all possible hypotheses. We do the same for some Czech words that are in nominative case but disambiguation labeled them with the accusative case. Such case is a typical source of errors in generating a zero subject as it fills the missing nominative slot of the governing verb's valency frame. To discover potentially spurious zero subjects, we also inspect if the verb has multiple arguments in accusative and if the argument in nominative is refused by the valency, as it is in the phrase "*Zdá se mi, že...* /It seems to me that.../". Furthermore, the unary features contain (deep) syntax features including its dependency relation,

a Cartesian product of selected namespaces. This massively extends the space of features. Such behavior can be controlled by Vowpal Wabbit's hyperparameters.

[3]Also in the form of tectogrammatical grammatemes, which may condense information from related auxiliary words.

semantic role, and formeme. We exploit the structure of the syntactic tree as well, extracting some features from the mention head's parent.

Many of these features are combined to binary variants by agreement and concatenation. Heuristics used for some anaphor types in the rule-based predecessors of Treex CR [Kučová and Žabokrtský, 2005, Nguy, 2006] gave birth to another pack of binary features. For instance, the feature indicating if a candidate is the subject of the anaphor's clause should target coreference of reflexive pronouns. Similarly, signaling whether a candidate governs the anaphor's clause should help with resolution of relative pronouns.

Lexical features. Lemmas of the mentions' heads and their parents are directly used as features. Such features may have an effect only if built from frequent words, though. By using them with an external lexical resource, this data sparsity problem can be reduced. Firstly, we used a long list of noun-verb collocations collected by [Nguy et al., 2009] on Czech National Corpus [syn, 2005]. Having this statistics, we can estimate how probable is that the anaphor's governing verb collocates with an antecedent candidate.

Another approach to fight data sparsity is to employ an ontology. Apart from an actual word, we can include all its hypernymous concepts from the hierarchy as features. We exploit WordNet [Fellbaum, 1998] and EuroWordNet [Vossen, 1998] for English and Czech, respectively.

To target proper nouns, we also extract features from tags assigned by named entity recognizers run during the pre-processing stage.

7.1.4 Cross-lingual Extension

Bilingually informed coreference resolution is an approach derived from monolingual CR. Both approaches address coreference in one target language at a time. However, bilingually informed CR exploits information not only from the target language but also from an additional auxiliary language. Particularly, the underlying data must contain texts in one language as well as its translations to the other one. In other words, bilingually informed CR requires parallel data. This requirement holds both for the training as well as the test data. The auxiliary-language side of the parallel data can be then exploited by various means, e.g. by an extended feature set or an advanced learning method. In our case, the cross-lingual information is exploited by the features accessing it through the alignment (as illustrated in Figure 7.2).

Our parallel data consists of English-Czech human translations, as introduced in Section 4.1. These are analyzed up to the tectogrammatical layer and aligned on a word level, with a special emphasis on alignment of coreferential expressions treated by a supervised method (see Section 6.2). Such data is then exploited by a feature set which in addition to the monolingual features describing the coreferential candidates in the target language contains also cross-lingual features focusing on the counterparts of the candidates from the aligned language (the auxiliary language). The system design that we implement for bilingually informed CR is exactly the same as the one we use in the monolingual approach. The only difference in our approaches to monolingual and cross-lingual CR therefore lies in the utilized feature set.

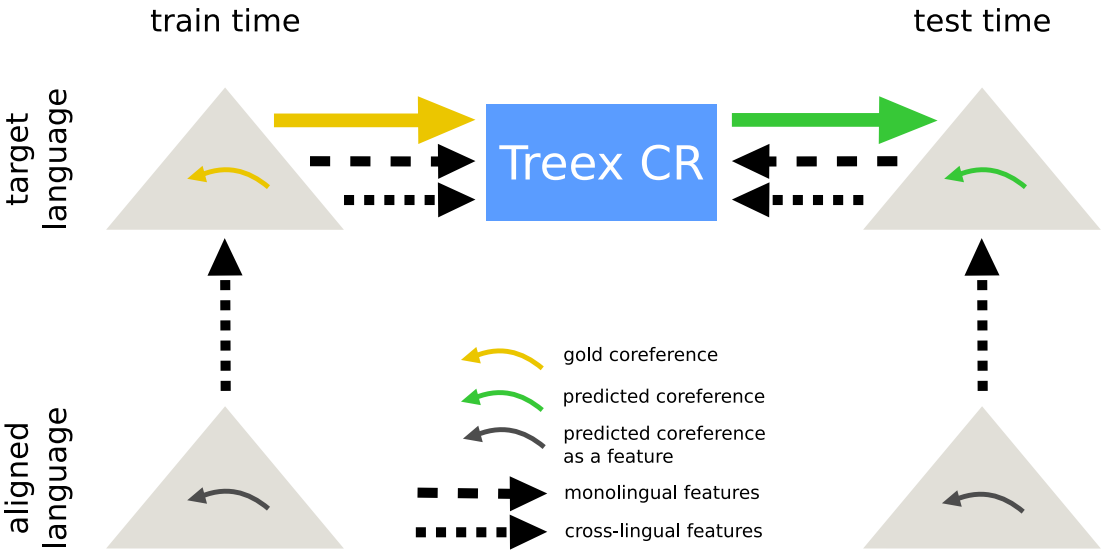

Figure 7.2: The workflow of Treex CR in its bilingually informed setting.

Cross-lingual Features. Our cross-lingual features describe the nodes aligned to the coreferential candidates in the target language. As elaborated in Section 7.1.3, monolingual features are always related to two nodes that may be in the end declared as coreferential – an anaphor candidate and an antecedent candidate. To construct the cross-lingual features, we follow the alignment links connected to these two nodes. For each of the two nodes, we take at most one of its aligned counterparts. In this way, we obtain at most two nodes aligned to the pair of potentially coreferential nodes. Having these two nodes from the aligned-language side of the parallel data, we can extract cross-lingual features consisting of unary and binary features as introduced in Section 7.1.3. Only unary features can be extracted in case a single node was found. Finally, if no aligned counterpart is found, we add no cross-lingual features for the given pair of coreferential candidates.

We extract two sets of cross-lingual features:

- *aligned_all*: it consists of all the features contained in a monolingual set for a given aligned language;
- *aligned_coref*: it consists of a single binary indicator feature, assigning the true value only if the two aligned nodes belong to the same coreferential entity. The coreference annotation in aligned language is expected to be a result of a automatic monolingual CR system for this language. We employ Treex CR and its monolingual models for English and Czech, but any CR system, even a rule-based one, could be used.

All cross-lingual features are prefixed with `align_` in order to avoid name collision with monolingual features.

We do not manually construct features combining both language sides. Nevertheless, such features are formed automatically by the machine-learning tool Vowpal Wabbit.

	Window size	Following nodes	Filtered nodes	Vowpal Wabbit
Relative pron.	current sent.	×	semantic nouns (see Section 2.4) and verbs	cost-sensitive one-against-all model with label-dependent features, logistic loss, L1 regularization: 5×10^{-8}, passes over data: 5, quadratic combination of anaphor and antecedent features
Reflexive pron. Refl. poss. pron.	current sent.	✓	semantic nouns	
Zeros in non-fin. cl.	current sent.	✓	semantic nouns and zeros	
Personal pron. Possessive pron. Zero subjects	current and previous sent.	×	semantic nouns in the 3rd or undefined person	

Table 7.1: Hyperparameters of Treex CR models.

7.2 Monolingual Resolution

For each of the languages, we trained one monolingual system that consists of four models specialized at anaphor types belonging to the core of our research: (1) relative pronouns, (2) reflexive pronouns (and reflexive possessive pronouns in Czech), (3) zeros in non-finite clauses, and (4) personal and possessive pronouns (and zero subjects in Czech). There are three hyperparameters that are set individually for each of the models: (a) the size of the window from which antecedent candidates are selected, (b) an indicator if the window covers also the nodes following the anaphor, and (c) the morpho-syntactic filter that restricts these candidates. Other hyperparameters including those designated for the Vowpal Wabbit learning tool are identical across all the models. The hyperparameters' values were selected as a result of manual inspection and testing on the development test sets, mainly on the Czech ones. Exactly the same values are then used for English.[4] All of the hyperparameters are listed in Table 7.1.

Performance of Treex CR is compared with its predecessor CzEng CR (see Section 4.3.1) on both languages. In addition, we contrast them with the three Stanford systems for English presented in Section 4.3.2.

We carried out training and development testing of Treex CR on the corresponding sections of PDT for Czech, and PCEDT for English (as specified in Section 4.1). The testing of all the systems was conducted on two datasets for each of the languages: PDT and PCEDT evaluation test set for Czech, and PCEDT test set and the CoNLL 2012 test set for English.

All systems are evaluated using the Prague anaphora score on individual anaphor types. We also report total numbers aggregated over multiple anaphor types. However, the extent of included types varies for different tables that we are showing in the following sections.

7.2.1 Overall Evaluation Results

Table 7.2 shows overall scores for both Czech and English. The overall scores are aggregated over the mention types targeted by Treex CR for the particular

[4] A better performance might be achieved if all the hyperparameters are tuned specifically for each of the models. Nevertheless, we did not seek for the truly optimal solution, since the main scope of this work is rather cross-lingual techniques.

	Czech		English	
	PDT	PCEDT	PCEDT	CoNLL
Stanford				
deterministic	—	—	$^{63.98}_{23.33}$ 34.19	$^{60.07}_{61.21}$ 60.64
statistical	—	—	$^{77.09}_{25.43}$ 38.24	$^{72.58}_{69.69}$ 71.10
neural	—	—	$^{78.87}_{27.39}$ 40.66	$^{74.47}_{66.91}$ 70.49
CzEng CR	$^{65.65}_{48.13}$ 55.54	$^{64.38}_{44.87}$ 52.88	$^{72.19}_{44.73}$ 55.24	$^{66.52}_{60.73}$ 63.50
Treex CR	$^{69.71}_{62.82}$ 66.08	$^{68.67}_{61.55}$ 64.92	$^{71.13}_{62.62}$ 66.61	$^{67.29}_{63.98}$ 65.60

Table 7.2: Overall performance of all tested CR systems on the evaluation sets of the English and Czech datasets.

language, if coreference for these types is annotated in the test set. It means that on Czech data the scores capture all targeted types; Czech types of reflexive possessive pronouns and zero subjects are excluded for English PCEDT, and, finally, the types of relative pronouns and other zeros are excluded for the CoNLL test set.

Treex CR outperforms its predecessor by a large margin on both Czech evaluation datasets – by 11-12 points. Although we observe a increase of precision, the improvement can mostly be attributed to the increase in recall by more than 14 percentage points.

On English PCEDT, we observe about the same sharp difference of 11 F-score points. Nevertheless, this time all the credit is taken by the improvement in recall, as the precision even slightly dropped. The difference on CoNLL data is only 2 points in favor of Treex CR, which suggests that most of the improvement of the model is achieved on the mention types not covered by CoNLL.

As for the Stanford systems, the deterministic method is outperformed by both the statistical and the neural method. However, the latter two methods seem to be more equal on pronouns than expected. The neural system is better on the PCEDT test set, but worse on the CoNLL set.

Contrasting Treex CR and the Stanford systems on PCEDT data via the overall score would be unfair, as the Stanford systems do not address zeros and relative pronouns. It should be fair on the CoNLL test set, though. Here, the results suggest that our English monolingual Treex CR system performs halfway between the deterministic and the other two Stanford systems. Recalling that Stanford systems implement more advanced approaches and that the Treex CR hyperparameters could be optimized better, Treex CR achieves a decent resolution quality.

7.2.2 Fine-grained Evaluation Results on Czech

Table 7.3 focuses on performance of the Czech systems on individual anaphor types. Treex CR is able to gain across all the types. Apart from the category of Czech zeros in non-finite clauses, which has not been targeted by CzEng CR, the highest improvement can be seen for relative pronouns and zero subjects.

Mention type	PDT		PCEDT	
	CzEng CR	Treex CR	CzEng CR	Treex CR
Personal pron.	61.27 62.91 62.08	64.02 62.35 63.18	60.45 60.09 60.27	65.62 64.66 65.14
Possessive pron.	58.98 58.79 58.89	65.57 64.09 64.82	59.69 60.31 60.00	64.16 63.32 63.74
Refl. poss. pron.	84.15 80.00 82.02	83.20 82.27 82.73	84.85 80.62 82.68	78.68 78.06 78.37
Reflexive pron.	61.71 60.00 60.84	65.67 57.53 61.33	36.36 54.78 43.71	46.58 56.03 50.87
Zero subject	64.68 42.90 51.58	59.90 60.63 60.26	67.91 36.55 47.52	63.33 53.30 57.88
Zero in nonfin. cl.	0.00 6.20 0.00	68.48 30.68 42.38	0.00 8.29 0.00	70.82 40.06 51.18
Relative pron.	64.79 51.18 57.18	84.12 76.88 80.34	57.71 50.73 54.00	75.32 72.64 73.96
Total	65.65 48.13 55.54	69.71 62.82 66.08	64.38 44.87 52.88	68.67 61.55 64.92

Table 7.3: Performance of Czech systems measured on fine-grained categories in PDT and PCEDT.

Mention type	Stanford			CzEng CR	Treex CR
	deter.	stat.	neur.		
Personal pron.	63.03 61.66 62.34	74.67 66.60 70.40	78.25 71.21 74.57	75.40 65.17 69.91	75.25 68.77 71.86
Possessive pron.	66.77 64.13 65.42	81.37 71.24 75.97	80.08 77.44 78.74	79.67 77.85 78.75	79.29 78.76 79.03
Reflexive pron.	56.25 54.00 55.10	69.77 60.00 64.52	75.00 66.00 70.21	71.43 60.00 65.22	74.51 74.00 74.25
Demonstr. pron.	7.61 4.52 5.67	10.64 3.23 4.95	37.50 1.94 3.68	0.00 0.65 0.00	0.00 0.65 0.00
Zero in nonfin. cl.	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	60.88 18.56 28.44	64.11 51.20 56.93
Relative pron.	27.78 0.59 1.15	0.00 0.00 0.00	0.00 0.00 0.00	72.10 69.24 70.64	78.26 73.57 75.84
1st/2nd pers. pron.	56.62 59.90 58.21	68.18 66.41 67.28	73.20 58.07 64.77	0.00 0.00 0.00	0.00 1.29 0.00
Named entities	76.28 80.68 78.41	76.70 61.35 68.17	76.69 73.17 74.89	0.00 0.00 0.00	0.00 0.74 0.00
Nominal group	39.77 51.58 44.91	59.61 46.98 52.55	63.63 50.66 56.41	0.00 0.08 0.00	72.90 0.68 1.35
Other	3.66 1.53 2.16	10.20 0.92 1.69	6.58 0.61 1.12	0.00 0.00 0.00	0.00 2.76 0.00
Total	53.58 37.11 43.85	68.58 35.49 46.78	71.49 38.20 49.79	72.18 20.44 31.85	70.90 29.42 41.59

Table 7.4: Performance of the English systems measured on fine-grained categories in PCEDT.

Whereas the CzEng CR rule-based block for relative pronouns sought for an antecedent only using a syntactic pattern, Treex CR can effectively benefit from the combination of syntactic patterns and gender/number agreement. It also succeeds in identifying non-anaphoric examples, for instance interrogative pronouns, which use many same forms. Zero subjects benefit from a much better recall at the expense of lower precision. This is probably caused by a new strategy of addressing spurious zeros, which are now often coreferential with the expression playing the same role in the sentence. This strengthens for example the features on gender/number agreement and thus makes the resolver less conservative. On the contrary, the performance dropped on reflexive possessives in PCEDT. This might be a consequence of their joint modeling with basic reflexive pronouns.

