# COREFERENCE FROM THE CROSS-LINGUAL PERSPECTIVE

Michal Novák

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

# STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Michal Novák

## COREFERENCE FROM THE CROSS-LINGUAL PERSPECTIVE

Published by the Institute of Formal and Applied Linguistics
as the 18th publication in the series
Studies in Computational and Theoretical Linguistics. First edition, Prague 2018.



Reviewers: Prof. Dr. Manfred Stede (Department of Linguistics, University of Potsdam, Potsdam)
Ing. Alexandr Rosen, Ph.D. (Filozofická fakulta, Univerzita Karlova, Praha)

This book has been printed with the support of the grant GA16-05394S of the Grant Agency of the Czech Republic and of the institutional funds of Charles University.
Printed by Printo, spol. s r. o.



ISBN 978-80-88132-06-6

# Contents

## CONTENTS

# Acknowledgement

I would like to thank the supervisor of my dissertation thesis 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 book.

This study is based on projects investigated at the Institute of Formal and Applied Linguistics, Charles University in Prague. It was supported by the Grant Agency of the Czech Republic (project GA16-05394S). It has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2015071).

# 1
# Introduction

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

Before we start discussing the particular topics that this book 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). One of the recent research projects 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 work 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 two specific techniques: projection and bilingually informed resolution. Both of them 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 is 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 monograph 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 is often very different. 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 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 book, 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

[1] Many examples of the similar form can be encountered throughout the book. 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.

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 our Work

The aims of this monograph 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 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 Book

The book 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 book, including the approaches to monolingual as well as cross-lingual coreference resolution In Chapter 4 we introduce all the datasets employed throughout the book. In addition, we describe the pre-processing 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.

# 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 monograph – 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 improbable 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

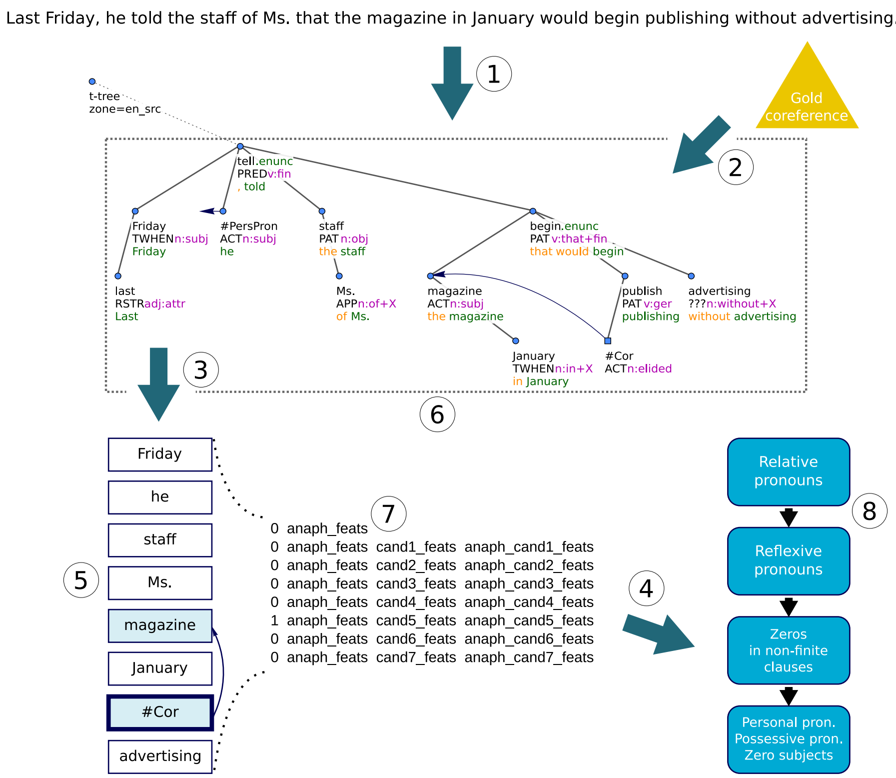


**Figure 7.1:** The architecture and the workflow of Treex CR in its monolingual setting. Note that the example tree is annotated automatically.

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.

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,[1] reflexive pronouns point to the subject of the clause, 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.[2] 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 sim-

[1] Note that the morphological persons may disagree if one of the mentions appears in a direct speech.

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

ple 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 distinction between grammatical and textual coreference made in the Prague tectogrammatics is motivated by the difference between linguistic means that get coreference relations realized (see Section 2.3). Whereas the grammar of the language plays an exclusive role in the former case, it is both the grammar and context in the latter case of coreference. The properties of coreferential relations are even more diverse, though. For instance, while the antecedent of a relative pronoun tends to lie a few words before the pronoun, a reflexive pronoun almost always refers to the subject of the clause it belongs to. By representing coreference of such expressions separately in multiple specialized models, such differences are expected to be highlighted by the weights associated with individual features more clearly than if a single joint model was used (as shown in Denis and Baldridge, 2008). Moreover, adjusting the abovementioned hyperparameters individually for each anaphor type allows for filtering out unlikely antecedent candidates. On the other hand, excessive granularity of the models would lead to the lack of generalization. We thus train specialized models that attempt to underline the specific properties of the individual mention types (as introduced in Section 2.4) and at the same time group the types with similar properties.

