# A Study on Bilingually Informed Coreference Resolution

Michal Novák

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Prague 1
mnovak@ufal.mff.cuni.cz

*Abstract:* Coreference is a basic means to retain coherence of a text that likely exists in every language. However, languages may differ in how a coreference relation is manifested on the surface. A possible way how to measure the extent and nature of such differences is to build a coreference resolution system that operates on a parallel corpus and extracts information from both language sides of the corpus. In this work, we build such a bilingually informed coreference resolution system and apply it on Czech-English data. We compare its performance with the system that learns only from a single language. Our results show that the cross-lingual approach outperforms the monolingual one. They also suggest that a system for Czech can exploit the additional English information more effectively than the other way round. The work concludes with a detailed analysis that tries to reveal the reasons behind these results.

## 1 Introduction

Cross-lingual techniques are becoming still more and more popular. Even though they do not circumvent the task of Coreference Resolution (CR), the research is mostly limited to cross-lingual projection. Other cross-lingual techniques remain a largely unexplored area for this task.

One of the yet neglected cross-lingual techniques is called *bilingually informed resolution*. It is an approach, in which decisions in a particular task are made based on the information from bilingual parallel data. Parallel texts must be available when a method is trained, but also at test time, that is when a trained model is applied to new data. In real-world scenarios, the availability of parallel data at test time requires the technique to apply a machine translation service to acquire them (MT-based bilingually informed resolution).

Nevertheless, for limited purposes it may pay off to use human-translated parallel data instead (corpus-based bilingually informed resolution). If it outperforms the monolingual approach, it may be used in building automatically annotated parallel corpora. Such corpora with more reliable annotation could be useful for corpus-driven theoretical research.[1] Furthermore, it can be also used for automatic processing. For instance, improved resolution on big parallel data might be leveraged in a weakly supervised manner to boost the models trained in a monolingual way.

The present work is concerned with corpus-based bilingually informed CR on Czech-English texts. Specifically, it focuses on resolution of pronouns and zeros, as these are the coreferential expressions whose grammatical and functional properties differ considerably across the languages. For instance, whereas in English most of non-living objects are referred to with pronouns in neuter gender (e.g. "*it*", "*its*"), genders are distributed more evenly in Czech. Information on Czech genders thus may be useful to filter out English candidates that are highly improbable to be coreferential with the pronoun. By comparison of its performance with a monolingual approach and by thorough analysis of the results, our work aims at discovering the extent and nature of such differences.

The paper is structured as follows. After mentioning related work (Section 2), we introduce a coreference resolver (Section 3), both its monolingual and cross-lingual variants. Section 4 describes the dataset used in experiments in Section 5. Before we conclude, the results of experiments are thoroughly analyzed (Section 6).

## 2 Related Work

Building a bilingually informed CR system requires a parallel corpus with at least the target-language side annotated with coreference. Even these days very few such corpora exist, e.g. Prague Czech-English Dependency Treebank 2.0 Coref [14], ParCor 1.0 [9] and parts of OntoNotes 5.0 [19].

It is thus surprising that the peak of popularity for such approach was reached around ten years before these corpora had been published. Harabagiu and Maiorano [10] present an heuristics-based approach to CR. The set of heuristics is expanded by exploiting the transitivity property of coreferential chains in a bootstrapping fashion. Moreover, they expand the heuristics even more, following mention counterparts in translations of source English texts to Romanian with coreference annotation. Mitkov and Barbu [13] adjust a rule-based pronoun coreference resolution system to work on a parallel corpus. After providing a linguistic comparison of English and French pronouns and their behavior in discourse, the authors distill their findings into a set of cross-lingual rules to be integrated into the CR system. In evaluation, they observe im-

[1] In case a cross-lingual origin of the annotation does not matter.

provements in resolution accuracy of up to 5 percentage points compared to the monolingual approach.

As for more recent works, the authors of [5] address the task of overt pronoun resolution in Chinese. Among the others they propose an MT-based bilingually informed approach. A model is built on Chinese coreference, exploiting Chinese features. These are augmented with English features, extracted from the Chinese texts machine-translated to English. It allows for taking advantage of English nouns' gender and number lists, which according to authors correspond to the distribution of genders and numbers over Chinese nouns.