Mention type	Stanford deter.	Stanford stat.	Stanford neur.	CzEng CR	Treex CR
Personal pron.	58.03 59.99 59.00	71.09 68.66 69.85	73.31 64.20 68.45	66.21 58.02 61.84	67.31 61.89 64.49
Possessive pron.	65.08 63.49 64.28	75.94 71.59 73.70	76.79 73.36 75.03	67.05 68.80 67.92	66.48 69.90 68.15
Reflexive pron.	70.90 72.52 71.70	81.89 79.39 80.62	81.25 79.39 80.31	69.09 58.02 63.07	73.91 64.89 69.11
Demonstr. pron.	7.51 10.28 8.68	11.01 5.61 7.43	21.05 3.74 6.35	0.00 0.00 0.00	0.00 0.00 0.00
1st/2nd pers. pron.	61.11 54.07 57.38	62.42 69.38 65.72	70.58 58.26 63.83	0.00 0.00 0.00	0.00 0.41 0.00
Named entities	60.25 59.25 59.75	69.54 60.47 64.69	68.65 57.88 62.80	0.00 0.00 0.00	0.00 0.00 0.00
Nominal group	27.82 39.78 32.74	49.32 37.34 42.50	59.23 38.35 46.55	0.00 0.00 0.00	89.34 0.92 1.81
Other	0.34 0.00 0.00	10.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00
Total	46.95 53.01 49.80	63.48 58.40 60.83	68.65 54.91 61.02	66.52 18.90 29.44	67.25 19.90 30.71

Table 7.5: Performance of English systems measured on fine-grained categories in CoNLL.

7.2.3 Fine-grained Evaluation Results on English

Tables 7.4 and 7.5 show the fine-grained evaluation results on the English part of PCEDT and CoNLL test set, respectively. This time, the tables show all types that are annotated for coreference in each of the dataset. The total numbers aggregate over all these types, and thus do not equal the overall scores presented in Table 7.2.

It is immediately obvious that the Stanford resolvers target different coreferential expressions than the two resolver based on tectogrammatics. The only types targeted by both are personal, possessive and reflexive pronouns. Other mention types are covered either by only one of these resolvers' groups, or none of them. For instance, it is surprising that demonstrative pronouns are barely treated with the Stanford tools. We suspect many of such pronouns do not in fact refer to an entity but to an event, which is beyond the scope of Stanford systems.

On both the datasets, Treex CR outperforms its predecessor CzEng CR on all the types these resolvers focus on. Nevertheless, the fine-grained evaluation reveals that the big gap between the overall scores on PCEDT should be mostly attributed to the mention types that are not represented as coreferential in the CoNLL dataset: relative pronouns and zeros. A dramatic improvement of 28 points observed on PCEDT's zeros is mainly caused by a leap in recall. This is the consequence of the pre-processing pipelines for the two resolvers which differ in the extent to which they reconstruct zeros (see Section 4.3.1). Table 4.5 in Section 4.2.1 shows that the current pipeline is able to restore more than 90% of the English zeros with a high precision. In contrast, the recall of the zero reconstruction heuristics in the CzEng pipeline is only 34%. The low recall of reconstruction then directly propagates to the low recall of coreference resolution.

Luckily, Treex CR managed to surpass Stanford systems (the neural one) on possessive and reflexive pronouns and the second best system (the statistical one) on personal pronouns in the PCEDT dataset. However, a completely different picture is painted on the CoNLL dataset. Treex CR is able to outperform only

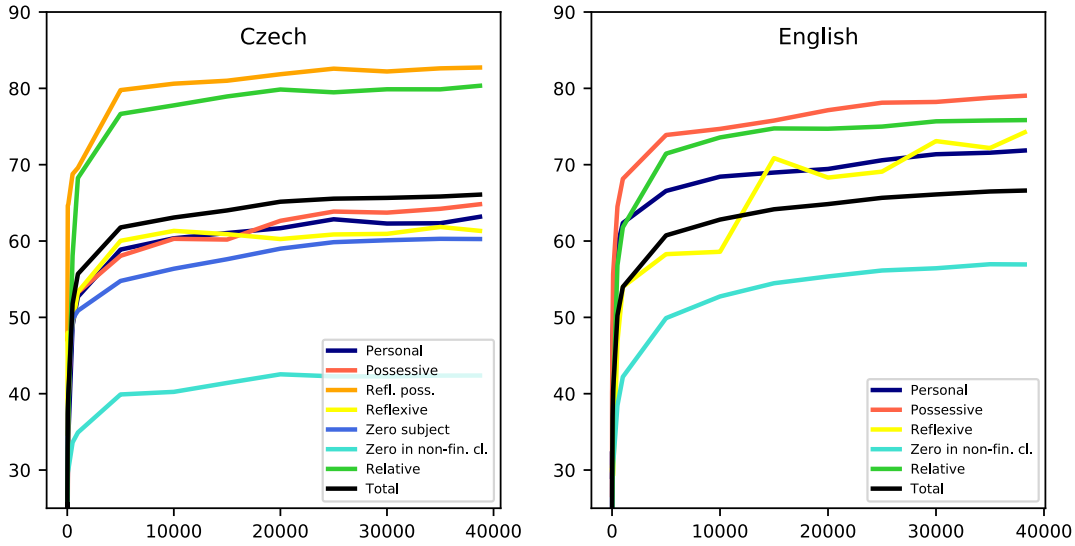

Figure 7.3: Learning curves of the Czech and the English monolingual CR system evaluated respectively on the PDT and PCEDT evaluation test set. The x-axis represents the number of sentences in the training data and the y-axis is the F-score.

the deterministic Stanford system there, and not even that in the case of reflexive pronouns. Since both the datasets come from a similar domain, even containing some overlapping documents (see Section 4.1.4), we suspect that the reason for this discrepancy lies in different standards for coreference annotation in PCEDT and OntoNotes (from which the CoNLL test set is sampled).

To the best of our knowledge, no analysis of how Stanford systems perform for individual anaphor types has been published so far. The deterministic approach seems to be outperformed on all mention types. The only exceptions are demonstrative pronouns, where the system achieve very low score anyway, and, quite surprisingly, named entities on the PCEDT dataset. The statistical method outperforms the other approaches in the category of pronouns in 1st and 2nd person consistently in both dataset. The neural system clearly dominates only on possessive pronouns and nominal groups in both datasets. Nevertheless, for the rest of the mention types discrepancies across the datasets similar to those mentioned above can be observed among the Stanford systems, too. Consequently, it makes it difficult to arrive at any clear conclusion on the performance of Stanford system on individual mention types.

7.2.4 Learning Curves

Figure 7.3 depicts the learning curves of the monolingual system for both Czech and English. The training data were randomly sampled from the full-size training set and evaluated on the evaluation test set. This was repeated three times and the scores were averaged.

A positive observation is that although slowly, especially the English curves

Mention type	PCEDT (Eval)		PCEDT (10-fcv)	
	monoling.	with EN	monoling.	with EN
Personal pron.	$^{66.54}_{67.24}$ 66.89	$^{70.33}_{66.81}$ 68.52	$^{64.33}_{61.81}$ 63.05	$^{67.07}_{63.58}$ 65.28
Possessive pron.	$^{68.91}_{67.55}$ 68.22	$^{73.97}_{73.09}$ 73.53	$^{72.41}_{71.92}$ 72.16	$^{75.74}_{74.69}$ 75.21
Refl. poss. pron.	$^{81.28}_{80.97}$ 81.13	$^{82.87}_{82.33}$ 82.60	$^{84.99}_{85.05}$ 85.02	$^{88.49}_{88.05}$ 88.27
Reflexive pron.	$^{62.24}_{50.00}$ 55.45	$^{60.00}_{50.00}$ 54.55	$^{66.86}_{56.66}$ 61.34	$^{66.96}_{55.54}$ 60.72
Zero subject	$^{73.25}_{52.93}$ 61.45	$^{77.60}_{54.95}$ 64.34	$^{70.55}_{57.42}$ 63.32	$^{75.72}_{59.52}$ 66.65
Zero in nonfin. cl.	$^{76.00}_{41.63}$ 53.79	$^{74.43}_{41.63}$ 53.39	$^{75.43}_{41.28}$ 53.36	$^{78.48}_{42.86}$ 55.44
Relative pron.	$^{80.35}_{79.34}$ 79.84	$^{81.80}_{80.29}$ 81.04	$^{81.62}_{79.92}$ 80.76	$^{83.51}_{81.67}$ 82.58
Total	$^{75.77}_{64.02}$ 69.40	$^{78.35}_{65.40}$ 71.29	$^{75.27}_{66.36}$ 70.53	$^{78.79}_{68.29}$ 73.17

Table 7.6: Comparison of the monolingual and the bilingually informed Treex CR on Czech. Scores were measured on the evaluation set of PCEDT, and on the full PCEDT excluding the evaluation set by 10-fold cross-validation.

are still growing, which is a promise of improving even more with more data. The ordering of anaphor types by performance of the system on them mostly does not change with growing size of the data. The only exception are reflexive pronouns in both languages. Especially for English, their curve is wilder than the others, exhibiting a big performance jump around 15,000 sentences. Recall from Section 2.4.1 that English reflexive pronouns occur in two distinct uses: basic and emphatic. Both of them are annotated for coreference in PCEDT, but their antecedents usually appear at different positions. We believe that the jump identifies the place where the model succeeded in learning to distinguish between them.

7.3 Bilingually Informed Resolution

In the following experiments, we train CR models using the cross-lingual features as presented in Section 7.1.4 in addition to the monolingual feature set. All the other settings remain the same as for the monolingual experiments (see Section 7.2). In other words, we build four specialized models with the hyperparameters defined as shown in Table 7.1.

The combination of employed datasets has slightly changed in comparison to the monolingual experiments. Cross-lingual experiments require a parallel corpus. All these experiments are therefore trained and tested on PCEDT, also for Czech.[5] Like in monolingual experiments, we train the models on the training set and evaluate them on the evaluation test set of PCEDT.

Nevertheless, due to the quantitative and qualitative analysis that we undertake in Section 7.4, we introduce another evaluation setup. Instead of the train–test split of the data, we run a 10-fold cross-validation on the full PCEDT data excluding the evaluation test section. The reason is that we wanted from the collected statistics to be as reliable as possible and offer enough examples,

[5]Note that the monolingual model for Czech was trained on PDT.

Mention type	PCEDT (Eval)		PCEDT (10-fcv)	
	monoling.	with CS	monoling.	with CS
Personal pron.	$^{75.25}_{68.77}$ 71.86	$^{78.17}_{69.61}$ 73.64	$^{75.57}_{71.09}$ 73.26	$^{78.12}_{72.60}$ 75.26
Possessive pron.	$^{79.29}_{78.76}$ 79.03	$^{80.34}_{79.57}$ 79.96	$^{79.43}_{78.89}$ 79.16	$^{81.45}_{80.95}$ 81.20
Reflexive pron.	$^{74.51}_{74.00}$ 74.25	$^{80.00}_{78.00}$ 78.99	$^{78.71}_{73.67}$ 76.11	$^{75.48}_{71.36}$ 73.36
Zero in nonfin. cl.	$^{64.11}_{51.20}$ 56.93	$^{65.93}_{51.76}$ 57.99	$^{65.95}_{57.13}$ 61.22	$^{67.70}_{58.21}$ 62.59
Relative pron.	$^{78.26}_{73.57}$ 75.84	$^{81.65}_{76.61}$ 79.05	$^{84.04}_{76.62}$ 80.16	$^{85.84}_{77.57}$ 81.50
Total	$^{71.13}_{62.62}$ 66.61	$^{73.29}_{63.61}$ 68.11	$^{72.68}_{66.42}$ 69.41	$^{74.61}_{67.70}$ 70.98

Table 7.7: Comparison of the monolingual and the bilingually informed Treex CR on English. Scores were measured on the evaluation set of PCEDT, and on the full PCEDT excluding the evaluation set by 10-fold cross-validation.

out of which we picked some to be presented in the thesis. At the same time, we wanted to avoid performing the analysis on the evaluation dataset by which we would inevitably collect too much information about the dataset.

Moreover, to estimate the upper bound for our approach, we utilized the PAWS section of PCEDT, which contains manual annotation of alignment between targeted coreferential expressions. Experiments on PAWS were also conducted using 10-fold cross-validation.[6]

7.3.1 Bilingually Informed vs. Monolingual

A central experiment in this chapter compares the bilingually informed approach on parallel data with the monolingual one. While the monolingual approach uses solely the target language features, the bilingually informed model combines them with both feature sets presented in Section 7.1.4 which capture counterparts in the aligned language. Coreference links in the aligned language have been resolved automatically by a monolingual CR model.

Tables 7.6 and 7.7 show the performance of both approaches on Czech and English, respectively, as a target language. They list the scores measured in a standard way on the evaluation test set of PCEDT, and by 10-fold cross-validation on the full PCEDT except for the evaluation set.

In overall, cross-lingual models succeed in exploiting additional knowledge from parallel data and perform better than the monolingual approach by 1.9 and 1.5 F-score points on Czech and English evaluation set, respectively. Scores achieved on the non-evaluation dataset are generally higher, also with a higher difference of 2.6 points on Czech. The results thus suggest that English is slightly more informative for Czech than vice versa.

The F-score improvement benefits mainly from a rise in precision, but recall also gets improved.

In both languages and consistently for both datasets, personal and possessive pronouns are the types that exhibit the greatest improvement. In Czech, the

[6] As PAWS is many times smaller than PCEDT, we increased the number of Vowpal Wabbit's passes over the data more or less proportionally from 5 to 225.

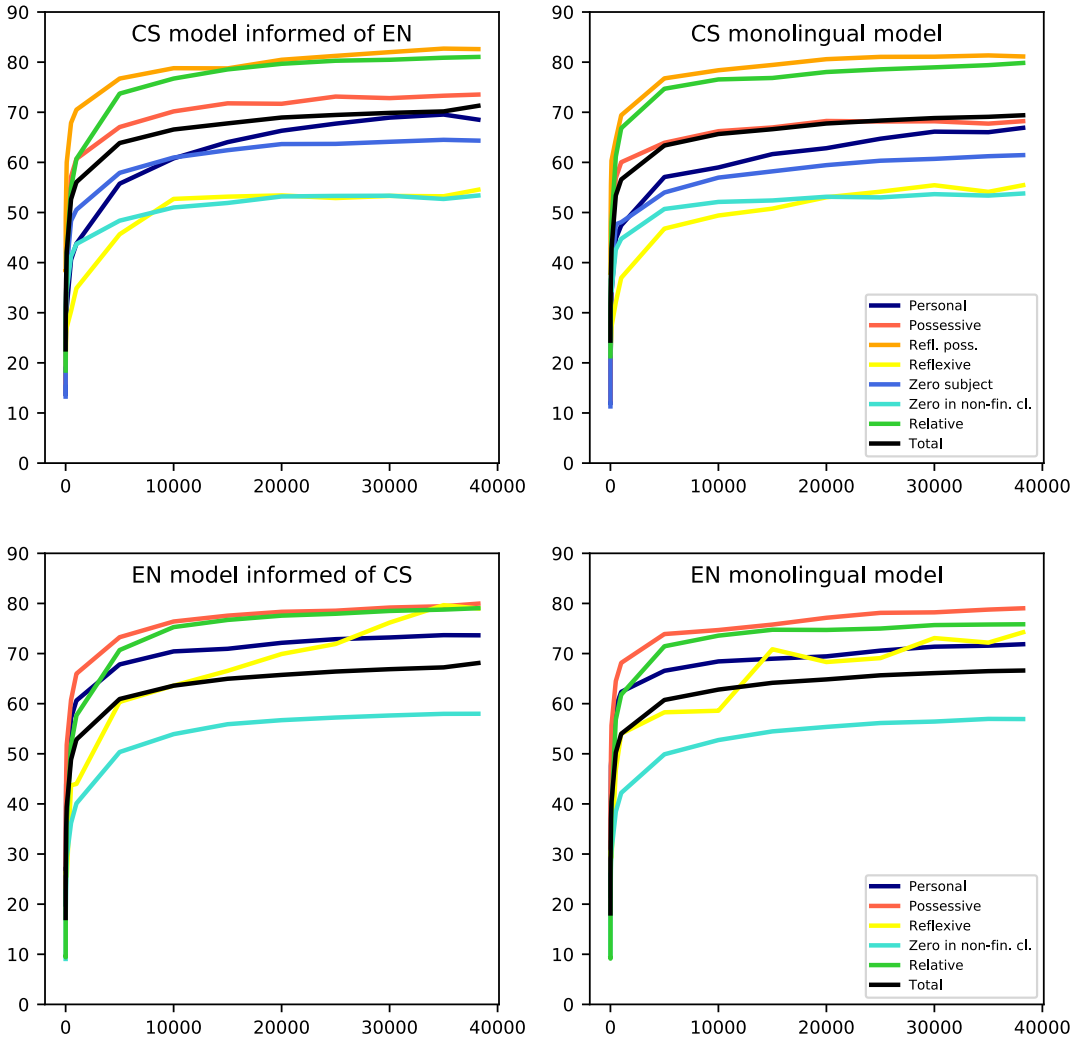

Figure 7.4: Comparison of learning curves calculated for both the bilingually informed and the monolingual system trained and evaluated on the training and the evalution set of PCEDT, respectively. The x-axis represents the number of sentences in the training data and the y-axis is the anaphora F-score.

top-scoring mention types include zero subjects, too. Nevertheless, there are some mention types, for which the differences vary across the datasets. English reflexive pronouns even exhibit contradicting results.