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). Nevertheless, in the present experiments the models are built independently of each other and can thus 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).[3]

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

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,[4] 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

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

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, 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 (CNC, 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

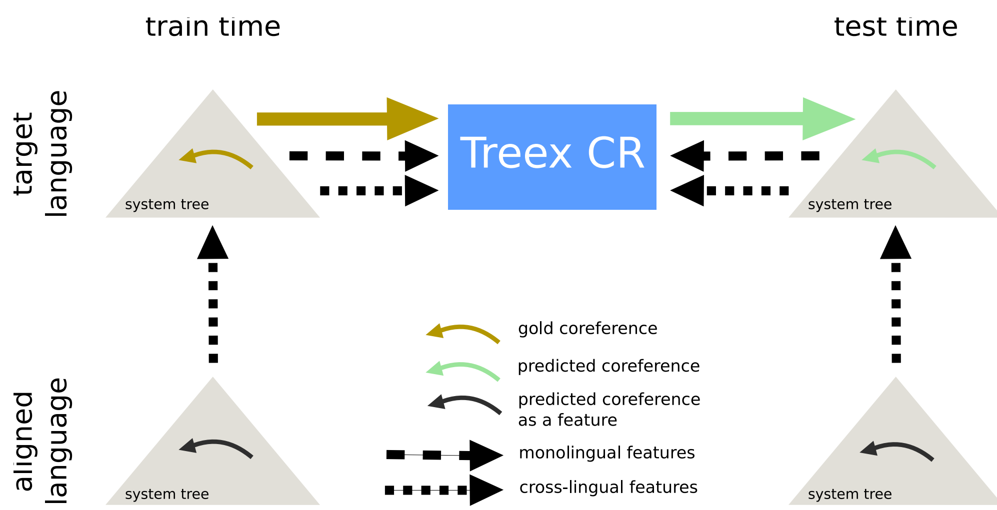


**Figure 7.2:** The workflow of Treex CR in its bilingually informed setting.

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.

**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.

| | 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: $5 \times 10^{-8}$, passes over data: 5, quadratic combination of anaphor and antecedent features |
| Reflexive pron.<br>Refl. poss. pron. | current sent. | ✓ | semantic nouns | |
| Zeros in non-fin. cl. | current sent. | ✓ | semantic nouns and zeros | |
| Personal pron.<br>Possessive pron.<br>Zero subjects | current and previous sent. | × | semantic nouns in the 3rd or undefined person | |

**Table 7.1:** Hyperparameters of Treex CR models.

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.

## 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

| | 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.

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.[5] 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 language,

[5] 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.

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. 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 | 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.

### 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

| Mention type | Stanford deter. | Stanford stat. | Stanford neur. | CzEng CR | Treex CR |
|---|---|---|---|---|---|
| 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.

| 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.

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 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), it would be interesting to find the reasons for this discrepancy.

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 C.1 in Appendix C 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 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.[6] 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, out of which we picked some to be presented in the book. 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.[7]

### 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

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

[7] 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.

| 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.

| 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.

| 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.

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 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 C.2 in Appendix C 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.

### 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.[8] 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 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,[9] 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

[8] 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.

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

| 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.

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-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:

| 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.

- 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[10] 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 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 mea-

[10] 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.

sured 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 A.1 in Section 4.4.3.

Let us start with a motivating example. Results in Tables 7.6 and 7.7 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 word order, which is more strict, 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***$_{\mathbf{m.pl}}$ | *skupin*$_{\text{f.pl}}$ | *a* | *sesmolil* | ***je***$_{\mathbf{mfn.pl}}$ | *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***$_{\mathbf{fmn.s}}$ | *o* | *společnost Jaguar*$_{\text{fm.s}}$ | *odráží* | *touhu*$_{\text{f.s}}$ | *pomoci*$_{\text{f.s}}$ | *zpestřit* |
| interest | GM-company's | in | Jaguar company | reflects | a desire | to help | diversify |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *produkty*$_{\text{m.p}}$ | ***této společnosti***$_{\mathbf{f.s}}$ | *na* | *trhu*$_{\text{m.s}}$ | *s* | *vozy*$_{\text{m.p}}$ | *.* | ***jeho***$_{\mathbf{mn.s}}$ | *série* | *Opel* |
| products | of this company | in | market | with | cars | . | 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*$_{\text{subj}}$ | ~~$\varnothing_{\textbf{subj}}$~~ | *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* | ~~$\varnothing_{\textbf{actor}}$~~ | *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 setting, 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.

# 9

# Conclusion

In this monograph, 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 contrastive research on 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 is interesting enough. However, it will be even more interesting as soon as the presented methods are applied 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.

# Summary

The subject of this monograph 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 this work.

# List of Figures

# List of Tables

# Bibliography


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.

Mona Baker and Harold Somers. *Corpus-based Translation Studies: The Challenges that Lie Ahead*, pages 175–186. John Benjamins Publishing Company, Amsterdam, Netherlands, 1996.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. *'Corpus Linguistics and Translation Studies: Implications and Applications'*, pages 233–250. John Benjamins Publishing Company, Amsterdam, Netherlands, 1993.

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.

Shoshana Blum-Kulka and Eddie A. Levenston. *Universals of Lexical Simplification*, pages 119–139. Longman, London & New York, 1983.

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.

CNC. *Czech National Corpus – SYN2005*. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic, 2005.

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.

Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642. Association for Computational Linguistics, 2016.

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 Multi-lingual Parallel Treebank with Anaphoric Relations. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 68–76, New Orleans, Louisiana, 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.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 517–527. Bălgarska akademija na naukite, Association for Computational Linguistics, 2013.

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.