Experiments of Novák and Žabokrtský [17], the first ones using bilingually informed CR on Czech-English data, are most relevant to the present work. With the focus on English personal pronouns only, their best cross-lingual configuration managed to outperform the monolingual CR by one F-score point. Taking advantage of a more developed version of their CR system, we extend their work in several directions. First, we explore the potential of such approach for a wider range of English coreferential expressions. Next, we perform experiments in the opposite direction, i.e. Czech CR informed by English. And finally, we provide a very detailed analysis of the results unveiling the nature of the cross-lingual aid.

## 3 Coreference Resolution System

For coreference resolution we adopt a more developed version of the resolver utilized in [17]. This new version builds on the monolingual Treex CR system [15], and augments it with the cross-lingual extension presented in [17]. The difference between the current system and the system in [17] lies mostly in that it can target a wider range of expressions, it exploits a richer feature set and the preprocessing stage analyzing the text to the tectogrammatical representation is of higher quality. Instead of listing all the changes, we briefly introduce the monolingual (Section 3.1) and the cross-lingual component (Section 3.2) of Treex CR from the scratch.[2]

### 3.1 Monolingual Resolution

Treex CR operates on the *tectogrammatical layer*. It is a layer of deep syntax based on the theory of Functional Generative Description [20]. The tectogrammatical representation of a sentence is a dependency tree with rich linguistic features consisting of the content words only. Furthermore, some surface ellipses are restored at this layer. It includes anaphoric zeros (e.g. zero subjects in Czech, unexpressed arguments of non-finite clauses in both English and Czech) that are introduced in the tectogrammatical layer with a newly established node.

[2]Please refer to [15] for more details on the monolingual component of the system.

The tectogrammatical layer is also the place, where coreference relations should be annotated. It is technically represented as a link between two coreferential nodes:[3] *the anaphor* (the referring expression) and *the antecedent* (the referred expression).

Each input text must be first automatically preprocessed up to this level of linguistic annotation. The CR system based on supervised machine learning then takes advantage of the information available in the annotation.

*Pre-processing.* The input text must undergo an analysis producing a tectogrammatical representation of its sentences before coreference resolution is carried out. We use pipelines for analysis of Czech and English available in the Treex framework [18]. The analysis starts with a rule-based tokenization, morphological analysis and part-of-speech tagging (e.g. [21] for Czech), dependency parsing to surface trees (e.g. MST parser [12] for English) and named entity recognition [22]. In addition, the NADA tool [3] is applied to help distinguish referential and non-referential occurrences of the English pronoun "*it*".

Tectogrammatical trees are created by a transformation from the surface trees. All function words are made hidden, morpho-syntactic information is transferred and semantic roles are assigned to tectogrammatical nodes [4]. On the tectogrammatical layer, certain types of ellipsis can be restored. The automatic pre-processing focuses only on restoring nodes that might be anaphoric. Such nodes are added by heuristics based on syntactic structures. The restored nodes include Czech zero subjects and both Czech and English zeros in non-finite clauses, e.g. zero relative pronouns, unexpressed arguments in infinitives, past and present participles.

*Model design.* Treex CR models coreference in a way to be easily optimized by supervised learning. Particularly, we use logistic regression with stochastic gradient descend optimization implemented in the Vowpal Wabbit toolkit.[4] Design of the model employs multiple concepts that have proved to be useful and simple at the same time.

Given an anaphor and a set of antecedent candidates, *mention-ranking* models [6] are trained to score all the candidates at once. On the one hand a mention-ranking model is able to capture competition between the candidates, but on the other hand features describe solely the actual mentions, not the whole clusters built up to the moment. Antecedent candidates for an anaphor (both positive and negative) are selected from the context window of a predefined size.

No anaphor detection stage precedes the coreference resolution. Unless another measure was taken, it would lead to all occurrences of the pronoun "*it*" labeled as referential, for instance. Nevertheless, the model determines

[3]A mention is determined only by its head in tectogrammatics. No mention boundaries are specified. Therefore, it is sufficient for a coreference link to determine only two nodes, the mentions' head nodes.

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

whether the anaphor is referential jointly with selecting its antecedent. This is ensured by adding a dummy candidate representing solely the anaphor itself. By selecting this candidate, the model claims that the anaphor is in fact non-referential.

Diverse properties of various types of coreferential relations (e.g. different referential scopes of personal and relative pronouns) encouraged us to model individual anaphor types separately. A specialized model is build for (1) personal and possessive pronouns in 3rd person (and zero subjects in Czech), (2) reflexive pronouns, (3) relative pronouns, and (4) zeros in non-finite clauses. Treex CR runs them in a sequence.