Learning curves. Figure 7.4 compares the learning curves calculated with the bilingually informed system as well as the monolingual system. We do not observe any substantial differences in the ordering of anaphor types by the systems' performance on them.

Let us now compare the overall F-scores of the two systems across different sizes of the training data. The comparison suggests that the information from the other language in the parallel corpus is equivalent to increasing the size of the data twice for English, and about 2.2-times for Czech.

Feature sets *aligned_all*	Feature sets *aligned_coref*	Czech	English
×	×	$^{75.77}_{64.02}$ 69.40	$^{71.13}_{62.62}$ 66.61
×	✓	$^{76.20}_{63.43}$ 69.23	$^{72.09}_{60.70}$ 65.90
✓	×	$^{77.57}_{66.88}$ 71.83	$^{72.06}_{64.94}$ 68.31
✓	✓	$^{78.35}_{65.40}$ 71.29	$^{73.29}_{63.61}$ 68.11

Table 7.8: Effect of combining the cross-lingual feature sets. Overall scores were measured on the evaluation set of PCEDT.

7.3.2 Contribution of Cross-lingual Feature Sets

Another experiment examines the partial contribution of the two sets of cross-lingual features: *aligned_all* and *aligned_coref*. Table 7.8 shows the overall performance of models based on combinations of the monolingual feature set with these two cross-lingual sets. Scores were measured on the evaluation set of PCEDT.

There are two messages that the results on both languages convey: (1) the *aligned_all* feature set seems to be forming the core of the bilingually informed approach, and (2) the *aligned_coref* feature set causes the scores to decrease a bit. Concerning the latter observation, the feature of aligned coreference seems to be positively effecting the precision (precision scores of the combination of all features is the highest) at the price of lowering the recall. However, the same experiments run on the development test data and by 10-fold cross-validation on the non-evaluation data suggest that the combination of all features is in fact outperforming the other settings. We therefore decided to use both cross-lingual feature sets in combination with monolingual features in all other bilingually informed experiments.

7.3.3 Alignment and Aligned Coreference Oracles

Performance of a bilingually informed system depends on quality of the following cross-lingual factors: (1) alignment, (2) coreference in the aligned language, (3) other tectogrammatical properties in the aligned language. This experiment demonstrates how much the cross-lingual method is possible to gain if quality of the first two factors reaches the quality of manual annotation, and thus attempts to set the upper bounds for resolvers in this configuration. Instead of using automatic annotation of alignment and coreference, we replace it by its manual alternatives.[7] Note that whereas improved coreference in the other language affects only a single feature, improved alignment may have an impact on all aligned features.

Manual coreference annotation in both Czech and English is available all over the PCEDT treebank. Performance of the cross-lingual method exploiting

[7]In fact, alignment is replaced only for selected coreferential expressions as specified in Section 6.1. It is one of the reasons why this should not be understood as an ultimate upper bound of alignment improvements for bilingually informed CR.

Auto / Manual		Czech		English	
Coref	Align	PAWS	PCEDT	PAWS	PCEDT
–	–	$^{62.80}_{52.73}$ 57.32	$^{75.77}_{64.02}$ 69.40	$^{59.46}_{50.43}$ 54.57	$^{71.13}_{62.62}$ 66.61
A	A	$^{63.63}_{52.56}$ 57.57	$^{78.35}_{65.40}$ 71.29	$^{62.45}_{51.65}$ 56.54	$^{73.29}_{63.61}$ 68.11
M	A	$^{65.01}_{53.55}$ 58.73	$^{80.73}_{67.45}$ 73.49	$^{64.10}_{52.87}$ 57.94	$^{75.04}_{65.21}$ 69.78
A	M	$^{68.02}_{55.37}$ 61.05	—	$^{64.32}_{54.15}$ 58.80	—
M	M	$^{70.36}_{57.27}$ 63.14	—	$^{66.45}_{55.95}$ 60.75	—

Table 7.9: Oracles of the current approach to bilingually informed CR, measured by alternating the manual/automatic annotation of alignment and aligned coreference.

manually aligned coreference thus can be measured on a standard scale. At the same time, alignment is manually annotated only within the PAWS section of PCEDT. Hence, effect of alignment can be precisely measured only on a small scale.

Table 7.9 shows the overall anaphora scores of the systems trained in all four combinations of manual and automatic alignment and aligned coreference. For comparison, we also report performance of the monolingual system in the top part of the table. Although the scores measured on PAWS are generally lower than on PCEDT,[8] an overall picture seems to be very similar. Results on PCEDT indicate that there is a room for improvement of CR in the target language that could be reached by increasing the quality of coreference in the aligned language. Results on PAWS show that increasing the quality of alignment (even only for coreferential expressions) is even more promising. A possible reason for this behavior might be that whereas quality of the aligned coreference affects only a single feature, quality of alignment links may result in a change of plenty of features. Moreover, higher quality of both the alignment and the aligned coreference seems to have a synergic effect, as indicated by the highest scores in the bottom line of Table 7.9. This performance gain is quite reasonable. The effect of improved coreference in the aligned language cannot express in its full power if the alignment between languages is not accurate enough.

7.4 Comparative Analysis of the Monolingual and Bilingually Informed CR

The results of experiments undoubtedly show the superiority of the cross-lingual CR over the monolingual one. Here, we delve more into the comparison of these two approaches. We inspect randomly sampled examples in an attempt to disclose what is behind the higher quality of the cross-lingual approach. In other words, what are the typical examples when the system takes advantage of the other language and, on the other hand, if there is a systematic case when the cross-

[8] A difference in score may be an artifact of different data sizes or different distributions of coreferential expressions there.

Mention type	Anaphoric				Non-anaphoric			
	Both ✓	Both ×	M > C	M < C	Both ✓	Both ×	M > C	M < C
Personal pron.	55.99	26.96	5.05	8.34	1.15	2.08	0.13	0.32
Possessive pron.	66.51	20.09	4.47	7.75	0.03	1.05	0.03	0.08
Refl. poss. pron.	82.45	9.59	2.64	4.27	0.11	0.89	0	0.05
Reflexive pron.	36.21	13.54	3.70	2.93	28.75	10.39	1.88	2.60
Zero subject	34.12	13.44	2.79	4.29	34.16	5.22	1.12	4.86
Zero in nonfin. cl.	68.54	12.62	2.94	5.24	3.82	6.08	0.42	0.32
Relative pron.	70.13	13.12	2.59	4.22	8.20	1.40	0.17	0.18
Total	53.76	14.20	3.00	4.73	17.96	3.52	0.61	2.22

Table 7.10: Comparison of resolution by the monolingual and the cross-lingual CR in Czech (M = Monolingual, C = Cross-lingual). The numbers are ratios (in %) of decision categories to which an anaphor candidate may fall.

lingual approach hurts. The analysis is carried out on the output of the systems run by 10-fold cross-validation on the complete PCEDT without its evaluation test section.

7.4.1 Quantitative Analysis

Let us start with a quantitative analysis of improvements and worsenings with respect to anaphoricity and type of the anaphor candidate. Tables 7.10 and 7.11 show for Czech and English, respectively, how often the cross-lingual system (denoted as C) is better than the monolingual (denoted as M). Each anaphor candidate falls to one of the four categories based on how C and M decided on the candidate:

- both decisions were the same and correct (Both ✓),
- both decisions were the same but incorrect (Both ×),
- negative decision change: M's decision was correct while C's decision was incorrect (M > C),
- positive decision change: M's decision was incorrect while C's decision was correct (M < C).

A decision is either assignment of the anaphor candidate to a coreferential entity[9] or labeling it as non-anaphoric. The tables also distinguish if the candidate is in fact anaphoric or non-anaphoric. Numbers in the tables represent proportions (in %) of these categories aggregated over all instances. Every row thus sums to 100%.

Distinguishing whether a mention that falls to a particular decision category is anaphoric or non-anaphoric allows us to directly relate this analysis to the

[9]Some of the anaphors that were assigned to the same entity (columns Both ✓ and Both ×) may have been in fact paired with different antecedents by each of the CR algorithms. As our anaphora score is agnostic to such changes, we do not distinguish such cases. In Tables 7.10 and 7.11, they are categorized as either Both ✓ or Both ×.

Mention type	Anaphoric				Non-anaphoric			
	Both ✓	Both ×	M > C	M < C	Both ✓	Both ×	M > C	M < C
Personal pron.	61.57	21.97	3.12	4.02	5.60	2.35	0.49	0.88
Possessive pron.	76.17	15.65	3.14	4.49	0.01	0.51	0.01	0.01
Reflexive pron.	69.78	15.00	7.17	5.22	0	2.83	0	0
Zero in nonfin. cl.	44.10	16.74	3.82	3.83	16.55	11.08	1.26	2.61
Relative pron.	58.06	10.46	2.12	2.94	23.53	1.82	0.26	0.80
Total	54.46	16.87	3.35	3.84	12.81	6.31	0.77	1.60

Table 7.11: Comparison of resolution by the monolingual and the cross-lingual CR in English (M = Monolingual, C = Cross-lingual). The numbers are ratios (in %) of decision categories to which an anaphor candidate may fall.

Prague anaphora scores shown in Tables 7.6 and 7.7. Note that while resolution on anaphoric mentions may have an effect on both the precision and the recall component of the anaphora score, resolution on non-anaphoric mentions affects only the precision.

Inspecting the overall distribution over decision categories, we observe that while in Czech 11% of decisions are changed, it accounts for 10% in English. More importantly, whereas we see over 64% of decisions changed positively in Czech, it corresponds to 55% of decisions in English. This accords with the evaluation scores measured on the examined dataset, where the cross-lingual system was able to outperform the monolingual system by 2.6 points in Czech, but only by 1.5 points in English.

Although in both languages around 2.5% of instances correspond to changed decisions on non-anaphoric mentions, the proportion of positive changes is substantially higher for Czech. Czech also exhibits a higher proportion of unchanged correct decisions than English.

The highest proportion of changed decisions is observed for personal pronouns (14% instances) and zero subjects in Czech (13%) and for reflexive pronouns in English (12%). Interestingly, whereas Czech personal pronouns and zero subjects are the mention types for which the cross-lingual system exhibits the largest improvement, English reflexive pronouns are the only mention type for which the resolution deteriorates with cross-lingual features. The systems' decisions differ the least for Czech reflexive possessive (7%) and English relative pronouns (6%). Here, we also observe a various effect on anaphora score. While the cross-lingual system's improvement is one of the smallest on Czech reflexive possessives, the small amount of changed decisions on relative pronouns suffices to achieve one of the biggest improvements among English coreferential expressions.

Basic reflexive pronouns in both languages are the only mention type, where the cross-lingual system is defeated more often than it wins, particularly on the anaphoric mentions. Although for Czech reflexive pronouns this excess of defeats is almost compensated by wins on non-anaphoric mentions, it is not sufficient. As a result, the cross-lingual system shows an anaphora score decrease for this category of mentions in both the Czech and English language (see Tables 7.6 and 7.7).

Apart from the Czech basic reflexives, Czech zero subjects and English zeros

are the only expressions, for which the cross-lingual system benefits more from the resolution of non-anaphoric mentions than of the anaphoric ones. Thanks to the resolution on non-anaphoric mentions, Czech zero subjects appear to lead also in the proportion of instances improved by the cross-lingual system (10%), compared to the proportion of the worsened ones (4%). And all these changes are reflected in the biggest improvement in terms of the anaphora F-score (see Table 7.6).

7.4.2 Qualitative Analysis

In the following, we scrutinize more closely what are the typical cases, where the cross-lingual system makes a different decision. For this analysis, we utilize the visual diagnostics provided by the Prague anaphora score as shown in Figure 4.3 in Section 4.4.3.

Let us start with a motivating example. Results in Tables 7.6 and 7.6 show that improvement of the bilingually informed system on Czech personal and possessive pronouns and zero subjects is much higher than on their English equivalents. This observation genuinely surprised us. We had expected the opposite. Our supposition was based on the fact that Czech grammatical gender is more evenly distributed over nouns. We assumed Czech gender could help filtering out the English antecedent candidates whose Czech counterparts do not match the pronoun's counterpart. Although this still may be true, obviously, there are even stronger factors that operate in the opposite direction – from English to Czech. And we examine them in the next paragraph.

Czech personal and possessive pronouns are the mention types that considerably benefit from the cross-lingual approach. The gender of the corresponding English pronoun appears to play an absolutely decisive role. Many times, gender of the Czech pronoun is masculine or feminine while gender of the English pronoun is neuter, as it is in Example 7.1. Recalling that the nature of gender in Czech and English differ (see Section 2.4.1), English pronoun's gender thus serves rather as an animacy feature, which cannot be reconstructed solely from the Czech pronoun. The correct antecedent is sometimes selected also with a help from the English pronoun's number.

(7.1)

Oponenti$_{\text{m.pl}}$	*soudce*$_{\text{m.sg}}$	*Borka*$_{\text{m.sg}}$	*zvolili*	***bojiště***$_{\textbf{n.sg}}$	*drželi*	***ho***$_{\textbf{mn.sg}}$
opponents	of judge	Bork	chose	the battlefield	held	it

Oponenti soudce Borka zvolili bojiště, drželi ho a udrželi si ho.

Mr. Bork's opponents chose the battlefield, held it and kept it.

The analysis also shows that English syntax, which is more strict and thus easier to reconstruct, often helps in determining the correct antecedent. Example 7.2 shows the case, where neither English gender nor number could affect the resolver's decision. The correct decision is rather a result of clear structure, where the objects in coordinated clauses very likely refer to the same entity.

(7.2)

kdo	*posbíral*	***plány***$_{\textbf{m.pl}}$	*skupin*$_{\text{f.pl}}$	*a*	*sesmolil*	***je***$_{\textbf{mfn.pl}}$	*do*	*iniciativy*
who	collected	plans	from groups	and	cobbled	them	into	an initiative

Van de Kamp je ten, kdo posbíral plány různých radikálních ekologických skupin a sesmolil je do jedné neohrabané iniciativy...

Mr. Van de Kamp is the one who collected the plans from the various radical environmental groups and cobbled them into a single unwieldy initiative...

Some of the possessive pronouns benefit from another syntax-related factor. Example 7.3 shows the case where the correct decision was very likely affected by the fact that the aligned English possessive pronoun ("*its* Opel line") is in a short context preceded by a construction with a possessive adjective ("*GM's* interest"). The possessivity factor also suppresses the unclear gender agreement in Czech ("*jeho* /its/" can be of masculine or neuter gender, whereas "*společnost* /company/" is of feminine gender and the gender of "*GM*" may be arbitrary).

(7.3)

zájem$_{\text{m.s}}$	***společnosti GM***$_{\textbf{fmn.s}}$	*o*	*společnost Jaguar*$_{\text{fm.s}}$	*odráží*	*touhu*$_{\text{f.s}}$			
interest	GM-company's	in	Jaguar company	reflects	a desire			
pomoci$_{\text{f.s}}$	*zpestřit*	*produkty*$_{\text{m.p}}$	***této společnosti***$_{\textbf{f.s}}$	*na*	*trhu*$_{\text{m.s}}$	*s*	*vozy*$_{\text{m.p}}$	.
to help	diversify	products	of this company	in	market	with	cars	.
jeho$_{\textbf{mn.s}}$	*série*	*Opel*						
its	line	Opel						

Zájem společnosti GM o společnost Jaguar odráží touhu pomoci zpestřit produkty této americké společnosti na rostoucím trhu s luxusními vozy. Jeho série Opel má zavedený image...

GM's interest in Jaguar reflects a desire to help diversify the U.S. company's products in the growing luxury-car segment of the market. Its Opel line has a solid image...

Zero subjects is another Czech mention type for which a large improvement of the cross-lingual approach is observed. Anaphoric zero subjects benefit from the aspects similar to those we mentioned for personal pronouns, e.g. gender and number of the anaphor, more strict syntactic constraints in English. English gender may be even more important here, as the gender of a zero subject is impossible to be recognized just from the form of the governing verb in the Czech sentence, if the verb is in present tense.

While inspecting a sample of changed decisions for English personal and possessive pronouns, we do not witness many examples of clear influence by Czech gender or number. As for the personal pronouns, influence of gender or number is most often combined with the pure fact that the English pronoun has an aligned counterpart in Czech. For many of such pronouns, the option that the pronoun is non-anaphoric can then be discarded. The strength of this aspect very likely accounts for the fact that the majority of decision changes with the highest confidence were in fact labeled as non-anaphoric by the monolingual system (e.g. in Example 7.4). Czech language side of the data thus helps correctly label these pronouns as anaphoric.

(7.4)

Compelled	***service***	*is*	*unconstitutional*	***It***	*is*	*also*	*unwise*	
Nucená	služba$_{f.s}$	je	protiústavní	$\varnothing_{f.s}$	Je	také	nerozumná	

Compelled service is unconstitutional. It is also unwise and unenforceable.

Nucená služba je protiústavní. Je také nerozumná a nevynutitelná.

Similarly, most of the improvements among English possessive pronouns do not result from additional information on gender and number from Czech. The cross-lingual system rather takes advantage of the cases where a reflexive possessive pronoun is a Czech counterpart of the English possessive pronoun (see Example 7.5), or the cases where the pronoun has no Czech counterpart at all. In all these cases, the subject of the clause in which the pronoun lies is a preferred antecedent.

(7.5)

Digital Equipment Corp.	*announced*	***its***	*line*	*of computers*	
společnost Digital Equipment Corp.	představila	svou	řadu	počítačů	

The hottest rivalry in the computer industry intensified sharply yesterday as Digital Equipment Corp. announced its first line of mainframe computers...

Nejžhavější rivalita v počítačovém průmyslu se včera notně přiostřila, když společnost Digital Equipment Corp. představila svou první řadu centrálních počítačů...

Back to the Czech zero subjects. Many of these mentions reconstructed during the automatic analysis are in fact spurious. It is usually a consequence of a parsing error, when the real subject of a clause is not recognized (e.g. the word "*společnosti* /companies/" in Example 7.6). This error subsequently propagates to a wrong decision of the monolingual resolver (the word "*zpráva* /report/" labeled as an antecedent). Any spurious zero subject may be correctly resolved in two ways: (1) labeling it as non-anaphoric, or (2) linking it to the expression that plays the same role in the sentence. We observe that 85% of the decisions corrected by the cross-lingual system are fixed in the former way. And a missing English counterpart of the spurious zero plays a significant role in such decisions.

(7.6)	*Avšak*	~~***zpráva***~~	*uvádí*	*že*	*společnosti*$_{\text{subj}}$	$\varnothing_{\sim\textbf{subj}}$	*platí*	*více*	*daní*
	But	the report	said	that	companies	–	are paying	more	taxes

Avšak zpráva uvádí, že ačkoliv společnosti platí více daní, mnoho jich stále platí méně, než činí zákonná sazba.

But even though companies are paying more taxes, many are still paying less than the statutory rate, the report said.

In a similar way, detection of English non-anaphoric zeros in non-finite clauses can be boosted by Czech features. If the zero is non-anaphoric, its governing clause usually remains non-finite in Czech or it turns into a nominal group. For instance, in Example 7.7 the entity which performs the act of "*hiring*" is not specified in the context of a given sentence, which is emphasized by the use of the noun "*nábor*" as a Czech translation of the participle. The automatically parsed structure of such cases is the same: since Czech non-subject zeros are rarely reconstructed by Treex linguistic pre-processing (see Section 4.2.1), there is usually no counterpart for the English zero to align with.

(7.7)	~~***Fear***~~	*of*	*AIDS*	*hinders*	$\varnothing_{\sim\textbf{actor}}$	*hiring*
	Strach	z	AIDS	komplikuje	–	nábor$_{noun}$

Fear of AIDS hinders hiring at few hospitals.

Strach z AIDS komplikuje nábor v několika nemocnicích.

In Section 2.4.2 we warned that the category of relative pronouns specified in terms of automatically set attributes may contain lots of pronouns that are in fact interrogative or fused. Such instances account for the majority of non-anaphoric English relative pronouns, correctly discovered by the cross-lingual system but not by the monolingual one.

Finally, we sought for the reasons of worsenings within a category of Czech and English reflexive pronouns. The worst changed decisions in Czech (made by the cross-lingual method and not by the monolingual one) are on the pronouns that ended up resolved as non-anaphoric. Most of the time these incorrectly labeled pronouns have no alignment to English, thus no cross-lingual features related to the anaphor can be activated. On the other hand, the English cross-lingual resolver adds the most serious mistakes by selecting a wrong antecedent. In these cases, the pronouns are most often aligned to their Czech counterparts and these counterparts are actually often correct. Yet, the choice of the English

antecedent seems to be random, regardless whether the Czech counterpart is labeled as coreferential with its correct antecedent, or the counterpart is any of the words *sám* or *samotný*, which should indicate emphatic use of the English reflexive pronoun.

7.5 Summary

In this chapter, we explored the possibilities of bilingually informed CR on Czech-English parallel data.

Firstly, we introduced Treex CR, the coreference resolver that targets the core expressions in both languages and is able to operate in a cross-lingual setting. It operates on the tectogrammatical layer, which allows it to address zeros and extract a rich feature set. In addition, it utilizes a sequence of mention-ranking models specialized at particular anaphor types. Its cross-lingual component enriches the features set with the features extracted from the nodes aligned to the anaphor and the antecedent candidates.

In its monolingual seeting, Treex CR outperforms the old approach used coreference annotation in CzEng 1.0 by a great margin. The improvement stems mainly from replacing heuristics with features weighed by machine learning, and by addressing some expressions that were not covered previously. Its comparison with the Stanford system shows a decent performance, which allows Treex CR to be used in further experiments. The fine-grained evaluation revealed inconsistent results on the two English datasets, though. The same holds for comparison of different approaches within the Stanford system. Since the domains of the two datasets barely differ, we presume that the annotation standards of the training data are a key factor in a resolver's performance.

With the bilingually informed setting, we managed to outperform the monolingual setting by 1.5 and 1.9 F-score points for English and Czech. The results thus suggest that English is more informative for Czech than vice versa. Learning curves showed that extracting information from the translations to the other language are equivalent to increasing the monolingual data twice. The analysis of individual cross-lingual features suggest that having the CR system for the aligned language is not necessary. The best results can be achieved without its output as a feature. We also showed that the potential of this method would be much higher if the alignment was even better.

As for the individual expression types, the biggest improvement is observed on personal and possessive pronouns in both languages, and zero subjects in Czech. The analysis revealed that the factors that mostly contribute to these improvements are inter alia:

- English pronouns which introduce the animacy information to the resolver of Czech pronouns
- English personal pronouns that help to identify Czech spurious zero subjects

Conversely, reflexive pronouns exhibit negative or contradictory results on different datasets.

8. Coreference Projection

In this chapter we experiment with the second cross-lingual approach to study coreference properties – coreference projection.

Projection techniques usually aim at acquiring linguistic resources for under-resourced languages. In our case, the motivation is different. We want to use a projection algorithm to measure coreference-based differences between the languages and thus contribute to building the language typology based on the ways how coreference and anaphora are expressed.

We first design the algorithm that performs the projection. As it is very simple, we expect that a more sophisticated algorithm (e.g. [Martins, 2015]) would perform better. However, thanks to its simplicity, it allows for tracing possible errors in projection easily. The error analysis can give us valuable information how the language relate or differ in this aspect.

We decided to carry out the projection experiments between the gold trees, projecting gold coreference links. The main motivation to start with such an upper bound[1] experiment were the results of the cross-lingual analysis in Chapter 5, which showed that English counterparts of Czech coreferential expressions are often heterogeneous and the same holds also in the opposite direction. As a result, these facts may have a negative effect on projection and the issues would be magnified if automatic annotation was used instead. Upper bound experiment should then indicate if it is worth trying to perform the same experiment between automatically annotated using the system coreference.

Having the gold projections, we train a coreference resolver on them. This is again an upper bound experiment of how much valuable information the CR system can adopt from the projected links.

The workflow of the projection experiments is sketched in Figure 8. Using a simple algorithm for projecting coreference between tectogrammatical trees that we propose in Section 8.1, we project gold coreference between gold trees of a parallel treebank. The resulting projections (no. (1) in the figure) are compared to the original gold coreference in the target language and thoroughly analyzed in Section 8.2. Finally, the projected links (or rather their monolingual projection to the target-language system trees) are exploited to train a CR system. Links produced by such a system (no. (2) in the figure) can be subsequently compared to a monolingual system trained on the original manual annotation of coreference as in Section 8.3.

8.1 Projection Mechanism

Our approach to coreference projection belongs to the corpus-based methods as introduced in Section 3.2.1. We work with manually translated word-aligned English-Czech parallel corpora, where we project coreference from one language side to the other. In fact, our approach is similar to the one adopted by multiple previous works [Postolache et al., 2006, de Souza and Orăsan, 2011, Wallin and Nugues, 2017, Grishina, 2017, i.a.]. Nevertheless, there is a substantial dif-

[1] Meaning the upper bound for the selected algorithm.

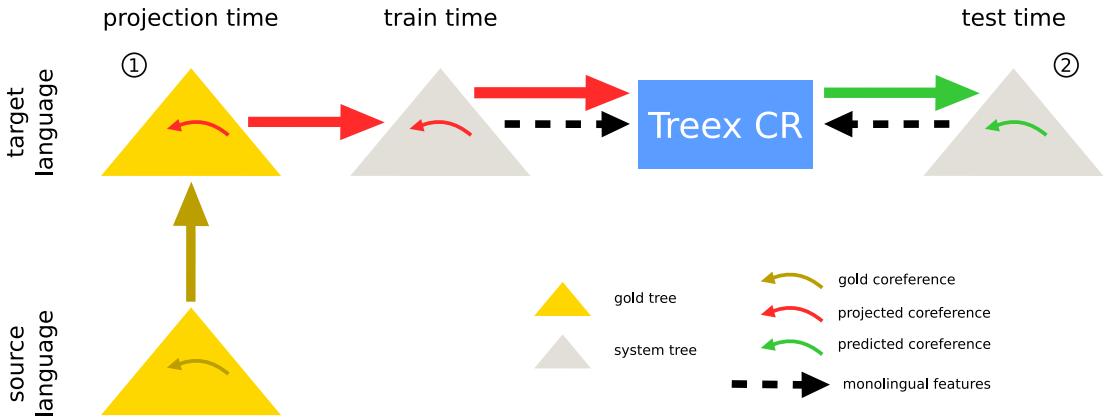

Figure 8.1: Workflow of our two projection experiments.

ference of our work compared to the others: our projection system operates on tectogrammatical representation.[2] It leads to three following consequences.

First, our system is able to address zero anaphora. Thorough cross-lingual analysis in Section 2.4 showed that many counterparts of Czech or English coreferential expressions are zeros. This likely holds for other pro-drop languages, too. Hence, it is surprising that the previous work on projection to Spanish [Rahman and Ng, 2012, Martins, 2015] or Portuguese [de Souza and Orăsan, 2011, Martins, 2015] did not accent this problem at all. In contrast, Ogrodniczuk [2013] is aware of dropped subjects in Polish and probably[3] allows making a coreference link from the inflectional ending of a governing verb, which signals morphological features of the dropped subject. In tectogrammatics, generated nodes serve this purpose instead.

Second, linguistic tools to carry out tectogrammatical analysis are required for the target language. Although in our work, the projection is conducted on gold trees, they will need to be replaced by automatically pre-processed trees in case of the real-world scenario. Tools for English and Czech are in fact well developed and integrated within the Treex framework [Popel and Žabokrtský, 2010] as discussed in Section 4.2.1. The pre-processing pipeline presented there is also not so difficult to be extended to other languages.[4]

Third, we do not define mention spans and project only heads of mentions. Focusing on projection, many of the previous works [Rahman and Ng, 2012, Postolache et al., 2006, Wallin and Nugues, 2017, i.a.] devote considerable space to answering the question of the proper strategy for determining boundaries of a projected mention. They seemingly alternate between two possible strategies, in which a target-language mention is defined: (1) as a minimal contiguous span that covers all words aligned to the source-language mention, or (2) by a target-language mention extractor and additional cross-lingual matching. In tectogram-

[2]To compare, our previous work [Novák et al., 2017] made use of surface-oriented projection.

[3]It is not completely clear from the paper.

[4]We did so for Russian and Polish to pre-process manually annotated trees in the PAWS corpus [Nedoluzhko et al., 2018]. Furthermore, a bunch of other languages got supported by introducing them to the TectoMT translation system [Dušek et al., 2015].

matics, the mention spans are not explicitly specified but they are implicitly understood to be the full subtree governed by the mention's head [Zikánová et al., 2015]. And the subtree may cover a large span of text, especially if it contains embedded clauses. This would inevitably caused problems of determining the target language mention span, particularly if the trees were automatically analyzed. We wanted to make our algorithm as simple as possible, so we decided to define mention only by its head.

```
   Input: SrcTrees = source language trees with coreference, TrgTrees
          = target language trees
   Output: TrgTrees = target language trees with projected coreference
 1 AllSrcChains ← GetCorefChains(SrcTrees) ;
 2 for SrcChain ∈ AllSrcChains do
 3 |   TrgLastAnte ← ∅;
 4 |   for SrcMention ∈ SrcChain do
 5 |   |   TrgMention ← GetAlignedAndInterlink(SrcMention,
   |   |    TrgTrees) ;
 6 |   |   if ∃ TrgMention then
 7 |   |   |   SrcAntes ← GetCorefNodes(SrcMention) ;
 8 |   |   |   TrgAntes ← GetAligned(SrcAntes, TrgTrees);
 9 |   |   |   if TrgAntes ≠ ∅ then
10 |   |   |   |   AddCorefNodes (TrgMention, TrgAntes) ;
11 |   |   |   end
12 |   |   |   else
13 |   |   |   |   AddCorefNodes (TrgMention, TrgLastAnte) ;
14 |   |   |   end
15 |   |   |   TrgLastAnte ← TrgMention;
16 |   |   end
17 |   end
18 end
```

Algorithm 1: Algorithm for coreference projection

Let us now describe our projection algorithm in greater details. In its description, we follow a pseudocode in Algorithm 1. The input of the algorithm are two lists of sentence-aligned and word-aligned tectogrammatical trees representing the same text in the source and the target language. First, a list of coreferential chains must be extracted from source trees (line 1). Every coreference chain is projected independently (lines 2-18) and processed mention by mention starting with the very first one, the one that has no outcoming link. For each mention, at the moment viewed as an anaphor, its counterpart in the target language is returned using the alignment (line 5). In case there are several nodes aligned to the anaphor, those which do not yet participate in a different chain are interlinked and only the very last mention is returned by the function GetAlignedAndInterlink. Unless no counterpart to the anaphor is found, counterparts of anaphor's direct antecedents are retrieved (line 7) and the algorithm adds a link between the anaphor's and antecedents' counterparts in the target language (line 10). If there are no antecedent's' counterparts, the last successfully projected anaphor from any of the previous iterations is used instead (line 13).