*Features.* The pre-processing stage enriches a raw text with a substantial amount of linguistic information. Feature extraction stage then uses this material to yield *features* consumable by the learning method. Features are always related to at most two nodes – an anaphor candidate and an antecedent candidate.

The features can be divided into three categories. Firstly, location and distance features indicate positions of the anaphor and the antecedent candidate in a sentence and their mutual distance in terms of words, clauses and sentences. Secondly, a big group of features reflects (deep) morpho-syntactic aspects of the candidates. It includes the mention head's part-of-speech tag and morphological features (e.g. gender, number, person, case), (deep) syntax features (e.g. dependency relation, semantic role) as well as some features exploiting the structure of the syntactic tree. Many of the features are combined by concatenation or by agreement, i.e. indicating whether the anaphor's value agrees with antecedent's one. Finally, lexical features focus on lemmas of the mentions' heads and their parents. These are used directly or through the frequencies collected in a large data of Czech National Corpus [1] indexed in a list of noun-verb collocations. Furthermore, all hypernymous concepts of a mention are extracted as features from ontologies (e.g. WordNet [7]) and named entity labels are also employed.

### 3.2 Cross-lingual Extension

The extension enables bilingually informed CR. Like the monolingual CR, it addresses coreference in one target language at a time. However, instead of data in single language, it must be fed with parallel data in two languages. Both language sides (Czech and English in this case) of the data must be first pre-processed with the pipelines analyzing the texts up to the diagrammatically layer. Furthermore, to facilitate the access to important information in the other language, the pre-processing stage also seeks for alignment between tectogrammatical nodes. The bilingually informed approach then augments the monolingual features with those accessing the other side of the parallel data. Design of the model remains the same as for the monolingual approach.

*Alignment.* It is central for our cross-lingual approach to have the English and Czech texts aligned on the level of tectogrammatical nodes. The alignment is based on unsupervised word alignment performed by MGIZA++ [8] trained on the data from CzEng 1.0 [4], and projected to the tectogrammatical layer. Furthermore, it is augmented with a supervised method [17] addressing selected coreferential expressions, including potentially anaphoric zeros.

*Features.* Cross-lingual features describe the nodes aligned to the coreferential candidates in the target language – the anaphor candidate and the antecedent candidate. To collect such nodes, we follow the alignment links connected to these two candidates. For each of the 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, for which we can extract cross-lingual features. If no aligned counterpart is found, no cross-lingual features are added.

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 indicator feature, assigning the true value only if the two aligned nodes belong to the same coreferential entity. This feature can be activated only if there exists a monolingual coreference resolver for the aligned 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.

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

## 4 Datasets

We employ Prague Czech-English Dependency Treebank 2.0 Coref [14, PCEDT 2.0 Coref] to train and test our CR systems. It is a Czech-English parallel corpus, consisting of almost 50k sentence pairs (more on its basic statistics is shown in the upper part of the Table 1). The English part of the treebank is based on texts from the Wall Street Journal collected for the Penn Treebank [11]. The Czech part was manually translated from English. All texts have been annotated at multiple layers of linguistic representation up to the tectogrammatical layer.

Although PCEDT 2.0 Coref has been extensively annotated by humans, we strip almost all manual annotations and replace it by the output of the pre-processing pipeline (see Sections 3.1 and 3.2). The only manually annotated information that we retain are the coreferential links.

We do not split the data into train and test sections. All the experiments are conducted using 10-fold cross-validation.

| Mention type | Czech | English |
|---|---|---|
| Sentences | 49,208 | 49,208 |
| Tokens | 1,151,150 | 1,173,766 |
| Tecto. nodes | 931,846 | 838,212 |
| Mentions (total) | 183,277 | 188,685 |
| Personal pron. | 3,038 | 14,887 |
| Possessive pron. | 3,777 | 9,186 |
| Refl. poss. pron. | 4,389 | — |
| Reflexive pron. | 1,272 | 484 |
| Zero subject | 16,875 | — |
| Zero in nonfin. cl. | 6,151 | 29,759 |
| Relative pron. | 15,198 | 8,170 |
| Other | 132,577 | 126,199 |

Table 1: Basic and coref. statistics of PCEDT 2.0 Coref.