Mention type	English to Czech	Czech to English
Personal pron.	$^{81.92}_{52.05}$ 63.65	$^{86.52}_{67.45}$ 75.80
Possessive pron.	$^{72.85}_{59.87}$ 65.73	$^{89.33}_{60.88}$ 72.41
Refl. poss. pron.	$^{80.21}_{68.42}$ 73.85	—
Reflexive pron.	$^{87.36}_{11.96}$ 21.04	$^{89.17}_{22.20}$ 35.55
Demonstr. pron.	$^{57.43}_{35.19}$ 43.64	$^{55.81}_{42.73}$ 48.40
Zero subject	$^{78.71}_{59.06}$ 67.49	—
Zero in nonfin. cl.	$^{78.75}_{52.96}$ 63.33	$^{83.78}_{34.34}$ 48.71
Relative pron.	$^{74.71}_{51.18}$ 60.75	$^{85.02}_{70.00}$ 76.78
1st/2nd pers. pron.	$^{67.97}_{57.08}$ 62.05	$^{83.21}_{58.55}$ 68.73
Named entities	$^{38.04}_{62.07}$ 47.17	$^{80.29}_{39.04}$ 52.54
Nominal group	$^{50.15}_{37.80}$ 43.11	$^{61.70}_{47.27}$ 53.53
Other	$^{20.73}_{17.68}$ 19.09	$^{22.82}_{26.90}$ 24.69
Total	$^{53.86}_{44.86}$ 48.95	$^{71.31}_{46.47}$ 56.27

Table 8.1: Anaphora scores of gold coreference projected on PCEDT.

8.2 Gold Projections

Statistics of cross-lingual correspondences presented in Chapter 5 give some clues of the ways how coreference projection as we designed it will perform. Inspired by Postolache et al. [2006] and Grishina and Stede [2015], we carry out projection of manually labeled (gold) coreference. This can be also considered as an upper bound for projection of system coreference on system trees, at least when using our method – such projecting then can be applied to any Czech-English parallel data. The research on projecting gold data can be also beneficial for building parallel coreference-annotated corpora. Imagine we are extending an already annotated coreference corpus with translations to a new language. This experiment can give us an estimate of how much manual post-editing is needed on projected coreference if the languages are as distant as English and Czech.[5]

Table 8.1 shows the Prague anaphora scores ($^{P}_{R}F$) of gold coreference projection in both language directions. The experiments were conducted on full PCEDT 2.0 Coref [Nedoluzhko et al., 2016a] with supervised alignment of coreferential expressions (see Section 6.2). Note that apart from coreference, most of linguistic information including tree topology is also gold. It may have a profound effect especially on zeros.

The main observation is that with the overall F-scores around 50%, coreference projection seems to be a difficult problem. Moreover, let us emphasize that due to its optimistic setting this experiment is supposed to set an upper bound for such an approach. The projection from English to Czech offers a low quality both in the precision and the recall. In the opposite direction, though, the precision is substantially higher than the recall and for some expressions reaches almost

[5]Some approaches that have been adopted for the task of dependency tree projection are sketched in [Rosa, 2018]. However, it cannot be guaranteed that these approaches would work out in the case of coreference.

Mention type	Czech		English	
	mentions (#)	aligned (%)	mentions (#)	aligned (%)
Personal pron.	3038	68.30	14887	85.46
Possessive pron.	3777	84.06	9186	75.42
Refl. poss. pron.	4389	87.08	—	
Reflexive pron.	1272	17.69	484	69.21
Demonstr. pron.	3429	79.29	1492	87.47
Zero subject	16875	84.31	—	
Zero in nonfin. cl.	6151	80.15	29759	49.90
Relative pron.	15198	86.88	8170	91.62
1st/2nd pers. pron.	4415	85.14	4557	84.46
Named entities	18874	79.60	36833	96.83
Nominal group	80124	72.99	68866	95.35
Other	25735	53.43	14451	73.54
Total	183277	73.88	188685	84.49

Table 8.2: Proportion of aligned mentions among all gold target-language mentions in PCEDT.

90%. It seems to be a promising result, if the gold projection is aimed to serve as pre-annotation before it is manually post-edited. In such an application scenario, high precision is preferred since then the human post-editors may focus only on the unresolved instances.

In both languages, coreference information is obviously best preserved for central pronouns (except for basic reflexives). It agrees with findings of Grishina and Stede [2017], where they observed higher precision for pronouns than for nominal groups. To find out the real justification for our low projection scores, we undergo a detailed analysis of factors that may directly influence it.

8.2.1 Error Analysis

There are three main factors that contribute to the quality of coreference projection: (1) alignment, (2) mention matching, and (3) antecedent selection. Every projection error can be classified to one of the three types, based on the factor that caused it. Let us inspect what the degree of influence on the final score is for each of the types and what the most typical reasons behind these three types of errors are. We take a recall-oriented viewpoint.

Proportion of aligned mentions. Our algorithm is not able to project coreference to a mention which is not aligned. Unaligned mentions thus cause errors of the first type. Table 8.2 shows this proportion for each mention type. Extremely low proportion of aligned mentions is observed for Czech basic reflexive pronouns. In the majority of cases, unaligned Czech basic reflexives are a result of not expressing the corresponding argument of the governing verb in English. For instance, the Czech translation of the verb *to rent* in Example 8.1 requires explicit reflexive pronoun to signal the meaning that Exxon will pay for using the

tower, not that Exxon will receive money as its owner.

(8.1) *Společnost Exxon* ***si*** *pronajme část stávající výškové budovy.*
Exxon **[to self]** will rent part of an existing tower.
Společnost Exxon si pronajme část stávající kancelářské výškové budovy.
Exxon will rent part of an existing office tower.

Czech personal pronouns are also less frequently aligned than the other mention types. Similarly to the previous case, the reason is often that some arguments of an English verb are not explicitly mentioned. In general, missing English counterparts are a result of compact formulation of English sentences, like in Example 8.2. Compact language is, in our view, an inherent property of English as well as a feature of the specific journalistic style used in Wall Street Journal (WSJ). Moreover, one should not neglect the factor of the so-called *Explicitation Hypothesis* as formulated by Blum-Kulka [1986]: the redundancy expressed by a rise of cohesive explicitness in the target-language text might be caused by the nature of the translation process itself. Problems with aligning Czech personal pronouns are surprising since the analysis in Chapter 5 showed that they usually map straightforwardly to English personal pronouns and supervised aligner appears to achieve high accuracy on these expressions (see Table 6.1 in Section 6.2). Moreover, it was Czech personal pronoun, on which the bilingually informed CR achieves one of the highest improvement in comparison to monolingual system (see Table 7.6 in Section 7.3). This issue should be further investigated.

(8.2) *s pocity,* ***které je od práce odrazují.***
feelings which discourage them from working.
Pro prodejce není úplně snadné vypořádat se s pocity, které je od práce odrazují.
It can be hard for a salesperson to fight off feelings **of discouragement.**

As for English, we can see lower scores for zeros in non-finite clauses and reflexive pronouns, again. The non-finite clauses mainly consist of past and present participles. All the missing Czech counterparts of zeros in the past participle are due to the participle being represented as an adjective in Czech, thus having no valency arguments annotated. However, the decision of the annotators how to annotate such cases often seems too arbitrary and thus inconsistent (see Examples 2.21 and 2.22 in Section 2.4.3). The reasons behind a missing Czech counterpart of a zero in the present participle are more diverse. The counterpart is often missing even for the governing verb, not just for its zero argument (see Example 8.3). As opposed to the previous case of explicitation, this is an example of implicitation in the English-to-Czech translation.

(8.3) *řada makléřských firem se vzdala — této strategie.*
a number of brokerage firms pulled back from **using** this strategy.
Program traders were publicly castigated following the 508-point crash Oct. 19, 1987, and a number of brokerage firms pulled back from using this strategy for a while.
Programoví obchodníci byli po propadu burzy o 508 bodů dne 19. října 1987 veřejně káráni a řada makléřských firem se načas této strategie vzdala.

Missing alignment for English reflexives stems from three prevailing reasons. In the first group, there is no counterpart at all. The second group has surface counterparts, however they are not represented in the tectogrammatical tree by their own nodes. This concerns Czech basic reflexive pronouns, which are often

Mention type	Czech		English	
	P	R	P	R
Personal pron.	99.69	93.88	98.68	94.02
Possessive pron.	99.78	98.24	99.80	94.66
Refl. poss. pron.	100.00	98.56	—	
Reflexive pron.	100.00	79.56	100.00	37.31
Demonstr. pron.	91.76	70.87	81.76	71.11
Zero subject	99.76	89.37	—	
Zero in nonfin. cl.	100.00	85.07	99.84	85.06
Relative pron.	99.05	80.48	97.29	90.26
1st/2nd pers. pron.	88.42	88.53	94.20	79.73
Named entities	50.09	87.51	91.24	48.61
Nominal group	75.08	72.76	79.22	59.27
Other	32.90	54.14	41.81	53.24
Total	70.86	77.51	83.51	65.20

Table 8.3: Mention matching scores measured only on aligned target-language mentions in PCEDT 2.0 Coref.

hard to distinguish whether they are tightly bound to a verb or they fill an argument of the verb. The last group are English reflexive pronouns in their emphatic use (see Example 2.14 in Section 2.4.1). As shown in Example 5.10 from Section 5.1, they are often translated as words "*sám, samotný* /alone/", for which the automatic alignment often fails.

Mention matching. A coreference relation cannot be correctly projected unless both the anaphor and the antecedent match a mention in the target language. Not matching a target-language mention is an error of the second type. To check what is the impact of mention matching, we measure it solely on aligned target language mentions and show the results in Table 8.3.

In agreement with findings of Grishina and Stede [2017], we observe that pronouns and zeros in the top part of the table clearly approach matching precision of 100% in both projection directions. At the same time, named entities, nominal and other coreferential expressions in the bottom part of the table exhibit drops in precision. Based on our experiments we can thus agree on what Grishina and Stede suggest – the precision score of mention matching (and projection, consequently) decreases with increasing length of the mention span. Even though, in our approach, no mention boundaries are defined, they are inherently present. The same issue, which is in Grishina and Stede's approach manifested by the incorrect matching of the boundaries, is in our approach manifested by the wrong selection of the mention's head. In either of the treatments of mentions, the accuracy of the mention matching could be improved by introducing at least some syntactic knowledge from the target language. This finding accords with the findings by Grishina and Stede [2015]. Moreover, our algorithm should specify the mention and its boundaries better (e.g. by including nodes from the full subtree

of the mention's head, or at least its subset) to keep more information in order to infer the target language mention.

The problem with mention matching can be illustrated on an example of errors caused by the specific translation of some named entities in PCEDT. Let us point you to the mention matching scores on named entities in Table 8.3. Whereas in Czech their precision is much lower than recall, these rates are very similar but swapped in English. A closer insight to the data gives us a clear explanation illustrated in Example 8.4. A modifier, such as "*společnost* /company/", "*firma* /firm/", "*trh* /market/" etc., is added to many named entities in Czech. It sounds more natural and is easier to comprehend, especially if you are not familiar with the Wall Street Journal domain. This modifier is in fact a head of the complete named entity and, more importantly, it is the node that may corefer with others. Since it has no counterpart in English, no coreference is transferred to English, which results in recall errors for corresponding named entities. In the opposite projection, the English coreference link that is connected directly to one of the words in a given named entity finds its Czech counterpart, which is not a head of the mention, though. Hence, the Czech counterpart is in fact not coreferential, which causes a precision error. And because the head of the mention, the true coreferential node, is a word like "*společnost*", the recall error incurred by not covering it falls into the category of nominal groups, not named entities.

(8.4)

Burmah	*announced*	*[that it]*	***SHV***	*held*
společnost Burmah	prohlásila,	že ho	**společnost SHV**	držela

The holding of 13.6 million shares is up from a 6.7% stake that Burmah announced SHV held as of last Monday.

Vlastnictví 13.6 milionu akcií je nárůst oproti 6.7% podílu, o kterém společnost Burmah prohlásila, že ho společnost SHV držela k minulému pondělí.

Another problem with the mention matching is observed on English reflexives, which exhibit a dramatic fall in recall. These errors are again incurred for instances that translate to Czech expressions "*sám, samotný* /alone/". ven if they are correctly aligned, these Czech expressions do not belong to those annotated for coreference. Therefore, no links can be projected.

Antecedent selection quality. If both the anaphor and the antecedent are correctly matched to some target-language mentions but these mentions belong to distinct chains, an error of the third type is incurred. Table 8.4 shows the anaphora scores calculated on the same data as used until now, but only on correctly matched mentions. It accounts for around 49% of all coreferential links in Czech and 53% in English.

All F-scores move around 90% and more. The only exception is the category of Czech demonstrative pronouns. The reasons behind its errors are various, including annotators' errors and alignment errors. But they are often caused by the relatively free nature of demonstrative pronouns, which can refer to nominal groups, predicates, larger segments as well as entities outside the text. The free nature then allows the annotators to mark different (but somehow related) mentions as antecedents, especially when a different syntax structure of the languages encourages it. For instance, in Example 8.5 both expressions "*the exchange*" and "*volume*" are in some sense possible as the antecedent of "*it*". The same holds for the Czech translation.

Mention type	English to Czech	Czech to English
Personal pron.	$^{95.34}_{93.92}$ 94.63	$^{95.20}_{90.59}$ 92.83
Possessive pron.	$^{95.64}_{94.76}$ 95.20	$^{95.45}_{90.61}$ 92.97
Refl. poss. pron.	$^{97.25}_{96.53}$ 96.89	—
Reflexive pron.	$^{95.60}_{92.68}$ 94.12	$^{93.86}_{89.92}$ 91.85
Demonstr. pron.	$^{84.79}_{82.40}$ 83.58	$^{91.41}_{88.25}$ 89.81
Zero subject	$^{95.71}_{94.78}$ 95.25	—
Zero in nonfin. cl.	$^{94.96}_{93.40}$ 94.17	$^{92.94}_{89.24}$ 91.05
Relative pron.	$^{95.43}_{91.95}$ 93.66	$^{97.35}_{93.80}$ 95.54
1st/2nd pers. pron.	$^{91.77}_{87.34}$ 89.50	$^{94.26}_{89.01}$ 91.56
Named entities	$^{95.82}_{94.15}$ 94.98	$^{95.03}_{91.88}$ 93.43
Nominal group	$^{90.52}_{87.46}$ 88.96	$^{90.60}_{84.37}$ 87.38
Other	$^{80.08}_{78.42}$ 79.24	$^{70.72}_{70.05}$ 70.38
Total	$^{92.91}_{90.69}$ 91.79	$^{92.73}_{88.31}$ 90.47

Table 8.4: Anaphora scores of gold coreference projected within PCEDT 2.0 Coref, measured only on correctly matched mentions.

(8.5) ***the exchange*** *run up* *volume* *of X contracts.* *later,* ***it*** *was* *Y.*
burza dosáhla **objemu** X smluv. později **to** bylo Y.
. . . and the options exchange had run up volume of 1.1 million contracts. A year later, it was 5.7 million.
. . . a opční burza dosáhla objemu 1.1 milionu smluv. O rok později to bylo 5.7 milionu.