| Mention type | Czech | | English | |
|---|---|---|---|---|
| | monoling | biling | monoling | biling |
| Personal | $^{63.84}_{61.24}$ 62.51 | $^{67.82}_{64.38}$ 66.06 | $^{76.34}_{71.37}$ 73.77 | $^{78.57}_{72.64}$ 75.49 |
| Possessive | $^{71.93}_{71.51}$ 71.72 | $^{75.73}_{74.85}$ 75.29 | $^{80.07}_{79.54}$ 79.81 | $^{81.46}_{81.00}$ 81.23 |
| Refl. poss. | $^{85.61}_{85.42}$ 85.52 | $^{87.70}_{87.04}$ 87.36 | — | — |
| Reflexive | $^{66.91}_{56.60}$ 61.33 | $^{67.24}_{55.66}$ 60.90 | $^{77.31}_{72.67}$ 74.92 | $^{75.88}_{71.01}$ 73.37 |
| Zero subj. | $^{73.18}_{55.46}$ 63.10 | $^{78.88}_{57.64}$ 66.61 | — | — |
| Zero nonfin. | $^{78.98}_{41.51}$ 54.42 | $^{81.52}_{42.63}$ 55.98 | $^{71.48}_{54.62}$ 61.92 | $^{73.31}_{54.75}$ 62.68 |
| Relative | $^{81.51}_{79.94}$ 80.72 | $^{83.48}_{81.62}$ 82.54 | $^{83.47}_{76.23}$ 79.69 | $^{85.76}_{77.13}$ 81.21 |
| Total | $^{76.83}_{65.17}$ 70.52 | $^{80.27}_{67.09}$ 73.09 | $^{75.93}_{65.26}$ 70.19 | $^{77.85}_{65.95}$ 71.41 |

Table 2: Anaphora scores of monolingual and bilingually informed coreference resolution.

As mentioned in Section 3.1, our CR system consists of four models targeting different types of mentions as anaphors. In evaluation, we split the anaphor candidates to even finer categories, namely: (1) personal pronouns, (2) possessive pronouns, (3) reflexive possessive pronouns, (4) reflexive pronouns, all four types of pronouns in the 3rd or ambiguous person, (5) zero subjects, (6) zeros in non-finite clauses, and (7) relative pronouns (the statistics of coreferential mentions is collected in the bottom part of Table 1). Driven by the findings in an analysis of Czech-English correspondences [16], these expressions are very interesting from a cross-lingual point of view, as they often transform to a different type or carry different grammatical properties, when translated. We assume this aspect is not so significant in case of nominal groups, for instance, which represent the majority of remaining mentions. The other types grouped under the category Other are demonstrative pronouns, pronouns in 1st and 2nd person etc. This category of anaphors is not targeted by our CR method.

## 5 Experiments

The following experiments compare the performance of the monolingual and bilingually informed system. Both systems are trained on the PCEDT dataset. All the design choices (except for the feature sets) and hyperparameter values are shared by both systems.

*Evaluation measure.* We expect different mention types to behave differently in the cross-lingual approach. Standard evaluation metrics (e.g. MUC [23], $B^3$ [2]), however, do not allow for scoring only a subset of mentions. Instead, we use the *anaphora score*, an anaphor-decomposable measure proposed by [15]. The score consists of three components: precision, recall, and F-score as a harmonic mean of the previous two. While precision expresses the success rate of a system averaged over all mentions labeled by the system as anaphoric, recall averages over all true anaphoric mentions. A decision on an anaphor candidate is correct if the system correctly labels it as non-anaphoric or the antecedent found by the system really belongs to the same entity as the anaphor. In the following tables, we use $^{\mathrm{P}}_{\mathrm{R}}$F to format the three components of the anaphora score.

*Bilingually informed vs. Monolingual CR.* Table 2 lists the anaphora scores measured on the output of 10-fold cross-validation. In overall, cross-lingual models succeed in exploiting additional knowledge from parallel data and perform better than the monolingual approach. The F-score improvement benefits mainly from a rise in precision, but recall also gets improved. In both languages, personal and possessive pronouns are the types that exhibit the greatest improvement. In Czech, the top-scoring mention types include zero subjects, too. Nevertheless, English as an aligned language seems to have a stronger impact on resolution in Czech (the difference between the systems is 2.5 F-score points) than Czech has on resolution in English (the difference of 1.2 F-score points).

## 6 Analysis of the Results

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. Firstly, we conduct a quantitative analysis of resolvers' decisions. It should show how many decision changes the switch to the cross-lingual approach introduces for individual mention types and what is the role of anaphoricity in these changes. Secondly, we inspect randomly sampled examples in a qualitative analysis. We attempt to disclose what are the typical examples when the system benefits from the other language and, on the other hand, if there is a systematic case when the cross-lingual approach hurts.

| Mention type | Anaph | | | | Non-anaph | | | |
|---|---|---|---|---|---|---|---|---|
| | 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 | | 0.05 |
| 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 3: 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.

### 6.1 Quantitative Analysis

Let us start with a quantitative analysis of improvements and worsenings with respect to anaphoricity and type of the anaphor candidate. Tables 3 and 4 show for Czech and English, respectively, how often the cross-lingual system (denoted as C) is better than the monolingual (denoted as M). Each anaphor candidate falls to one of the four categories based on how C and M decided on the candidate:

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

A decision is either assignment of the anaphor candidate to a coreferential entity[5] 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%.

Conditioning on anaphoricity allows us to directly relate this analysis to the anaphora scores shown in Table 2. 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.

Changed decisions account for around 10% in both Czech and English. More importantly, whereas we see over 7% of decisions changed positively in Czech, it corresponds to 5.5% of decisions in English. This accords with the extents of improvement observed on anaphora score. In Czech, a difference between improved and worsened decisions is only a bit higher for anaphoric mentions. It means that the positive effect of English on resolution of Czech anaphoric mentions is about on par with its effect on resolution on non-anaphoric mentions. But conversely, Czech helps more in resolution of non-anaphoric mentions.

Let us zoom in to the individual mention types. The highest proportion of changed decisions appears for personal pronouns and zero subjects in Czech (14% instances) and for reflexive pronouns in English (12%). Interestingly, its effect on anaphora score cannot be more different. Czech personal pronouns and zero subjects are the mention types where the cross-lingual approach improves the anaphora score the most. On the other hand, 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 resolution of Czech reflexive possessives is hardly improved by English features, the small amount of changed decisions on English relative pronouns suffices to achieve one of the biggest improvements among English coreferential expressions.

Anaphora scores in Table 2 have already shown that basic reflexive pronouns are the only mention type, where the cross-lingual approach falls behind the monolingual one. The quantitative analysis of changed decisions confirms it, especially for anaphoric occurrences.

The gains of the Czech cross-lingual system on non-anaphoric mentions can be attributed mostly to zeros. Also thanks to the resolution on non-anaphoric mentions, the highest margin between the proportion of improved and worsened instances (5%) is observed on Czech zero subjects. It leads to one of the biggest improvement in terms of the anaphora F-score (see Table 2).

### 6.2 Qualitative Analysis

In the following, we scrutinize more closely what are the typical cases, where the cross-lingual system makes a different decision.

[5] 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.

| Mention type | Anaph | | | | Non-anaph | | | |
|---|---|---|---|---|---|---|---|---|
| | 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 | | 2.83 | | |
| 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 4: 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.

Let us start with a motivating example. Results in Table 2 show that improvement of the bilingually informed system on Czech personal and possessive pronouns and zero subjects is twice as high 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.

Czech personal and possessive pronouns are the mention types that considerably benefit from the cross-lingual approach. 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 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.

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

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

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

The analysis also shows that English syntax, which is more strict and thus easier to reconstruct, often helps in determining the correct antecedent. Example 2 shows the case, where neither English gender nor number could affect the resolver's decision. The correct decision is rather a result of a clear structure, where the syntactic objects in coordinated clauses very likely refer to the same entity.

(2) *kdo* *posbíral* ***plány***$_{\text{m.p}}$ *skupin*$_{\text{f.p}}$ *a* *sesmolil* ***je***$_{\text{mfn.p}}$
who collected plans from groups and cobbled them
*do* *iniciativy*
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 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"). Not only the possessed objects does not have to be the same, but 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).

(3) *zájem*$_{\text{m.s}}$ ***společnosti GM***$_{\text{fm.s}}$ *o* *společnost Jaguar*$_{\text{fm.s}}$
interest GM-company's in Jaguar company
*odráží* *touhu*$_{\text{f.s}}$ *pomoci*$_{\text{f.s}}$ *zpestřit* *produkty*$_{\text{m.p}}$
reflects a desire to help diversify products
***této společnosti***$_{\text{f.s}}$ *na* *trhu*$_{\text{m.s}}$ *s* *vozy*$_{\text{m.p}}$ . ***jeho***$_{\text{mn.s}}$
of this company in market with cars . its
*série* *Opel*
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: gender and number of the anaphor, more strict syntactic constraints in English etc. English gender may be even more important here, as the gender of a subject zero is impossible to be recognized just from the form of the governing verb, 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 most confident decision changes were in fact labeled as non-anaphoric by the monolingual system (e.g. in Example 4). Czech language side of the data thus help correctly label these pronouns as anaphoric.