8.2.2 Effect of Alignment Quality

Until this point, the projection experiments were all conducted on PCEDT 2.0 Coref aligned by the supervised alignment. We will now alternate the alignment, using the original (see Section 4.2.3) and also the manual alignment (on the PAWS section only, see Section 6.1). By doing this, we can investigate the impact of alignment on coreference projection. In other words, we undertake an extrinsic evaluation of alignment quality, which will complement its intrinsic evaluation and the coreference-related correspondence scores presented in Section 6.2.

Figure 8.2 shows the effect of alignment style on quality of coreference projection for selected anaphor types as well as in total. It depicts the mention-matching F-score and the Prague anaphora F-scores, embracing all three factors that we examined in Section 8.2.1. For many mention types (e.g. Czech zero subjects or English relative pronouns shown in the figure), the rising supervision of the alignment method affects both scores by a very similar margin. Improvements in alignment hence appear to contribute mostly to better mention matching. Nonetheless, there are types for which improved alignment helps the antecedent selection, as well. For instance, for Czech personal pronouns the mention-matching F-score decreases for supervised alignment, while the anaphora score increases. The explanation lies in spurious links made by the original alignment. That is, some pronouns that in fact have no counterpart in English were spuriously aligned to a

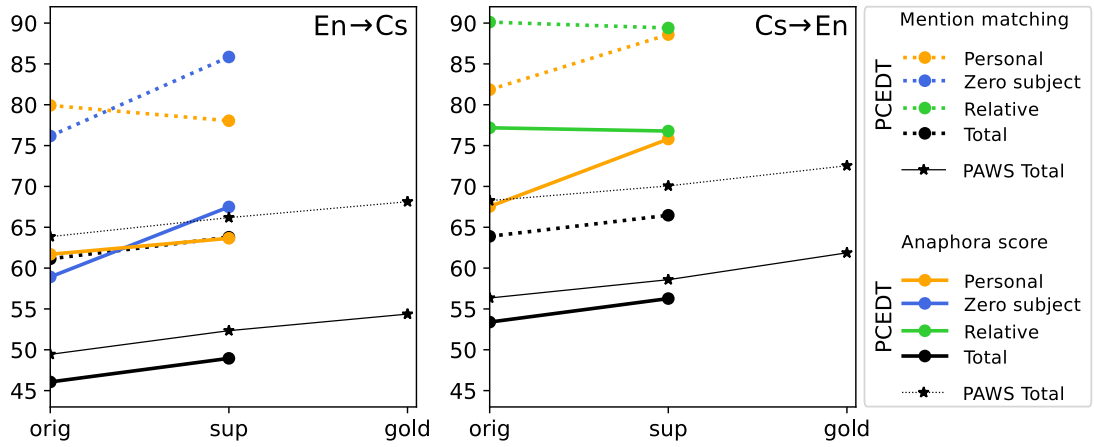

Figure 8.2: Effect of alignment style on projection quality, measured by Prague anaphora F-score and mention matching F-score. The scores are measured on full PCEDT as well as just on its PAWS section.

mention referring to a different entity. English relative pronouns are on the other hand the only expression, for which the transition from the original to the supervised alignment caused a decrease in quality of projection, though the marginal one.

Overall scores show that the supervised alignment improves the projection by 2-3% in absolute. Projection through the supervised alignment thus performs just in the middle of the projection qualities through the original and the manual alignment. It seems to be a marginal improvement, yet the scores on the data with manually aligned suggest that we reached 50% error reduction of alignment errors. Nevertheless, even in the unrealistic scenario of using the manual alignment, the performance of projection is still far away from being satisfactory.

8.3 Resolver Trained on Projected Gold Coreference

In this experiment, we exploit the gold coreference links projected through the supervised alignment acquired in the previous experiment to train a coreference resolver that can be applied to any target-language text. As a coreference resolver, we employ Treex CR using exactly the same setting of hyperparameters as for the monolingual resolution (see Section 7.2). We thus solely replace the "gold" annotation of coreference relations in the training data.

Table 8.5 shows the performance of a resolver trained on PCEDT training set and evaluated on its evaluation test set.[6] We observe a dramatic drop of 12 F-score points for Czech coreference, but for English it is even more – 21 points. The results thus suggest that the Czech resolver can leverage the coreference

[6]Note that the total numbers for the projections of gold coreference differ from those in Table 8.1, because here they are aggregated only over the anaphor types targeted by Treex CR.

Mention type	Czech		English	
	Projected	Resolved	Projected	Resolved
Personal pron.	$^{81.92}_{52.05}$ 63.65	$^{68.57}_{17.67}$ 28.10	$^{86.52}_{67.45}$ 75.80	$^{86.84}_{43.87}$ 58.29
Possessive pron.	$^{72.85}_{59.87}$ 65.73	$^{57.45}_{40.90}$ 47.78	$^{89.33}_{60.88}$ 72.41	$^{84.80}_{45.43}$ 59.16
Refl. poss. pron.	$^{80.21}_{68.42}$ 73.85	$^{65.17}_{60.97}$ 63.00	—	—
Reflexive pron.	$^{87.36}_{11.96}$ 21.04	$^{0.00}_{0.00}$ 0.00	$^{89.17}_{22.20}$ 35.55	$^{100.00}_{4.00}$ 7.69
Zero subject	$^{78.71}_{59.06}$ 67.49	$^{68.52}_{40.79}$ 51.14	—	—
Zero in nonfin. cl.	$^{78.75}_{52.96}$ 63.33	$^{69.12}_{30.20}$ 42.04	$^{83.78}_{34.34}$ 48.71	$^{80.44}_{10.76}$ 18.98
Relative pron.	$^{74.71}_{51.18}$ 60.75	$^{80.82}_{44.77}$ 57.62	$^{85.02}_{70.00}$ 76.78	$^{84.45}_{54.04}$ 65.90
Total	$^{77.43}_{55.23}$ 64.47	$^{70.31}_{41.35}$ 52.07	$^{85.83}_{50.66}$ 63.72	$^{84.51}_{28.74}$ 42.90

Table 8.5: Anaphora scores of the CR system trained on gold coreference in PCEDT 2.0 Coref projected via supervised alignment. The column "Projected" shows quality of the gold coreference projection described in Section 8.2, and the column "Resolved" is performance of the CR system trained on the projections.

projected from English more than the English resolver from Czech. Recall that we observed very similar behavior for bilingually informed CR (see Section 7.3). In other words, there are two different cross-lingual approaches that suggest the same: *With respect to coreference, English is more informative for Czech than vice versa.*

In projection of the gold coreference in Section 8.2, we noticed that the projection from Czech to English achieves relatively high precision and suffers from recall. If we train a CR model from such data, the recall decreases even more, but the precision remains nearly the same, making thus the gap between the two wider than 55 percentage points. Due its high performance around 85% and the recall for the most frequent pronouns still around 50%, it might be found useful for annotators as an automatic pre-processing step. In Czech, the precision and recall of projected gold links were less distant from each other. Modeling the projected links lowered both the precision and the recall and widened the gap, since the recall decreased more.

On individual categories, the biggest fall of 35 points is attributed to Czech personal pronouns. It might be justified by a heterogeneous nature of Czech personal pronouns. They can appear in any morphological case, be bound with any preposition and due to zero subjects the personal pronouns in subject position, which have a strong tendency to corefer with a subject from a previous sentence, account for only a small proportion in this type. Combined with the noisy training data, it will probably throw the model into a considerable confusion. Furthermore, note that the model for personal pronouns is shared also with possessive pronouns and zero subject, which may have an effect on the resulting score.

Conversely, the smallest decrease of only 3 points is experienced on Czech relative pronouns. The properties of Czech relative pronouns seem to be so clear that the noise in the training data causes just a minimal additional harm to the model.

The category of reflexive pronouns is both small (see Table 4.3 in Section 4.1) and heterogeneous. The consequence is that the Czech model learned to resolve every such pronoun as non-anaphoric and the English model is very close to that point, achieving only 4% in recall.

8.3.1 Projected vs. Monolingual Coreference

Looking at the mediocre results of coreference projection, it would be interesting to know how much target-language training data we need to achieve a similar score with a coreference resolver. Figure 8.3 shows learning curves of the system trained on projected links and the monolingual system trained on the original coreference annotation. In all the configurations, the training examples were sampled from the PCEDT training set and its evaluation test set served to measure the performance. This was repeated in three rounds and scores associated with the same size of the training set were averaged.

Let us inspect how many sentences it is necessary to annotate with coreference relations in order to achieve the same performance (in terms of the F-score) as we can obtain by training from almost 40,000 sentences with projected coreference. Averaging over all focused types, annotating less than 500 sentences should in fact suffice. It holds for both Czech and English and also for the majority of the individual mention types. The only Czech mention type for which it might pay off performing projection are zero subjects. Otherwise, 1,000–5,000 sentences need to be annotated to get similar performance. In English, the only expression type that is worth projecting are relative pronouns, which would also require to annotate nearly 5,000 sentences in case of training a native model. On the other hand, the models for reflexive pronouns are really poor in both languages and one would outperform them with just some tens of sentences. A few tens of annotated sentences would be also sufficient for English zeros in non-finite clauses and Czech personal pronouns.

8.4 Summary

In this chapter we introduced the second approach to cross-lingual treatment of coreference – coreference projection. We explored its possibilities in two experiments: projection in completely "gold" scenario, and learning a CR on projections from the first experiment.

The first experiment was, to the best of our knowledge, the first properly analyzed experiment of this kind on Czech-English texts. We performed annotation projection on manually translated and almost entirely manually annotated data (the supervised alignment was the only exception). The reason for such overly optimistic setting was to track the impact of correspondences collected in Chapter 5 on the projections performed by a simple algorithm, without introducing noise from the automatic methods.

Taking into account the optimistic setting, the performance we observed was not as optimistic. However, our factored error analysis revealed interesting facts manifested in three types of projection errors.

First, many errors are caused by linguistic differences that can be a result of inherent nature of the languages, but also of the translation. This mainly

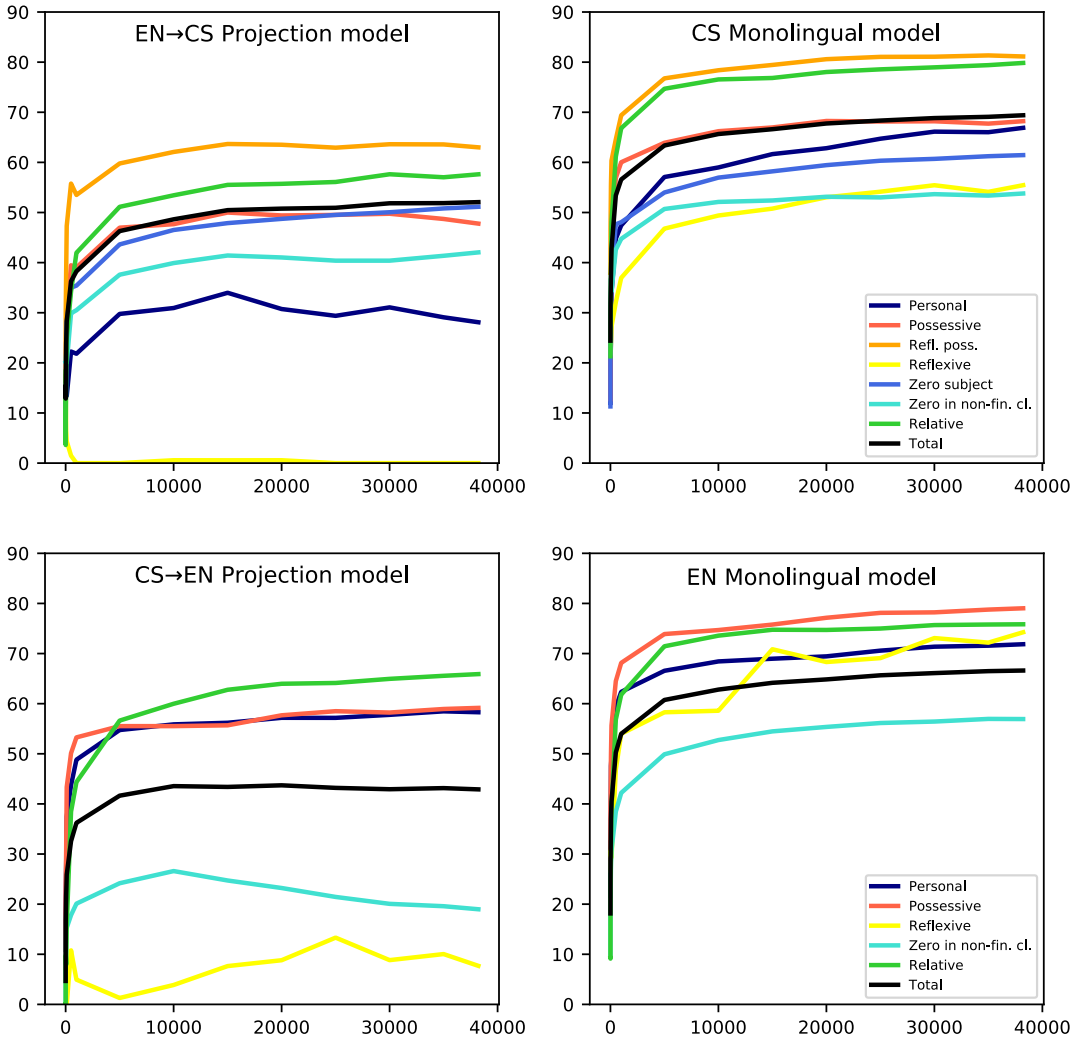

Figure 8.3: Learning curves of the projection models in comparison with the monolingual models on the PCEDT evaluation test set. The x-axis represents the number of sentences in the training data and the y-axis is the anaphora F-score.

concerns missing counterparts of Czech reflexive and personal pronouns. It is interesting especially in the latter case, as the statistics on correspondence, and the performance of supervised alignment and bilingually informed CR suggest that Czech personal pronouns can be aligned easily.

The second type of errors result from the annotation differences, such as the unclear boundaries between deverbative adjectives and participles (see Section 2.4.3), non-anaphoricity of the Czech expressions "*sám, samotný*", or unclear antecedents of some demonstrative pronouns.

The reason for the last type of errors was a simplistic treatment of mentions by our algorithm. Modifying the algorithm to incorporate the dependent members of the mention head and the syntactic information from the target language should improve its performance. However, we definitely do not expect the modification

to fix all the errors related to mention matching.

The experiment with training a CR system on projections shows that the quality of the coreference annotation deteriorates by 12–21 F-score points. A very similar margin was observed between the CR system trained on projection and the monolingual CR system trained on original manual annotations of coreference.

9. Conclusion

In this thesis, we presented two computational approaches to study the properties of coreference from the cross-lingual perspective: the bilingually informed coreference resolution and coreference projection. The motivation of our work was twofold.

- We wanted to contribute to our project, which concerns with contrasting languages (currently English, Czech, Russian and Polish) with respect to how they express coreference. The aim of our work was to find out if we can adopt the two cross-lingual computational methods in order to quantify the similarities and differences of the languages.

  The results of *bilingually informed resolution* confirmed that this method can take advantage of differences between languages. Our experiments disclosed that English is more informative for Czech than vice versa. For instance, English can help filter out the antecedent candidates based on their animacy property and identify spurious zero subjects in Czech.

  Coreference projection also highlighted the most important linguistic and annotation-style differences, where projection between Czech and English fails, even though some of its errors resulted from an overly simplistic nature of our projection algorithm. Nevertheless, the models trained on projected links showed that Czech is able to leverage projections from English more than vice versa.

  In essence, there are two completely different cross-lingual methods showing that *English is more informative for Czech than vice versa*. Even this observation should be interesting enough. However, it will start to be even more interesting as soon as we apply the presented methods also to other language pairs within the PAWS corpus. The results can bring us more information about coreference-related differences within the family of Slavic languages.

- We also wanted to explore bilingually informed resolution as a means to obtain automatic coreference annotation on parallel corpora.

  Our experiments revealed that the *bilingually informed resolution outperforms the monolingual approach* for both combinations of Czech and English. Therefore, applying them on parallel corpora should result in their better annotation.

  Parallel corpora automatically annotated with coreference can then serve as an additional source of data for semi-supervised machine learning techniques, and in this way push the information collected by a bilingually informed system to a monolingual coreference resolver. Our experiments can be also viewed as a proof of concept that the methods exploiting the differences of languages can successfully work also for coreference resolution. Consequently, the differences can be in the future approached by more sophisticated learning methods.

As a side product of this work, we managed to improve the monolingual resolver for Czech. In addition, we designed a method based on supervised learning that targets selected coreferential expressions and produces the alignments of much better quality than the traditional approaches. Finally, we collected a dataset of manually annotated correspondences between Czech and English coreferential expressions that can be used for further empirical or computational linguistic studies.

Bibliography

Czech National Corpus – SYN2005. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic, 2005.

Mariana S. C. Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. Aligning Opinions: Cross-Lingual Opinion Mining with Dependencies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, pages 408–418, Stroudsburg, PA, USA, 2015. The Association for Computer Linguistics.

Chinatsu Aone and Scott William Bennett. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 122–129, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87, 1988.

Amit Bagga and Breck Baldwin. Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473, 2014.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013.

Shane Bergsma and David Yarowsky. NADA: A Robust System for Non-referential Pronoun Detection. In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, pages 12–23, Berlin, Heidelberg, 2011. Springer-Verlag.

Anders Björkelund and Jonas Kuhn. Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Shoshana Blum-Kulka. Shifts of Cohesion and Coherence in Translation. In *Interlingual and intercultural communication*, pages 17–35, Tübingen, Germany, 1986. Gűnter Narr.

Andreea Bodnari. *Joint Multilingual Learning for Coreference Resolution*. Thesis, Massachusetts Institute of Technology, 2014.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. CzEng 1.0, 2011.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May 2012. ELRA, European Language Resources Association.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6, 2016.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

David Burkett and Dan Klein. Two Languages Are Better Than One (for Syntactic Parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 877–886, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

David Burkett, John Blitzer, and Dan Klein. Joint Parsing and Alignment with Weakly Synchronized Grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–135, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Eugene Charniak and Micha Elsner. EM Works for Pronoun Anaphora Resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, 2009. Association for Computational Linguistics.

Chen Chen and Vincent Ng. Linguistically Aware Coreference Evaluation Metrics. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1366–1374, Nagoya, Japan, 2013. Asian Federation of Natural Language Processing.

Chen Chen and Vincent Ng. Chinese Overt Pronoun Resolution: A Bilingual Approach. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1615–1621, Québec City, Québec, Canada, 2014. AAAI Press.

Kevin Clark and Christopher D. Manning. Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, 2015. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 643–653, Berlin, Germany, 2016. Association for Computational Linguistics.

Michael Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Aron Culotta, Michael Wick, and Andrew McCallum. First-Order Probabilistic Models for Coreference Resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, Rochester, New York, 2007. Association for Computational Linguistics.

František Daneš and Karel Hausenblas. Přivlastňovací zájmena osobní a zvratná ve spisovné češtině. *Slavica Pragensia*, 4:191–202, 1962.

Hal Daumé, III, John Langford, and Daniel Marcu. Search-based Structured Prediction. *Machine Learning*, 75(3):297–325, 2009.

Robert De Beaugrande and Wolfgang U. Dressler. *Introduction to Text Lingustics*. Longman, London, UK, 1981.

José G. C. de Souza and Constantin Orăsan. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, pages 59–69, Berlin, Heidelberg, 2011. Springer-Verlag.

Pascal Denis and Jason Baldridge. Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester, New York, 2007a. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. A Ranking Approach to Pronoun Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1588–1593, San Francisco, CA, USA, 2007b. Morgan Kaufmann Publishers Inc.

Pascal Denis and Jason Baldridge. Specialized Models and Ranking for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Greg Durrett and Dan Klein. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA, 2013. Association for Computational Linguistics.

Greg Durrett and Dan Klein. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *TACL*, 2:477–490, 2014.

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada, 2012. Association for Computational Linguistics.

Ondřej Dušek, Jan Hajič, and Zdeňka Urešová. Verbal Valency Frame Detection and Selection in Czech and English. In *The 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

Ondřej Dušek, Luís Gomes, Michal Novák, Martin Popel, and Rudolf Rosa. New Language Pairs in TectoMT. In *Proceedings of the 10th Workshop on Machine Translation*, pages 98–104, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. Latent Trees for Coreference Resolution. *Computational Linguistics*, 40(4): 801–835, 2014.

Jenny Rose Finkel and Christopher D. Manning. Enforcing Transitivity in Coreference Resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 45–48, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. Semi Supervised Preposition-Sense Disambiguation using Multilingual Data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2718–2729, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Yulia Grishina. Combining the output of two coreference resolution systems for two source languages to improve annotation projection. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 67–72, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 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, 2015. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. Multi-source annotation projection of coreference chains: Assessing strategies and testing opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 41–50, Valencia, Spain, 2017. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 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)*, pages 3191–3198, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

Aria Haghighi and Dan Klein. Coreference Resolution in a Modular, Entity-centered Model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. Prague Dependency Treebank 2.0, 2006.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan
v Snaidauf, Emanuel Beška, Jakub Kracmar, and Kamila Hassanová. Prague arabic dependency treebank 1.0, 2009.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank 2.0, 2011.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey, 2012. European Language Resources Association.

Jan Hajič, Petr Pajas, Pavel Ircing, Jan Romportl, Nino Peterek, Miroslav Spousta, Marie Mikulová, Martin Grůber, and Milan Legát. Prague DaTabase of Spoken Czech 1.0, 2017.

Eva Hajičová. Focussing – A Meeting Point of Linguistics and Artificial Intelligence. In *Artificial Intelligence 2. Methodology, Systems, Applications*, pages 311–321, Amsterdam, The Netherlands, 1987. North-Holland.

Eva Hajičová, Petr Kuboň, and Vladislav Kuboň. Hierarchy of Salience and Discourse Analysis and Production. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, pages 144–148, Stroudsburg, PA, USA, 1990. Association for Computational Linguistics.

Eva Hajičová, Jarmila Panevová, et al. Coreference in the grammar and in the text 1. *Prague Bulletin of Mathematical Linguistics*, 44:3–20, 1985.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, UK, 1976.

Sanda M. Harabagiu and Steven J. Maiorano. Multilingual Coreference Resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 142–149, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

Martin Wittorff Haulrich. *Data-Driven Bitext Dependency Parsing and Alignment*. PhD thesis, Copenhagen Business School, 2012.

Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

Jerry R. Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, 1997.

Gordana Ilic Holen. Critical Reflections on Evaluation Practices in Coreference Resolution. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 1–7, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping Parsers via Syntactic Projection Across Parallel Texts. *Natural Language Engineering*, 11(3):311–325, 2005.

Hamidreza Kobdani. *A Modular Framework for Coreference Resolution*. PhD thesis, Universität Stuttgart, 2012.

Lucie Kučová and Zdeněk Žabokrtský. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora. In *Lecture Notes in Artificial Intelligence, Proceedings of the 8th International Conference, TSD 2005*, volume 3658 of *Lecture Notes in Computer Science*, pages 93–98, Berlin / Heidelberg, 2005. Springer.

Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, and Oliver Čulo. Annotation of Coreference in the Prague Dependency Treebank. Technical Report TR-2003-19, 2003.

Emmanuel Lassalle and Pascal Denis. Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2274–2280, Austin, Texas, 2015. AAAI Press.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916, 2013.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

Els Lefever and Veronique Hoste. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Els Lefever and Véronique Hoste. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Els Lefever, Véronique Hoste, and Martine De Cock. ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 317–322, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A Mention-synchronous Coreference Resolution Algorithm Based on the Bell Tree. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, pages 136–143, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Penn Treebank 3, 1999.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of the Twelfth EAMT Conference*, pages 102–111, Hamburg, Germany, 2008. HITEC e.V.

André F. T. Martins. Transferring Coreference Resolvers with Posterior Regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, pages 1427–1437, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

Sebastian Martschat. *Structured Representations for Coreference Resolution*. PhD thesis, University of Heidelberg, 2017.

Sebastian Martschat and Michael Strube. Latent Structures for Coreference Resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418, 2015.

Joseph F. McCarthy. *A Trainable Approach to Conference Resolution for Information Extraction*. PhD thesis, Amherst, MA, USA, 1996.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural*

Language Processing, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Ševčíková, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Technical Report 3.1, 2007.

Ruslan Mitkov. *Anaphora Resolution*. Longman, London, 2002.

Ruslan Mitkov and Catalina Barbu. Using Bilingual Corpora to Improve Pronoun Resolution. *Languages in contrast*, 4(2), 2003.

MUC-6. *Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1995.

MUC-7. *Proceedings of the 7th Conference on Message Understanding*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1998.

Anna Nedoluzhko. *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Česká republika, 2011.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Prague Czech-English Dependency Treebank 2.0 Coref, 2016a.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Coreference in prague czech-english dependency treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris, France, 2016b. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, and Maciej Ogrodniczuk. PAWS: A Multilingual 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, 2018. Association for Computational Linguistics.

Václav Němčík. Anaphora Resolution. Master's thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2006.

Václav Němčík. The Saara Framework: An Anaphora Resolution System for Czech. In *RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing*, pages 49–54, Brno, Czech Republic, 2009. Masaryk University.

Vincent Ng. Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Vincent Ng. Supervised Ranking for Pronoun Resolution: Some Recent Improvements. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1081–1086. AAAI Press / The MIT Press, 2005.

Vincent Ng. Unsupervised Models for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 640–649, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Vincent Ng. Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4877–4884, San Francisco, California, USA, 2017. AAAI Press.

Vincent Ng and Claire Cardie. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, Stroudsburg, PA, USA, 2002a. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. Identifying Anaphoric and Non-anaphoric Noun Phrases to Improve Coreference Resolution. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7, Stroudsburg, PA, USA, 2002b. Association for Computational Linguistics.

Giang Linh Nguy. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, MFF UK, Prague, Czech Republic, 2006. In Czech.

Giang Linh Nguy and Magda Ševčíková. Unstated Subject Identification in Czech. In *WDS'11 Proceedings of Contributed Papers, Part I*, pages 149–154, Praha, Czechia, 2011. Matfyzpress.

Giang Linh Nguy and Zdeněk Žabokrtský. Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, pages 77–81, Lagos (Algarve), Portugal, 2007. CLUP-Center for Linguistics of the University of Oporto.

Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285, London, UK, 2009. The Association for Computational Linguistics.

NIST. Automatic Content Extraction. Technical report, 2003. http://www.nist.gov/speech/tests/ace/index.htm.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia, 2016.

Michal Novák. Machine Learning Approach to Anaphora Resolution. Master's thesis, MFF UK, Prague, Czech Republic, 2010.

Michal Novák. Coreference Resolution System Not Only for Czech. In *Proceedings of the 17th conference ITAT 2017: Slovenskočeský NLP workshop (SloNLP 2017)*, volume 1885 of *CEUR Workshop Proceedings*, pages 193–200, Praha, Czechia, 2017. CreateSpace Independent Publishing Platform.

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

Michal Novák and Zdeněk Žabokrtský. Resolving Noun Phrase Coreference in Czech. *Lecture Notes in Computer Science*, 7099:24–34, 2011.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Translation of "It" in a Deep Syntax Framework. In Bonnie L. Webber, editor, *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofija, Bulgaria, 2013a. Bălgarska akademija na naukite, Omnipress, Inc.

Michal Novák, Zdeněk Žabokrtský, and Anna Nedoluzhko. Two Case Studies on Translating Pronouns in a Deep Syntax Framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1037–1041, Nagoya, Japan, 2013b. Asian Federation of Natural Language Processing, Asian Federation of Natural Language Processing.

Michal Novák, Dieke Oele, and Gertjan van Noord. Comparison of coreference resolvers for deep syntax translation. In Bonnie L. Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier, editors, *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 17–23, Lisboa, Portugal, 2015. Association for Computational Linguistics, Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Projection-based coreference resolution using deep syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics (ACL).

Václav Novák and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. volume 4629, pages 92–98, Berlin / Heidelberg, 2007. Springer.

Franz J. Och and Hermann Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Maciej Ogrodniczuk. Translation- and Projection-Based Unsupervised Coreference Resolution for Polish. In *Language Processing and Intelligent Information Systems*, number 7912, pages 125–130, Berlin / Heidelberg, 2013. Springer.

Sebastian Padó and Mirella Lapata. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research*, 36(1):307–340, 2009.

Jarmila Panevová. Koreference gramatická nebo textová? In *Etudes de linguistique romane et slave*, Krakow, Poland, 1992. École normale supérieure de Cracovie, Département d'études romanes.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Petr Piťha. *Posesivní vztah v češtině*. AVED, Praha, 1992.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2016.

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler, and Radek Ocelák. Prague Discourse Treebank 1.0, 2012.

Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

Oana Postolache, Dan Cristea, and Constantin Orăsan. Transferring Coreference Chains through Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 889–892, Genoa, Italy, 2006. European Language Resources Association.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 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, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted

Coreference in OntoNotes. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012*, pages 1–40, Jeju, Korea, 2012. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, 1985.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A Multi-pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, pages 968–977, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. Translation-based Projection for Multilingual Coreference Resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 968–977, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

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

Rudolf Rosa. *Discovering the structure of natural language sentences by semi-supervised methods*. PhD thesis, Charles University, Faculty of Mathematics and Physics, Praha, Czechia, 2018.

Rudolf Rosa, Ondřej Dušek, David Mareček, and Martin Popel. Using parallel features in parsing of machine-translated sentences for correction of grammatical errors. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 39–48, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 2011. PMLR.

Petr Sgall. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic, 1967.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, Netherlands, 1986.

David A. Smith and Noah A. Smith. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 49–56, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, 2001.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 67–74, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, 2014a. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, 2014b. Association for Computational Linguistics.

Ondřej Svoboda. Functions of the Czech reflexive marker se/si. Master's thesis, Faculty of Humanities, Leiden University, Leiden, The Netherlands, 2014.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan T. McDonald, and Joakim Nivre. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *TACL*, 1:1–12, 2013.

Jörg Tiedemann. Emerging Language Spaces Learned From Massively Multilingual Corpora. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, DHN 2018, Helsinki, Finland, March 7-9, 2018.*, pages 188–197. CEUR-WS.org, 2018.

Don Tuggener. Coreference Resolution Evaluation for Higher Level Applications. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 231–235. The Association for Computer Linguistics, 2014.

Don Tuggener. *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich, 2016.

Don Tuggener and Manfred Klenner. A Hybrid Entity-Mention Pronoun Resolution Model for German Using Markov Logic Networks. In *Proceedings of the 12th Edition of the Konvens Conference*, pages 21–29, Hildesheim, Germany, 2014. Universitätsbibliothek Hildesheim.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, and Viktor Trón an d Viktor Nagy. Parallel Corpora for Medium Density Languages. In *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, volume 292, pages 247–258, Amsterdam, The Netherlands & Philadelphia, PA, USA, 2005. John Benjamins Publishing Company.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. BART: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Kateřina Veselovská, Giang Linh Nguy, and Michal Novák. Using Czech-English parallel corpora in automatic identification of "it". In *The Fifth Workshop on Building and Using Comparable Corpora*, İstanbul, Turkey, 2012. European Language Resources Association.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

Piek Vossen. Introduction to EuroWordNet. *Computers and the Humanities, Special Issue on EuroWordNet*, 32(2–3), 1998.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Alexander Wallin and Pierre Nugues. Coreference Resolution for Swedish and German using Distant Supervision. In *Linköping Electronic Conference Proceedings*, volume 131, Linköping, Sweden, 2017. Linköping University Electronic Press.