(4) *Compelled* ***service*** *is* *unconstitutional* ***<u>It</u>*** *is* *also*
Nucená služba$_{f.s}$ je protiústavní $\varnothing_{f.s}$ Je také
*unwise*
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 5), or the cases where the pronoun has no Czech counterpart at all. In all these cases, the syntactic subject of the clause in which the pronoun lies is a preferred antecedent.

(5) ***Digital Equipment Corp.*** *announced* ***<u>its</u>*** *line*
společnost Digital Equipment Corp. představila svou řadu
*of computers*
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 expressions reconstructed during the automatic analysis are in fact superfluous. 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 6). This error subsequently propagates to a wrong decision of the monolingual resolver (the word "*zpráva* /report/" labeled as an antecedent). Any superfluous 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 superfluous zero plays a significant role in such decisions.

(6) *Avšak* ***~~zpráva~~*** *uvádí* *že* *společnosti*$_{\text{subj}}$ $\underline{\varnothing}_{\text{subj}}$
But the report said that companies –
*platí* *více* *daní*
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 noun phrase. For instance, in Example 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, there is usually no counterpart for the English zero to align with.

(7) ***~~Fear~~*** *of* *AIDS* *hinders* $\underline{\varnothing}_{\text{actor}}$ *hiring*
Strach z AIDS komplikuje – nábor$_{noun}$

Fear of AIDS hinders hiring at few hospitals.

Strach z AIDS komplikuje nábor v několika nemocnicích.

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 decisions made by the cross-lingual method in Czech 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 makes 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 Conclusion

This work conducts experiments on bilingually informed coreference resolution on Czech-English data. Comparing this cross-lingual approach to a monolingual resolver, we discovered that English helps more in resolution of Czech expressions than vice versa. A quantitative analysis shows that while English facilitates resolution of both Czech anaphoric and non-anaphoric mentions, Czech primarily helps to identify non-anaphoric mentions. The qualitative analysis reveals main reasons for improvements and worsenings all over the mention types. The most surprising finding is that the information on English gender seems to be improving resolution of Czech coreference more than vice versa. The animacy feature hidden in English gender appeared to be stronger than more even distribution of Czech genders across nouns.

## Acknowledgments

This project has been funded by the Czech Science Foundation grant GA-16-05394S. This work has been also sup-

ported and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.

## References

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

[2] A. Bagga and B. 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.

[3] S. Bergsma and D. 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.

[4] O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May 2012. ELRA, European Language Resources Association.

[5] C. Chen and V. 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.

[6] P. Denis and J. 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, 2007. Morgan Kaufmann Publishers Inc.

[7] C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[8] Q. Gao and S. 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.

[9] L. Guillou, C. Hardmeier, A. Smith, J. Tiedemann, and B. 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).

[10] S. M. Harabagiu and S. 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.

[11] M. Marcus, B. Santorini, M. A. Marcinkiewicz, and A. Taylor. Penn Treebank 3, 1999.

[12] R. McDonald, F. Pereira, K. Ribarov, and J. 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.

[13] R. Mitkov and C. Barbu. Using Bilingual Corpora to Improve Pronoun Resolution. *Languages in contrast*, 4(2), 2003.

[14] A. Nedoluzhko, M. Novák, S. Cinková, M. Mikulová, and J. Mírovský. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris, France, 2016. European Language Resources Association.

[15] M. 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.

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

[17] M. Novák and Z. Žabokrtský. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

[18] M. Popel and Z. 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.

[19] S. Pradhan, A. Moschitti, N. Xue, H. T. Ng, A. Björkelund, O. Uryupina, Y. Zhang, and Z. 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.

[20] P. Sgall, E. Hajičová, and J. Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, Netherlands, 1986.

[21] J. Straková, M. Straka, and J. 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, 2014. Association for Computational Linguistics.

[22] J. Straková, M. Straka, and J. 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, 2014. Association for Computational Linguistics.

[23] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. 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.