Ralph Weischedel and Ada Brunstein. BBN Pronoun Coreference and Entity Type Corpus, 2005.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, 2015. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. An NP-cluster Based Approach to Coreference Resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–232, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Amir Zeldes and Dan Simonson. Different Flavors of GUM: Evaluating Genre and Sentence Type Effects on Multilayer Corpus Annotation Quality. In *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, pages 68–78, Berlin, Germany, 2016. Association for Computational Linguistics.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015.

List of Figures

List of Tables

List of Publications

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. CzEng 1.0, 2011.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May 2012. ELRA, European Language Resources Association.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6, 2016.

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada, 2012. Association for Computational Linguistics.

Ondřej Dušek, Luís Gomes, Michal Novák, Martin Popel, and Rudolf Rosa. New Language Pairs in TectoMT. In *Proceedings of the 10th Workshop on Machine Translation*, pages 98–104, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Prague Czech-English Dependency Treebank 2.0 Coref, 2016a.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Coreference in prague czech-english dependency treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris, France, 2016b. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, and Maciej Ogrodniczuk. PAWS: A Multilingual 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, 2018. Association for Computational Linguistics.

Michal Novák. Machine Learning Approach to Anaphora Resolution. Master's thesis, MFF UK, Prague, Czech Republic, 2010.

Michal Novák. Coreference Resolution System Not Only for Czech. In *Proceedings of the 17th conference ITAT 2017: Slovenskočeský NLP workshop (SloNLP 2017)*, volume 1885 of *CEUR Workshop Proceedings*, pages 193–200, Praha, Czechia, 2017. CreateSpace Independent Publishing Platform.

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

Michal Novák and Zdeněk Žabokrtský. Resolving Noun Phrase Coreference in Czech. *Lecture Notes in Computer Science*, 7099:24–34, 2011.

Michal Novák and Zdeněk Žabokrtský. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Translation of "It" in a Deep Syntax Framework. In Bonnie L. Webber, editor, *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofija, Bulgaria, 2013a. Bălgarska akademija na naukite, Omnipress, Inc.

Michal Novák, Zdeněk Žabokrtský, and Anna Nedoluzhko. Two Case Studies on Translating Pronouns in a Deep Syntax Framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1037–1041, Nagoya, Japan, 2013b. Asian Federation of Natural Language Processing, Asian Federation of Natural Language Processing.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Projection-based coreference resolution using deep syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics (ACL).

Kateřina Veselovská, Giang Linh Nguy, and Michal Novák. Using Czech-English parallel corpora in automatic identification of "it". In *The Fifth Workshop on Building and Using Comparable Corpora*, İstanbul, Turkey, 2012. European Language Resources Association.

Attachments

A Distributions of Coreferential Expressions and Their Counterparts

This attachment shows in four bar charts a less detailed but much larger-scaled distributions than those presented in tables in Chapter 5. The distributions are measured on the following data: (1) PAWS with gold trees and manual alignment, (2) PAWS with system (auto) trees and supervised alignment, (3–4) PCEDT 2.0 Coref in the same variants as PAWS, and (5–9) CzEng 1.0 with supervised alignment for each of the four selected domains and in total for all domains in the corpus (not only the four selected ones).

We split the analysis of coreferential expressions' counterparts into two parts: (1) the analysis of relative frequencies of source-language coreferential expressions, and (2) the analysis of distribution over other-language counterparts aligned to these coreferential expressions. All these distributions are visualized in sets of bar charts.

A.1 Distributions of Coreferential Expressions

Relative frequencies of source-language coreferential expressions are plotted in Figures 9.1 and 9.2 for English and Czech, respectively. Both figures consist of several bar charts associated with a given source-language mention type. Each bar chart consists of eight bars corresponding to eight datasets as specified above. The height of a bar represents a proportion (in %) of nodes of a given type in a given dataset among all tectogrammatical nodes. The scale of the vertical axis differs for every bar chart. That is, relative frequencies of two different expressions within the same dataset are unlikely the same, even if the corresponding bars are visually of the same height.

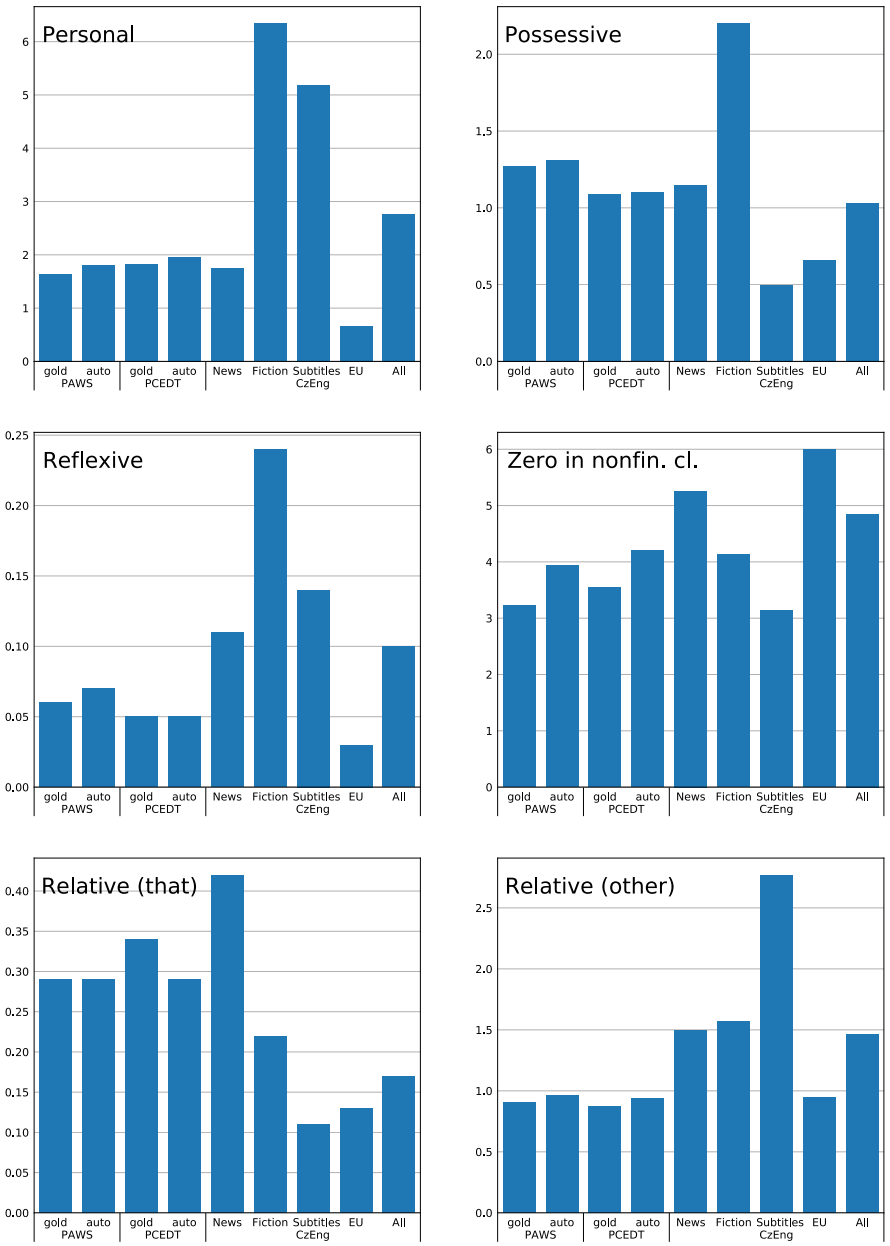

Figure 9.1: Distribution of English coref. expressions across various datasets and domains.

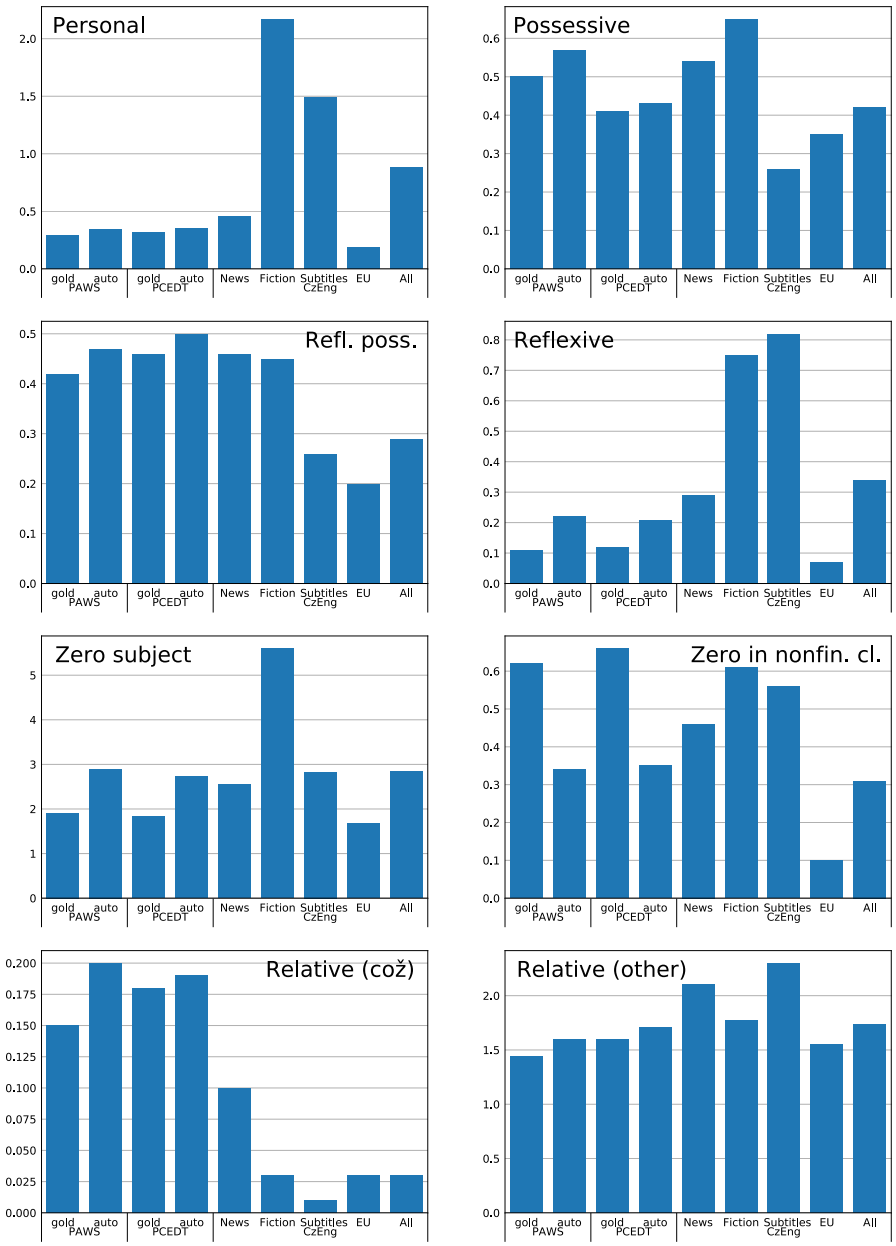

Figure 9.2: Distribution of Czech coref. expressions across various datasets and domains.

A.2 Distributions of Expressions' Counterparts

Distribution over counterparts aligned to the source-language coreferential expressions is depicted in bar charts in Figures 9.3 and 9.4 for English and Czech as source languages, respectively. Like in the previous figures, individual bar charts are associated with mention types and individual bars with the eight datasets. However, in Figures 9.3 and 9.4 all the bars are of the same height and they are partitioned into multiple colored segments. This partitioning corresponds to the distribution of source-language expressions with respect to mention types of their counterparts in the aligned lanugage. The partitioning may also contain a special segment attributed to the fact that no aligned counterpart has been found. Colored segments in the partitionings of different bars in a chart are sorted by the same order, which is their relative frequency within the concatenation of all available datasets.

Figure 9.3: Distribution of Czech expressions aligned to English coref. expressions across various datasets and domains.

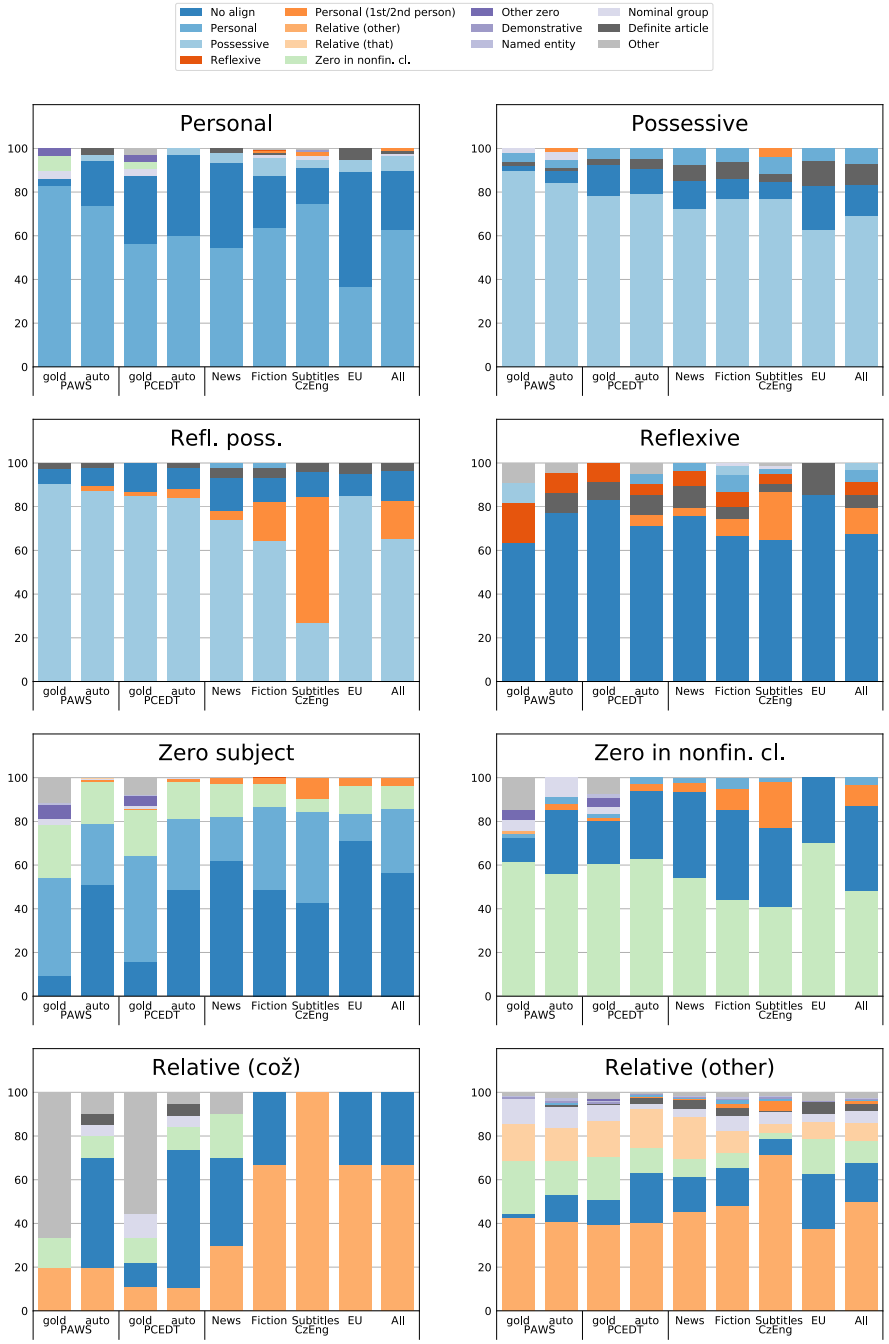

Figure 9.4: Distribution of English expressions aligned to Czech coref. expressions across various datasets and domains.